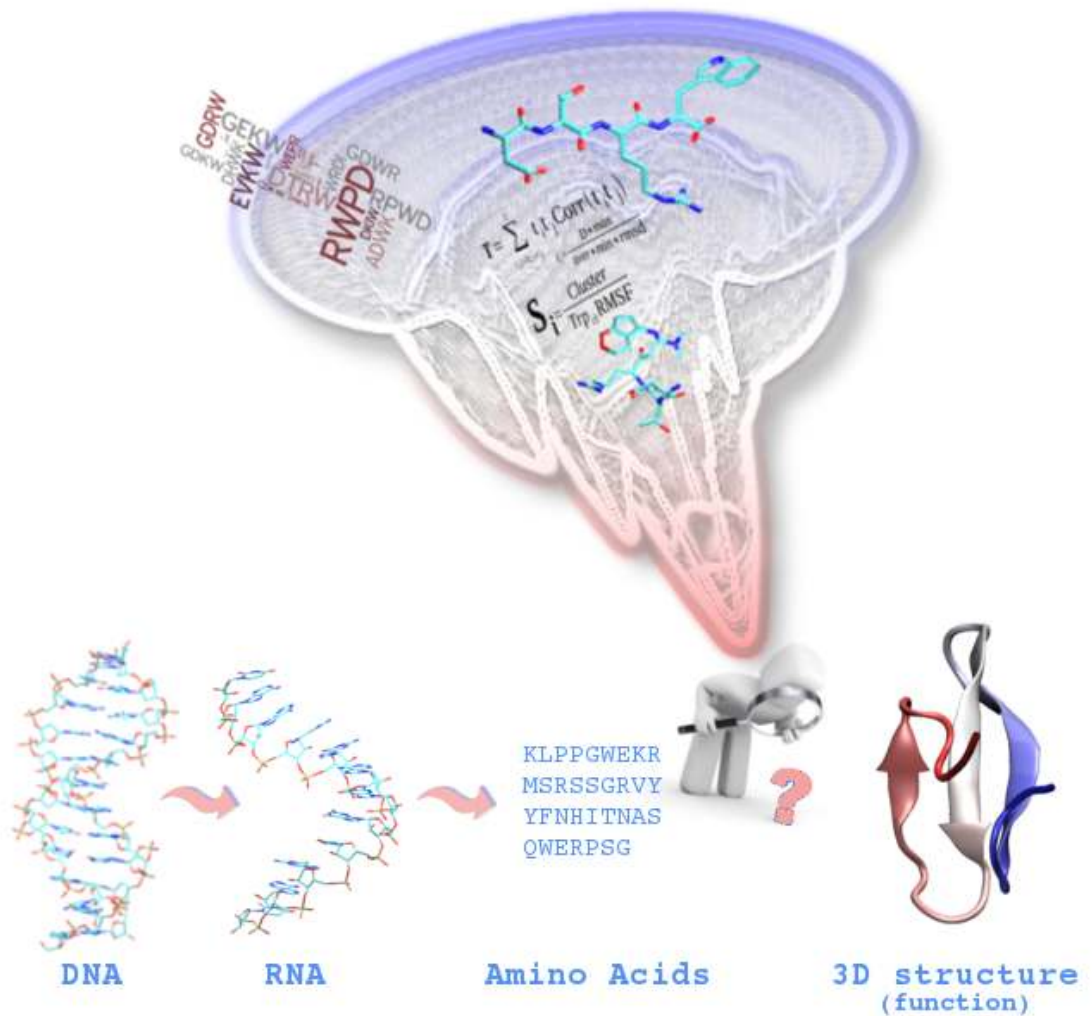


# Δομικές και Υπολογιστικές Μελέτες Πεπτιδίων



**ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ**  
**ΤΜΗΜΑ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ**  
**ΚΑΙ ΓΕΝΕΤΙΚΗΣ**



**Δομικές Και Υπολογιστικές**  
**Μελέτες Πεπτιδίων**

**ΠΑΝΑΓΙΩΤΑ Σ. ΓΕΩΡΓΟΥΛΙΑ**

**ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ**

---

ΕΚΠΟΝΗΘΗΚΕ ΣΤΟ ΕΡΓΑΣΤΗΡΙΟ ΔΟΜΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΙΚΗΣ ΒΙΟΛΟΓΙΑΣ,  
ΤΟΥ ΤΜΗΜΑΤΟΣ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ ΚΑΙ ΓΕΝΕΤΙΚΗΣ,  
ΤΟΥ ΔΗΜΟΚΡΙΤΕΙΟΥ ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΘΡΑΚΗΣ

**ΑΛΕΞΑΝΔΡΟΥΠΟΛΗ 2012**

## ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ

Νικόλαος Μ. Γλυκός, Επίκουρος Καθηγητής Υπολογιστικής και Δομικής Βιολογίας, Δ.Π.Θ.

## ΤΡΙΜΕΛΗΣ ΣΥΜΒΟΥΛΕΥΤΙΚΗ ΕΠΙΤΡΟΠΗ

Νικόλαος Γλυκός, Επίκουρος Καθηγητής Υπολογιστικής και Δομικής Βιολογίας, Δ.Π.Θ.

Ραφαήλ Σανδάλτζόπουλος, Αναπληρωτής Καθηγητής Μοριακής Βιολογίας, Δ.Π.Θ.

Αγλαΐα Παππά, Επίκουρος Καθηγήτρια Φυσιολογίας και Μοριακής Φαρμακολογίας, Δ.Π.Θ.

## ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Νικόλαος Γλυκός, Επίκουρος Καθηγητής Υπολογιστικής και Δομικής Βιολογίας, Δ.Π.Θ.

*με γνωστικό αντικείμενο “Υπολογιστική και Δομική Βιολογία”*

Ραφαήλ Σανδάλτζόπουλος, Αναπληρωτής Καθηγητής Μοριακής Βιολογίας, Δ.Π.Θ.

*με γνωστικό αντικείμενο “Μοριακή Βιολογία”*

Αγλαΐα Παππά, Επίκουρος Καθηγήτρια Φυσιολογίας και Μοριακής Φαρμακολογίας, Δ.Π.Θ.

*με γνωστικό αντικείμενο “Φυσιολογία Οργανισμών με έμφαση στους Μοριακούς Μηχανισμούς Δράσης Φαρμάκων”*

Ιωάννης Καραφυλλίδης, Καθηγητής Ηλεκτρονικής και Τεχνολογίας Συστημάτων Πληροφορικής, Δ.Π.Θ.

*με γνωστικό αντικείμενο “Σχεδιασμός, Μοντελοποίηση και Προσομοίωση Μικροηλεκτρονικών και Νανοηλεκτρονικών Στοιχείων Ολοκληρωμένων Κυκλωμάτων και Συστημάτων”*

Αναστασία Πολίτου, Επίκουρος Καθηγήτρια Βιολογικής Χημείας, Παν/μιο Ιωαννίνων

*με γνωστικό αντικείμενο “Βιολογική Χημεία”*

Γεώργιος Συρακούλης, Επίκουρος Καθηγητής Ηλεκτρονικής και Τεχνολογίας Συστημάτων Πληροφορικής, Δ.Π.Θ.

*με γνωστικό αντικείμενο “Ηλεκτρονικά Συστήματα με έμφαση σε Εργαλεία Αυτοματοποίησης Σχεδιασμού τους”*

Αθανάσιος Σταυρακούδης, Λέκτορας Υπολογιστικών Προσομοιώσεων, Παν/μιο Ιωαννίνων

*με γνωστικό αντικείμενο “Υπολογιστική Προσομοίωση με έμφαση στις Οικονομικές και Οικονομετρικές Εφαρμογές”*

## ΠΕΡΙΕΧΟΜΕΝΑ

<b>Κεφάλαιο 1 ΕΙΣΑΓΩΓΗ</b>	<b>1</b>
1.1 Το πρόβλημα της αναδίπλωσης των πρωτεϊνών	2
1.2 Αναδίπλωση πεπτιδίων	14
1.3 Σκοπός της παρούσας μελέτης	21
<b>Κεφάλαιο 2 ΥΠΟΛΟΓΙΣΤΙΚΕΣ ΜΕΘΟΔΟΙ</b>	<b>22</b>
2.1 Πρωτόκολλο προσομοίωσης	23
2.2 Αυτοματοποίηση της μεθόδου μέσω ενός Perl script	25
2.3 Συναρτήσεις εκτίμησης της αναδιπλωσιμότητας	33
2.4 Μέθοδοι ανάλυσης των τροχιακών	60
<b>Κεφάλαιο 3 ΤΕΤΡΑΠΕΠΤΙΔΙΑ</b>	<b>66</b>
3.1 Επιλογή τετραπεπτιδικών αλληλουχιών	67
3.2 Σχεδιασμός, αριθμός και διάρκεια προσομοιώσεων	70
3.3 Επιλογή 130 υποψήφιων δυνητικά αναδιπλούμενων τετραπεπτιδίων	76
3.4 Επιλογή 36 υποψήφιων δυνητικά αναδιπλούμενων τετραπεπτιδίων	86
3.5 Επιλογή 4 υποψήφιων δυνητικά αναδιπλούμενων τετραπεπτιδίων	91
3.6 Μελέτη της αναδίπλωσης των RWPD, DTRW, RPWD, EVKW σε τέσσερις θερμοκρασίες	110
3.7 Μελέτη της αναδίπλωσης των RWPD, DTRW με τρία force fields	152
<b>Κεφάλαιο 4 ΠΕΝΤΑΠΕΠΤΙΔΙΑ</b>	<b>174</b>
4.1 Επιλογή πενταπεπτιδικών αλληλουχιών	175
4.2 Σχεδιασμός, αριθμός και διάρκεια προσομοιώσεων	179
4.3 Επιλογή 480 υποψήφιων δυνητικά αναδιπλούμενων πενταπεπτιδίων	183
4.4 Επιλογή 32 υποψήφιων δυνητικά αναδιπλούμενων πενταπεπτιδίων	188
4.5 Μελέτη 32 πενταπεπτιδίων με τέσσερα force fields	192
4.6 Μελέτη της αναδίπλωσης 8 πενταπεπτιδίων με το AMBER99SB-ILDN force field	199
4.7 Μελέτη της επίδρασης της θερμοκρασίας στην αναδίπλωση των RDKWP και NEWRD	210
4.8 Μελέτη της cis/trans ισομερείωσης του πεπτιδικού δεσμού της προλίνης στο RDKWP μέσω adaptive tempering	220
<b>Κεφάλαιο 5 ΣΥΖΗΤΗΣΗ</b>	<b>230</b>
<b>Κεφάλαιο 6 ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ</b>	<b>238</b>
<b>Κεφάλαιο 7 ΠΑΡΑΡΤΗΜΑ</b>	<b>264</b>



**ΛΙΣΤΑ ΕΙΚΟΝΩΝ**

Εικόνα 2.1	Εξέλιξη στο χρόνο των αποστάσεων μεταξύ όλων των πιθανών ζευγών ατόμων Ca που δεν συνδέονται μέσω πεπτιδικού δεσμού ενός τετραπεπτιδίου (WEMK)	35
Εικόνα 2.2	Εξέλιξη στο χρόνο των αποστάσεων μεταξύ όλων των πιθανών ζευγών ατόμων Ca που δεν συνδέονται μέσω πεπτιδικού δεσμού ενός τετραπεπτιδίου (SKWD)	36
Εικόνα 2.3	Εξέλιξη στο χρόνο των αποστάσεων μεταξύ όλων των πιθανών ζευγών ατόμων Ca που δεν συνδέονται μέσω πεπτιδικού δεσμού ενός τετραπεπτιδίου (RDWP)	37
Εικόνα 2.4	Εξέλιξη στο χρόνο των αποστάσεων μεταξύ όλων των πιθανών ζευγών ατόμων Ca που δεν συνδέονται μέσω πεπτιδικού δεσμού ενός τετραπεπτιδίου (DRNW)	38
Εικόνα 2.5	Εξέλιξη στο χρόνο των ατομικών αποστάσεων των “γρήγορων αναδιπλωτών”	44
Εικόνα 2.6	Εξέλιξη στο χρόνο των ατομικών αποστάσεων των “αργών αναδιπλωτών”	45
Εικόνα 2.7	Εξέλιξη στο χρόνο των ατομικών αποστάσεων των πεπτιδίων που περιέχουν προλίνη	46
Εικόνα 2.8	Γραφική απεικόνιση δισδιάστατων πινάκων RMSD μεταξύ όλων των πιθανών δομών του τροχιακού ενός τετραπεπτιδίου (WEMK)	48
Εικόνα 2.9	Γραφική απεικόνιση δισδιάστατων πινάκων RMSD μεταξύ όλων των πιθανών δομών του τροχιακού ενός τετραπεπτιδίου (SKWD)	50
Εικόνα 2.10	Γραφική απεικόνιση δισδιάστατων πινάκων RMSD μεταξύ όλων των πιθανών δομών του τροχιακού ενός τετραπεπτιδίου (RDWP)	51
Εικόνα 2.11	Γραφική απεικόνιση δισδιάστατων πινάκων RMSD μεταξύ όλων των πιθανών δομών του τροχιακού ενός τετραπεπτιδίου (DRNW)	52
Εικόνα 2.12	Αντιπροσωπευτικές γραφικές παραστάσεις των κατανομών που προκύπτουν από την εφαρμογή του αλγόριθμου των “επεκτεινομένων παραθύρων”	54
Εικόνα 2.13	Δενδρογράμμα των κατανομών των πεπτιδίων, χρησιμοποιώντας τους γραμμικούς συντελεστές συσχέτισης του Πίνακα 2.2	56
Εικόνα 2.14	Γραφική απεικόνιση των πινάκων RMSD των 130 τετραπεπτιδίων του δεύτερου κύκλου των προσομοιώσεων με βάση τον αλγόριθμο των “επεκτεινομένων παραθύρων”	57
Εικόνα 3.1	Γραφική απεικόνιση των αμινοξέων των οποίων επιβάλλαμε την παρουσία κατά τον σχεδιασμό των τετραπεπτιδικών αλληλουχιών	68
Εικόνα 3.2	Word-cloud των 1440 τετραπεπτιδίων με βάση τη συνάρτηση TF1	77
Εικόνα 3.3	Ιστογράμματα κατανομής των βαθμολογιών με βάση τις συναρτήσεις TF1 και TF2 για τα 1440 τετραπεπτιδία	78
Εικόνα 3.4	Γραφική απεικόνιση των τετράγωνων συμμετρικών πινάκων των διαφορών των βαθμολογιών των 1440 τετραπεπτιδίων με βάση τις συναρτήσεις TF1 και TF2	80
Εικόνα 3.5	Cluster analysis των βαθμολογιών των πεπτιδίων όπου οι απόλυτες διαφορές των βαθμολογιών αντιμετωπίζονται ως αποστάσεις για την κατασκευή δενδρογραμμάτων	81
Εικόνα 3.6	Διάγραμμα Venn των πεπτιδίων που ανήκουν στους αναδιπλωτές βάσει cluster analysis	82
Εικόνα 3.7	Διάγραμμα Venn των 130 τετραπεπτιδίων που επιλέχθηκαν από τον πρώτο κύκλο προσομοιώσεων	84
Εικόνα 3.8	Κατάταξη των 130 τετραπεπτιδίων με βάση τη βαθμολογία της συνάρτησης TF3	88
Εικόνα 3.9	Κατάταξη των πινάκων RMSD για τα 130 τετραπεπτιδία με βάση τον αλγόριθμο των “επεκτεινομένων παραθύρων”	89
Εικόνα 3.10	Γραφική απεικόνιση των πινάκων RMSD των 15 τετραπεπτιδίων που ορίσαμε ως test data-set	90
Εικόνα 3.11	Γραφικές παραστάσεις των κατανομών που προκύπτουν από τον αλγόριθμο των “επεκτεινομένων παραθύρων” για τα 36 τετραπεπτιδία του τρίτου κύκλου προσομοιώσεων	94
Εικόνα 3.12	Συνοπτική γραφική αναπαράσταση των πινάκων RMSD των 36 τετραπεπτιδίων	96
Εικόνα 3.13	Συνοπτική παρουσίαση των ενεργειακών τοπίων από την ανάλυση Dihedral-PCA των 36 τετραπεπτιδίων	98

Εικόνα 3.14	Συνοπτική παρουσίαση των ενεργειακών τοπίων από την ανάλυση Cartesian-PCA των 36 τετραπεπτιδίων	100
Εικόνα 3.15	Εντροπία κατά Shannon της κατανομής των τριών κυρίαρχων principal components από την ανάλυση PCA των 36 τετραπεπτιδίων	102
Εικόνα 3.16	Συνοπτική παρουσίαση των δισδιάστατων κατανομών της εξέλιξης της γυρεοσκοπικής ακτίνας ως προς την εξέλιξη της απόστασης μεταξύ του N- τελικού και του C- τελικού άκρου για τα 36 τετραπεπτιδία	103
Εικόνα 3.17	Συνοπτική παρουσίαση αντιπροσωπευτικών δομών (σε υπέρθεση) του κυρίαρχου cluster, όπως προέκυψε από την ανάλυση Cartesian-PCA για τα 36 τετραπεπτιδία.	105
Εικόνα 3.18	Κατάταξη των 36 τετραπεπτιδίων με βάση τη βαθμολογία της συνάρτησης (4)	106
Εικόνα 3.19	Γραφική απεικόνιση των πινάκων RMSD των προσομοιώσεων διάρκειας 100ns των 15 τετραπεπτιδίων που ορίσαμε ως test data-set	107
Εικόνα 3.20	Διάγραμμα ροής της πορείας που ακολουθήσαμε, ξεκινώντας από το σύνολο των 1.440 τετραπεπτιδίων για να καταλήξουμε στα τέσσερα δυνητικά αναδιπλούμενα τετραπεπτιδία	109
Εικόνα 3.21	Γραφική απεικόνιση του ενιαίου πίνακα RMSD μετά την ένωση των τροχιακών των τεσσάρων θερμοκρασιών του πεπτιδίου RWPD	112
Εικόνα 3.22	Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RWPD για το τροχιακό των 283K	113
Εικόνα 3.23	Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RWPD για το τροχιακό των 298K	114
Εικόνα 3.24	Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RWPD για το τροχιακό των 320K	114
Εικόνα 3.25	Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RWPD για το τροχιακό των 340K	115
Εικόνα 3.26	Αντιπροσωπευτικές δομές (σε stereo αναπαράσταση) από τα cluster των τροχιακών των τεσσάρων θερμοκρασιών για το πεπτιδίο RWPD	117/118
Εικόνα 3.27	Γραφική απεικόνιση του ενιαίου πίνακα RMSD μετά την ένωση των τροχιακών των τεσσάρων θερμοκρασιών του πεπτιδίου RPWD	119
Εικόνα 3.28	Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RPWD για το τροχιακό των 283K	120
Εικόνα 3.29	Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RPWD για το τροχιακό των 298K	120
Εικόνα 3.30	Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RPWD για το τροχιακό των 320K	121
Εικόνα 3.31	Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RPWD για το τροχιακό των 340K	121
Εικόνα 3.32	Αντιπροσωπευτικές δομές (σε stereo αναπαράσταση) από τα cluster των τροχιακών των τεσσάρων θερμοκρασιών για το πεπτιδίο RPWD	122/123
Εικόνα 3.33	Γραφική απεικόνιση του ενιαίου πίνακα RMSD μετά την ένωση των τροχιακών των τεσσάρων θερμοκρασιών του πεπτιδίου DTRW	125
Εικόνα 3.34	Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου DTRW για το τροχιακό των 283K	126
Εικόνα 3.35	Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου DTRW για το τροχιακό των 298K	126
Εικόνα 3.36	Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου DTRW για το τροχιακό των 320K	127
Εικόνα 3.37	Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου DTRW για το τροχιακό των 340K	127
Εικόνα 3.38	Αντιπροσωπευτικές δομές (σε stereo αναπαράσταση) από τα cluster των τροχιακών των τεσσάρων θερμοκρασιών για το πεπτιδίο DTRW	128/129

Εικόνα 3.39	Γραφική απεικόνιση του ενιαίου πίνακα RMSD μετά την ένωση των τροχιακών των τεσσάρων θερμοκρασιών του πεπτιδίου EVKW	130
Εικόνα 3.40	Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου EVKW για το τροχιακό των 283K	131
Εικόνα 3.41	Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου EVKW για το τροχιακό των 298K	131
Εικόνα 3.42	Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου EVKW για το τροχιακό των 320K	132
Εικόνα 3.43	Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου EVKW για το τροχιακό των 340K	132
Εικόνα 3.44	Αντιπροσωπευτικές δομές (σε stereo αναπαράσταση) από τα cluster των τροχιακών των τεσσάρων θερμοκρασιών για το πεπτίδιο EVKW	133/134
Εικόνα 3.45	Κατανομές των δίεδρων γωνιών για το πεπτίδιο RWPD για τα τροχιακά των τεσσάρων θερμοκρασιών	137
Εικόνα 3.46	Κατανομές των δίεδρων γωνιών για το πεπτίδιο RPWD για τα τροχιακά των τεσσάρων θερμοκρασιών	138
Εικόνα 3.47	Κατανομές των δίεδρων γωνιών για το πεπτίδιο DTRW για τα τροχιακά των τεσσάρων θερμοκρασιών	139
Εικόνα 3.48	Κατανομές των δίεδρων γωνιών για το πεπτίδιο EVKW για τα τροχιακά των τεσσάρων θερμοκρασιών	140
Εικόνα 3.49	Χαρακτηριστική κίνηση που περιγράφεται από τον eigenvector με την υψηλότερη τιμή eigenvalue της ανάλυσης Cartesian-PCA	141
Εικόνα 3.50	Κατανομές των διαφορών πληθυσμών από διαμορφώσεις για τα τέσσερα πεπτίδια ως συνάρτηση της απόστασης μεταξύ των άκρων	142
Εικόνα 3.51	Ιστογράμματα κατανομής των RMSD όλων των δομών κάθε τροχιακού από την αντιπροσωπευτική δομή του κυρίαρχου cluster	145
Εικόνα 3.52	Εκτιμώμενη ελεύθερη ενέργεια αναδίπλωσης (KJ/mol) ως συνάρτηση της θερμοκρασίας (K) διεξαγωγής της προσομοίωσης	146
Εικόνα 3.53	Εξέλιξη στο χρόνο της προσομοίωσης των διαφορών τύπων β-στροφών	149
Εικόνα 3.54	Γραφική απεικόνιση του ενιαίου πίνακα RMSD μετά την ένωση των τροχιακών των 3 force fields για τα πεπτίδια RWPD και DTRW	155
Εικόνα 3.55	Γραφική απεικόνιση των Πινάκων 3.7-3.8 των αντιπροσωπευτικών δομών για το πεπτίδιο RWPD και τα αντίστοιχα δενδρογράμματα	157
Εικόνα 3.56	Συγκεντρωτικά αποτελέσματα για το πεπτίδιο RWPD για το AMBER force field	159
Εικόνα 3.57	Συγκεντρωτικά αποτελέσματα για το πεπτίδιο RWPD για το CHARMM force field	160
Εικόνα 3.58	Συγκεντρωτικά αποτελέσματα για το πεπτίδιο RWPD για το OPLS force field	161
Εικόνα 3.59	Αντιπροσωπευτικές δομές (σε stereo αναπαράσταση) από τα κυρίαρχα cluster των τροχιακών των δύο τετραπεπτιδίων RWPD και DTRW για τα τρία force fields	162
Εικόνα 3.60	Γραφική απεικόνιση των Πινάκων 3.9-3.10 των αντιπροσωπευτικών δομών για το πεπτίδιο DTRW και τα αντίστοιχα δενδρογράμματα	165
Εικόνα 3.61	Συγκεντρωτικά αποτελέσματα για το πεπτίδιο DTRW για το AMBER force field	166
Εικόνα 3.62	Συγκεντρωτικά αποτελέσματα για το πεπτίδιο DTRW για το CHARMM force field	167
Εικόνα 3.63	Συγκεντρωτικά αποτελέσματα για το πεπτίδιο DTRW για το OPLS force field	169
Εικόνα 3.64	Eigenspace overlap μεταξύ του πρώτου και δεύτερου μη επικαλυπτόμενου μισού κάθε ανεξάρτητης προσομοίωσης με τα τρία force fields για τα πεπτίδια RWPD και DTRW από την ανάλυση PCA	171
Εικόνα 3.65	Eigenspace overlap μεταξύ των force fields ως συνάρτηση των 50 eigenvalues με τα υψηλότερα eigenvalues από την ανάλυση Cartesian-PCA για τα πεπτίδια RWPD και DTRW	171
Εικόνα 3.66	Κατανομή των τιμών RMSD των δομών ολόκληρου του τροχιακού από την αντιπροσωπευτική δομή του κυρίαρχου cluster όπως ορίστηκε μέσω της ανάλυσης Cartesian-PCA	173

Εικόνα 4.1	Word-cloud των 7.200 πενταπεπτιδίων, όπου το μέγεθος της αλληλουχίας είναι ενδεικτικό της βαθμολογίας που έλαβε με βάση τη συνάρτηση TF2	184
Εικόνα 4.2	Κατανομή της βαθμολογίας των 7.200 πενταπεπτιδίων με βάση τη συνάρτηση TF2	186
Εικόνα 4.3	Cluster analysis των βαθμολογιών των πεπτιδίων της Εικόνας 4.1, όπου οι απόλυτες διαφορές των βαθμολογιών αντιμετωπίζονται ως αποστάσεις για την κατασκευή δενδρογράμματος	186
Εικόνα 4.4	Κατάταξη των πινάκων RMSD για τα 480 πενταπεπτίδια με βάση τον αλγόριθμο των “επεκτεινομένων παραθύρων”	189
Εικόνα 4.5	Κατάταξη των πινάκων RMSD για τα 480 πενταπεπτίδια με βάση τη συνάρτηση TF3	190
Εικόνα 4.6	Διάγραμμα Venn των 32 πενταπεπτιδίων που επιλέχθηκαν για τον επόμενο κύκλο προσομοιώσεων	191
Εικόνα 4.7	Συνοπτική γραφική αναπαράσταση των πινάκων RMSD των 32 πενταπεπτιδίων και για τα τέσσερα force fields	194
Εικόνα 4.8	Word-cloud των 8 πενταπεπτιδίων, όπου το μέγεθος της αλληλουχίας είναι ενδεικτικό της συμφωνίας (η ασυμφωνίας) στις προβλέψεις των τεσσάρων force fields για κάθε πεπτίδιο	196
Εικόνα 4.9	Σύγκριση των τεσσάρων force fields των 16 πενταπεπτιδίων της λίστας top16A με το CHARMM27, με το οποίο διεξήχθησαν οι προσομοιώσεις στο στάδιο των top480A	197
Εικόνα 4.10	Συγκεντρωτικά αποτελέσματα για το πεπτίδιο DPWRE από προσομοίωση διάρκειας 1μς με το force field AMBER99SB-ILDN	200
Εικόνα 4.11	Συγκεντρωτικά αποτελέσματα για το πεπτίδιο ECKRW από προσομοίωση διάρκειας 1μς με το force field AMBER99SB-ILDN	201
Εικόνα 4.12	Συγκεντρωτικά αποτελέσματα για το πεπτίδιο ELRKW από προσομοίωση διάρκειας 1μς με το force field AMBER99SB-ILDN	202
Εικόνα 4.13	Συγκεντρωτικά αποτελέσματα για το πεπτίδιο NEWRD από προσομοίωση διάρκειας 1μς με το force field AMBER99SB-ILDN	203
Εικόνα 4.14	Συγκεντρωτικά αποτελέσματα για το πεπτίδιο RDKWP από προσομοίωση διάρκειας 1μς με το force field AMBER99SB-ILDN	205
Εικόνα 4.15	Συγκεντρωτικά αποτελέσματα για το πεπτίδιο RELWK από προσομοίωση διάρκειας 1μς με το force field AMBER99SB-ILDN	206
Εικόνα 4.16	Συγκεντρωτικά αποτελέσματα για το πεπτίδιο REWDV από προσομοίωση διάρκειας 1μς με το force field AMBER99SB-ILDN	207
Εικόνα 4.17	Συγκεντρωτικά αποτελέσματα για το πεπτίδιο REWID από προσομοίωση διάρκειας 1μς με το force field AMBER99SB-ILDN	208
Εικόνα 4.18	Γραφική απεικόνιση του ενιαίου πίνακα RMSD μετά την ένωση των τροχιακών των 2 και 4 θερμοκρασιών του πεπτιδίου NEWRD και RDKWP	211
Εικόνα 4.19	Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου NEWRD για το τροχιακό των 298K	212
Εικόνα 4.20	Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου NEWRD για το τροχιακό των 320K	213
Εικόνα 4.21	Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RDKWP για το τροχιακό των 298K	214
Εικόνα 4.22	Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RDKWP για το τροχιακό των 320K	215
Εικόνα 4.23	Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RDKWP για το τροχιακό των 340K	216
Εικόνα 4.24	Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RDKWP για το τροχιακό των 360K	217
Εικόνα 4.25	Γραφική απεικόνιση (σε stereo αναπαράσταση) των 2 διακριτών δομών του πεπτιδίου RDKWP	218
Εικόνα 4.26	Γραφική απεικόνιση του ενιαίου πίνακα RMSD των δύο τροχιακών trans και cis του πεπτιδίου RDKWP	223

Εικόνα 4.27	Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RDKWP ξεκινώντας από <i>trans</i> διαμόρφωση	225
Εικόνα 4.28	Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RDKWP ξεκινώντας από <i>cis</i> διαμόρφωση	226
Εικόνα 4.29	Διαγράμματα Ramachadran για τα εσωτερικά κατάλοιπα (2-4) του πεπτιδίου RDKWP για το <i>trans</i> και <i>cis</i> τροχιακό	227
Εικόνα 4.30	Διαγράμματα Ramachadran για τα cluster 1 και 2 που προέκυψαν από την ανάλυση Dihedral-PCA για το <i>trans</i> τροχιακό	228
Εικόνα 4.31	Διαγράμματα Ramachadran για τα cluster 1 και 2 που προέκυψαν από την ανάλυση Dihedral-PCA για το <i>cis</i> τροχιακό	229
Εικόνα 5.1	Κατανομές των μέσων βαθμολογιών και rmsd (error-bars) για κάθε αμινοξύ και κάθε θέση στην αλληλουχία για το σύνολο των 5.760 προσομοιώσεων των τετραπεπτιδίων	233
Εικόνα 5.2	Κατανομές των μέσων βαθμολογιών και rmsd (error-bars) για κάθε αμινοξύ και κάθε θέση στην αλληλουχία για το σύνολο των 130 τετραπεπτιδίων	234

**ΛΙΣΤΑ ΠΙΝΑΚΩΝ**

Πίνακας 2.1	Συνοπτικός πίνακας στατιστικών μέτρων χαρακτηριστικών των κατανομών των βαθμολογιών των ατομικών αποστάσεων με βάση τις συναρτήσεις TF1 και TF2	43
Πίνακας 2.2	Γραμμικοί συντελεστές συσχέτισης της κατάταξης των πεπτιδίων κατά την ανάπτυξη του αλγόριθμου των “επεκτεινομένων παραθύρων”	55
Πίνακας 2.3	Παράμετροι που εξετάσθηκαν ως δυνητικοί εκτιμητές της αναδιπλωσιμότητας των πεπτιδίων της παρούσας μελέτης	58
Πίνακας 3.1	Αριθμός πιθανών τετραπεπτιδικών αλληλουχιών και περιοριστικές παράμετροι στην επιλογή αμινοξικών καταλοίπων	69
Πίνακας 3.2	Συνοπτικός πίνακας των δοκιμαστικών προσομοιώσεων στο πεπτίδιο-μοντέλο RWTDDQ	73
Πίνακας 3.3	Συγκεντρωτικός πίνακας των προσομοιώσεων που πραγματοποιήσαμε στο σύνολο των 1.440 τετραπεπτιδίων	74
Πίνακας 3.4	Συνοπτικός πίνακας στατιστικών μέτρων χαρακτηριστικών των κατανομών των βαθμολογιών της Εικόνας 3.3	79
Πίνακας 3.5	Μέση τιμή των ατομικών διακυμάνσεων (RMSFs) για τέσσερα σύνολα ατόμων για τα τέσσερα τετραπεπτίδια και για τα τροχιακά των τεσσάρων θερμοκρασιών	116
Πίνακας 3.6	Αμινοξέα με υψηλή συχνότητα παρατήρησης σε μοτίβα β-στροφής στην PDB και όπως προσδιορίστηκαν από τις συναρτήσεις εκτίμησης της αναδιπλωσιμότητας μέσω των προσομοιώσεων μοριακής δυναμικής που πραγματοποιήσαμε	150
Πίνακας 3.7	Τιμές RMSD μεταξύ όλων των αντιπροσωπευτικών δομών όλων των cluster όπως προκύπτουν από την ανάλυση Cartesian-PCA και Dihedral-PCA, χρησιμοποιώντας τα άτομα του πεπτιδικού σκελετού, για το πεπτίδιο RWPD και για τα τρία force fields	156
Πίνακας 3.8	Τιμές RMSD μεταξύ όλων των αντιπροσωπευτικών δομών όλων των cluster όπως προκύπτουν από την ανάλυση Cartesian-PCA και Dihedral-PCA, χρησιμοποιώντας όλα τα βαριά άτομα, για το πεπτίδιο RWPD και για τα τρία force fields	156
Πίνακας 3.9	Τιμές RMSD μεταξύ όλων των αντιπροσωπευτικών δομών όλων των cluster όπως προκύπτουν από την ανάλυση Cartesian-PCA και Dihedral-PCA, χρησιμοποιώντας τα άτομα του πεπτιδικού σκελετού, για το πεπτίδιο DTRW και για τα τρία force fields	164
Πίνακας 3.10	Τιμές RMSD μεταξύ όλων των αντιπροσωπευτικών δομών όλων των cluster όπως προκύπτουν από την ανάλυση Cartesian-PCA και Dihedral-PCA, χρησιμοποιώντας όλα τα βαριά άτομα, για το πεπτίδιο DTRW και για τα τρία force fields	164
Πίνακας 4.1	Αριθμός πιθανών πενταπεπτιδικών αλληλουχιών και περιοριστικές παράμετροι	177
Πίνακας 4.2	Συγκεντρωτικός πίνακας των προσομοιώσεων που πραγματοποιήσαμε στο σύνολο των 7.200 πενταπεπτιδίων.	180

## ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω θερμά τον μέντορα και δάσκαλό μου, Δρ Νικόλαο Μ. Γλυκό, για τις ατέλειωτες ώρες που αφιέρωσε προσπαθώντας να μου μεταλαμπαδεύσει τον ενθουσιασμό του για την επιστήμη και την έρευνα. Ήταν δίπλα μου, αδιάκοπα, σε όλα τα χρόνια των προπτυχιακών και μεταπτυχιακών μου σπουδών με κύριο μέλημα να μου διδάξει τον κριτικό τρόπο σκέψης.

Ευχαριστώ ιδιαίτερος τα μέλη της τριμελούς συμβουλευτικής επιτροπής, Δρ. Σανδαλτζόπουλο και Δρ. Παππά, που μαζί με τον Δρ. Γλυκό, με στήριξαν με όλα τα μέσα και ήταν δίπλα μου στο εγχείρημα αυτό.

Ευχαριστώ τους Δρ Καραφυλλίδη, Δρ. Πολίτου, Δρ. Συρακούλη, Δρ. Σταυρακούδη για τη συμμετοχή τους στην επταμελή εξεταστική επιτροπή, τα πολύτιμα σχόλια τους και τις εύστοχες παρατηρήσεις τους.

Ευχαριστώ όλα τα μέλη του εργαστηρίου Δομικής και Υπολογιστικής Βιολογίας αλλά και των γειτονικών εργαστηρίων για το ευχάριστο και φιλικό περιβάλλον κατά τη συστέγασή μας.

Ευχαριστώ το Τμήμα Μοριακής Βιολογίας και Γενετικής που μου έδωσε την ευκαιρία και την υλικοτεχνική υποστήριξη να καταρτιστώ και να εκπαιδευτώ ως επιστήμονας.

Τέλος, να εκφράσω την ευγνωμοσύνη μου στην οικογένεια μου για την αμέριστη ηθική και οικονομική υποστήριξη κατά τις πολυετείς σπουδές μου.

## ΠΕΡΙΛΗΨΗ

Το πρόβλημα της αναδίπλωσης των πρωτεϊνών αποτελεί το Άγιο Δισκοπότηρο για τους σύγχρονους Βιοχημικούς και Δομικούς Βιολόγους. Ένας ιδιαίτερα ελκυστικός υποψήφιος για τη μελέτη της αναδίπλωσης είναι τα πεπτιδία λόγω του μικρού τους μεγέθους, το ρόλο που θεωρείται ότι διαδραματίζουν κατά την έναρξη της αναδίπλωσης και τη δυνατότητα που προσφέρουν για άμεση σύγκριση μεταξύ θεωρίας και πειράματος. Η παρούσα διατριβή εστιάζεται στον ορθολογικό σχεδιασμό (πεπτιδική μηχανική) μικρών πεπτιδίων με σταθερή δομή και επιθυμητές μοριακές ιδιότητες με προφανείς φαρμακολογικές εφαρμογές. Ο πρωταρχικός στόχος είναι η εύρεση πεπτιδίων μικρού μοριακού βάρους που μπορούν να υιοθετήσουν σταθερή δομή σε υδατικά διαλύματα. Αρχικά γίνεται *ab initio* σχεδιασμός των αλληλουχιών ενώ όλη η διαδικασία γίνεται με συστηματικό τρόπο μέσω προγραμμάτων. Στην επόμενη φάση οι πεπτιδικές αλληλουχίες αναδιπλώνονται και χαρακτηρίζονται δομικά *in silico* μέσω προσομοιώσεων μοριακής δυναμικής. Όλες οι προσομοιώσεις πραγματοποιούνται με αναλυτική παρουσία του διαλύτη, συνθήκες περιοδικής οριοθέτησης και πλήρη υπολογισμό των ηλεκτροστατικών αλληλεπιδράσεων χρησιμοποιώντας state-of-the-art αλγόριθμους. Η εκτίμηση της αναδιπλωσιμότητας των πεπτιδίων γίνεται μέσω συναρτήσεων χρησιμοποιώντας μεταξύ άλλων ατομικές αποστάσεις, ατομικές διακυμάνσεις και πίνακες RMSD μεταξύ διαδοχικών δομών του τροχιακού. Η χρονική διάρκεια των προσομοιώσεων επιλέγεται κατάλληλα ώστε να εγγυάται σε κάθε κύκλο των αποκλεισμό των ασταθών πεπτιδίων, ενώ αυτά που λαμβάνουν υψηλή βαθμολογία μελετώνται περαιτέρω με προσομοιώσεις μεγαλύτερης διάρκειας και διαφορετικά force fields. Η πορεία που ακολουθήσαμε οδήγησε στην ταυτοποίηση ενός μικρού αριθμού τετραπεπτιδίων και πενταπεπτιδίων με ισχυρή αναδιπλωσιμότητα. Τα τετραπεπτιδία είναι περισσότερο ασταθή και με συχνότερα γεγονότα αναδίπλωσης/αποδιάταξης, ενώ δύο πενταπεπτιδία σχηματίζουν σταθερή δομή για σημαντικό ποσοστό του χρόνου προσομοίωσης και για τα συγκεκριμένα force fields. Περαιτέρω πειραματικές πλέον μελέτες αναμένεται να δείξουν την εγκυρότητα ή μη των προβλέψεών μας, αναδεικνύοντας την ικανότητα των σύγχρονων υπολογιστικών εργαλείων αλλά και το περιθώριο βελτίωσής τους.



## **ABSTRACT**

A computational solution to the protein folding problem has evolved to the Holy Grail of modern Biochemistry and Structural Biology. Peptides have been an attractive candidate for such studies due to their small size and the opportunity that they offer to directly compare theory with experiment. The present thesis is focused on developing a novel method for the rational design and engineering of small peptides with desired molecular properties, which has direct implications in drug discovery. The primary aim is to identify sufficiently small peptides that can still form stable structures in aqueous solutions. The first step includes *ab initio* design of peptides using automated methods to select peptide sequences and is entirely implemented through programs written in C and Perl. The second step includes the *in silico* folding and structure characterization using molecular dynamics simulations. The simulations are performed using explicit representation of the solute and state-of-the-art algorithms for the production of NpT runs with periodic boundary conditions and full treatment of the electrostatics. The evaluation and sorting of peptides is achieved through a number of scoring functions, that include terms based on interatomic vector distances, atomic fluctuations and RMSD matrices between successive frames of a trajectory. The length of the simulations is such that can guarantee the successful exclusion of unfolded peptides. Highly-scored peptides are studied further through longer simulations and using different empirical force fields. Our method concluded to only a handful of tetra- and pentapeptides that have a good prognosis for being foldable. The tetrapeptides are more flexible and disordered whilst we were able to point out two pentapeptides with strong foldability prognosis. Experimental validation of our computational predictions shall lend support to our methodology or shall reveal the deficiencies of the current generation theoretical tools, that can only lead to their improvement.

We shall not cease from exploration,  
and the end of all our exploring  
will be to arrive where we started  
and know the place for the first time.

Thomas Stearns Eliot 1898-1965



# Κεφάλαιο 1

# ΕΙΣΑΓΩΓΗ



*"The most beautiful and most profound emotion we can experience is the sensation of the mystical. It is the sower of all true science. He to whom this emotion is a stranger, who can no longer stand rapt in awe, is as good as dead. That deeply emotional conviction of the presence of a superior reasoning power, which is revealed in the incomprehensible Universe, forms my idea of God."*  
*Anfinsen quoting Einstein*



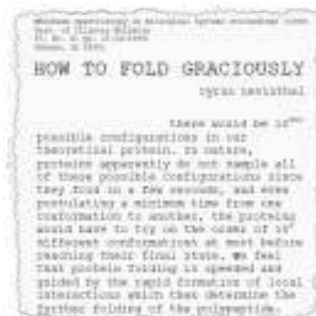
## 1.1 Το πρόβλημα της αναδίπλωσης των πρωτεϊνών

Το πρόβλημα της αναδίπλωσης των πρωτεϊνών θεωρείται από την επιστημονική κοινότητα ως το "Άγιο Δισκοπότηρο" της σύγχρονης Βιοχημείας και Δομικής Βιολογίας. Έχει αποδώσει το ένα τέταρτο των βραβείων Nobel Χημείας (Seringhaus et al., 2007) ενώ το 2005, το περιοδικό Science, το συμπεριέλαβε στη λίστα των 125 μεγαλύτερων άλυτων επιστημονικών προβλημάτων (Editorial, 2005). Από την εποχή των ευρημάτων των δύο ιστορικών προπατόρων μας, Christian B. Anfinsen και Cyrus Levinthal έχει σημειωθεί σημαντική έρευνα, χωρίς όμως την αναμενόμενη πρόοδο σε σχέση με τα θεμελιώδη ευρήματά τους:



*'Όλη η απαραίτητη πληροφορία για την αναδίπλωση μιας πρωτεΐνης στη φυσική δομή της εμπεριέχεται στην αμινοξική της αλληλουχία.*  
*(Anfinsen, 1973)*

*Εάν κατά την αναδίπλωσή της, μία πρωτεΐνη επισκεπτόταν όλες τις πιθανές διαμορφώσεις ( $10^{300}$  για αλυσίδα 100 αμινοξέων  $\rightarrow 1.6 \times 10^{52}$  years) δεν θα ήταν δυνατό να αναδιπλωθεί σε λίγα δευτερόλεπτα.*  
*(Levinthal, 1969)*



Η αναδίπλωση και αποδιάταξη των πρωτεϊνικών αλυσίδων, δύο θεμελιώδεις διαδικασίες, είναι γνωστές εδώ και 80 χρόνια (Mirsky & Pauling, 1936). Οι παράγοντες που ενορχηστρώνουν τη διαδικασία της αναδίπλωσης έχουν εδραιωθεί μέσα από πολύχρονες μελέτες και ανήκουν σε δύο ομάδες με αντίθετη δράση: οι μη ομοιοπολικές αλληλεπιδράσεις (ηλεκτροστατικές αλληλεπιδράσεις, δυνάμεις Van der Waals, δεσμοί υδρογόνου) συναγωνίζονται την απώλεια της εντροπίας διαμόρφωσης λόγω στερικών παρεμποδίσεων (Dill, 1990). Η τελική διαμόρφωση που θα πάρει μία πολυπεπτιδική αλυσίδα όμως δεν καθορίζεται μόνο από την ενέργεια (matter of stability) αλλά και από άλλους παράγοντες εξειδίκευσης της εκάστοτε διαμόρφωσης (matter of specificity) (Lattman et al., 1993).

Το πρόβλημα της αναδίπλωσης ανάγεται σε τρία αλληλένδετα ζητήματα: (1) Υπάρχει κώδικας αναδίπλωσης; (2) Ποιός είναι ο μηχανισμός της αναδίπλωσης; (3) Μπορούμε εμείς να προβλέψουμε τη φυσική (native) δομή από την αμινοξική αλληλουχία; (Dill et al., 2007, 2008).

Για την επίλυση του προβλήματος της αναδίπλωσης των πρωτεϊνών, έχουν διατυπωθεί πολυάριθμα μοντέλα. Σε μία ιστορική αναδρομή των διαφόρων αυτών μηχανισμών (Ivarsson et al., 2008), αναφέρουμε τους επικρατέστερους:



#### Το μοντέλο diffusion-collision

Τοπικά στοιχεία δευτεροταγούς δομής σχηματίζονται με βάση την πρωτοταγή δομή (framework model) αλλά ανεξάρτητα από την τριτοταγή. Στη συνέχεια τα στοιχεία αυτά διαχέονται μέχρι να συγκρουστούν οπότε και συναρμολογούνται, δημιουργώντας την τελική δομή. Το μοντέλο αυτό υποστηρίζεται από την πειραματική παρατήρηση της ταχείας δημιουργίας των στοιχείων δευτεροταγούς δομής και συνάδει με περιπτώσεις κινητικής two-state (Karplus et al., 1994, DeMarco et al., 2004). Τα ελικοειδή δεμάτια φαίνεται να ακολουθούν ένα μοντέλο hierarchical diffusion-collision, όπου η δευτεροταγής δομή σχηματίζεται πρώτη και η περαιτέρω συγκρότηση προχωρά με ιεραρχικό τρόπο (Myers et al., 2001), ή τμηματικό μέσω συγκρότησης δομικών υπομονάδων, των foldons, ή μέσω τοπομερών, μη-αναδιπλωμένων ενδιαμέσων με χαρακτηριστικά φυσικής τοπολογίας (Debe et al., 1999).



#### Το μοντέλο nucleation-condensation

Σύμφωνα με το μοντέλο αυτό ο σχηματισμός τριτοταγούς και δευτεροταγούς δομής είναι



συζευγμένος και ο σχηματισμός της μίας είναι τόσο συνέπεια όσο και αιτία του σχηματισμού της άλλης (Wetlaufer, 1973, 1990). Δημιουργείται ένας πυρήνας αναδίπλωσης γύρω από τον οποίο δημιουργείται το σύνολο δομών της μεταβατικής κατάστασης TSE (Transition State Ensemble) και ταυτόχρονα αρχίζει η δημιουργία της φυσικής δομής. Το μοντέλο αυτό υποστηρίζεται από θεωρητικές μελέτες και μελέτες πρωτεϊνικής μηχανικής και συνάδει με περιπτώσεις κινητικής three-state (Itzhaki et al., 1995, Fersht, 1997).

Μελέτες σε ένα σύνολο πρωτεϊνών με υποθετική two-state συμπεριφορά έδειξε ότι η κύρια διαφορά μεταξύ των two-state και multi-state είναι στη σχετική σταθερότητα των μερικώς αναδιπλωμένων ενδιάμεσων TSE (Sanchez et al., 2003). Έτσι, οι δύο αυτοί μηχανισμοί δεν είναι τελείως διακριτοί, αλλά ακραίες καταστάσεις ενός κοινού μηχανισμού, όπου η πρωτεΐνη ακολουθεί το ένα ή άλλο άκρο ανάλογα με την εγγενή σταθερότητα των στοιχείων δευτεροταγούς δομής (Gianni et al., 2003).



#### Το μοντέλο hydrophobic-collapse

Το μοντέλο αυτό χρησιμοποιείται για την περιγραφή των πρώιμων σταδίων της αναδίπλωσης και θεωρεί ότι η κινητήριος δύναμη κατά την αναδίπλωση είναι οι υδρόφοβες αλληλεπιδράσεις που είναι αποτέλεσμα της αλληλεπίδρασης μεταξύ των υδρόφοβων πλευρικών ομάδων των αμινοξέων και των υδρόφιλων μορίων του νερού. Δημιουργείται ο υδρόφοβος πυρήνας (η ενδιάμεση αυτή κατάσταση ονομάζεται molten globule) και μετά η πρωτεΐνη οργανώνεται γύρω από αυτόν. Ο μηχανισμός αυτός είναι συμβατός συνήθως με μικρές σφαιρικές πρωτεΐνες (Dill, 1985).



#### Το μοντέλο Zippering & Assembly (ZA)

Το γεγονός ότι ο κώδικας της αναδίπλωσης εκτείνεται τόσο τοπικά όσο και γενικότερα στην αλληλουχία, με αποτέλεσμα η δευτεροταγής δομή να είναι τόσο η αιτία όσο και το αποτέλεσμα της τριτοταγούς δομής (Dill, 1990), οδήγησε σε ένα νέο μηχανισμό αναδίπλωσης, zippering and assembly (ZA) (Dill, 1993). Σύμφωνα με αυτό το μηχανισμό, οι πρωτεΐνες λύνουν το παράδοξο του Levinthal μέσω ενός μηχανισμού “διαίρει και βασίλευε”: αντί η αναζήτηση να γίνεται σε καθολικό επίπεδο, γίνεται τοπικά σε επίπεδο μικρών πεπτιδικών τμημάτων, τα οποία επεκτείνονται και συναρμολογούνται για να δώσουν την τελική φυσική δομή (Dill et al., 2007, 2008, Ozkan et al., 2007).



### Το μοντέλο funneled energy-landscape

Όλα τα προηγούμενα μοντέλα έχουν τόσο κοινά στοιχεία όσο και διαφορές, και αναπτύχθηκαν ώστε να ταιριάζουν σε συγκεκριμένα πειραματικά ευρήματα, υποδεικνύοντας την έλλειψη ενός καθολικού μηχανισμού που να καλύπτει όλο το φάσμα των πρωτεϊνών (Daggett et al., 2003). Οι νεότερες απόψεις άρχισαν να τείνουν προς την ύπαρξη παράλληλων μικροσκοπικών διαδικασιών που μοιάζουν με διάχυση και περιλαμβάνουν πολλαπλά μονοπάτια (Caflisch, 2004). Έτσι η ιδέα του ενός μονοπατιού αντικαταστάθηκε από μία στατιστική περιγραφή της ενέργειας της πρωτεΐνης μέσω ενός ενεργειακού τοπίου (energy landscape) (Onuchic et al., 1997, 2004): η αναδίπλωση διοχετεύεται (funnel) προς μία σταθερή κατάσταση μέσα από πολλές οδούς διαμόρφωσης στο χώρο και μέσω ενός συνόλου μερικώς αναδιπλωμένων δομών, οι οποίες βρίσκονται σε υψηλή ενεργειακή στάθμη και ονομάζονται transition state ensemble, TSE (Onuchic et al., 1996). Έτσι η αναδίπλωση νοείται ως η μετάβαση από την αταξία στην τάξη και όχι από μία δομή σε μία άλλη (Dill et al., 1997).



Το ενεργειακό τοπίο προκύπτει από την ελεύθερη ενέργεια κάθε διαμόρφωσης συναρτήσει των βαθμών ελευθερίας. Ο κάθετος άξονας αντιπροσωπεύει την 'εσωτερική' ελεύθερη ενέργεια (το άθροισμα όλων των αλληλεπιδράσεων, δεσμοί υδρογόνου, υδρόφοβες αλληλεπιδράσεις, ενέργεια διαλυτοποίησης κ.τ.λ. πλην της εντροπίας διαμόρφωσης) και εξαρτάται από τη θερμοκρασία και τον περιβάλλοντα διαλύτη. Οι ποικίλοι πλευρικοί άξονες είναι οι συντεταγμένες των διαφόρων διαμορφώσεων. Κάθε μία από τις πιθανές διαμορφώσεις είναι ένα σημείο στο πολυδιάστατο ενεργειακό τοπίο, του οποίου η μορφή μοιάζει με βουνά και κοιλάδες: οι διαμορφώσεις με υψηλή ενέργεια βρίσκονται στις κορυφές και οι διαμορφώσεις με χαμηλή ενέργεια σε κοιλάδες. Κατά την έναρξη της αναδίπλωσης η πρωτεΐνη τείνει να επιλέγει τις διαμορφώσεις που θα μειώσουν την ενέργεια αλλά ταυτόχρονα πηγαίνει και προς άλλες διαμορφώσεις λόγω της κίνησης Brown (Onuchic et al. 1997). Ο funnel-like χαρακτήρας των διαγραμμάτων δίνει μία εσφαλμένη αντίληψη ότι η προοδευτική μείωση του αριθμού των διαμορφώσεων οδηγεί στη φυσική δομή, αλλά ο κύριος καθοριστής της ταχύτητας της αναδίπλωσης είναι το ενεργειακό τοπίο (landscape) το οποίο περιλαμβάνει τη δυναμική ενέργεια (ευνοϊκή) και την εντροπία διαμόρφωσης (μη-ευνοϊκή) οι οποίες μειώνονται κατά την πορεία προς τη φυσική δομή και η ευαίσθητη ισορροπία μεταξύ



των δύο οδηγεί στη δημιουργία ενεργειακού φράγματος (free-energy barrier) και στην κινητική συμπεριφορά των δύο σταδίων (two-state folding) (Karplus, 2011).

Μία συγκεκριμένη κατηγορία πρωτεϊνών (ultra-fast folders) που αναδιπλώνονται χωρίς την ύπαρξη ενεργειακού φράγματος λόγω ισχυρού bias προς τη φυσική δομή, χαρακτηρίζονται από 'downhill folding' και έχουν ιδιαίτερη σημασία λόγω του γεγονότος ότι η αναδίπλωση γίνεται εξαιρετικά γρήγορα (at the speed limit) (Kubelka et al., 2004). Η αναδίπλωση σε αυτήν την περίπτωση προχωρά μέσω ενός συνόλου προσωρινών διαμορφώσεων με μεγάλο εύρος χρόνων αναδίπλωσης (Bryngelson et al., 1995).

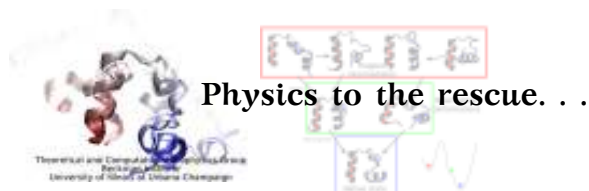
Το μεγαλύτερο εμπόδιο για τη συγκρότηση μίας ολοκληρωμένης εικόνας της αναδίπλωσης είναι οι περιορισμένες γνώσεις μας για τη φύση της μη-αναδιπλωμένης κατάστασης (unfolded state) η οποία ίσως να μην περιλαμβάνει τον τεράστιο αριθμό διαμορφώσεων (1.3-1.4 conformers per torsional degree of freedom) που προβλέπεται βάσει του αριθμού των βαθμών ελευθερίας (vanGunsteren et al., 2001). Πειραματικά δεδομένα (NMR, SAXS, SANS) συνηγορούν προς μία όχι τόσο τυχαία και χαστική φύση της μη-αναδιπλωμένης κατάστασης, έτσι ώστε η πιθανότερη εξήγηση για το παράδοξο του Levinthal να είναι ότι δεν υπάρχει όντως κάτι παράδοξο, αλλά μία εσφαλμένη αντίληψη (misconception). Υπάρχουν στοιχεία 'φυσικής' τοπολογίας και βιολογικής λειτουργίας στις ξεδιπλωμένες δομές, ακόμα και παρουσία αποδιατακτικών παραγόντων που αποτρέπουν τις υδρόφοβες αλληλεπιδράσεις (Plaxco et al., 1997, Plaxco et al., 2001, Shortle et al., 2001). Νεώτερες απόψεις υποστηρίζουν την ετερογένεια της μη-αναδιπλωμένης κατάστασης με στοχαστικές μεταπηδήσεις μεταξύ μετασταθερών καταστάσεων και παρουσιάζουν μία εικόνα, όπου η φυσική δομή λειτουργεί ως κόμβος (hub) άμεσα προσιτός σε όλες τις ενδιαμέσες καταστάσεις οι οποίες αλληλομετατρέπονται πιο συχνά με τη φυσική δομή παρά μεταξύ τους (Bowman et al., 2010, Bowman et al., 2011, Caflisch et al., 2012).



Ιστορικά, η πρόγνωση της τρισδιάστατης δομής των πρωτεϊνών από την αμινοξική τους αλληλουχία στηριζόταν σε εμπειρικές μεθόδους (code-based), όπως οι αλγόριθμοι πρόγνωσης δευτεροταγούς δομής (Chou et al., 1978). Στη δεκαετία του '90 σημειώθηκε ραγδαία πρόοδος

στον τομέα της βιοπληροφορικής με την ανάπτυξη βάσεων δεδομένων και αλγορίθμων που βασίζονταν στην ομολογία των αλληλουχιών (homology modelling) και τη μεταξύ τους στοίχιση (multiple sequence alignment) (Chothia et al., 1986), ή την αναγνώριση μοτίβων αναδίπλωσης απουσία ομολογίας με πειραματικά προσδιορισμένες δομές (threading, fold recognition) (Bowie et al., 1991, Jones et al., 1992). Πρωταγωνιστής σε αυτή την προσπάθεια είναι το CASP (Critical Assessment of Techniques for Protein Structure Prediction) (Moult et al., 1995, Das et al., 2007, Moult et al., 2009), μία προσπάθεια αξιολόγησης των μεθόδων πρόβλεψης δομών. Ένα ακόμα ορόσημο είναι η κατασκευή ενός μοντέλου 'from scratch' για την πρόγνωση της δομής ή το σχεδιασμό νέων πρωτεϊνών ([Rosetta@Home](#)) με την εθελοντική χρήση προσωπικών υπολογιστών ανά την υφήλιο: η πρωτεϊνική ακολουθία “σπάει” σε μικρά κομμάτια τα οποία αντιστοιχίζονται σε πανομοιότυπα τμήματα πρωτεϊνών γνωστής δομής, τα οποία στη συνέχεια συρράπτονται με τρόπο ώστε να ελαχιστοποιηθεί η ελεύθερη ενέργεια (Rohl et al., 2004). Η ιδέα αυτή εξελίχθηκε ακόμα περισσότερο με την εθελοντική βοήθεια των ίδιων των χρηστών μέσω ενός ηλεκτρονικού παιχνιδιού, του [Foldit](#) (Cooper et al., 2010, Khatib et al., 2011).

Τα κύρια μειονεκτήματά των εμπειρικών αυτών μεθόδων συνοψίζονται στην αδυναμία να εντοπίσουν αλλαγές στη στερεοδιαμόρφωση, που είναι μείζονος λειτουργικής σημασίας (αλλοστερικός έλεγχος, καταλυτικά κέντρα ενζύμων, διάυλοι ιόντων), προσφέρουν μόνο στατικές δομές, ενώ έχουν ένα εγγενές bias προς τη δομή της πρωτεΐνης που χρησιμοποιήθηκε ως πρότυπο (template). Έτσι, ενώ η αποτελεσματικότητά τους έχει βελτιωθεί σημαντικά από τον εμπλουτισμό της βάσης δεδομένων PDB, προβλέποντας δομές με απόκλιση της τάξης των 2-6Å από τις πειραματικά προσδιορισμένες κατά περίπτωση, η αξιοπιστία τους παραμένει περιορισμένη (Shell et al., 2009). Η ακρίβεια των *ab initio* μεθόδων πρόγνωσης δομής, με κυρίαρχο εκπρόσωπο το Rosetta, εκτιμήθηκε με τη βοήθεια των πρωτεϊνών NBP (Never Born Peptides) (Minervini et al., 2008). Οι NBPs είναι πρωτεΐνες των οποίων η αλληλουχία προκύπτει τυχαία από το συνδυασμό των 20 αμινοξέων και δεν έχει οδηγήσει μέσω της δράσης της εξελικτικής πίεσης σε λειτουργικές πρωτεΐνες (Chiarabelli et al. 2006). Η μελέτη τους αναμένεται να καταδείξει τις ιδιότητες των αλληλουχιών που καθιστούν τις πρωτεΐνες λειτουργικές και τις διαχωρίζουν από τις υπόλοιπες, μία έρευνα απαραίτητα υπολογιστική που είναι δύσκολο να διεξαχθεί με καθαρά πειραματικές μεθόδους (Minervini et al., 2008).



Αργότερα, ήρθαν στο προσκήνιο οι 'φυσικές' μέθοδοι (energy-based) με πρωταγωνιστή τις προσομοιώσεις μοριακής δυναμικής (Karplus et al., 2002). Οι μέθοδοι αυτές βασίζονται στη θερμοδυναμική ιδέα που διατύπωσε ο Anfinsen το 1973: η βιολογικά ενεργή κατάσταση (native fold) έχει την ελάχιστη ελεύθερη ενέργεια. Με αυτό ως δεδομένο, η πρόγνωση μίας δομής από την αμινοξική της αλληλουχία θα μπορούσε να δοθεί από κατάλληλες συναρτήσεις περιγραφής της δυναμικής ενέργειας του συστήματος και όχι μέσω εμπειρικών αλγόριθμων αναδίπλωσης (Šali et al., 1994).

Οι προσομοιώσεις μοριακής δυναμικής αντιμετωπίζουν το πεπτιδικό μόριο και τον περιβάλλοντα διαλύτη ως κλασσικά σωματίδια που αλληλεπιδρούν μέσω μία εμπειρικής περιγραφής της δυναμικής ενέργειας του συστήματος, το force field. Η εξέλιξη της δυναμικής ενέργειας του συστήματος γίνεται μέσω αριθμητικής ολοκλήρωσης των εξισώσεων κίνησης του Νεύτωνα, με διακριτοποίηση σε βήματα της τάξης των femtoseconds ( $10^{-15}$ sec).

Οι προσομοιώσεις μοριακής δυναμικής είναι η μόνη μέθοδος που μπορεί να προσφέρει σε ατομική διακριτικότητα όχι μόνο τη φυσική δομή, αλλά και πληροφορία για το μηχανισμό της αναδίπλωσης και τις μεταβάσεις μεταξύ καταστάσεων. Προσφάτως οδήγησαν στην επιτυχή αναδίπλωση σε ατομική διακριτικότητα ενός εύρους πρωτεϊνών (Shaw et al., 2010, Lindorff-Larsen et al., 2011, Lindorff-Larsen et al., 2012), δικαιολογώντας τον ισχυρισμό ότι αποτελούν ένα είδος υπολογιστικού μικροσκοπίου (Dror et al., 2012). Τα μεγαλύτερα μειονεκτήματα των προσομοιώσεων είναι οι ανακρίβειες των force fields (βλέπε Ενότητα 3.7) και η ανάγκη υπέρογκης υπολογιστικής δύναμης προκειμένου να γίνει διερεύνηση όλων των πιθανών διαμορφώσεων (sufficient configurational sampling) ώστε να έχουμε πλήρη περιγραφή του ενεργειακού τοπίου (Freddolino et al., 2010).

Το 1998, οι Duan και Kollman πραγματοποίησαν έναν υπολογιστικό άθλο για την εποχή, με μία προσομοίωση αναδίπλωσης διάρκειας 1μs της 36-καταλοίπων villin headpiece με αναλυτική παρουσία του διαλύτη (explicit solvent), δίνοντας μία δομή με απόκλιση 4.5Å από αυτήν που προσδιορίστηκε με NMR (Duan et al., 1998). Λίγα χρόνια αργότερα ο Pande και οι συνεργάτες του με το [Folding@Home](#) (Shirts et al., 2000) κατάφεραν να την αναδιπλώσουν με απόκλιση



Αναπαράγεται άνευ αδειάς  
από TCBG Workshop,  
Beckman Institute, UIUC  
<http://www.ks.uiuc.edu/>

μόλις 1.7Å (Zagrovic et al., 2002). Την ίδια εποχή τρεις ανεξάρτητες ομάδες, αναδίπλωσαν το μήκους 20 καταλοίπων πεπτιδίο Trp-cage σε απόκλιση 1Å (Simmerling et al., 2002, Pitera et al., 2003, Chowdhury et al., 2003). Η επικράτεια WW της πρωτεΐνης Pin1 αποτέλεσε επίσης σύστημα-μοντέλο για πρωτεΐνες που περιέχουν μόνο β-φύλλα, αποκαλύπτοντας με τον τρόπο αυτό την εσφαλμένη προτίμηση (bias) κάποιων force fields προς ελικοειδείς διαμορφώσεις (Freddolino et al., 2008, 2009).

Η συνεχής βελτίωση των παραμέτρων των force fields (βλέπε Ενότητα 3.7), των αλγόριθμων και των προγραμμάτων που χρησιμοποιούνται για τις προσομοιώσεις (Bowers et al., 2006, Fredollino et al., 2008, Hess et al., 2008) έχει αυξήσει σημαντικά την αξιοπιστία των προσομοιώσεων μοριακής δυναμικής ως εργαλείο μελέτης της πρωτεϊνικής αναδίπλωσης. Οι πιο σημαντικές προσπάθειες που οδήγησαν σε εξαιρετική επιμήκυνση του χρόνου αναδίπλωσης και στις πιο αξιοσημείωτες προσομοιώσεις αναδίπλωσης πρωτεϊνών είναι:



Το [Folding@Home](#) (Shirts et al., 2000) είναι ένα παράδειγμα distributed grid computing που τρέχει μέσω της προστασίας οθόνης (screensavers) και στο οποίο συμμετέχουν πάνω από 300.000 εθελοντές από όλο τον κόσμο, παράγοντας εκατοντάδες ή και χιλιάδες σύντομα τροχιακά (Ferst et al., 2002,

Larson et al., 2003, Beberg et al., 2009) (ανάλογο με το [Seti@Home](#) που αναλύει ραδιοκύματα ψάχνοντας για ενδείξεις εξωγήινης νοημοσύνης).

Ο Anton είναι ένα ολοκληρωμένο και ειδικά σχεδιασμένο κύκλωμα (ASICs, application-specific integrated circuits) για την εκτέλεση προσομοιώσεων μοριακής δυναμικής από το εργαστήριο του [D.E.Shaw](#), ικανό να παράγει 17.000ns υπολογιστικού χρόνου την ημέρα για συστήματα (πρωτεΐνης σε υδατικό διάλυμα) άνω των 20.000 ατόμων (Shaw et al., 2009).



Το υψηλό κόστος των προηγούμενων είναι απαγορευτικό για τη χρήση από το ευρύτερο επιστημονικό κοινό. Εν αντιθέσει, οι συστοιχίες υπολογιστών (High Performance Computing Clusters) και τελευταία η πρόδος των [GPGPUs](#) (General Purpose Graphics Processing Units) έδωσαν σημαντική ώθηση στην απόδοση ανά κόμβο (10-1000 φορές μεγαλύτερη απόδοση). Η χρήση των GPGPUs αναμένεται να εκτοξευθεί τα επόμενα χρόνια λόγω της υψηλής ζήτησης για καλύτερη απόδοση των γραφικών από την παγκόσμια αγορά ηλεκτρονικών παιχνιδιών (Stone et al., 2007).

Οι παραπάνω τρόποι οδήγησαν σε δραματική αύξηση της υπολογιστικής ισχύος καθιστώντας

εφικτές τις προσομοιώσεις αναδίπλωσης πεπτιδίων και μίνι-πρωτεϊνών (Ferrara et al., 2000, Snow et al., 2002, Gnanakaran et al., 2003, Ensign et al., 2007, Freddolino et al., 2008, Matthes et al., 2009, Mittal et al., 2010, Shaw et al., 2010). Παράλληλα, αναπτύχθηκαν και διάφοροι αλγόριθμοι που επιταχύνουν τεχνητώς τη διερεύνηση των διαμορφώσεων (configurational sampling). Ειδικές περιπτώσεις αποτελούν οι καθοδηγούμενες προσομοιώσεις (Steered Molecular Dynamics) στις οποίες μία εξωτερική δύναμη δίνει ώθηση στο σύστημα προς συγκεκριμένες διαμορφώσεις (Isralewitz et al., 2001). Πολλές μέθοδοι (accelerated dynamics) (Voter, 1997) αφορούν την τροποποίηση του ενεργειακού τοπίου του συστήματος ώστε να ξεπεραστούν τα ενεργειακά φράγματα, όπως umbrella sampling (Torrie et al., 1977), replica-exchange (Sugita et al., 1999), aMD (Hamelberg et al., 2004), adaptive-tempering (Zhang et al., 2010).

Οι προσομοιώσεις μοριακής δυναμικής είναι το αποτέλεσμα του συγκεκριμένου πολλών ακόμα επιστημονικών πεδίων εκτός από την Πληροφορική.

Η Χημεία έδωσε τη συνάρτηση που περιγράφει τη δυναμική ενέργεια του συστήματος:

Αποτελείται από δύο τμήματα, τις δεσμικές ή εσωτερικές αλληλεπιδράσεις (μωβ πλαίσιο) και τις μη δεσμικές (πράσινο πλαίσιο). Οι δεσμικές αλληλεπιδράσεις περιλαμβάνουν όρους για τους δεσμούς (bonds), τις γωνίες δεσμών (angles) και τις δίεδρες που περιγράφουν τις αλληλεπιδράσεις μεταξύ ομοιοπολικά συνδεδεμένων ατόμων με

The diagram illustrates the potential energy function  $U(\vec{R})$  for a molecular system. It is divided into two main regions: a purple region labeled 'heuristic' and a green region labeled 'from physics'. The purple region contains terms for bonds ( $\sum_{bonds} k_b^{bond} (r_i - r_0)^2$ ), angles ( $\sum_{angles} k_a^{angle} (\theta_i - \theta_0)^2$ ), and dihedrals ( $\sum_{dihedrals} k_d^{dih} [1 + \cos(n_i \phi_i + \delta_i)]$ ). The green region contains terms for non-bonded interactions ( $\sum_{i < j} \sum_{l < m} k_{ij,lm} \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 + \sum_{i < j} \sum_{l < m} \frac{q_i q_l}{\epsilon_0 r_{ij}}$ ). A 'Parameters:' box on the left points to the various constants in the equation. Small molecular models illustrate the different interaction types.

Αναπαράγεται άνευ αδείας  
από TCBG Workshop,  
Beckman Institute, UIUC  
<http://www.ks.uiuc.edu/>

αρμονικά δυναμικά (harmonic potentials). Το μειονέκτημα της αντιμετώπισης των δεσμικών αλληλεπιδράσεων μέσω αρμονικού δυναμικού είναι ότι δεν επιτρέπει την αλλαγή δεσμών μεταξύ των ατόμων ώστε να επιτρέπει την περιγραφή χημικών αντιδράσεων, αλλά η απλοποίηση αυτή επιταχύνει τις προσομοιώσεις κατά τρεις έως τέσσερις τάξεις μεγέθους. Το τμήμα αυτό είναι παρόμοιο στα διάφορα force fields και η παραμετροποίησή του έχει προκύψει από συνδυασμό κβαντικής μηχανικής και φασματοσκοπικών δεδομένων σε χημικές ενώσεις-μοντέλα (Freddolino et al., 2010).

Οι μη δεσμικές αλληλεπιδράσεις περιλαμβάνουν δύο όρους, ένα για την περιγραφή των ηλεκτροστατικών αλληλεπιδράσεων (Coulombic potential) και ένα για τις δυνάμεις van der Waals που περιλαμβάνουν τις δυνάμεις διασποράς (ασθενείς ελκτικές δυνάμεις μεταξύ ενός



παροδικού και ενός επαγόμενου-παροδικού διπόλου ή δυνάμεις London) και τις ισχυρές απωστικές δυνάμεις λόγω αλληλεπικάλυψης των ηλεκτρονικών νεφών σε αποστάσεις μικρότερες της ακτίνας van der Waals και αναφέρεται συνολικά ως δυναμικό Lennard-Jones 6-12. Στο τμήμα αυτό εντοπίζονται και οι μεγαλύτερες διαφορές μεταξύ των force fields, ως προς τον τρόπο που έγινε η παραμετροποίηση, ειδικά στα φορτία (partial charges) των ατόμων (Freddolino et al., 2010).

Η Φυσική μας έδωσε τις εξισώσεις κίνησης:

και τα Μαθηματικά την επίλυση των εξισώσεων κίνησης σε διακριτά βήματα μέσω του

αλγόριθμου Verlet (Izaguirre et al., 1999):

Οι θέσεις και επιταχύνσεις των ατόμων σε χρόνο  $t$  και οι θέσεις σε χρόνο  $t-\delta t$  χρησιμοποιούνται για τον υπολογισμό των θέσεων σε χρόνο  $t+\delta t$ .

Ο αλγόριθμος της προσομοίωσης

περιγράφεται στο διπλανό σχεδιάγραμμα:

Οι αρχικές συντεταγμένες (στο χρόνο  $t$ ) είναι αυτές που δίνει ο χρήστης (με προέλευση από κρυσταλλική/NMR δομή, ή μοντελοποίηση ή

όπως στην περίπτωση μας από κάποιο πρόγραμμα που παράγει συντεταγμένες μίας πεπτιδικής αλυσίδας με καθορισμένη αλληλουχία και

συγκεκριμένη διαμόρφωση). Οι αρχικές ταχύτητες προκύπτουν τυχαία από την κατανομή Boltzmann. Η προσομοίωση εξελίσσεται στο χρόνο με την επανάληψη της διαδικασίας υπολογισμού των δυνάμεων που ασκούνται στα άτομα και την επίλυση των εξισώσεων κίνησης βάσει των επιταχύνσεων που προκύπτουν από τις νέες δυνάμεις.

Οι περισσότεροι κώδικες προσομοιώσεων μοριακής δυναμικής είναι αρκετά πολύπλοκοι στην πραγματικότητα. Για την επίλυση των εξισώσεων κίνησης περιλαμβάνουν 2 βήματα (predictor και corrector) και πολλά επιπρόσθετα βήματα μεταξύ άλλων για τον έλεγχο της θερμοκρασίας και της πίεσης, τον υπολογισμό ενέργειας και την έξοδο (output) του προγράμματος. Η επίλυση των εξισώσεων κίνησης σε διακριτά χρονικά διαστήματα έχει σαν αποτέλεσμα να απαιτείται ο

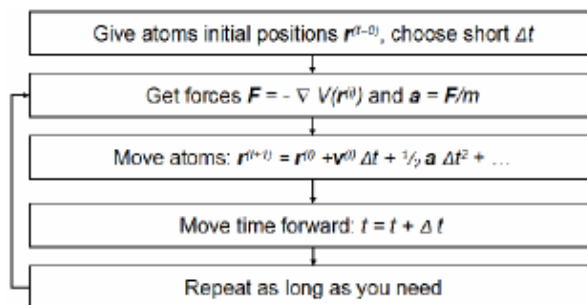
$$m_i \frac{d^2 \vec{r}_i}{dt^2} = \vec{F}_i = -\vec{\nabla} U(\vec{R})$$

$$\begin{aligned} \vec{r}(t + \delta t) &\approx \vec{r}(t) + \vec{v}(t)\delta t + \frac{1}{2}\vec{a}(t)\delta t^2 \\ \vec{r}(t - \delta t) &\approx \vec{r}(t) - \vec{v}(t)\delta t + \frac{1}{2}\vec{a}(t)\delta t^2 \end{aligned} +$$

“Verlet algorithm”



$$\vec{r}(t + \delta t) \approx 2\vec{r}(t) - \vec{r}(t - \delta t) + \vec{a}(t)\delta t^2$$

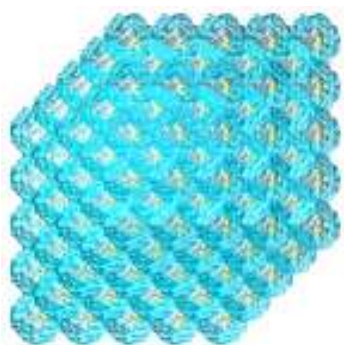


Αναπαράγεται άνευ αδείας  
από TCBC Workshop,  
Beckman Institute, UIUC  
<http://www.ks.uiuc.edu/>

προσδιορισμός δυνάμεων για την τήρηση των περιορισμών (bond geometry constraints) στο τέλος κάθε διακριτού βήματος, το οποίο γίνεται μέσω του αλγόριθμου SHAKE (Ryckaert et al., 1977).

Για την διεξαγωγή των προσομοιώσεων χρειαζόμαστε (1) ένα αρχείο τύπου PDB που περιέχει τις συντεταγμένες (x,y,z) των ατόμων, (2) ένα αρχείο παραμέτρων που περιλαμβάνει όλα τα δεδομένα της παραμετροποίησης του εκάστοτε force field για την εκτίμηση των δυνάμεων και της δυναμικής ενέργειας (bond strength, bond length κ.τ.λ.), (3) ένα αρχείο τοπολογίας που περιλαμβάνει επίσης δεδομένα της παραμετροποίησης του force field (atom types, atom masses, partial charges, connectivity) και (4) ένα αρχείο τύπου PSF (protein structure file) το οποίο περιλαμβάνει όλη τη δομική πληροφορία του συστήματος.

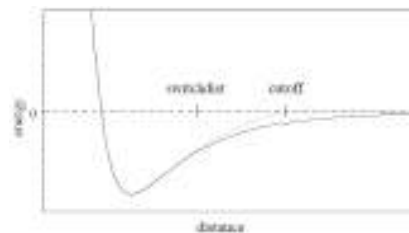
Όλες οι προσομοιώσεις της παρούσας διατριβής έχουν διεξαχθεί σε υδατικά διαλύματα (Rhee et al., 2004) με αναλυτική περιγραφή για τον διαλύτη (explicit solvent). Για την αναπαράσταση του νερού χρησιμοποιούμε το μοντέλο TIP3P (Jorgensen et al., 1983) βάσει του οποίου έγινε η παραμετροποίηση των περισσότερων force fields και κυρίως του CHARMM με το οποίο έγινε η πλειοψηφία των προσομοιώσεων της παρούσας εργασίας. Επίσης το μοντέλο αυτό αναπαριστά καλύτερα την πυκνότητα του υγρού νερού προσεγγίζοντας ικανοποιητικότερα τις συνθήκες NpT που χρησιμοποιούμε. Ωστόσο άλλα μοντέλα όπως τα TIP4P και SPC αναπαράγουν καλύτερα τις δομικές ιδιότητες του νερού, η χρήση τους όμως μένει περιορισμένη καθώς πρέπει να γίνει εκ νέου παραμετροποίηση των force fields (με εξαίρεση τις νεότερες εκδόσεις των AMBER force fields) (Zielkiewicz, 2005).



Οι προσομοιώσεις γίνονται με περιοδικές οριακές συνθήκες (periodic boundary conditions) όπου τοποθετούνται και προς τις τρεις διαστάσεις πανομοιότυπες μονάδες με το προς μελέτη σύστημα. Έτσι αποφεύγονται τα προβλήματα στα όρια του κουτιού καθώς όλοι οι υπολογισμοί αφορούν μόνο το κεντρικό κουτί. Εάν κάποιο άτομο διαφύγει από τα υποθετικά όρια του κουτιού, το είδωλο του από το γειτονικό κουτί θα πάρει τη θέση του. Το μέγεθος του

υποθετικού κουτιού είναι τέτοιο ώστε να εγγυάται την αποφυγή σύγκρουσης του εμβαπτιζόμενου μορίου με τα γειτονικά του είδωλα (Haile, 1997).

Το περισσότερο υπολογιστικά χρονοβόρο κομμάτι των προσομοιώσεων είναι ο υπολογισμός της δύναμης που ασκείται



Αναπαράγεται άνευ αδείας  
από <http://www.ks.uiuc.edu/>

σε κάθε σωματίδιο διότι προκύπτει αθροιστικά (pairwise additive interactions) από όλα τα γειτονικά άτομα. Για το λόγο αυτό οι δυνάμεις αντιμετωπίζονται ως δύο κατηγοριών, short-range και long-range. Η έκφραση Lennard-Jones 6-12 της δυναμικής ενέργειας περιγράφει αποτελεσματικά τις ελκτικές και απωστικές δυνάμεις, όπως προαναφέρθηκε. Επειδή οι δυνάμεις αυτές εξασθενούν σημαντικά σε μεγάλες αποστάσεις, χρησιμοποιείται μία συνάρτηση για την περικοπή (truncation) του δυναμικού από ένα κατώφλι απόστασης και πάνω (μέσω switching function). Η περικοπή αυτή σε συνδυασμό με τη μέθοδο PME (Darden et al., 1993) για τον υπολογισμό των ηλεκτροστατικών αλληλεπιδράσεων οδηγεί σε σημαντική μείωση του υπολογιστικού κόστους. Επιπλέον οι διακυμάνσεις των long-range αλληλεπιδράσεων είναι αργές και δεν απαιτείται ο υπολογισμός τους σε κάθε βήμα (αλλά κάθε 2-3 timesteps). Το μέγιστο όμως βήμα (timestep) με το οποίο προχωρούν οι προσομοιώσεις δεν μπορεί να είναι μεγαλύτερο από 1-2fs, λόγω της ταλάντωσης των δεσμών (bond stretching).

Μελλοντική ενσωμάτωση επιπλέον παραμέτρων στα force fields ώστε να λαμβάνονται υπόψη φαινόμενα όπως η διεύθυνση του υδρογονικού δεσμού (Kortemme et al., 2003) και η ατομική πόλωση (Halgren et al., 2001) αναμένεται να βελτιώσουν αισθητά την εγκυρότητα των force fields (Harder et al., 2006).

*“ When an old and distinguished person speaks to you,  
listen to him carefully and with respect ~ but do not believe him.  
Never put your trust into anything but your own intellect.  
Your elder, no matter whether he has gray hair or has lost his hair,  
no matter whether he is a Nobel laureate ~ may be wrong.  
The world progresses, year by year, century by century,  
as the members of the younger generation  
find out what was wrong among the things that their elders said.  
So you must always be skeptical ~ always think for yourself. ”*  
*Linus Pauling*



*“ I like people. I like animals, too  
—whales and quail, dinosaurs and dodos.  
But I like human beings especially,  
and I am unhappy that the pool of human germ plasm,  
which determines the nature of the human race,  
is deteriorating”*  
Linus Pauling



## 1.2 Αναδίπλωση πεπτιδίων

**Η** αναδίπλωση των πεπτιδίων, λόγω του μικρού μεγέθους τους ως σύστημα, έχει αποτελέσει μοντέλο για τη θεωρητική και πειραματική μελέτη της δυναμικής και των μηχανισμών των πρώιμων γεγονότων κατά την αναδίπλωση μεγαλύτερων πρωτεϊνικών συστημάτων (Gnanakaran et al., 2003). Η αναδίπλωση των μικρών πεπτιδίων λαμβάνει χώρα στην κλίμακα των nano/micro-second (Lapidus et al., 2000), γεγονός που επιτρέπει τη γεφύρωση και άμεση σύγκριση μεταξύ θεωρίας και πειράματος καθιστώντας τα εξαιρετικούς υποψήφιους για τη σύγκριση, επαλήθευση και βελτίωση των force fields (Snow et al., 2002, Gnanakaran et al., 2003, Matthes et al., 2009, Lindorff-Larsen et al., 2012).

Η κατανόηση των μηχανισμών αναδίπλωσης έδωσε γένεση στο σχεδιασμό μη βιολογικών πολυμερών, τα λεγόμενα foldamers (Gellman, 1998) με πρακτικές και θεραπευτικές εφαρμογές από τη βιοϊατρική και τη νανοτεχνολογία έως διάφορους κλάδους της μοριακής βιολογίας (Kirshenbaum et al., 1999). Η αξία των πεπτιδίων στη φαρμακευτική βιομηχανία κερδίζει

### DrugBank



<http://www.drugbank.ca/>

συνεχώς έδαφος τα τελευταία χρόνια. Τα βιοφάρμακα της παρούσας γενιάς έχουν μεγάλο μοριακό βάρος, μειωμένη σταθερότητα και χαμηλή βιοδιαθεσιμότητα ενώ ο ενέσιμος τρόπος χορήγησης (παρεντερική οδός) είναι ακριβός και δυσάρεστος για τον ασθενή (Antonosova et al., 2009). Η χρήση βιοπεπτιδίων έχει σημαντικά πλεονεκτήματα σε σχέση με τις συνθετικές ουσίες όσον αφορά την ειδικότητα δράσης και τη χαμηλή τοξικότητα, αλλά η χρήση τους παραμένει περιορισμένη λόγω της επιρρεπείας τους σε πρωτεόλυση και ταχεία απομάκρυνση μέσω των νεφρών (Kliger, 2010). Ο αποτελεσματικότερος

θεραπευτικός παράγοντας λοιπόν, θεωρείται ότι πρέπει να είναι μικρού μοριακού βάρους, να μιμείται επιτυχώς την επιθυμητή βιολογική δράση, να έχει χαμηλό κόστος παραγωγής, να έχει στοχευμένη δράση και να μπορεί να χορηγηθεί με εύκολο τρόπο, για παράδειγμα διά της στοματικής οδού (orally) (Edwards et al., 1999). Δύο προσεγγίσεις φαίνονται πολλά υποσχόμενες: (1) η ανακάλυψη βιολογικά ενεργών πεπτιδίων μέσω ενός κύκλου πρόβλεψης-επιλογής από το ίδιο το *in silico* πεπτιδίωμα (peptidome) – όλα τα πιθανά πεπτίδια μεταξύ δύο πρωτεολυτικών θέσεων σε μία πρωτεΐνη και μεταξύ κάθε τέτοιας θέσης και του τερματικού άκρου της πρωτεΐνης (Ueki et al., 2007) – και (2) ο *de novo* σχεδιασμός βάσει ενός μοτίβου δομής ή αλληλουχίας (Kliger, 2010).

Η συμβολή των προσομοιώσεων μοριακής δυναμικής ως εργαλείο για το σχεδιασμό φαρμάκων αναγνωρίζεται ολοένα και περισσότερο λόγω της πληθώρας των επιτυχημένων προσπαθειών που έχουν σημειωθεί τελευταία (Borhani et al., 2012). Οι προσομοιώσεις Monte Carlo έχουν χρησιμοποιηθεί με επιτυχία για τον υπολογισμό δεικτών όπως η ενέργεια αλληλεπίδρασης λόγω Coulomb μεταξύ διαλυμένης ουσίας και διαλύτη (ESXC) και η προσβάσιμη επιφάνεια (SASA), ο λόγος των οποίων εμφανίζει υψηλή συσχέτιση με παραμέτρους όπως τα logS και logP, δείκτες της ικανότητας του φαρμάκου να έχει σημαντική συγκέντρωση στο αίμα και να μπορεί να διανεμηθεί στους ιστούς (Jorgensen et al., 2000). Για το λόγο αυτό έχουν αναπτυχθεί πολλά προγράμματα για πρόβλεψη της τιμής τους με βάση τη δομή με ικανοποιητική ακρίβεια (Jorgensen et al., 2002). Τα προγράμματα της παρούσας γενιάς που χρησιμοποιούνται για το σχεδιασμό φαρμάκων βασίζονται σε γνωστές δομές της περιοχής πρόσδεσης της πρωτεΐνης και σάρωση βιβλιοθηκών πεπτιδίων ή σχεδιασμό συνθετικών μορίων (docking), για παράδειγμα η συνάρτηση DrugScore με το πρόγραμμα DOCK (Gohlke et al., 2000).

Τα τελευταία χρόνια βλέπουμε μία συσσώρευση συγκριτικών θεωρητικών και πειραματικών μελετών σε πεπτίδια βιολογικού ενδιαφέροντος. Για παράδειγμα μία κατηγορία πεπτιδίων με άμεση εφαρμογή στο σχεδιασμό φαρμάκων είναι τα CPPs (cell-penetrating peptides) (Zorko et al., 2005). Τα πεπτίδια αυτά έχουν μήκος μικρότερο από 30 κατάλοιπα, είναι αμφιπαθή με θετικό καθαρό φορτίο και έχουν την ιδιότητα να εισχωρούν (με ενδοκυττάρωση) στις κυτταρικές μεμβράνες και να μεταφέρουν έτσι το φορτίο τους (Tréhin et al., 2004). Τα κυκλικά πεπτίδια, μήκους από λίγα έως και εκατοντάδες αμινοξέα, λόγω της εξαιρετικής τους αντοχής στην πρωτεόλυση μπορούν να λειτουργήσουν ως δομικά ικρίωματα για το σχεδιασμό φαρμάκων που μπορούν να χορηγηθούν δια του στόματος (Craik et al., 2006). Η χρησιμότητα των πεπτιδίων για

θεραπευτικούς σκοπούς διαφαίνεται στην πατέντα για τη χρήση ενός τετραπεπτιδίου (Lys-Glu-Asp-Trp-NH<sub>2</sub>) που ρυθμίζει τα επίπεδα της γλυκόζης του αίματος στο σακχαρώδη διαβήτη τύπου 1 και 2 και οδήγησε σε αύξηση της ανταπόκρισης των ιστών στην ινσουλίνη και σε μερική αποκατάσταση της λειτουργίας του παγκρέατος των ασθενών (Khavinson et al., 2009). Το πεπτίδιο αυτό φαίνεται να έχει μιμητική δράση, να είναι ανθεκτικό στη γαστρεντερική πρωτεόλυση και να δρα μέσω ενεργοποίησης του υποκινητή του γονιδίου της πρόδρομης μορφής, δρώντας ανταγωνιστικά στην πρόσδεση της alloxan (τοξικό ανάλογο της γλυκόζης) λόγω στερικής παρεμπόδισης από την πλευρική ομάδα της τρυπτοφάνης (Khavinson, 2005).

Προκειμένου ένα πεπτίδιο να είναι δυνητικά χρήσιμο από φαρμακολογική άποψη θα πρέπει να σχηματίζει σταθερή δομή, να είναι ευδιάλυτο σε υδατικά διαλύματα και να έχει όσον το δυνατό χαμηλό μοριακό βάρος (Keller et al., 2006). Η αναζήτηση για το ελάχιστο μήκος πεπτιδικής αλληλουχίας που να μπορεί να υιοθετήσει μία σταθερή αναδίπλωση, αναπόφευκτα μας παραπέμπει σε στοιχεία δευτεροταγούς δομής στις ανώτερες δομές μεγάλων πρωτεϊνών.

Τα τετραπεπτίδια και τα πενταπεπτίδια έχουν υποδειχθεί κριτικής σημασίας σε μελέτες πρόγνωσης δευτεροταγούς δομής (Feng et al., 2008). Αποτελούν τις δομικές μονάδες σε βασικά στοιχεία δευτεροταγούς δομής, όπως η α-έλικα αλλά και δομών θηλιάς/στροφής που διαδραματίζουν καίριο ρόλο κατά την έναρξη της αναδίπλωσης (Lewis et al., 1971, Zimmerman et al., 1977). Ως εκ τούτου έχουν θεωρηθεί ως ο κώδικας της πρωτεϊνικής αναδίπλωσης (Rackovsky, 1993) αναπαριστώντας τις “δομικές λέξεις” για την ανάδειξη σχέσεων ανάμεσα στη δομή και την αλληλουχία μίας πολυπεπτιδικής αλυσίδας (Meus et al., 2006). Στατιστική ανάλυση των προτιμήσεων των δίεδρων φ/ψ γωνιών σε τετραπεπτιδικά τμήματα οδήγησε στη χρησιμοποίησή τους ως δομικούς λίθους για το σχεδιασμό νέων πρωτεϊνών (Dallüge et al., 2007). Επίσης, υπάρχει πληθώρα βιβλιογραφικών αναφορών σύμφωνα με τις οποίες υπάρχει ένα κρίσιμο μήκος αλληλουχίας γύρω στα έξι κατάλοιπα το οποίο περιέχει βιολογικά χρήσιμη λειτουργία: μπορεί να διατηρήσει ένα τοπικό μοτίβο αναδίπλωσης ανεξάρτητα από την περιβάλλουσα αλληλουχία (Compani et al., 1998), να αποτελέσει ένα υδρόφοβο πυρήνα αναδίπλωσης (hydrophobicity nucleation site), ακόμα και να προκαλέσει αλλεργική αντίδραση μέσω μηχανισμού μοριακής μίμησης (Hemmer et al., 2000). Μήκος μεγαλύτερο των έξι καταλοίπων, μαζί με τη σχετική του θέση στην αλληλουχία και το καθαρό φορτίο του πολυπεπτιδίου, θεωρούνται καθοριστικοί παράγοντες μείωσης της διαλυτότητας και προώθησης της συσσωμάτωσης και κατακρήμνισης της πρωτεϊνικής αλυσίδας (Chiti et al., 2002, Schwartz et al., 2006, Zbilut et al., 2006).

Η μελέτη τετραπεπτιδικών και πενταπεπτιδικών μη-επικαλυπτόμενων αλληλουχιών που προκύπτουν από τις καταχωρημένες στην Pfam πρωτεΐνες, έδειξε ότι η αλληλουχίες τους είναι τυχαίες και υπόκεινται σε λίγους περιορισμούς λόγω της δομής των πρωτεϊνών στις οποίες ανήκουν (Lavelle et al., 2009). Το γεγονός αυτό συνάδει με την αντίληψη ότι η αλληλουχία (πρωτοταγής δομή) καθορίζει τη δευτεροταγή δομή και όχι το αντίθετο (Anfinsen, 1973). Οκταπεπτίδια που προκύπτουν από τον κατακερματισμό πρωτεϊνικών αλυσίδων δείχνουν διαφορετική αναδιπλωσιμότητα με ένα σημαντικό ποσοστό από αυτά να δείχνει προτίμηση προς μία δομή, γεγονός που τα καθιστά πιθανά κομβικά σημεία κατά την αναδίπλωση (Ho et al., 2006).

Πεπτίδια ιδιαίτερου βιολογικού ενδιαφέροντος έχουν μελετηθεί μέσω προσομοιώσεων μοριακής δυναμικής και φαίνεται να διατηρούν δομικά στοιχεία. Προσομοιώσεις αναδίπλωσης του πενταπεπτιδίου YPGDV δείχνουν ότι υιοθετεί το μοτίβο β-στροφής τύπου II (50%) σε συμφωνία με πειραματικά NMR δεδομένα (Wu et al., 2000). Η διαπίστωση ότι τα πεπτίδια RVEW και CSVTC έχουν δομή σε υδατικά διαλύματα, όπως φαίνεται τόσο από θεωρητικά (προσομοιώσεις) όσο και από πειραματικά δεδομένα (NMR) οδήγησε στην πρόταση ότι μπορεί να αποτελούν θέσεις έναρξης της αναδίπλωσης (nucleation sites) (Simmerling et al., 1995). Πεπτίδια που περιέχουν προλίνη (*cis*-διαμόρφωση) σε γειτνίαση με κάποιο αρωματικό αμινοξύ τείνουν να σχηματίζουν ιδιαίτερα συμπαγείς δομές, όπως τα πεπτίδια SYPFDV (β-στροφή τύπου VIa) και AYPYD (β-στροφή τύπου VIb) (Demchuk et al., 1997). Το βιοενεργό πεπτίδιο GDNP σχεδιάστηκε βάσει της ελαστικής και μελετήθηκε εκτενώς μέσω προσομοιώσεων, CD και NMR δείχνοντας μία ισορροπία μεταξύ δομών β-στροφής τύπου VIII και περισσότερο εκτεταμένων διαμορφώσεων (Fuchs et al., 2006). Ακόμα και τόσο μικρά πεπτίδια μπορούν να σχηματίσουν αμυλοειδείς δομές, όπως για παράδειγμα το DFNKF, με σημαντικές εμπλοκές σε ασθένειες όπως Alzheimer, διαβήτης τύπου II, και νόσοι που οφείλονται σε prions (Flöck et al., 2006). Επικαλυμμένα νανοσωματίδια με το πενταπεπτίδιο CREKA (tumor-homing) προσδένονται ειδικά σε πρωτεΐνες θρόμβωσης στην περιοχή των αγγείων του όγκου παρέχοντας ένα τρόπο απεικόνισης του όγκου αλλά και στοχευμένης μεταφοράς φαρμάκων (Zanuy et al., 2008). Τα πεπτίδια χρησιμοποιούνται ευρέως στη νανοτεχνολογία και δη στη νανοϊατρική (κατασκευή ιστών, κυτταροκαλλιέργειες, αναγεννητική ιατρική, στοχευμένη μεταφορά φαρμάκων) καθώς είναι βιο-συμβατά, βιο-αποικοδομήσιμα, μη τοξικά και ανταποκρίνονται σε μεταβολές του περιβάλλοντος όπως pH, θερμοκρασία, συγκέντρωση άλατος (Adhikari et al., 2011).

Βλέπουμε λοιπόν πως η μελέτη πεπτιδίων με σταθερή δομή συνδέεται άμεσα τόσο με το σχεδιασμό φαρμάκων και την πεπτιδική μηχανική, όσο και με τη θεωρητική επιβεβαίωση και βελτιστοποίηση των 'φυσικών' μεθόδων, δηλαδή των προσομοιώσεων μοριακής δυναμικής με εμπειρικά παραμετροποιημένα force fields ως εργαλεία για την επίλυση του προβλήματος της αναδίπλωσης.

Ο σχεδιασμός αλληλουχιών με σταθερή δομή είναι άρρηκτα συνδεδεμένος με την κατανόηση των μηχανισμών αναδίπλωσης και της δράσης της φυσικής επιλογής (Kuhlman et al., 2004) και οδήγησε στην ιδέα της 'αναδιπλωσιμότητας' (foldability). Η αναδιπλωσιμότητα είναι εγγενές χαρακτηριστικό των αλληλουχιών που σχηματίζουν συμπαγείς φυσικές δομές και ο διαχωρισμός τους (three-body correlations, lower Shannon entropy) από τις τυχαίες αλληλουχίες (scrambled sequences) διαφαίνεται από τα πρώτα στάδια της αναδίπλωσης (Sosnick et al., 2002). Ένα αξιοσημείωτο παράδειγμα είναι η chignolin, ένα σχεδιασμένο πεπτίδιο μήκους μόλις 10 αμινοξέων με σταθερή αυτόνομη δομή β-φουρκέτας σε υδατικά διαλύματα και θερμοδυναμικά χαρακτηριστικά ανάλογα των πρωτεϊνών (Honda et al., 2004). Λίγα χρόνια αργότερα το κρίσιμο μήκος των 10 αμινοξέων για μίνι-πρωτεΐνες (microproteins) μειώθηκε σε 7-10 κατάλοιπα (Kier et al., 2008).

Στο διαδίκτυο υπάρχει, όπως προαναφέραμε, μία πληθώρα *in silico* εργαλείων για την πρόγνωση δομής από την αμινοξική αλληλουχία με μεθόδους βιοπληροφορικής, ωστόσο μόνο τρία από αυτά είναι προσαρμοσμένα για μικρού μήκους αλληλουχίες: [PepLook](#) (Thomas et al., 2009), [Pepstr](#), [Robetta](#). Το PepLook είναι σχεδιασμένο για την πρόβλεψη της δομής, της σταθερότητας και της ικανότητας πρόσδεσης πεπτιδίων μήκους 5-30 καταλοίπων σε υδατικό περιβάλλον, σε υδρόφοβο και σε μεμβράνη και βασίζεται σε τυχαίο συνδυασμό δίεδρων φ/ψ γωνιών. Το Pepstr, ο παλιότερος χρονολογικά αλγόριθμος, πραγματοποιεί πρόγνωση δευτεροταγούς δομής με έμφαση στα μοτίβα β-στροφής σε πεπτίδια μήκους 7-25 καταλοίπων και χρησιμοποιεί AMBER force fields για τη βελτιστοποίηση των δομών. Το Robetta είναι σχεδιασμένο για πρωτεΐνες και όχι για μικρά πεπτίδια, αλλά υπάρχει η δυνατότητα να χειριστεί κανείς και πεπτίδια μήκους 20-25 καταλοίπων μέσω της μεθόδου Rosetta (ομόλογη μοντελοποίηση ή *de novo* προσδιορισμός απουσία ομολογίας και μετέπειτα συρραφή των κομματιών) που έχουμε προαναφέρει. Μία ενδελεχής σύγκριση των τριών αυτών μεθόδων, σε πεπτίδια μήκους 16-27 καταλοίπων, έδειξε πως το Robetta είναι αποτελεσματικό μόνο σε πρωτεΐνες, ενώ τα Pepstr και PepLook δίνουν περισσότερο αξιόπιστα αποτελέσματα για πεπτίδια, και μάλιστα με το PepLook επιχειρείται να

προβλεφθεί και η τάση της πεπτιδικής αλληλουχίας να μην υιοθετεί σταθερή δομή (disordered) (Thomas et al., 2006).

Από τη μικρή αυτή ανασκόπηση της σύγχρονης βιβλιογραφίας προκύπτει ότι η μελέτη των πεπτιδίων έχει κρίσιμη σημασία και πολλαπλές εφαρμογές:



Ο ορθολογικός σχεδιασμός βιοενεργών πεπτιδίων με σταθερή δομή και μικρό μοριακό βάρος είναι ένα εξαιρετικά ενεργό πεδίο με ευρύ φάσμα εφαρμογών από τη διάγνωση έως το φαρμακολογικό σχεδιασμό και τη θεραπεία. Ο *ab initio* σχεδιασμός αλληλουχιών εμπλέκει αναπόφευκτα τις σύγχρονες υπολογιστικές μεθόδους για το χειρισμό τέτοιου όγκου πληροφορίας. Ωστόσο τα εργαλεία της βιοπληροφορικής φαίνεται ότι έχουν αποδώσει το μέγιστο των δυνατοτήτων τους οι οποίες αποδείχθηκαν περιορισμένες.



Τα πεπτίδια λόγω του μικρού μεγέθους τους αποτέλεσαν σύστημα-μοντέλο για τη μελέτη του προβλήματος της αναδίπλωσης μέσω προσομοιώσεων μοριακής δυναμικής. Χρησιμοποιούνται κατά κόρον για τη βελτιστοποίηση και παραμετροποίηση των force fields και αποτελούν το μέσο για την επικύρωσή τους.



Το μικρό μέγεθος του συστήματος επιτρέπει την πραγματοποίηση συστηματικής διερεύνησης και εκτεταμένων προσομοιώσεων με μικρό υπολογιστικό κόστος.



Ο μικρότερος χρόνος της αναδίπλωσης των πεπτιδίων σε σχέση με τις πρωτεΐνες καθιστά δυνατή τη γεφύρωση του χάσματος μεταξύ θεωρίας και πειράματος και επιτρέπει την άμεση σύγκριση μεταξύ των πειραματικά προσδιορισμένων τιμών και αυτών που προβλέπουν οι προσομοιώσεις μοριακής δυναμικής.

Πλέον υπάρχει μία πληθώρα πειραματικών τεχνικών για την παρακολούθηση γεγονότων αναδίπλωσης αλλά και τη διερεύνηση της αναδιπλωμένης και μη-αναδιπλωμένης κατάστασης πρωτεϊνών και πεπτιδίων: πυρηνικός μαγνητικός συντονισμός (NMR) (Roder, 1995, Plaxco et al., 1996, Peter et al., 2001, Feenstra et al., 2002), SAXS (Small-Angle X-ray scattering), SANS (Small-Angle neutron scattering), infrared, circular dichroism and fluorescence spectroscopy (Greenfield et al., 2006, Amadei et al., 2010), temperature-jump (Yang et al., 2003), μεταλλαξιγένεση ( $\phi, \psi$  values) (Sosnick et al., 2004), FRET (Förster resonance energy transfer) (Schuler et al., 2002, Merchant et al., 2007, Allen et al., 2009), hydrogen-exchange (Maity et al., 2005), πρωτεϊνική μηχανική (Itzhaki et al., 1995), laser initiated folding (Jones et al., 1993), ultrafast mixing experiments (Chan et al., 1997).

Όλες αυτές οι τεχνικές συνεισφέρουν η κάθε μία και όλες μαζί στην κατανόηση του μηχανισμού

αναδίπλωσης αλλά είναι κρίσιμο να υπάρχει κατανόηση της πειραματικής διαδικασίας ώστε να αποδίδεται σωστή ερμηνεία στο αποτέλεσμα και να γίνεται μία επικοδομητική σύγκριση μεταξύ θεωρητικών και πειραματικών δεδομένων. Χαρακτηριστικό παράδειγμα του κινδύνου που ελλοχεύει, είναι η μελέτη της αναδίπλωσης της υπομονάδας μήκους 35 καταλοίπων της villin headpiece μέσω φθορισμού με τη μέθοδο laser temperature-jump, που όπως φάνηκε από προσομοιώσεις μοριακής δυναμικής, ανιχνεύει τη δημιουργία δομής γύρω από την τρυπτοφάνη (στην οποία συνδέεται ο ανιχνευτής) και όχι της πλήρους δομής, η οποία συμβαίνει μεταγενέστερα (Cellmer et al., 2010). Είναι σημαντικό λοιπόν να υπάρχει συνεργεία και αλληλοσυμπλήρωση μεταξύ θεωρίας και πειράματος λόγω της διαφορετικής χωρικής και χρονικής διακριτικότητας του καθενός (Matysiak et al., 2007).

*“ The best way to have a good idea  
is to have lots of ideas  
and throw the bad ones away. ”*  
*Linus Pauling*

*“ Reason, Observation, and Experience,  
the Holy Trinity of Science. ”*

*Robert G. Ingersoll*



## 1.3 Σκοπός της παρούσας μελέτης

Σκοπός της παρούσας διατριβής είναι η ταυτοποίηση δυνητικά αναδιπλούμενων πεπτιδίων χρησιμοποιώντας προσομοιώσεις μοριακής δυναμικής. Το μήκος των πεπτιδίων διατηρήθηκε στα τέσσερα και πέντε κατάλοιπα λόγω της ιδιαίτερης σημασίας του μήκους αυτού τόσο στο μηχανισμό της πρωτεϊνικής αναδίπλωσης όσο και στη διαλυτότητα του πεπτιδίου για μετέπειτα φαρμακολογικές μελέτες. Το μεγάλο πλήθος των πιθανών αλληλουχιών (*ab initio* design) που προκύπτουν (160.000 τετραπεπτίδια και 3.200.000 πενταπεπτίδια) καθιστά επιτακτική την αυτοματοποίηση του σχεδιασμού και της επιλογής των πεπτιδίων. Η αναδίπλωση και ο χαρακτηρισμός της δομής γίνεται *in silico* χρησιμοποιώντας προσομοιώσεις μοριακής δυναμικής. Η εκτίμηση της δυνητικής αναδιπλωσιμότητας με ειδικά σχεδιασμένες συναρτήσεις θα αποκλείσει πεπτίδια που δεν υιοθετούν σταθερή δομή. Πεπτίδια που έχουν περάσει επιτυχώς τα στάδια αυτά και έχουν θετική πρόγνωση για υψηλή αναδιπλωσιμότητα αποτελούν υποψήφιους για τεχνική σύνθεση σε στερεά φάση και περαιτέρω πειραματική μελέτη. Ο απώτερος στόχος της παρούσας εργασίας είναι η παραγωγική σύγκριση μεταξύ των θεωρητικών προγνωστικών και των πειραματικών αποτελεσμάτων που δε μπορεί παρά να οδηγήσει στη βελτίωση και των δύο προσεγγίσεων.

Στα κεφάλαια που ακολουθούν περιγράφεται η περιπλάνηση στο κόσμο των τετραπεπτιδίων και πενταπεπτιδίων προς αναζήτηση αναδιπλούμενων αλληλουχιών.

*“ It is a lot better to come from an evolved monkey than from a fallen angel. ”*

*Marcellin Boule*



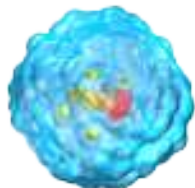


# Κεφάλαιο 2

# ΥΠΟΛΟΓΙΣΤΙΚΕΣ ΜΕΘΟΔΟΙ

*“ We’re in the matrix. We think we see everything, but we don’t really know if Google’s showing us all. ”*

<http://computationalbiologynews.blogspot.com/2007/09/bioinformatics-quotes.html>



## 2.1 Πρωτόκολλο προσομοίωσης

**Στα** πλαίσια αυτής της διατριβής πραγματοποιήσαμε ένα εκτεταμένο σύνολο προσομοιώσεων (15.210), ο υπολογιστικός χρόνος των οποίων ανέρχεται σε 272.46μs. Όλες οι προσομοιώσεις πραγματοποιούνται σε συνθήκες περιοδικής οριοθέτησης, με αναλυτική παρουσία του διαλύτη (explicit solvent) και πλήρη υπολογισμό των ηλεκτροστατικών αλληλεπιδράσεων χρησιμοποιώντας το πρόγραμμα NAMD (Kale et al., 1999).

Η πλειοψηφία των προσομοιώσεων έγινε με συστηματικό τρόπο μέσω της οργανωμένης δέσμης ενεργειών (Perl script) που παρουσιάζεται στην επόμενη ενότητα (Ενότητα 2.2). Για τις ανάγκες των προσομοιώσεων αυτών δημιουργήσαμε ένα ενιαίο πρωτόκολλο (Παράρτημα #13, all.namd) όπου συμπεριλαμβάνονται τα στάδια της ελαχιστοποίησης της ενέργειας, της σταδιακής ανόδου της θερμοκρασίας και της εξισορρόπησης που αντιστοιχεί στην παραγωγική φάση. Για το μικρό (συγκριτικά) πλήθος των 36 προσομοιώσεων που ήταν μεγαλύτερης διάρκειας δημιουργήσαμε δύο ξεχωριστά πρωτόκολλα, ένα για την διαδικασία της ελαχιστοποίησης της ενέργειας και της σταδιακής ανόδου της θερμοκρασίας (Παράρτημα #14, heat.namd) και ένα για την διαδικασία της εξισορρόπησης που συνιστά και την παραγωγική φάση (Παράρτημα #15, equi.namd). Σε αυτό το μικρότερο πλήθος των προσομοιώσεων διαφοροποιείται τόσο η τελική θερμοκρασία της προσομοίωσης, όσο και το force field που χρησιμοποιήσαμε (λεπτομέρειες αναφέρονται στις αντίστοιχες ενότητες όπου αναλύονται τα αποτελέσματα της εκάστοτε προσομοίωσης). Για το μεγάλο πλήθος των προσομοιώσεων που γίνονται συστηματικά χρησιμοποιείται το force field CHARMM22 (MacKerell et al., 1998) ενώ σε μετέπειτα μεγάλης διάρκειας προσομοιώσεις εξετάζονται και άλλα γνωστά force fields, όπως CHARMM-CMAP (MacKerell et al., 2004), OPLS-AA (Jorgensen et al., 1996, Kaminski et al., 2001), AMBER99SB (Hornak, et al., 2006,

Wickstrom et al., 2009) και AMBER99SB-ILDN (Lindorff-Larsen et al., 2010).

Εδώ παρουσιάζουμε το πρωτόκολλο της προσομοίωσης (Παράρτημα #13, all.namd) όπως εφαρμόστηκε στα πενταπεπτίδια, στην ενιαία του μορφή, που αφορά την πλειοψηφία των προσομοιώσεων που πραγματοποιήθηκαν. Μικρές διαφορές στις παραμέτρους του πρωτοκόλλου αναφέρονται κατά περίπτωση στις αντίστοιχες ενότητες όπου αναλύονται τα αποτελέσματα. Τα βήματα που ακολουθούνται στο πρωτόκολλο έχουν ως εξής:

- Ελαχιστοποίηση της ενέργειας (energy minimization) του συστήματος για 500 βήματα διατηρώντας σταθερές τις θέσεις των ατόμων του πεπτιδικού σκελετού και στη συνέχεια για ακόμα 500 βήματα χωρίς περιορισμούς θέσης.
- Ακολουθεί μία φάση (heating) κατά την οποία αυξάνεται σταδιακά η θερμοκρασία με βήμα 20K μέχρι τελικής θερμοκρασίας 320K που διαρκεί 32ps.
- Στη συνέχεια γίνεται εξισορρόπηση (equilibration) του συστήματος για 1000 βήματα κρατώντας σταθερές τις θέσεις των ατόμων Ca.
- Συνεχίζεται η εξισορρόπηση κατά την παραγωγική πλέον φάση (production phase) χωρίς κανένα περιορισμό για 10.000.000 βήματα (20ns).

Η παραγωγική φάση γίνεται πάντοτε σε συνθήκες NpT όπου για τη διατήρηση της θερμοκρασίας (συνήθως 320K, εκτός αν αναφέρεται κάποια άλλη) και της πίεσης (1atm) χρησιμοποιούνται Nosé-Hoover Langevin Dynamics και η μέθοδος ελέγχου Langevin piston barostat, ενώ οι εξισώσεις ταχύτητας επιλύονται με τον αλγόριθμο Verlet (Izaguirre et al., 1999), σύμφωνα με το πρόγραμμα NAMD. Το βήμα των προσομοιώσεων είναι 2fs (1fs =  $10^{-15}$ s) και οι ατομικές συντεταγμένες για τη δημιουργία του τροχιακού σώζονται κάθε 200 βήματα. Οι μη-δεσμικές αλληλεπιδράσεις υπολογίζονται κάθε 2 βήματα και οι ηλεκτροστατικές αλληλεπιδράσεις κάθε 4 βήματα με τη μέθοδο PME (Particle-Mesh Ewald) (Darden et al., 1993). Ο αλγόριθμος SHAKE (Ryckaert et al., 1977) χρησιμοποιείται για τον περιορισμό όλων των δεσμών, συμπεριλαμβανομένων και των πρωτονίων.

*“ Unix is user-friendly. It’s just very selective about who its friends are. ”*

<http://www.gdargaud.net/Humor/QuotesProgramming.html>

*“ Sometimes it pays to stay in bed on Monday,  
rather than spending the rest of the week  
debugging Monday’s code. ”*

*Christopher Thompson*




## 1.1 Αυτοματοποίηση της μεθόδου μέσω ενός Perl script

Ο μεγάλος αριθμός των υποψήφιων πεπτιδικών αλληλουχιών και το πλήθος των προσομοιώσεων που έπρεπε να πραγματοποιηθούν κατέστησε απαραίτητη την αυτοματοποίηση της όλης διαδικασίας. Τα βασικά βήματα που ακολουθούμε είναι:

- ☞ έλεγχος της κατάστασης και της διαθεσιμότητας της συστοιχίας υπολογιστών ([Norma computing cluster](#))
- ☞ τυχαία επιλογή μίας πεπτιδικής αλληλουχίας από το αρχικό σύνολο
- ☞ ετομασία του συστήματος της προσομοίωσης
- ☞ υποβολή της προσομοίωσης για εκτέλεση
- ☞ εκτίμηση της “αναδιπλωσιμότητας” μέσω συναρτήσεων (target functions) (Ενότητα 2.3)

Αυτή η δέσμη ενεργειών οργανώθηκε μέσω της γλώσσας [Perl](#) σε ένα αυτοτελές και αυτόνομο script το οποίο τρέχει (κάθε μισή ώρα) στον κεντρικό υπολογιστή και μας ενημερώνει μέσω e-mail για σημαντικά γεγονότα όπως κάθε φορά που υποβάλλεται ή ολοκληρώνεται η προσομοίωση μίας καινούργιας πεπτιδικής αλληλουχίας. Παρακάτω ακολουθεί μία αναλυτική περιγραφή των διαφόρων τμημάτων του script, ενώ ο κώδικας αυτούσιος (για την περίπτωση των πενταπεπτιδίων) περιλαμβάνεται στο Παράρτημα (#10, systematic.pl).




 Δήλωση μεταβλητών

Στην αρχή του script δηλώνουμε μία σειρά από μεταβλητές όπως το όνομα χρήστη (\$USER) το οποίο χρειαζόμαστε για να ελέγξουμε τον αριθμό εργασιών που έχει υποβάλει ο εκάστοτε χρήστης που θέλει να χρησιμοποιήσει το script. Ο επιθυμητός αριθμός διαδικασιών που εκκρεμούν (\$target\_PD) και ο μέγιστος επιτρεπόμενος αριθμός διαδικασιών σε εκκρεμότητα (\$max\_pending) μας επιτρέπουν να ελέγξουμε την ομαλή κίνηση στη συστοιχία των υπολογιστών. Μία ακόμα μεταβλητή (\$jobs\_per\_perft) μας επιτρέπει να επιλέξουμε πόσες επαναλήψεις θα πραγματοποιήσουμε για κάθε πεπτίδιο (στη δική μας περίπτωση έχουμε 1/πενταπεπτίδιο και 4/τετραπεπτίδιο). Επίσης ορίζουμε το όνομα του αρχείου ("2GO") που περιέχει τη λίστα με τις πεπτιδικές αλληλουχίες, το οποίο χρησιμοποιείται μετέπειτα στην υπορουτίνα που διαλέγει τυχαία το προς προσομοίωση πεπτίδιο. Τέλος, ορίζουμε και κάποιες καθολικές σταθερές (WIDTH, RMS\_CUTOFF) στις οποίες αναθέτουμε σταθερές αριθμητικές τιμές και οι οποίες χρησιμοποιούνται στις συναρτήσεις.

 Κατάσταση της συστοιχίας υπολογιστών (cluster)

Η γνώση της κατάστασης στην οποία βρίσκονται οι υπολογιστές είναι μείζονος σημασίας για την ομαλή διεξαγωγή ενός τόσο μεγάλου πλήθους προσομοιώσεων. Θα πρέπει λοιπόν να γνωρίζουμε πόσοι υπολογιστές είναι εκχωρημένοι, πόσοι είναι σε κατάσταση αδράνειας ή σε λειτουργία. Για παράδειγμα έχουμε ορίσει ότι αν υπάρχουν διαθέσιμοι λιγότεροι από 12 πυρήνες (δηλαδή 3 τετραπύρρηνοι υπολογιστές) από το σύνολο των 32, το πρόγραμμα διακόπτεται γιατί έχουμε χάσει πάνω από το 50% των υπολογιστών και συνεπώς συντρέχει κάτι σοβαρό που πρέπει να ελέγξουμε.

 Υποβολή των προσομοιώσεων

Για να διασφαλίσουμε την ομαλή λειτουργία και να αποφύγουμε την υπερφόρτωση του δικτύου των υπολογιστών ελέγξουμε τόσο τον αριθμό των προσομοιώσεων που τρέχουν, όσο και τον αριθμό αυτών που εκκρεμούν. Έτσι, εάν ο αριθμός των εν αναμονή προσομοιώσεων του συνόλου των χρηστών που είναι καταχωρημένοι στο σύστημα είναι μεγαλύτερος από 8 (\$max\_pending),












τότε το πρόγραμμα τερματίζεται χωρίς να υποβάλει καινούργια προσομοίωση. Αυτό θα συνεχιστεί μέχρις ότου να μείνουν λιγότερες από 4 ( $\$target\_PD$ ) διεργασίες εν αναμονή.

Σε αυτήν την περίπτωση το πρόγραμμα ακολουθεί επαναληπτικά τα ακόλουθα βήματα μέχρις ότου ο αριθμός των εν αναμονή διεργασιών να γίνει και πάλι 8:

- ☞ καλείται η υπορουτίνα *get\_peptide()*, η οποία μας επιστρέφει μέσω μιας μεταβλητής ( $\$peptide$ ) το όνομα του πεπτιδίου, δηλαδή την αμινοξική του αλληλουχία.
- ☞ καλείται η υπορουτίνα *prepare\_MD\_files()*, η οποία προετοιμάζει το σύστημα της προσομοίωσης και επιστρέφει ένα φάκελο με το όνομα του πεπτιδίου που περιέχει όλα τα απαραίτητα αρχεία για να τρέξει η προσομοίωση. Η υποβολή των προσομοιώσεων γίνεται μέσω του SLURM (Simple Linux Utility for Resource Management), ενός ανοιχτού λογισμικού που διαχειρίζεται συστοιχίες υπολογιστών Linux και προγραμματισμένες διεργασίες (Jette et al. 2003, Yoo et al. 2003, Balle et al. 2007, Layton 2009). Για το λόγο αυτό, στο φάκελο κάθε πεπτιδίου υπάρχει και ένα εκτελέσιμο αρχείο (shell script) το οποίο περιλαμβάνει μία δέσμη ενεργειών για το SLURM για την υποβολή της προσομοίωσης. Όλες οι προσομοιώσεις, όπως έχει προαναφερθεί, πραγματοποιούνται με το NAMD (Nanoscale Molecular Dynamics), ένα ελεύθερο λογισμικό προσομοιώσεων μοριακής δυναμικής σχεδιασμένο για υψηλή απόδοση και υψηλό παραλληλισμό (Phillips et al., 2005). Το NAMD χρειάζεται επίσης ένα αρχείο το οποίο περιλαμβάνει ένα σύνολο από παραμέτρους και ονόματα αρχείων απαραίτητα για την πραγματοποίηση της προσομοίωσης (Παράρτημα, #13 all.namd). Στο εκτελέσιμο αρχείο για την υποβολή της προσομοίωσης μέσω του SLURM συμπεριλαμβάνεται η εντολή του Unix, *sleep*, η οποία προκαλεί καθυστέρηση για τόσο χρόνο όσο ορίζει ο χρήστης. Ο λόγος είναι ότι το NAMD χρησιμοποιεί την ώρα του υπολογιστή για να αναθέσει μία τυχαία τιμή σε κάθε διεργασία (προσομοίωση) που εκτελεί. Με τον τρόπο αυτό διασφαλίζουμε ότι κάθε διεργασία θα λάβει τη δική της αριθμητική τιμή (seed number) και έτσι δε θα πάρουμε το ίδιο αποτέλεσμα όταν κάνουμε αντίγραφα της ίδιας προσομοίωσης.
- ☞ Όταν μία προσομοίωση έχει ολοκληρωθεί επιτυχώς (και μόνο τότε), μεταφέρεται σε άλλο φάκελο προς επεξεργασία και ανάλυση των αποτελεσμάτων της προσομοίωσης.

### Επεξεργασία δεδομένων και εφαρμογή των συναρτήσεων (target functions)

Το πρόγραμμα ελέγχει για την παρουσία ολοκληρωμένων διεργασιών προτού αρχίσει να υποβάλλει καινούργιες. Στην περίπτωση αυτή πρέπει να γίνει ανάλυση των δεδομένων και ακολούθως η διαγραφή των μη απαραίτητων αρχείων καθώς ο όγκος των δεδομένων των προσομοιώσεων από τέτοιο πλήθος πεπτιδίων είναι απαγορευτικός για να μπορούν να αποθηκευτούν όλα στο σκληρό δίσκο. Για κάθε λοιπόν πεπτιδίο για το οποίο έχει ολοκληρωθεί επιτυχώς η προσομοίωση γίνεται επίκληση τριών συναρτήσεων (Ενότητα 2.3) και του προγράμματος CARMA (Glykos, 2006) το οποίο χρησιμοποιείται για το μεγαλύτερο μέρος των αναλύσεων μέσω της ακόλουθης διαδικασίας:

-  χρήση του προγράμματος CARMA για την αφαίρεση μεταθέσεων/περιστροφών και τη δημιουργία ενός αρχείου DCD μόνο για την πρωτεΐνη (αφαίρεση νερού και ιόντων)
-  υπολογισμός της εξέλιξης στο χρόνο της ευκλείδειας απόστασης μεταξύ των ατόμων Ca 1-5 (1-4 για τα τετραπεπτίδια)
-  καλείται η υπορουτίνα *Target\_Function()*, η οποία εφαρμόζεται στην παραπάνω απόσταση και μας επιστρέφει μία τιμή (score)
-  υπολογισμός της εξέλιξης στο χρόνο της ευκλείδειας απόστασης μεταξύ των ατόμων Ca 2-5 (2-4 για τα τετραπεπτίδια)
-  καλείται η υπορουτίνα *Target\_Function()*, η οποία εφαρμόζεται στην παραπάνω απόσταση και μας επιστρέφει μία τιμή (score)
-  υπολογισμός της εξέλιξης στο χρόνο της ευκλείδειας απόστασης μεταξύ των ατόμων Ca 1-4 (1-3 για τα τετραπεπτίδια)
-  καλείται η υπορουτίνα *Target\_Function()*, η οποία εφαρμόζεται στην παραπάνω απόσταση και μας επιστρέφει μία τιμή (score)
-  καλείται η υπορουτίνα *Correlation\_Function()*, η οποία επιστρέφει μία τιμή (score) που βασίζεται στο γραμμικό συντελεστή συσχέτισης μεταξύ των τριών ζευγών αποστάσεων και τις τιμές (score) των αντίστοιχων αποστάσεων από την συνάρτηση *TargetFunction*
-  υπολογισμός της εξέλιξης στο χρόνο της γυροσκοπικής ακτίνας (radius of gyration) του πεπτιδίου χρησιμοποιώντας όλα τα βαριά άτομα
-  καλείται η υπορουτίνα *Target\_Function()*, η οποία εφαρμόζεται στην παραπάνω γυροσκοπική ακτίνα (radius of gyration) και μας επιστρέφει μία τιμή (score)
-  χρήση του προγράμματος CARMA για την πραγματοποίηση ανάλυσης PCA στο



Καρτεσιανό σύστημα (principal component analysis in Cartesian space)

- ☞ υπολογισμός της εντροπίας κατά Shannon της κατανομής των τριών πρώτων χαρακτηριστικών ανυσμάτων (eigenvectors) με τις υψηλότερες χαρακτηριστικές τιμές (eigenvalues)
- ☞ υπολογισμός των αριθμών από ομάδες δομών (clusters) όπως προκύπτουν από την ανάλυση Cartesian-PCA
- ☞ υπολογισμός του αριθμού από στιγμιότυπα (frames) που ομαδοποιούνται στην πολυπληθέστερη ομάδα δομών (prominent cluster)
- ☞ εάν η πολυπληθέστερη ομάδα δομών συνιστά περισσότερο από το 10% (2500/25000 frames) του τροχιακού τότε γίνονται επιπλέον υπολογισμοί. Αυτοί περιλαμβάνουν τον υπολογισμό μέσης δομής για την ομάδα, υπολογισμό των μέσων ατομικών διακυμάνσεων (average rmsf) για όλα τα βαριά άτομα, για τα άτομα του σκελετού, για τα άτομα της πλευρικής ομάδας της τρυπτοφάνης και για τα άτομα των πλευρικών ομάδων των υπόλοιπων καταλοίπων. Επίσης δημιουργείται ένα αρχείο PDB το οποίο περιλαμβάνει διαδοχικά, με σταθερό βήμα (100), στιγμιότυπα μαζί με τις υπολογισμένες σε σχέση με τη μέση δομή, ατομικές διακυμάνσεις στη στήλη των ατομικών θερμικών παραγόντων (B-factors).
- ☞ υπολογισμός ενός πίνακα RMSD που περιέχει τις τιμές rmsd, χρησιμοποιώντας όλα τα βαριά άτομα, μεταξύ όλων των πιθανών δομών του τροχιακού χρησιμοποιώντας βήμα 250 (διαφορετικά ο πίνακας θα είναι τόσο μεγάλος που δε θα φτάνει η μνήμη του υπολογιστή για τον επεξεργαστεί).
- ☞ καλείται η υπορουτίνα *Expand\_Windows()*, η οποία βαθμολογεί τον παραπάνω δισδιάστατο πίνακα και μας επιστρέφει μία τιμή (score).

Όλες οι τιμές (score) που επιστρέφονται από τις διάφορες υπορουτίνες αποθηκεύονται σε ένα ενιαίο πίνακα του οποίου κάθε στήλη αντιπροσωπεύει:

- I. Όνομα πεπτιδίου
- II. Βαθμολογία του δισδιάστατου πίνακα RMSD
- III. Βαθμολογία της απόστασης μεταξύ ατόμων Ca 1-5 (1-4 αντίστοιχα για τα τετραπεπτίδια)
- IV. Βαθμολογία της απόστασης μεταξύ ατόμων Ca 1-4 (1-3 αντίστοιχα για τα τετραπεπτίδια)
- V. Βαθμολογία της απόστασης μεταξύ ατόμων Ca 2-5 (2-4 αντίστοιχα για τα τετραπεπτίδια)

- VI. Βαθμολογία με βάση τη γραμμική συσχέτιση των τριών αποστάσεων
- VII. Βαθμολογία της γυροσκοπικής ακτίνας ( $R_g$ )
- VIII. Εντροπία της κατανομής των τριών κυρίαρχων principal components (PCs)
- IX. Αριθμός από ομάδες δομών (clusters) όπως προέκυψαν από την ανάλυση PCA
- X. Αριθμός από στιγμιότυπα (frames) της κυρίαρχης ομάδας δομών (prominent cluster)
- XI. Μέση τετραγωνική ρίζα των ατομικών διακυμάνσεων (rmsf) για όλα τα βαριά άτομα για την κυρίαρχη ομάδα δομών
- XII. Μέση τετραγωνική ρίζα των ατομικών διακυμάνσεων (rmsf) για όλα τα άτομα της πλευρικής ομάδας της τρυπτοφάνης για την κυρίαρχη ομάδα δομών
- XIII. Μέση τετραγωνική ρίζα των ατομικών διακυμάνσεων (rmsf) για όλα τα άτομα του σκελετού για την κυρίαρχη ομάδα δομών
- XIV. Μέση τετραγωνική ρίζα των ατομικών διακυμάνσεων (rmsf) για όλα τα άτομα των υπόλοιπων πλευρικών ομάδων για την κυρίαρχη ομάδα δομών

Τα μόνα αρχεία τα οποία διατηρούνται είναι αυτά στα οποία εφαρμόζονται οι υπορουτίνες, για μετέπειτα διορθώσεις και ελέγχους των συναρτήσεων, και το αρχείο PDB με τις αντιπροσωπευτικές δομές. Το αρχείο με τα αποτελέσματα παίρνει τη μορφή που φαίνεται ακολούθως:

YEKDW	5.507	5.040	3.391	53.390	5.917	9.558	7	2703	1.892	0.847	1.747
WEMRD	3.128	4.611	2.224	19.360	4.166	8.749	2	6167	1.261	0.640	1.736
DWRYK	3.147	2.749	5.322	17.468	3.143	9.590	4	2441			
EWKDH	7.063	3.280	2.857	32.374	8.216	9.250	9	5264	1.812	0.864	2.575
ERQWK	3.242	3.480	4.210	26.924	4.904	9.014	1	10107	1.615	0.813	2.311
WIDKE	2.186	2.309	2.529	7.010	3.151	9.519	5	4735	1.628	0.694	1.845
RWEHD	4.678	1.654	5.192	3.703	5.719	7.790	2	10506	0.902	0.418	0.927
QWDER	2.967	3.678	3.449	8.332	6.282	8.859	5	4953	1.940	0.777	2.401
DFEWK	3.416	2.925	3.437	13.241	4.204	9.094	4	5700	1.609	0.660	1.821
AKWDE	5.214	3.143	3.915	30.803	5.108	8.688	2	8436	1.335	0.651	1.989
RWDHK	3.127	2.929	3.196	19.603	5.286	9.363	4	4760	1.505	0.734	2.002
IRDWK	4.321	4.264	3.359	14.372	4.479	8.773	3	9847	1.940	0.867	2.560
WIDRE	2.935	2.645	2.690	13.979	4.008	9.649	12	1343			
EWFDK	6.663	2.188	4.609	42.341	5.317	9.372	7	3156	1.219	0.423	1.423
WKDEV	11.915	4.171	6.112	100.073	5.860	8.911	4	4906	1.113	0.589	1.522
TEWRD	3.917	3.789	3.512	34.176	4.461	9.216	5	2291			
RKEFW	1.941	3.176	2.522	9.060	3.953	9.567	6	1489			

### Υπορουτίνες (subroutines)

Οι υπορουτίνες μας επιτρέπουν να ομαδοποιήσουμε και να οργανώσουμε μία ομάδα εντολών την οποία θέλουμε να την επαναλάβουμε αυτούσια παραπάνω από μία φορές. Στη δική μας περίπτωση, και όπως έχουν ήδη αναφερθεί, υπάρχουν τέτοιες διαδικασίες που τις επαναλαμβάνουμε περιοδικά, όπως η επιλογή της πεπτιδικής αλληλουχίας (*get\_peptide()*), η προετοιμασία του συστήματος της προσομοίωσης (*prepare\_MD\_files()*), καθώς και μία σειρά από μαθηματικές πράξεις που περιλαμβάνουν τον υπολογισμό μέσου όρου (*aver()*), και τις τρεις συναρτήσεις εκτίμησης της αναδιπλωσιμότητας, *Target\_Function()*, *Correlation\_Function()*, *Expand\_Windows()*. Οι συναρτήσεις εκτίμησης της αναδιπλωσιμότητας, είναι μαθηματικές εξισώσεις σχεδιασμένες να αξιολογούν ένα γεγονός αναδίπλωσης και τη δημιουργία σταθερής δομής. Ο τρόπος σύλληψης, τα κριτήρια που εξετάζουν και η ερμηνεία τους αναλύονται στην επόμενη ενότητα (Ενότητα 2.3). Οι άλλες δύο υπορουτίνες περιγράφονται ακολούθως:

#### υπορουτίνα *get\_peptide()*

Η υπορουτίνα αυτή χρησιμοποιείται για τη διατήρηση και τον έλεγχο, έμμεσα, της λίστας των πεπτιδικών αλληλουχιών. Η λίστα με τις πεπτιδικές αλληλουχίες βρίσκεται στο αρχείο με όνομα “2GO”, το μήκος του οποίου μας επιτρέπει να ελέγχουμε πόσα πεπτίδια έχουν μείνει στη λίστα, πόσα έχουν επεξεργασθεί και εάν υπάρχει κάποιο πεπτίδιο για το οποίο δεν έχει ολοκληρωθεί με επιτυχία η προσομοίωση ή η μετέπειτα επεξεργασία των δεδομένων. Μία ενσωματωμένη στην Perl συνάρτηση, η *rand()* μας επιτρέπει να διαλέξουμε ένα τυχαίο αριθμό που παίρνει τιμή από το μηδέν μέχρι το τρέχον μήκος της λίστας. Η πεπτιδική αλληλουχία που αντιστοιχεί σε αυτήν τη θέση της λίστας (που αντιστοιχεί στον εκάστοτε τυχαία επιλεγμένο αριθμό) είναι αυτή που μας επιστρέφεται από την υπορουτίνα. Όλες οι υπόλοιπες αλληλουχίες τυπώνονται εκ νέου στο ίδιο αρχείο, μήκους τώρα κατά 1 πεπτίδιο λιγότερο.

#### υπορουτίνα *prepare\_MD\_files()*

Η υπορουτίνα αυτή χρησιμοποιείται για να ετοιμάσει το σύστημα της προσομοίωσης και όλα τα υπόλοιπα απαραίτητα αρχεία. Σε πρώτη φάση, το όνομα του πεπτιδίου χωρίζεται στα ξεχωριστά γράμματα που αντιστοιχούν σε κάθε αμινοξύ σύμφωνα με τον κανόνα του ενός γράμματος. Αυτό μπορεί να δημιουργεί ασυμβατότητα με προγράμματα που χρησιμοποιούνται στη συνέχεια, τα οποία κάνουν χρήση της ονοματολογίας των τριών γραμμάτων για τα αμινοξέα,

επιτρέπει όμως την λειτουργία του κώδικα ανεξάρτητα από το μήκος του πεπτιδίου (τετραπεπτίδια, πενταπεπτίδια, κ.τ.λ.). Στη συνέχεια χρησιμοποιείται το πρόγραμμα [Ribosome](#) (Srinivasan) για τη δημιουργία των ατομικών συντεταγμένων της εκάστοτε πεπτιδικής αλυσίδας σε εκτεταμένη μορφή και με ελεύθερα άκρα. Όλες οι απαραίτητες παράμετροι του προγράμματος περιέχονται σε ένα εκτελέσιμο αρχείο εντολών (ribosome.script) το οποίο δημιουργούμε αυτόματα για κάθε αλληλουχία και περιλαμβάνει:

- ☞ τον τίτλο, που αντιστοιχεί στο όνομα του πεπτιδίου
- ☞ την προεπιλεγμένη διαμόρφωση, στην περίπτωση αυτή, την εκτεταμένη (extended)
- ☞ μία σειρά από εντολές res που δηλώνουν την αμινοξική ακολουθία, με δομή κατάλοιπο ανά γραμμή

Το πρόγραμμα Ribosome χρησιμοποιεί την ονοματολογία με τρία γράμματα για τα αμινοξέα. Για να το αντιμετωπίσουμε αυτό, χρησιμοποιούμε την Perl, για να περιγράψουμε την αντιστοιχία μεταξύ του κώδικα του ενός και των τριών γραμμάτων. Στη συνέχεια χρησιμοποιούμε πάλι μέσω ενός εκτελέσιμου αρχείου εντολών (moleman.sh) το πρόγραμμα MOLEMAN (Kleywegt et al., 2001) για να ευθυγραμμίσουμε τους αδρανειακούς άξονες του πεπτιδίου με το ορθοκανονικό σύστημα και να το μετατοπίσουμε ώστε το κέντρο βάρους του να συμπίσει με την αρχή των αξόνων. Ακολουθώντας, εκτελείται ένα αρχείο εντολών (Παράρτημα, #11 pshgen.sh) το οποίο κάνει χρήση του προγράμματος PSFGEN από τη διανομή του NAMD (Kale et al., 1999), για να προσθέσει τις συντεταγμένες των πρωτονίων και να παραγάγει το αρχείο PSF (protein structure file). Στο σημείο αυτό γίνεται και ανάθεση της ιστιδίνης στην πλήρως πρωτονιωμένη μορφή (HSP) προς αποφυγή αυθαιρεσίας ως προς την επιλογή ατόμου αζώτου για την πρωτονίωση (HSE/HSD). Η προσθήκη νερού και ιόντων γίνεται (Παράρτημα, #12 VMD script) μέσω του προγράμματος VMD (Humphrey et al., 1996). Το κουτί της προσομοίωσης έχει σχήμα κολοβού (περικομμένου) οκταέδρου για την εφαρμογή συνθηκών περιοδικής οριοθέτησης και σχηματίζεται από έναν κύβο αρχικών διαστάσεων  $28 \times 28 \times 28 \text{ \AA}^3$ , ώστε η ελάχιστη απόσταση μεταξύ γειτονικών ειδώλων να είναι (αρχικά)  $\sim 6 \text{ \AA}$ . Το κουτί γεμίζει με προ-εξισορροπημένα μόρια νερού τύπου TIP3P (Jorgensen, et al., 1983, Zielkiewicz, 2005) όπου εμβαπτίζεται το πεπτίδιο, αφαιρώντας όλα τα μόρια νερού που βρίσκονται σε απόσταση μικρότερη από  $1.8 \text{ \AA}$  από αυτό. Το τελικό σύστημα αποτελείται από  $\sim 900$  άτομα, εκ των οποίων τα  $\sim 90$  είναι άτομα του πεπτιδίου και 1 ιόν νατρίου. Η ισοδύναμη συγκέντρωση NaCl ενός διαλύματος με ίδιες αναλογίες μορίων νερού και ιόντων είναι 30mM.

*“ The number of the beast — vi vi vi. ”*

<http://www.gdargaud.net/Humor/QuotesProgramming.html>

*“ Profanity is the one language  
all programmers know best. ”*

<http://www.gdargaud.net/Humor/QuotesProgramming.html>



## 1.3 Συναρτήσεις εκτίμησης της αναδιπλωσιμότητας

Η μελέτη ενός μεγάλου αριθμού πεπτιδίων μέσω προσομοιώσεων μοριακής δυναμικής για την εύρεση αναδιπλούμενων αλληλουχιών καθιστά επιτακτική την ανάγκη εύρεσης ενός τρόπου ανίχνευσης και συστηματικής εκτίμησης της “αναδιπλωσιμότητας” μέσω συναρτήσεων.

Υπάρχει μία πληθώρα παραμέτρων που θεωρούνται ισχυροί υποψήφιοι λόγω της ικανότητάς τους να ξεχωρίζουν την αναδιπλωμένη κατάσταση στο γενικό πληθυσμό διαμορφώσεων που απαρτίζουν το ενεργειακό τοπίο των πεπτιδίων και των πρωτεϊνών. Μεταξύ των πιο συχνά χρησιμοποιούμενων κατά την ανάλυση των προσομοιώσεων (Freddolino et al., 2010) είναι η προβολή του τροχιακού σε μία τρισδιάστατη επιφάνεια που ορίζεται από τους τρεις principal components της ανάλυσης PCA (Principal Component Analysis), η εξέλιξη στο χρόνο της τιμής του RMSD από μία δομή αναφοράς (αρχική δομή, μέση δομή, native δομή) ή της τιμής Q (το ποσοστό των native αλληλεπιδράσεων), η ομαδοποίηση δομών (cluster analysis) με κινητικά ή γεωμετρικά κριτήρια (Keller et al., 2010, Cossio et al., 2011), ο αριθμός των ομάδων δομών και η κατοχή τους σε χρόνο προσομοίωσης, ο αριθμός των δεσμών υδρογόνου και η ενέργεια των μη-δεσμικών αλληλεπιδράσεων (Smith et al., 2002). Ωστόσο, καμία από αυτές τις παραμέτρους μεμονωμένα δεν είναι αρκετή για τη συστηματική εύρεση της αναδιπλωμένης κατάστασης (Taly et al., 2008), με αποτέλεσμα να χρειάζονται συναρτήσεις (scoring functions) που προκύπτουν από συνδυασμούς αυτών (Cootes et al., 2000).

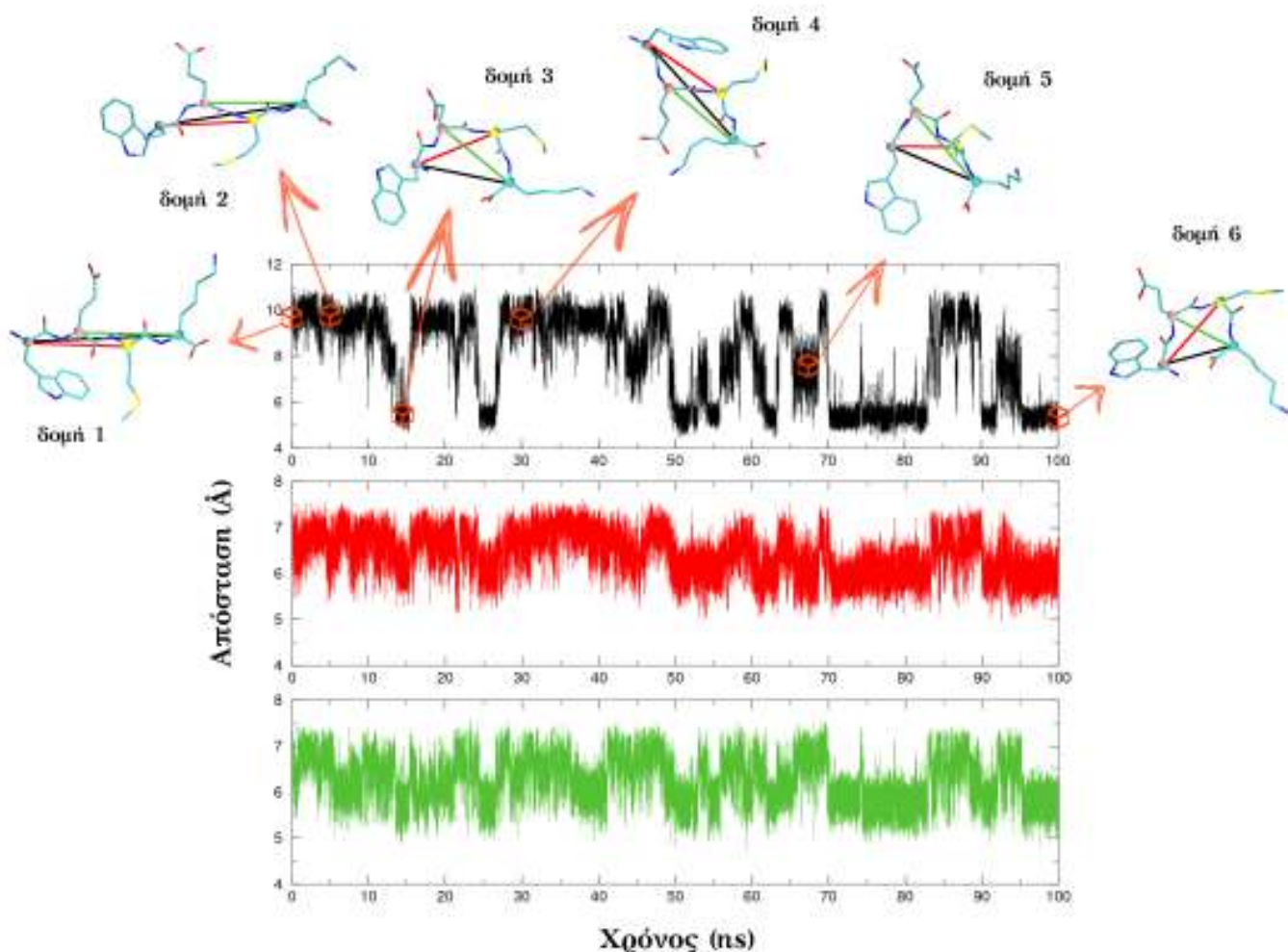
Στα πλαίσια της παρούσας εργασίας επιλέχθηκαν και μελετήθηκαν, όπως περιγράφεται ακολούθως, μία πληθώρα παραμέτρων για την ικανότητά τους να βαθμολογούν τις

τετραπεπτιδικές και τις πενταπεπτιδικές αλληλουχίες ως προς την αναδιπλωσιμότητά τους. Η απουσία *a priori* γνώσης της αναδιπλωμένης δομής των πεπτιδίων, περιορίζει τις επιλογές μας ως προς τις παραμέτρους που μπορούμε να αξιοποιήσουμε. Επιπλέον, δεν γνωρίζουμε εάν πεπτίδια τόσο μικρού μήκους έχουν την κλασική two-state συμπεριφορά ή χαρακτηρίζονται από περισσότερες από μία διακριτές διαμορφώσεις (Berezhkovskii et al., 2011).

Έτσι μελετήσαμε μεταξύ άλλων την εξέλιξη στο χρόνο της απόστασης μεταξύ των άκρων του πεπτιδίου (N-C distance), της ενέργειας των μη δεσμικών αλληλεπιδράσεων και των διακυμάνσεων του κυρίαρχου principal component (eigenvector) από την ανάλυση PCA. Η διακριτική ικανότητα των δύο τελευταίων φάνηκε περιορισμένη. Από την άλλη, ο υπολογισμός της εξέλιξης στο χρόνο των ατομικών αποστάσεων, είναι υπολογιστικά εύχρηστος και προσφέρει άμεση πληροφορία για τη δημιουργία δομής θηλιάς. Ένα ακόμα ισχυρό πλεονέκτημα είναι η δυνατότητα άμεσης σύγκρισης με πειραματικά δεδομένα (Schuetz et al., 2010) μέσω FRET (single molecule - Förster resonance energy transfer).

Προκειμένου να καταδειχθεί το πληροφοριακό περιεχόμενο των ατομικών αποστάσεων και η δυναμική χρησιμότητά τους ως παράμετρος εκτίμησης της δημιουργίας γεγονότος αναδίπλωσης παραθέτουμε στις Εικόνες 2.1-2.4 αποτελέσματα για τέσσερα αντιπροσωπευτικά πεπτίδια από τροχιακά διάρκειας 100ns.

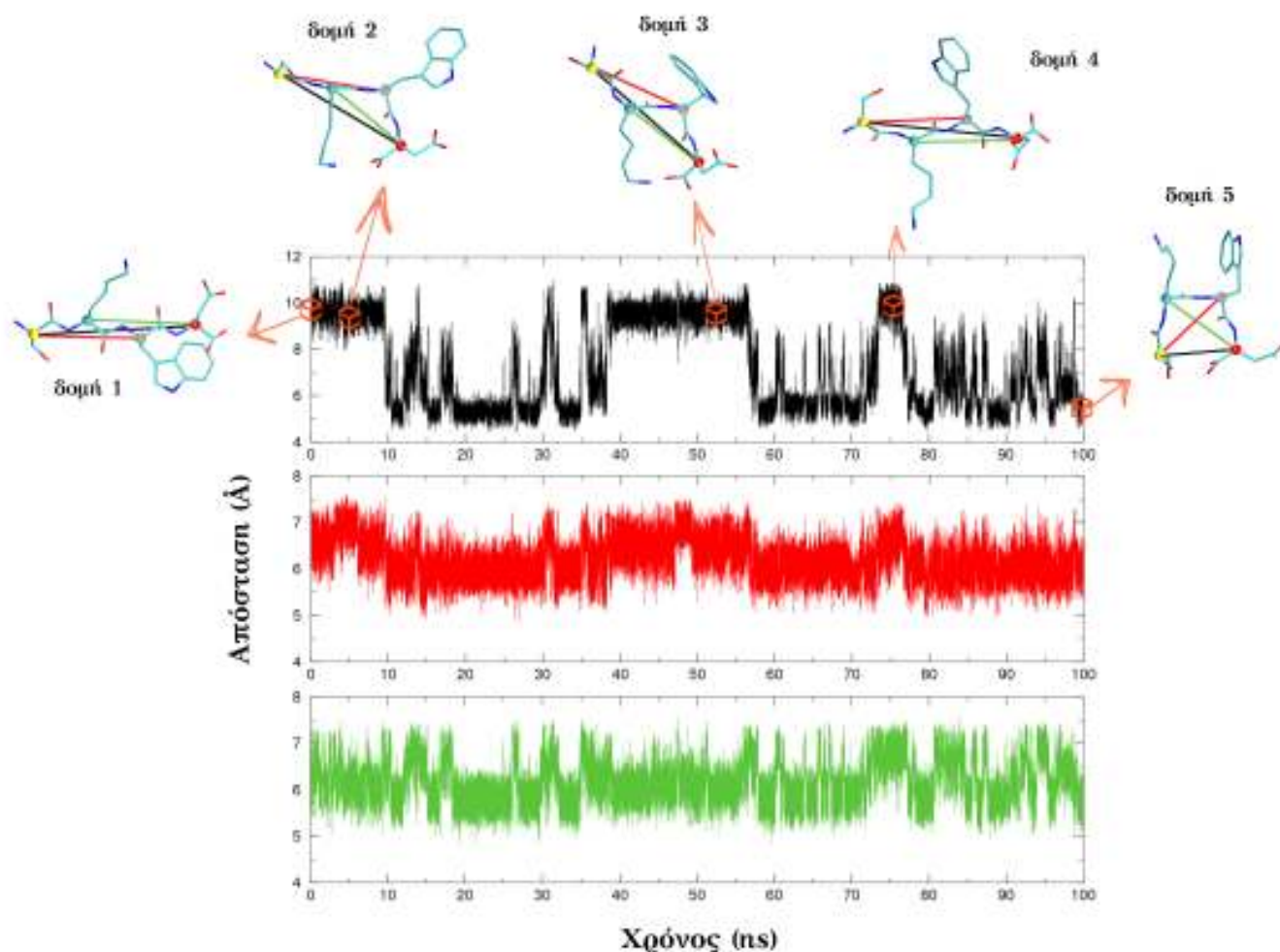
Στην Εικόνα 2.1 απεικονίζονται οι ατομικές αποστάσεις (1-4Dist, 1-3Dist, 2-4Dist) για το πεπτίδιο WEMK, το οποίο δείχνει την αντιπροσωπευτική εικόνα ενός ασταθούς πεπτιδίου. Το πεπτίδιο ξεκινάει από την εκτεταμένη διαμόρφωση (δομή 1), όπου και οι τρεις ατομικές αποστάσεις έχουν υψηλή τιμή, γύρω στα 10Å για τις αποστάσεις 1-4Dist (μεσολαβούν 2 Ca άτομα) και γύρω στα 7Å για τις αποστάσεις 1-3Dist/2-4Dist (μεσολαβεί 1 Ca άτομο). Το πρώτο γεγονός αναδίπλωσης συμβαίνει περίπου στα 14ns (δομή 3) όπου βλέπουμε πτώση της τιμής των αποστάσεων γύρω στα 5.5Å. Η τιμή της απόστασης 1-4Dist ωστόσο από μόνης της δεν είναι αντιπροσωπευτική, καθώς βλέπουμε ότι το πεπτίδιο σχηματίζει και δομές (δομή 4) όπου η τιμή της απόστασης 1-4Dist είναι στα 9.5Å αλλά οι άλλες δύο αποστάσεις παραμένουν γύρω στα 6Å. Ακόμα παρατηρούμε δομές (δομή 5) όπου η τιμή της απόστασης 1-4Dist είναι στα 7.5Å και των άλλων δύο αποστάσεων στα 6-6.5Å. Βλέπουμε λοιπόν πως η πληροφορία που χρειάζεται να εκμαιεύσουμε από τις ατομικές αποστάσεις δεν είναι απλά μία απόλυτος τιμή. Χρειαζόμαστε την καταγραφή της γενικότερης συμπεριφοράς της απόστασης στη διάρκεια του χρόνου η οποία περιλαμβάνει και τις διάφορες μεταπτώσεις της τιμής της απόστασης.



Εικόνα 2.1 Εξέλιξη στο χρόνο των αποστάσεων μεταξύ όλων των πιθανών ζευγών ατόμων Ca που δεν συνδέονται μέσω πεπτιδικού δεσμού ενός τετραπεπτιδίου (WEMK). Η απόσταση μεταξύ του πρώτου και τέταρτου Ca ατόμου (*1-4Dist*) απεικονίζεται με μαύρο χρώμα, μεταξύ του πρώτου και τρίτου (*1-3Dist*) με κόκκινο χρώμα και μεταξύ του δεύτερου και τέταρτου με πράσινο χρώμα (*2-4Dist*). Πάνω από τις γραφικές παραστάσεις επισημαίνονται αντιπροσωπευτικά στιγμιότυπα-δομές του τροχιακού όπου με γραμμές αντίστοιχου χρώματος υποδεικνύονται οι τρεις ατομικές αποστάσεις.

Η διάρκεια της εκάστοτε μετάπτωσης είναι επίσης μείζονος σημασίας καθώς δείχνει τη διάρκεια παραμονής στην τρέχουσα διαμόρφωση. Στην Εικόνα 2.2 απεικονίζονται οι ατομικές αποστάσεις (*1-4Dist*, *1-3Dist*, *2-4Dist*) για το πεπτιδίο SKWD, το οποίο δείχνει την αντιπροσωπευτική εικόνα ενός πεπτιδίου που αναδιπλώνεται αλλά τα γεγονότα αναδίπλωσης είναι πολλαπλά και περιορισμένης διάρκειας.

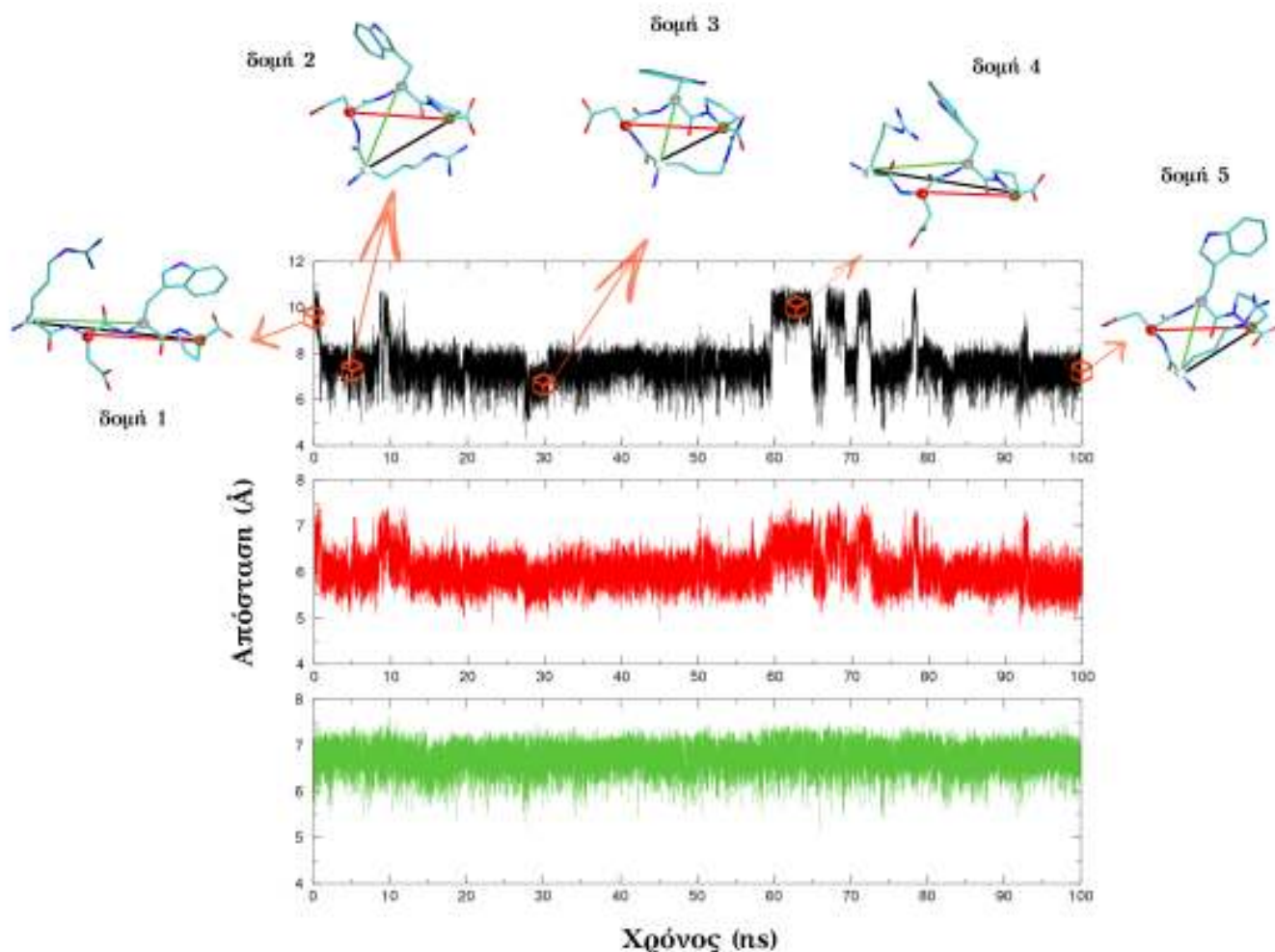




Εικόνα 2.2 Εξέλιξη στο χρόνο των αποστάσεων μεταξύ όλων των πιθανών ζευγών ατόμων Ca που δεν συνδέονται μέσω πεπτιδικού δεσμού ενός τετραπεπτιδίου (SKWD) σε αντιστοιχία με την Εικόνα 2.1.

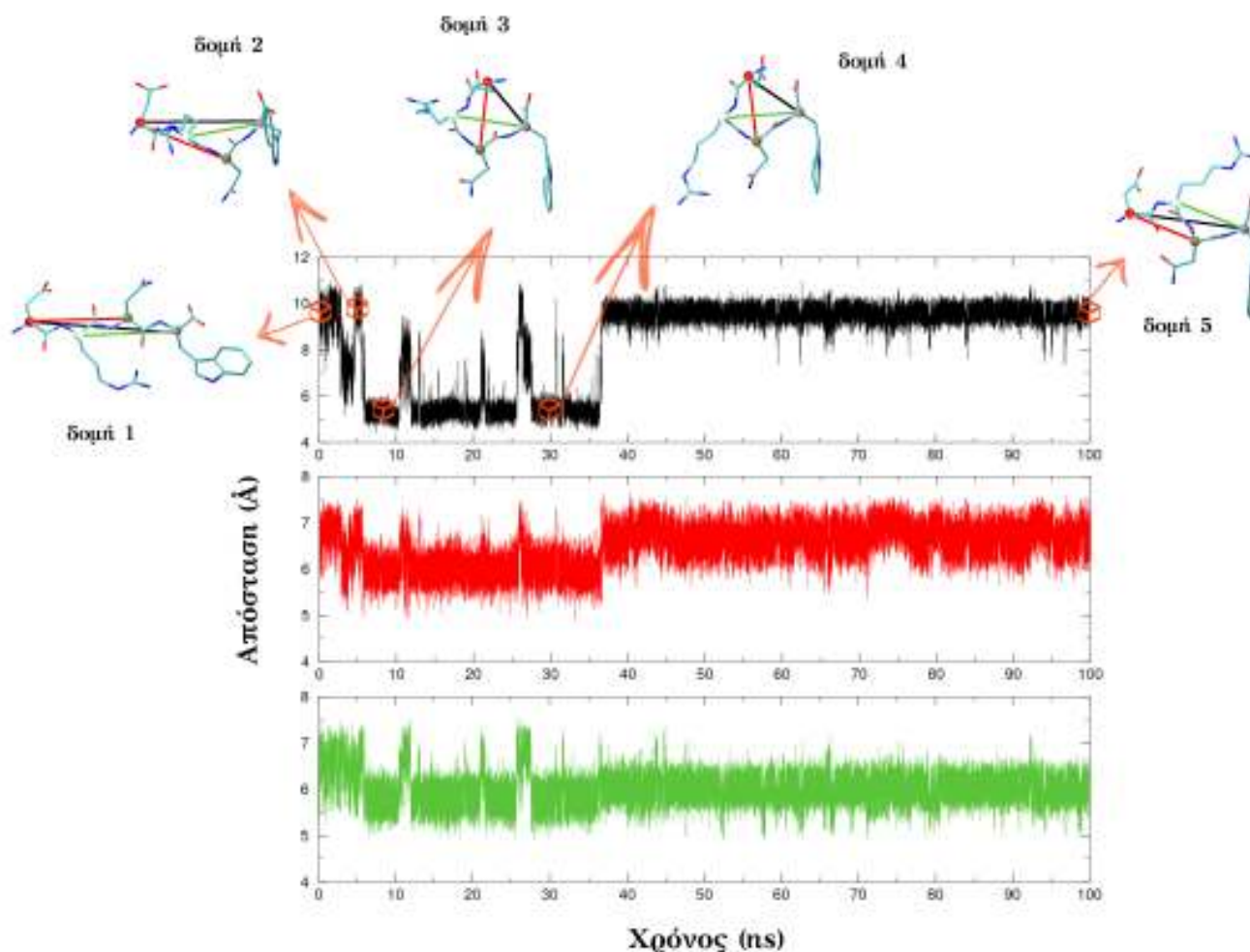
Ξεκινώντας πάλι από την εκτεταμένη διαμόρφωση (δομή 1) η πρώτη μετάπτωση των τιμών των ατομικών αποστάσεων γίνεται περίπου στα 5ns (δομή 2) και είναι περισσότερο αισθητή στις αποστάσεις 1-3Dist/2-4Dist. Στο πεπτίδιο αυτό βλέπουμε καθαρά πως η πτώση και μόνο της τιμής της απόσταση μεταξύ των άκρων (1-4Dist) δεν είναι ενδεικτική της δημιουργίας δομής. Εάν συγκρίνουμε τις δομές 3 και 5 βλέπουμε τη δημιουργία παρόμοιας δομής η οποία όμως προκύπτει από τη συμμετοχή διαφορετικών ατόμων με αποτέλεσμα οι ατομικές αποστάσεις να είναι πολύ διαφορετικές. Από την άλλη, οι δομές 3 και 4 έχουν παρόμοιες τιμές για τις ατομικές αποστάσεις αλλά είναι εμφανές ότι η δομή 4 προέκυψε από γεγονός αποδιάταξης.





Εικόνα 2.3 Εξέλιξη στο χρόνο των αποστάσεων μεταξύ όλων των πιθανών ζευγών ατόμων Ca που δεν συνδέονται μέσω πεπτιδικού δεσμού ενός τετραπεπτιδίου (RDWP) σε αντιστοιχία με την Εικόνα 2.1.

Τα φαινόμενα αυτά παρατηρούνται γιατί η δημιουργία δομής είναι αποτέλεσμα συμμετοχής όλων των ατόμων, τόσο του πεπτιδικού σκελετού όσο και των πλευρικών ομάδων. Έτσι, ενώ η εξέλιξη των ατομικών αποστάσεων μεταξύ ατόμων Ca αποτελεί ένδειξη για μεταβολή στη δομή, για την εκτίμηση της ίδιας της δομής με συστηματικό τρόπο ίσως να χρειάζεται κάποιος άλλος δείκτης. Στην Εικόνα 2.3 απεικονίζονται οι ατομικές αποστάσεις (1-4Dist, 1-3Dist, 2-4Dist) για το πεπτίδιο RWPD, για το οποίο η παρουσία της προλίνης στην αλληλουχία του αναμένεται να επηρεάσει τις τιμές που παρατηρούμε για τις ατομικές αποστάσεις στη διάρκεια του χρόνου.



Εικόνα 2.4 Εξέλιξη στο χρόνο των αποστάσεων μεταξύ όλων των πιθανών ζευγών ατόμων Ca που δεν συνδέονται μέσω πεπτιδικού δεσμού ενός τετραπεπτιδίου (DRNW) σε αντιστοιχία με την Εικόνα 2.1.

Η παρουσία της προλίνης στην τέταρτη θέση περιορίζει αισθητά την απόσταση 2-4 η οποία παραμένει σταθερή γύρω στα 6-7Å. Οι άλλες δύο αποστάσεις φαίνεται να έχουν συγχρονισμένες μεταβολές αλλά με μικρότερο εύρος και η σταθεροποίηση τους γύρω από μία μέση τιμή με μικρές διακυμάνσεις συνιστά ένδειξη για τη δημιουργία δομής.

Στην Εικόνα 2.4 απεικονίζονται οι ατομικές αποστάσεις (1-4Dist, 1-3Dist, 2-4Dist) για το πεπτίδιο DRNW, το οποίο δείχνει την αντιπροσωπευτική εικόνα ενός πεπτιδίου με περισσότερες από μία διακριτές διαμορφώσεις στις οποίες παραμένει για σημαντικό ποσοστό του χρόνου προσομοίωσης.

Με βάση τις μεταπτώσεις που ακολουθούν και οι τρεις ατομικές αποστάσεις βλέπουμε δημιουργία δομής τόσο για μικρές τιμές αποστάσεων (δομή 3 και δομή 4) όσο και για υψηλές τιμές (δομή 5). Αν παρατηρήσουμε τη συσχέτιση μεταξύ και των τριών αποστάσεων φαίνεται να υπάρχουν τουλάχιστον δύο διακριτές διαμορφώσεις με τη δεύτερη (δομή 5) να διατηρείται για περισσότερο από 50% του τροχιακού. Στην πρώτη ομάδα διαμορφώσεων με παρόμοιες τιμές ατομικών αποστάσεων οι δομές είναι παρόμοιες σε επίπεδο πεπτιδικού σκελετού αλλά διαφέρουν στη διαμόρφωση των πλευρικών ομάδων.

Οι ατομικές αποστάσεις λοιπόν, θα μπορούσαν να χρησιμοποιηθούν ως δυνητικοί εκτιμητές της αναδιπλωσιμότητας των πεπτιδίων αξιολογώντας τη δημιουργία γεγονότος αναδίπλωσης, θέτοντας τα ακόλουθα κριτήρια:

- ☐ Μεταβολή της απόστασης των N-C άκρων από  $\sim 10\text{\AA}$  (αρχική, εκτεταμένη διαμόρφωση) σε  $\sim 4\text{-}6\text{\AA}$  (δημιουργία δομής θηλιάς). Η μεταβολή μπορεί να είναι απότομη ή σταδιακή, μικρής ή μεγάλης κλίσης ή ακόμα να περιλαμβάνει 1 ή 2 “σκαλοπάτια”. Όσο πιο μεγάλη και απότομη η αλλαγή και όσο πιο νωρίς συμβεί στο τροχιακό, τόσο υψηλότερη θα είναι η βαθμολογία.
- ☐ Εφόσον παρατηρηθεί η πτώση της τιμής της απόστασης των N-C άκρων, θα πρέπει να παραμείνει σε χαμηλές τιμές με μικρές αποκλίσεις ως το τέλος της προσομοίωσης. Η μικρή απόκλιση είναι ενδεικτική της σταθεροποίησης της τιμής της απόστασης γύρω από μία μέση τιμή. Όσο μεγαλύτερη η παραμονή της απόστασης σε χαμηλή τιμή και όσο μικρότερες οι αποκλίσεις γύρω από αυτή τη μέση χαμηλή τιμή, τόσο υψηλότερη θα είναι η βαθμολογία.
- ☐ Ταχείες και επαναλαμβανόμενες μεταβολές της απόστασης των N-C άκρων μεταξύ υψηλών και χαμηλών τιμών είναι ενδεικτικές πολλαπλών γεγονότων αναδίπλωσης/αποδιάταξης και θα οδηγούν σε χαμηλή βαθμολογία.

Προκειμένου να προχωρήσουμε σε μία συστηματική αξιολόγηση των πεπτιδίων με μία συνάρτηση (target function), τα κριτήρια αυτά θα πρέπει να αναχθούν σε παραμέτρους που θα οδηγήσουν σε μία μαθηματική εξίσωση. Ένα μέτρο θα πρέπει να δείχνει τις απότομες μεταπτώσεις κι έτσι ορίζουμε μία παράμετρο (**D**) που είναι η διαφορά μεταξύ της μέγιστης και της ελάχιστης παρατηρούμενης τιμής. Όσο μεγαλύτερη η διαφορά, τόσο πιο απότομη είναι η μεταβολή στην απόσταση, εξ ου και η παράμετρος **D** θα βρίσκεται στον αριθμητή της εξίσωσης. Η μέση τιμή της απόστασης είδαμε ότι από μόνη της δεν έχει διακριτική ικανότητα ωστόσο

μπορεί να χρησιμοποιηθεί μέσα στην εξίσωση για τον αποκλεισμό πεπτιδίων που έχουν παραμείνει σε εκτεταμένη διαμόρφωση και δεν δείχνουν αναδίπλωση. Στα πεπτίδια αυτά, η απόσταση εκτός από υψηλή τιμή έχει και μεγάλες διακυμάνσεις. Το εύρος των διακυμάνσεων εκφράζεται μέσω του rmsd (root-mean-square deviation). Έτσι η μέση τιμή (**aver**) και η διακύμανση (**rmsd**) θα βρίσκονται στον παρονομαστή της εξίσωσης για τον αποκλεισμό πεπτιδίων στα οποία δεν παρατηρήθηκε γεγονός αναδίπλωσης ή παρατηρούνται πολλές και διαδοχικές μεταπτώσεις.

Από τις γραφικές παραστάσεις είδαμε ότι οι μεταπτώσεις μεταξύ των τιμών των αποστάσεων δεν είναι πάντοτε απότομες, αλλά συνήθως είναι σταδιακές και μπορεί να περιλαμβάνουν και ενδιάμεσα σκαλοπάτια. Η διακύμανση (rmsd) που υπολογίζεται για ολόκληρο το τροχιακό όπως προαναφέραμε αποκλείει τα πεπτίδια με πολλαπλές μεταπτώσεις. Όμως το εύρος των διακυμάνσεων τοπικά είναι μείζονος σημασίας. Έτσι υπολογίζουμε μία σειρά από μέσες τιμές και rmsd για κυλιόμενα παράθυρα σταθερού πλάτους (WIDTH). Το παράθυρο ορίστηκε “αυθαίρετα” στην τιμή 500 καθώς είδαμε ότι αντιπροσωπεύει αποτελεσματικά τις μεταβάσεις και κατατάσσει σωστά τα πεπτίδια ως προς την εξέλιξη των ατομικών αποστάσεων σε τροχιακά μικρής διάρκειας στα οποία θέλουμε να εφαρμόσουμε τη συνάρτηση. Εάν για παράδειγμα ο αριθμός από frames του τροχιακού είναι 12.500, αυτό σημαίνει 25 κυλιόμενα παράθυρα και ο υπολογισμός αφορά ένα παράθυρο διάρκειας 0.4ns. Από το σύνολο των τιμών rmsd για τα κυλιόμενα παράθυρα χρησιμοποιούμε το μέγιστο rmsd (**max**) και το ελάχιστο rmsd (**min**) στον αριθμητή και τον παρονομαστή αντίστοιχα.

Η συνάρτηση (Target Function 1, TF1) περιγράφεται από την ακόλουθη μαθηματική εξίσωση, ενώ ο κώδικας που χρησιμοποιείται για να γίνει ο υπολογισμός παραθέτεται στο Παράρτημα (#10, systematic.pl, subroutine *The Target Function*):

$$t_i = \frac{D * \max}{\text{aver} * \min * \text{rmsd}} \quad (\text{TF1})$$

όπου:

**D** = η απόλυτη διαφορά μεταξύ της μέγιστης και της ελάχιστης παρατηρούμενης τιμής της παραμέτρου (εδώ της απόστασης)

**max** = η μέγιστη τιμή του rmsd που υπολογίζεται για κυλιόμενα παράθυρα της παραμέτρου

**min** = η ελάχιστη τιμή του rmsd που υπολογίζεται για κυλιόμενα παράθυρα της παραμέτρου

**aver** = η μέση τιμή που υπολογίζεται για το σύνολο των τιμών της παραμέτρου

**rmsd** = το rmsd που υπολογίζεται για το σύνολο των τιμών της παραμέτρου

Η συνάρτηση αυτή εφαρμόζεται σε μονοδιάστατα δεδομένα και είναι ενσωματωμένη στο Perl script με τη μορφή υπορουτίνας, η οποία καλείται συνολικά 4 φορές για τις τρεις ατομικές αποστάσεις (1-4Dist, 1-3Dist, 2-4Dist) και για τη γυροσκοπική ακτίνα (Ενότητα 2.2).

Από τα αποτελέσματα των αντιπροσωπευτικών τετραπεπτιδίων που παρουσιάστηκαν (Εικόνες 2.1-2.4) είδαμε ότι ιδιαίτερα σημαντικό χαρακτηριστικό είναι ο συγχρονισμός των μεταβολών και των τριών αποστάσεων για την ανίχνευση ενός γεγονότος αναδίπλωσης. Για το λόγο αυτό, δημιουργήσαμε και μία δεύτερη εξίσωση (Παράρτημα #10, systematic.pl, subroutine Linear Corr. Coef. between distances) που υπολογίζει το γραμμικό συντελεστή συσχέτισης των τριών ατομικών αποστάσεων ανά ζεύγη και υπολογίζει μία νέα βαθμολογία βασισμένη και στις τρεις αποστάσεις. Η κλήση της υπορουτίνας γίνεται μόνο μία φορά, και αφού έχει εφαρμοσθεί η συνάρτηση TF1 και στις τρεις αποστάσεις και έχει υπολογισθεί η αντίστοιχη βαθμολογία. Για τη νέα συνάρτηση (Target Function 2, TF2) θα πρέπει να πληρούνται όλα τα κριτήρια που προαναφέρθηκαν για τη συνάρτηση TF1 και επιπλέον το κριτήριο της ταυτόχρονης και συντονισμένης μεταβολής και στις τρεις αποστάσεις.

Η συνάρτηση περιγράφεται από την ακόλουθη μαθηματική εξίσωση, ενώ ο κώδικας που χρησιμοποιείται για να γίνει ο υπολογισμός παραθέτεται στο Παράρτημα (#10, systematic.pl, subroutine *Linear Corr. Coef. Between distances*):

$$T = \sum_{i,j=0,i \neq j}^2 t_i t_j \text{Corr}(t_i t_j) \quad (\text{TF2})$$

όπου:

$t_{i,j}$  = η βαθμολογία της συνάρτησης TF1 για την απόσταση μεταξύ ατόμων Ca

$i,j$  = ο αύξων αριθμός της απόστασης μεταξύ ατόμων Ca, όπως υπολογίζονται και δηλώνονται μέσα στο Perl script

$\text{Corr}(t_i t_j)$  = ο γραμμικός συντελεστής συσχέτισης κατά Pearson μεταξύ δύο αποστάσεων ( $i,j$ )

Πιο αναλυτικά, αρχικά γίνεται η δίλωση κάποιων τοπικών μεταβλητών και η ανάγνωση των δεδομένων (μία στήλη από αριθμούς κινητής υποδιαστολής) από τα τρία αρχεία, που αντιστοιχούν στην εξέλιξη στο χρόνο των τριών ατομικών αποστάσεων, 1-4Dist, 1-3Dist, 2-4Dist. Για τον υπολογισμό του γραμμικού συντελεστή συσχέτισης κατά Pearson χρησιμοποιείται ο ακόλουθος τύπος:

$$\text{Corr}(t_i, t_j) = \frac{\text{COV}(t_i, t_j)}{\sigma_{t_i} \cdot \sigma_{t_j}}$$

όπου:

$\text{COV}(t_i, t_j)$  = η συνδιακύμανση των μεταβλητών  $t_i$  και  $t_j$

$\sigma_{t_i}, \sigma_{t_j}$  = η τυπική απόκλιση των μεταβλητών  $t_i$  και  $t_j$  αντίστοιχα

Για τον σωστό υπολογισμό του γραμμικού συντελεστή συσχέτισης χρειάζεται να γίνει πέρασμα των τιμών δύο φορές, την πρώτη για να γίνει ο υπολογισμός της μέσης τιμής και μία δεύτερη για τον υπολογισμό των τυπικών αποκλίσεων. Εδώ χρησιμοποιήθηκε ένας πιο γρήγορος αλγόριθμος, όπου γίνεται μόνο ένα πέρασμα των τιμών αλλά διατηρείται ικανοποιητική αριθμητική ακρίβεια. Επί της ουσίας, στην υπορουτίνα ο συντελεστής συσχέτισης υπολογίζεται κατευθείαν και για τα τρία ζεύγη τιμών που αντιστοιχούν στις τρεις ατομικές αποστάσεις και επιστρέφεται η τελική τιμή της συνάρτησης TF2 που προκύπτει από το άθροισμα τριών όρων (για τα τρία πιθανά ζεύγη ατομικών αποστάσεων) καθένας από τους οποίους είναι το γινόμενο του γραμμικού συντελεστή συσχέτισης μεταξύ δύο αποστάσεων και των τιμών της συνάρτησης TF1 για τις αποστάσεις αυτές.

Από την εφαρμογή των βασιζόμενων σε ατομικές αποστάσεις συναρτήσεων TF1 και TF2 στο σύνολο των πεπτιδίων της παρούσας μελέτης (8.640 πεπτιδικές αλληλουχίες) αναδείχθηκαν κάποιες αδυναμίες όταν εφαρμόζονται σε σύντομα τροχιακά όπως αυτά του πρώτου κύκλου (Κεφάλαιο 3, Ενότητα 3.2) προσομοιώσεων των τετραπεπτιδίων (διάρκειας 5ns). Τα πεπτίδια ομαδοποιούνται σε 3 κατηγορίες, τους γρήγορους-αναδιπλωτές, τους αργούς-αναδιπλωτές και τους μη-αναδιπλωτές. Ευνοούνται με υψηλή βαθμολογία οι γρήγοροι αναδιπλωτές, ενώ οι σταδιακές και ομαλές μεταπτώσεις δεν ευνοούνται βαθμολογικά. Επίσης δεν βαθμολογούνται σωστά οι πεπτιδικές αλληλουχίες που περιλαμβάνουν προλίνη. Ο λόγος είναι ότι λόγω των ασυνήθιστων φ/ψ γωνιών του μινοξέος αυτού οι τιμές των ατομικών αποστάσεων παραμένουν

υψηλές (7-9Å) και ως εκ τούτου δυσδιάκριτες από αυτές των πεπτιδίων που παρέμειναν σε εκτεταμένη διαμόρφωση (Ενότητα 3.3). Οι παρατηρήσεις αυτές γίνονται κατανοητές μέσω των γραφικών παραστάσεων (Εικόνες 2.5-2.7) από ενδεικτικά πεπτίδια του πρώτου κύκλου προσομοιώσεων των τετραπεπτιδίων (Κεφάλαιο 3, Ενότητα 3.3).

Στην Εικόνα 2.5 βλέπουμε την εξέλιξη στο χρόνο των τριών ατομικών αποστάσεων για αντιπροσωπευτικά τετραπεπτίδια τα οποία έδωσαν τις υψηλότερες και χαμηλότερες βαθμολογίες με βάση τις συναρτήσεις TF1 και TF2 (Πίνακας 2.1). Βλέπουμε στο πεπτίδιο DMWR ότι οι συντονισμένες μεταβολές μεταξύ και των τριών αποστάσεων σε συνδυασμό με τις μικρές διακυμάνσεις οδηγούν στην υψηλότερη βαθμολογία [66.6] με βάση τη συνάρτηση TF2 (12.4 με βάση τη συνάρτηση TF1). Από την άλλη, στο πεπτίδιο WSDK, οι μεταβολές μεταξύ των αποστάσεων 1-3Dist και 2-4Dist είναι συντονισμένες μεταξύ τους αλλά όχι και με την απόσταση 1-4Dist (η οποία ασκεί τη μεγαλύτερη επιρροή στη συνάρτηση (2) λόγω των πιο απότομων μεταβολών) με αποτέλεσμα να έχει την υψηλότερη βαθμολογία [17.8] με βάση τη συνάρτηση TF1 αλλά ελαφρώς χαμηλότερη βαθμολογία [61.9] με βάση τη συνάρτηση TF2 σε σύγκριση με το DMWR. Η σταδιακή πτώση της τιμής της απόστασης 1-4Dist σε συνδυασμό με τις μεγάλες και συχνές διακυμάνσεις οδηγούν το πεπτίδιο REWA στις τελευταίες θέσεις της βαθμολογίας [0.05] με βάση τη συνάρτηση TF2. Όπως βλέπουμε λοιπόν, οι συναρτήσεις που βασίζονται στις ατομικές αποστάσεις ευνοούν ιδιαίτερω τα πεπτίδια που ανήκουν στην κατηγορία των γρήγορων αναδιπλωτών (fast-folders).

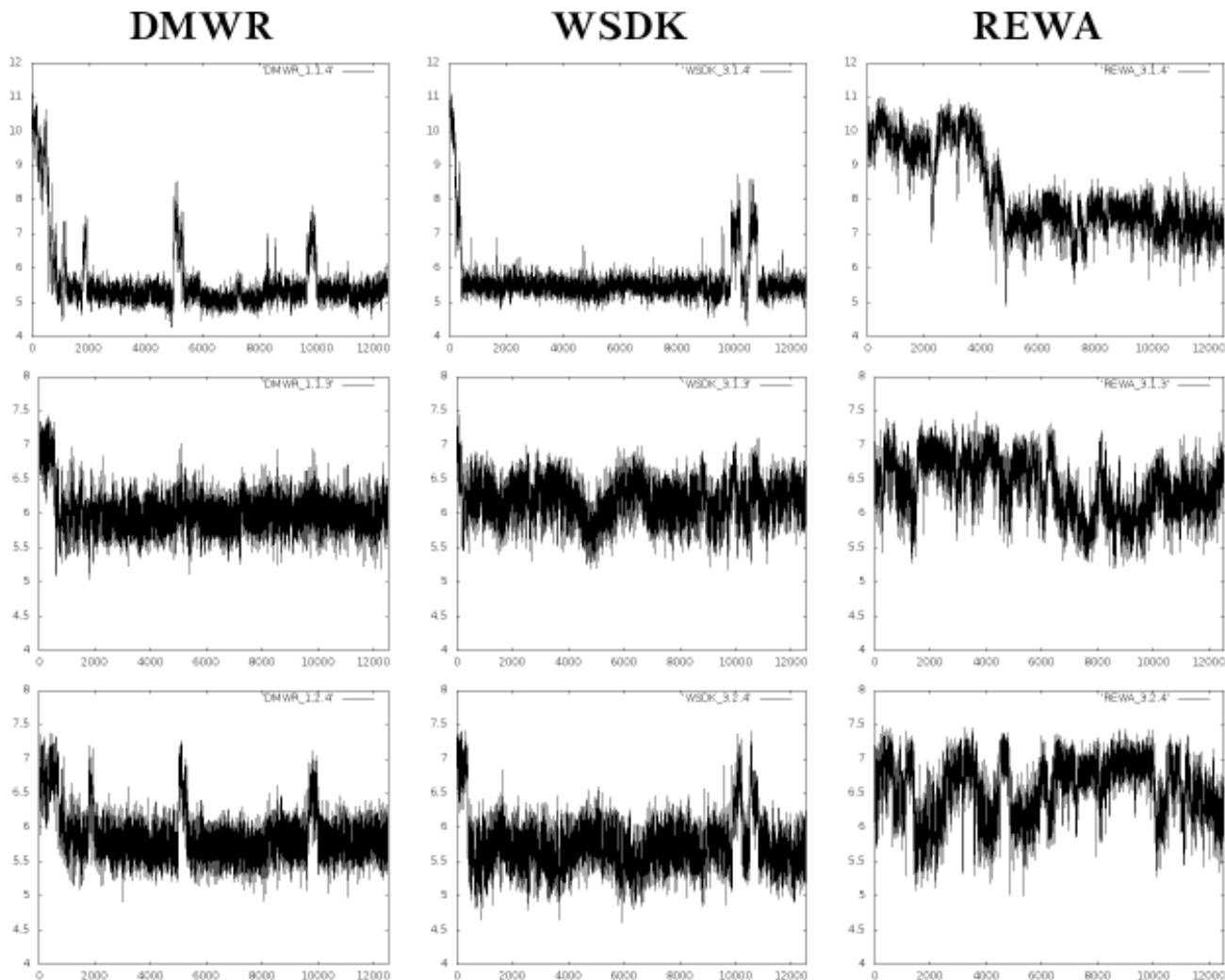
	Μέσος Όρος	Τυπική Απόκλιση	Μικρότερη Τιμή	Μεγαλύτερη Τιμή
TF1	3.2248	1.3713	0.8100	17.7850
TF2	8.7092	6.3102	0.0459 (-2.4500)	66.5720

Πίνακας 2.1 Συνοπτικός πίνακας στατιστικών μέτρων χαρακτηριστικών των κατανομών των βαθμολογιών των ατομικών αποστάσεων με βάση τις συναρτήσεις TF1 και TF2.

Τα πεπτίδια που ανήκουν στους αργούς αναδιπλωτές (slow-folders) όπου το γεγονός αναδίπλωσης συμβαίνει προς το τέλος του χρόνου της προσομοίωσης (Εικόνα 2.6), λαμβάνουν από τις συναρτήσεις TF1 και TF2 ενδιάμεσες βαθμολογίες. Το πεπτίδιο DKWA έχει πάρει από τις υψηλότερες βαθμολογίες [19.2] με βάση τη συνάρτηση TF2 σε σχέση με τους υπόλοιπους “αργούς αναδιπλωτές”, αλλά η βαθμολογία του στο γενικό σύνολο των πεπτιδίων παραμένει



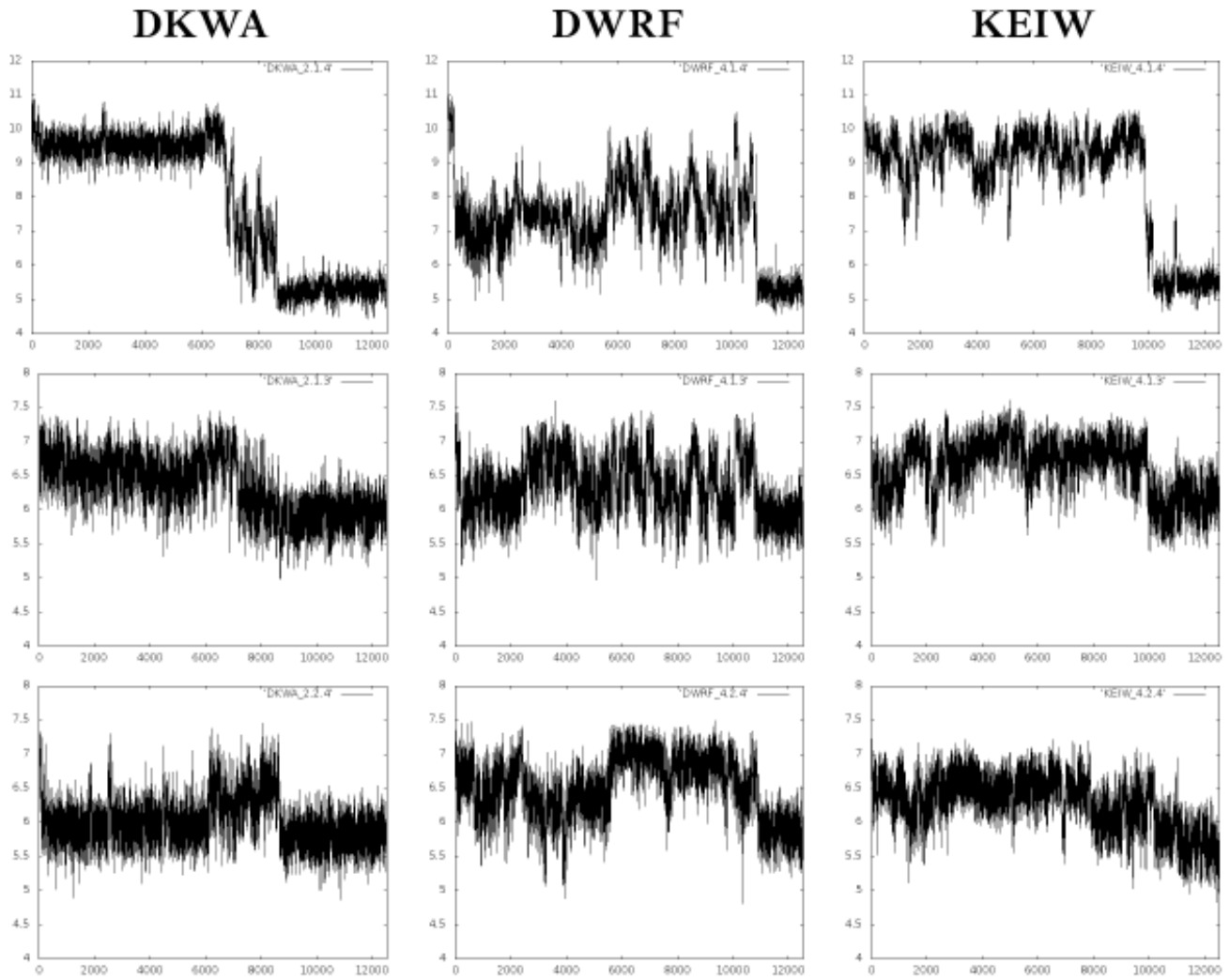
χαμηλή όπως βλέπουμε στον Πίνακα 2.1, μόλις λίγο μεγαλύτερη από 1σ πάνω από το μέσο όρο.



Εικόνα 2.5 Αντιπροσωπευτικά πεπτίδια του πρώτου κύκλου των προσομοιώσεων τα οποία ανήκουν στη κατηγορία των γρήγορων αναδιπλωτών (fast-folders) και τα οποία έδωσαν, την υψηλότερη βαθμολογία με βάση συνάρτηση TF2 [DMWR], την υψηλότερη βαθμολογία με βάση τη συνάρτηση TF1 [WSDK] και τη χαμηλότερη βαθμολογία με βάση τη συνάρτηση TF2 [REWA], αντιστοίχως.

Η σταδιακή πτώση της τιμής της απόστασης 1-4Dist δίνει επίσης ενδιαμέση βαθμολογία [7.6]. Το πεπτίδιο DWRF θα περιμέναμε να πάρει παρόμοιες βαθμολογίες με το DKWA και μάλιστα ελαφρώς υψηλότερη για την απόσταση 1-4Dist. Ωστόσο η βαθμολογία του είναι χαμηλότερη [5.9], υποδεικνύοντας τη μειωμένη διακριτική ικανότητα των συναρτήσεων αυτών για πεπτίδια με σταδιακές μεταπτώσεις στις ατομικές αποστάσεις. Αυτό διαφαίνεται και με το πεπτίδιο

ΚΕΙW το οποίο έλαβε επίσης βαθμολογία κοντά στο μέσο όρο (20.9 για τη συνάρτηση TF2 και

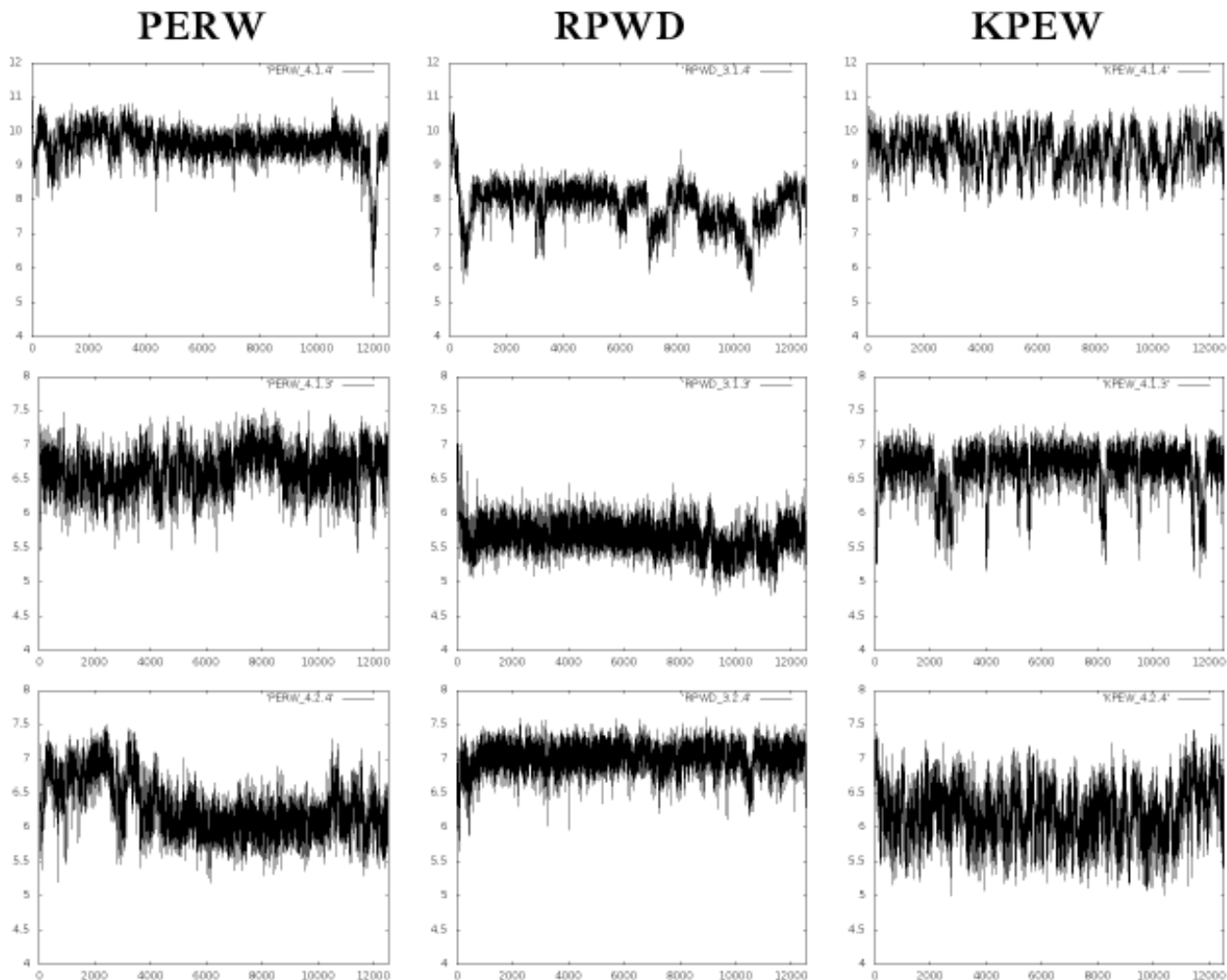


Εικόνα 2.6 Αντιπροσωπευτικά πεπτιδία του πρώτου κύκλου των προσομοιώσεων τα οποία ανήκουν στη κατηγορία των αργών αναδιπλωτών (slow-folders) και τα οποία που έδωσαν, την υψηλότερη βαθμολογία με βάση συνάρτηση TF2 [DKWA], την υψηλότερη βαθμολογία με βάση τη συνάρτηση TF1 [DWRF] και τη χαμηλότερη βαθμολογία με βάση τη συνάρτηση TF2 [KEIW], αντιστοίχως.

4.7 για τη συνάρτηση TF1). Η συμπεριφορά αυτή είδαμε ωστόσο, ότι είναι χαρακτηριστική στις βαθμολογίες με βάση τις ατομικές αποστάσεις και τα πεπτιδία αυτά αποτελούν περισσότερο τον γενικό κανόνα παρά τις εξαιρέσεις.

Η κατάσταση περιπλέκεται ακόμα περισσότερο όταν συναντάμε στην αλληλουχία του πεπτιδίου

την προλίνη. Όπως έχει προαναφερθεί, οι δύο αυτές συναρτήσεις αδυνατούν να διαχωρίσουν,



Εικόνα 2.7 Αντιπροσωπευτικά πεπτίδια του πρώτου κύκλου των προσομοιώσεων τα οποία περιλαμβάνουν προλίνη στην αλληλουχία τους και τα οποία έδωσαν, την υψηλότερη βαθμολογία με βάση συνάρτηση TF1 [PERW], την υψηλότερη βαθμολογία με βάση τη συνάρτηση TF2 [RPWD] και τη χαμηλότερη βαθμολογία με βάση τη συνάρτηση TF2 [KPEW], αντιστοίχως.

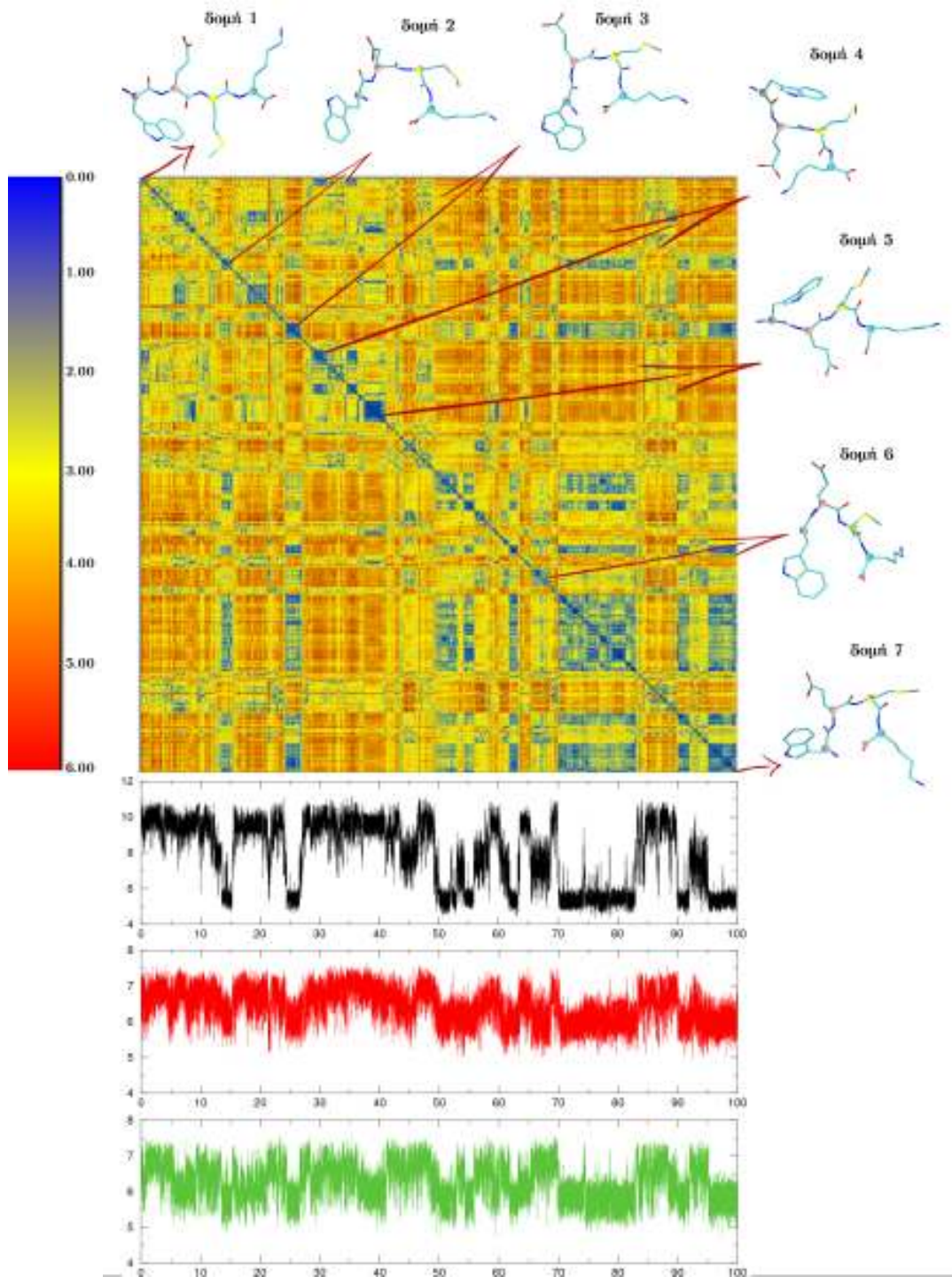
στην παρούσα μορφή τους, τους αναδιπλωτές εκείνους των οποίων οι τιμές των ατομικών αποστάσεων είναι υψηλές αλλά με μικρές διακυμάνσεις. Έτσι βλέπουμε στην Εικόνα 2.7 ότι το πεπτίδιο PERW έλαβε την υψηλότερη βαθμολογία [7.1] με βάση τη συνάρτηση TF1 σε σχέση με τα πεπτίδια που έχουν προλίνη, χωρίς ωστόσο να δείχνει αναδίπλωση με βάση τις γραφικές παραστάσεις της εξέλιξης στο χρόνο των τριών αποστάσεων (Εικόνα 2.7). Το πεπτίδιο RPWD έλαβε την υψηλότερη βαθμολογία [30.1] με βάση τη συνάρτηση TF2 σε σχέση με τα υπόλοιπα

πεπτιδία αυτής της κατηγορίας (6.8 με βάση τη συνάρτηση TF1). Το πεπτιδίο KPEW βρίσκεται δικαίως, με βάση τις γραφικές παραστάσεις, στις τελευταίες θέσεις της βαθμολογικής κατάταξης (1.2 με τη συνάρτηση TF1 και 1.5 με τη συνάρτηση TF2).

Οι βασισμένες σε ατομικές αποστάσεις συναρτήσεις TF1 και TF2 που σχεδιάστηκαν με σκοπό να ανιχνεύουν γεγονότα αναδίπλωσης φέρουν όπως δείξαμε τόσο πλεονεκτήματα όσο και μειονεκτήματα, τα οποία διαφαίνονται όταν εξετάζουμε την κατάταξη του συνόλου των πεπτιδίων. Το γεγονός αυτό μπορεί να οφείλεται σε τρεις λόγους: (1) αδυναμία της δικής μας συνάρτησης να αποδώσει σωστά την πληροφορία των γραφημάτων των αποστάσεων (2) αδυναμία της ίδιας της παραμέτρου των ατομικών αποστάσεων ως αποτελεσματικός κριτής της δημιουργίας γεγονότος αναδίπλωσης και (3) περιορισμένος χρόνος προσομοίωσης.

Όπως προαναφέρθηκε, για την αξιολόγηση της δημιουργίας σταθερής δομής χρειάζεται μία δεύτερη παράμετρος. Η εξέλιξη στο χρόνο της τιμής του RMSD από μία δομή αναφοράς (αρχική δομή, μέση δομή, native δομή) είναι μεταξύ των πιο κοινών μέτρων που χρησιμοποιούνται κατά την ανάλυση των προσομοιώσεων μοριακής δυναμικής. Απουσία όμως γνώσης *a priori* των αναδιπλωμένων διαμορφώσεων των πεπτιδίων, το πληροφοριακό τους περιεχόμενο είναι περιορισμένο. Εναλλακτικά, θα ήταν επιθυμητό να μπορούμε να συγκρίνουμε το σύνολο των διαμορφώσεων που παρατηρούνται στη διάρκεια του τροχιακού και ιδανικά να μπορούμε και να το βαθμολογήσουμε. Έτσι, χρησιμοποιήσαμε δισδιάστατους πίνακες RMSD μεταξύ διαδοχικών δομών του τροχιακού (βλέπε Ενότητα 2.4). Πρόκειται για ένα τετράγωνο συμμετρικό πίνακα του οποίου οι δύο άξονες αντιστοιχούν σε διαδοχικά στιγμιότυπα-δομές του τροχιακού και οι εσωτερικές θέσεις του πίνακα συμπληρώνονται με το RMSD μεταξύ των εκάστοτε δύο δομών. Έτσι η διαγώνιος του πίνακα είναι μηδενική. Ο υπολογισμός του RMSD αφορά όλα τα βαριά άτομα (δηλαδή όλα τα άτομα πλην των πρωτονίων), και συνεπώς είναι ενδεικτικός της κινητικότητας τόσο του σκελετού όσο και των πλευρικών ομάδων του πεπτιδίου. Σκοπός είναι η αξιολόγηση της δημιουργίας σταθερής δομής και ακολούθως η κατάταξη των πεπτιδίων βάσει αυτής. Στις Εικόνες 2.8-2.11 που ακολουθούν παρουσιάζονται οι πίνακες RMSD όπως υπολογίστηκαν για τα τέσσερα αντιπροσωπευτικά τετραπεπτιδία που χρησιμοποιήσαμε κατά την παρουσίαση των συναρτήσεων TF1 και TF2.

Στην Εικόνα 2.8 βλέπουμε τη γραφική απεικόνιση του πίνακα RMSD για το πεπτιδίο WEMK, το οποίο όπως έχει συζητηθεί είναι ένα ασταθές πεπτιδίο. Ο χρωματισμός του πίνακα ακολουθεί την αύξηση της τιμής του RMSD όπως φαίνεται στο ένθετο αριστερά.



Εικόνα 2.8 Γραφική απεικόνιση διδιάστατων πινάκων RMSD μεταξύ όλων των πιθανών δομών του τροχιακού ενός τετραπεπτιδίου (WEMK). Οι τρεις ατομικές αποστάσεις επισημαίνονται από κάτω για λόγους σύγκρισης χρησιμοποιώντας τον χρωματικό κώδικα της Εικόνας 2.1 Γύρω από τον πίνακα RMSD επισημαίνονται αντιπροσωπευτικά στιγμιότυπα-δομές του τροχιακού.

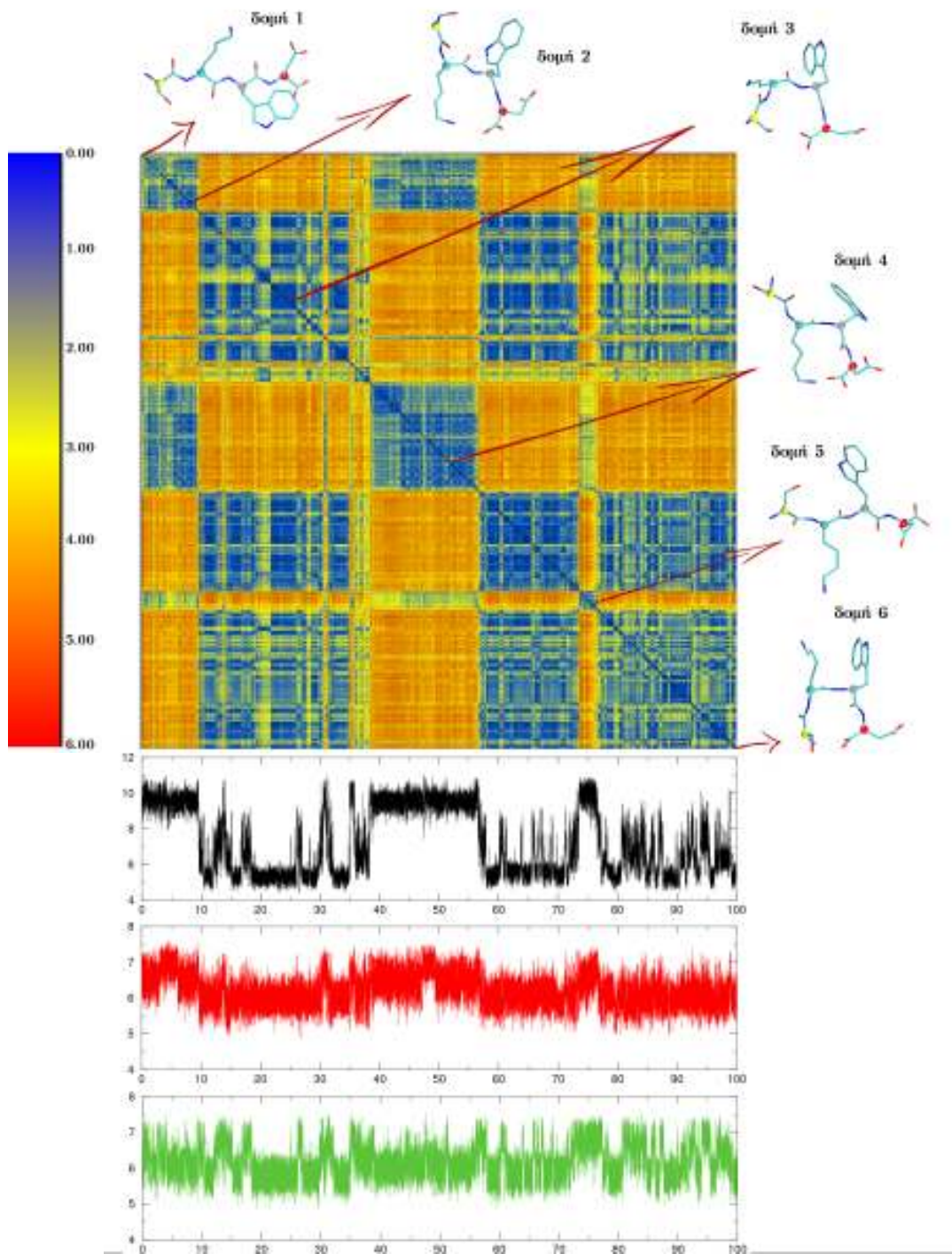
Με βάση τη χρωματική απεικόνιση του πίνακα RMSD, το ζητούμενο μας, που είναι η δημιουργία ενός μεγάλου, σταθερού και συμπαγούς cluster δομών με μικρό μεταξύ τους RMSD, οπτικοποιείται με την παρατήρηση ενός συμπαγούς μπλε τετραγώνου ( $RMSD < 2.0 \text{ \AA}$ ) επί της διαγωνίου. Βλέπουμε για παράδειγμα στην Εικόνα 2.8 πως τα cluster δομών που δημιουργούνται είναι σποραδικά και μικρής διάρκειας. Η δημιουργία cluster συνοδεύεται πάντα από συγχρονισμένες μεταβολές στις τρεις ατομικές αποστάσεις. Όμως στις ατομικές αποστάσεις η δημιουργία δομής ήταν δυσδιάκριτη με βάση τις τιμές των αποστάσεων καθώς δεν συνοδεύεται απαραίτητα από πτώση της τιμής (δομές 2, 3, 7) αλλά συναντάμε δομές και σε υψηλές τιμές (δομές 4, 5). Με τους πίνακες RMSD η δημιουργία δομής είναι όχι μόνο εμφανής, αλλά μπορούμε να παρατηρήσουμε εύκολα τη διάρκεια παραμονής στην τρέχουσα διαμόρφωση, το χρόνο εμφάνισης της στη διάρκεια του τροχιακού αλλά και τυχόν επανεμφάνισή της. Η επανεμφάνιση μίας διαμόρφωσης διακρίνεται από τις οριζόντιες και κάθετες γραμμές εκτός της διαγωνίου (cross-vectors) όπως για παράδειγμα μεταξύ των δομών 2, 3 και 7 με τις δομές 3 και 7 να συνδέονται με μικρότερες τιμές RMSD (περισσότερο μπλε).

Στην Εικόνα 2.9 βλέπουμε τα αντίστοιχα αποτελέσματα για το πεπτιδίο SKWD το οποίο δείχνει αναδίπλωση αλλά με πολλαπλά γεγονότα αναδίπλωσης/αποδιάταξης. Το πεπτιδίο αυτό έλαβε μία ενδιάμεση βαθμολογία με βάση τις συναρτήσεις TF1 [5.2] και TF2 [15.1] και ο πίνακας RMSD μας δείχνει το λόγο. Το πεπτιδίο περνάει από δύο κύριες διαμορφώσεις, δομές 2/4 και δομές 3/6 και μία τρίτη παροδική, τη δομή 5. Ωστόσο όλα τα cluster εμφανίζουν μεγάλη διασπορά, το οποίο φαίνεται και από το εύρος των διακυμάνσεων των ατομικών αποστάσεων.

Μία από τις κύριες αδυναμίες των βασισμένων σε ατομικές αποστάσεις συναρτήσεων TF1 και TF2 είναι η λανθασμένη βαθμολόγηση των πεπτιδίων που περιέχουν προλίνη όπως το RDWP (4.4 με βάση τη συνάρτηση TF1 και 17.1 με βάση τη συνάρτηση TF2). Η αδυναμία αυτή ξεπερνιέται με τους πίνακες RMSD όπως διαφαίνεται στην Εικόνα 2.10. Το πεπτιδίο RDWP σχηματίζει ένα αρκετά συμπαγές cluster δομών (δομές 2/3/5) για μεγάλο ποσοστό του χρόνου προσομοίωσης ενώ περνάει παροδικά και από άλλες διαμορφώσεις (όπως η δομή 4).

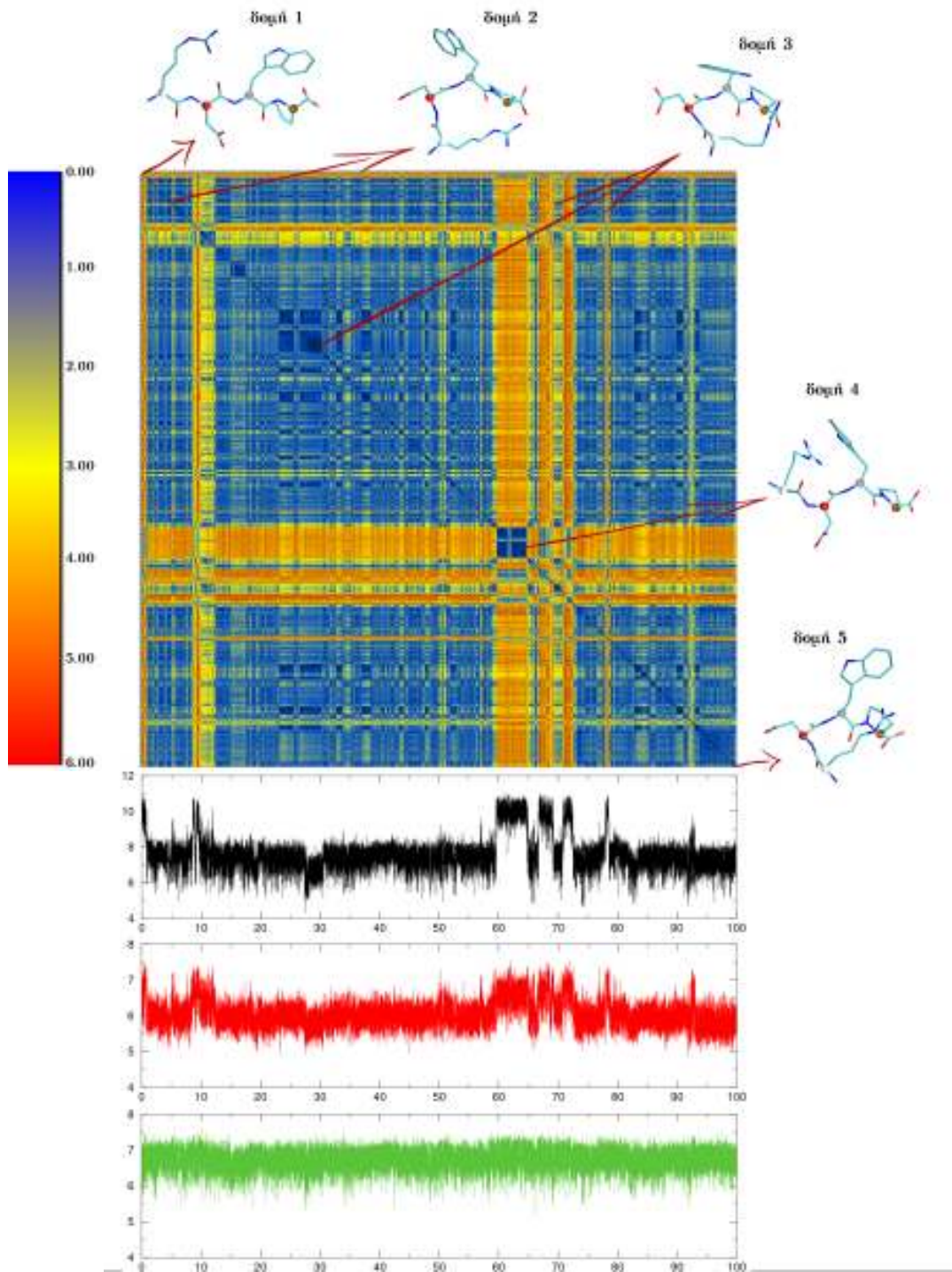
Το πεπτιδίο DRNW (Εικόνα 2.11) περνάει από μία σειρά διαμορφώσεων (δομές 2/3/4/5) και καταλήγει σε ένα συμπαγές cluster δομών (δομές 6 και 7). Ωστόσο, βαθμολογήθηκε χαμηλά από τις συναρτήσεις TF1 [4.0] και TF2 [13.7] επιβεβαιώνοντας το γεγονός ότι οι ατομικές αποστάσεις διακρίνουν τα γεγονότα αναδίπλωσης αλλά δεν αξιολογούν σωστά τη δημιουργία σταθερής δομής.



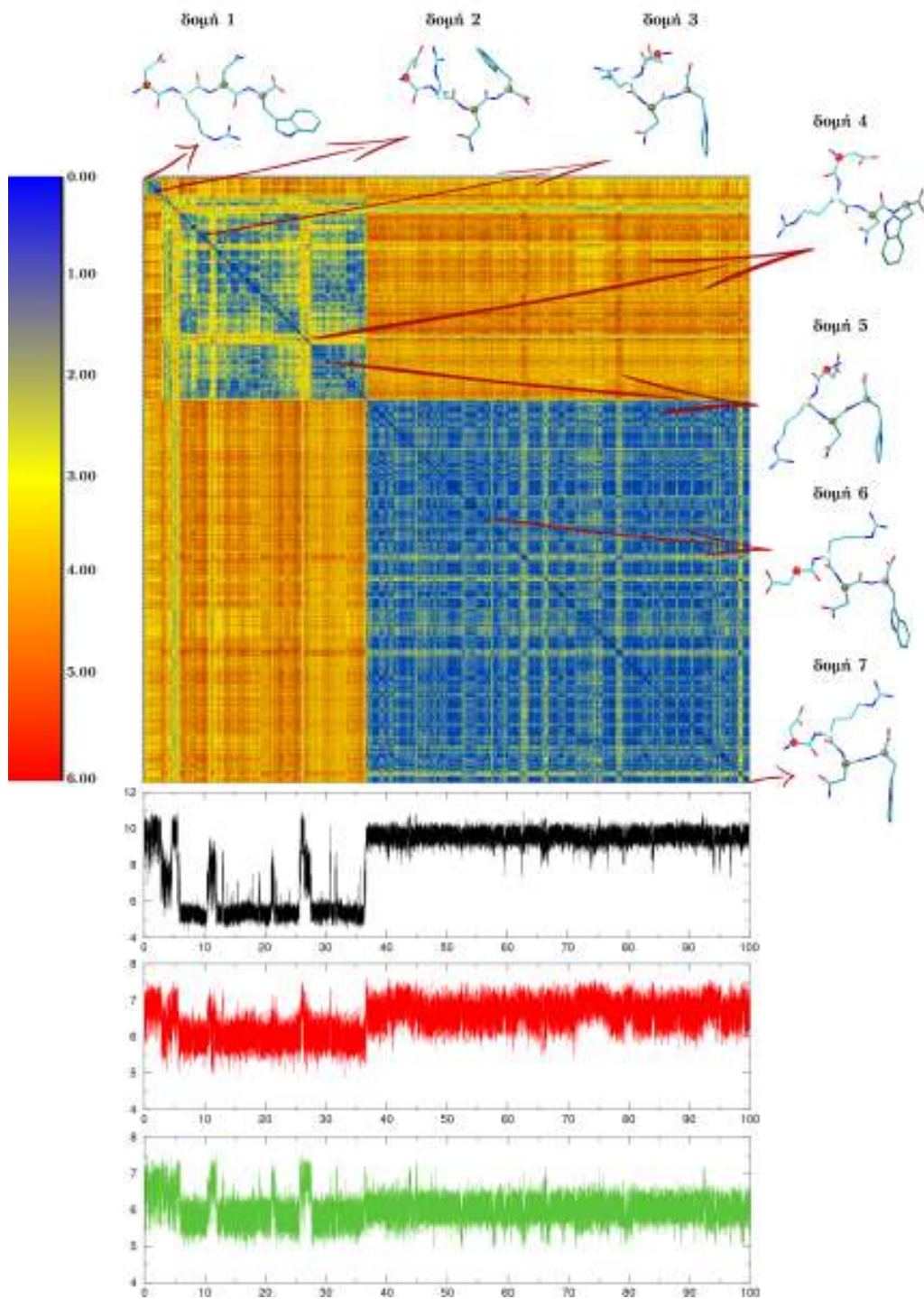


Εικόνα 2.9 Γραφική απεικόνιση δισδιάστατων πινάκων RMSD μεταξύ όλων των πιθανών δομών του τροχιακού ενός τετραπεπτιδίου (SKWD). Οι τρεις ατομικές αποστάσεις επισημαίνονται από κάτω για λόγους σύγκρισης χρησιμοποιώντας τον χρωματικό κώδικα της Εικόνας 2.1 Γύρω από τον πίνακα RMSD επισημαίνονται αντιπροσωπευτικά στιγμιότυπα-δομές του τροχιακού.





Εικόνα 2.10 Γραφική απεικόνιση δισδιάστατων πινάκων RMSD μεταξύ όλων των πιθανών δομών του τροχιακού ενός τετραπεπτιδίου (RDWP). Οι τρεις ατομικές αποστάσεις επισημαίνονται από κάτω για λόγους σύγκρισης χρησιμοποιώντας τον χρωματικό κώδικα της Εικόνας 2.1 Γύρω από τον πίνακα RMSD επισημαίνονται αντιπροσωπευτικά στιγμιότυπα-δομές του τροχιακού.



Εικόνα 2.11 Γραφική απεικόνιση διδιάστατων πινάκων RMSD μεταξύ όλων των πιθανών δομών του τροχιακού ενός τετραπεπτιδίου (DRNW). Οι τρεις ατομικές αποστάσεις επισημαίνονται από κάτω για λόγους σύγκρισης χρησιμοποιώντας τον χρωματικό κώδικα της Εικόνας 2.1 Γύρω από τον πίνακα RMSD επισημαίνονται αντιπροσωπευτικά στιγμιότυπα-δομές του τροχιακού.

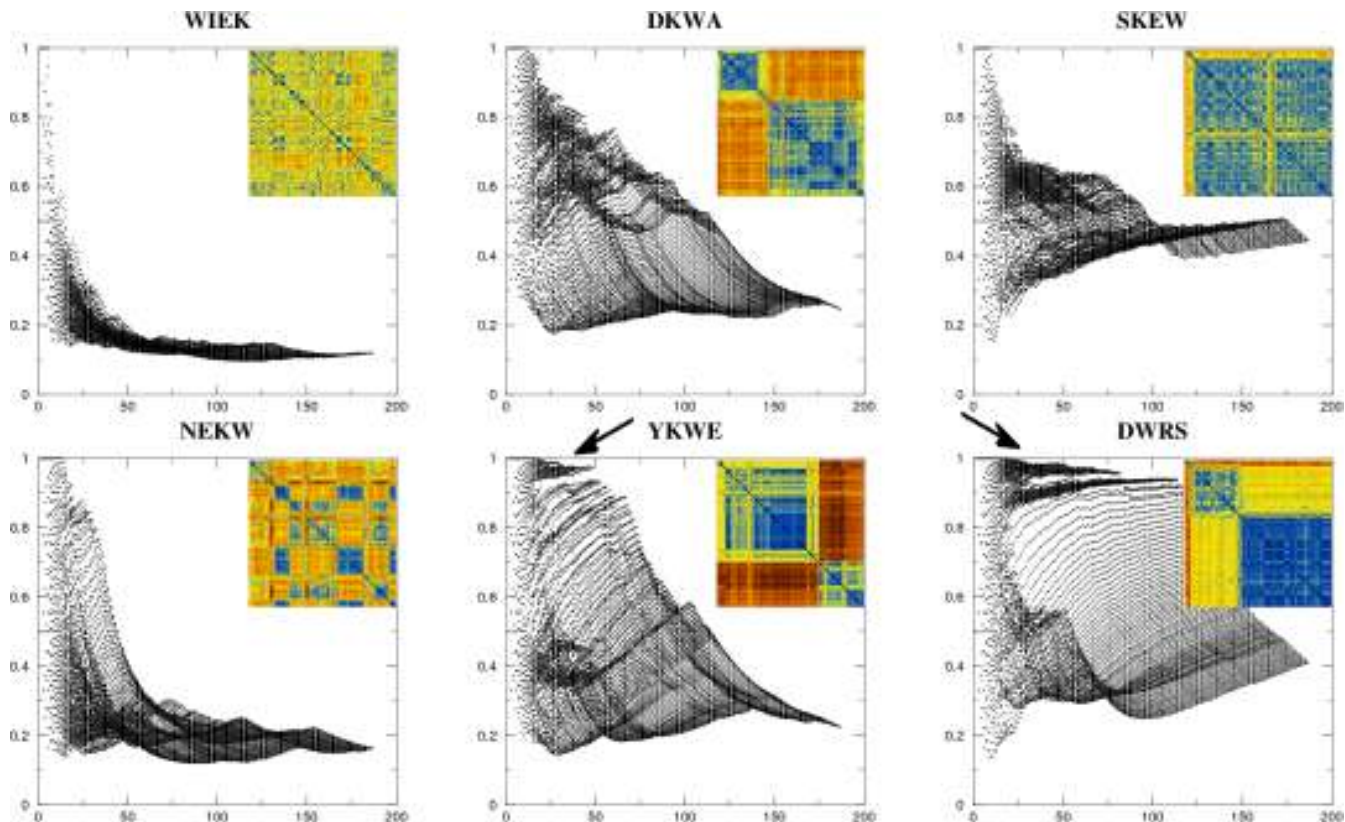
Το κριτήριο λοιπόν με βάση το οποίο θέλουμε να κατατάξουμε τα πεπτίδια είναι ως προς τη δημιουργία σταθερής δομής, δηλαδή την παρουσία ενός μεγάλου, σταθερού και συμπαγούς cluster δομών με μικρό RMSD ή πιο απλά τη δημιουργία ενός μεγάλου και συμπαγούς μπλε τετραγώνου επί της διαγωνίου. Ο αλγόριθμος που περιγράφεται ακολούθως ονομάστηκε “επεκτεινόμενα παράθυρα” για να περιγράψει την κεντρική ιδέα: σε ένα διδιάστατο πίνακα, κινούμενοι πάνω στη διαγώνιο, δημιουργούμε υποθετικά τετράγωνα ποικίλου, αυξομειούμενου μεγέθους για τα οποία μετράμε το πλήθος των μπλε pixels, δηλαδή των καταχωρήσεων του πίνακα με τιμή RMSD μικρότερη από 2.0Å. Ο αλγόριθμος περιγράφεται ως εξής:

- ☐ Ανάγνωση των τιμών του πίνακα RMSD, πλήθους  $N \times N$ .
- ☐ Μετατροπή σε δυαδική μορφή, όπου 0 αντιστοιχεί σε  $\text{RMSD} \geq 2.0\text{\AA}$  και 1 αντιστοιχεί σε  $\text{RMSD} < 2.0\text{\AA}$ .
- ☐ Υπολογισμός των επί της εκατό (%) καταχωρήσεων με τιμή 1 για κάθε υποθετικό συμμετρικό τετράγωνο διάστασης  $(N-m) \times (N-m)$ , όπου  $N$  είναι η διάσταση του αρχικού πίνακα και το  $m$  παίρνει τιμές  $2 < m < N-2$ .
- ☐ Για την κατανομή των επί της εκατό (%) μπλε pixels ως προς την διάσταση κάθε υποθετικού τετραγώνου του προηγούμενου βήματος, υπολογισμός της διάμεσης τιμής (median) και της τιμής με τη μεγαλύτερη συχνότητα εμφάνισης, η αλλιώς επικρατούσας τιμής (mode).
- ☐ Επιστροφή ως βαθμολογίας του γινόμενου της επικρατούσας επί τη διάμεσο τιμή.

Στην Εικόνα 2.12 βλέπουμε γραφικές παραστάσεις των κατανομών της διάστασης του υποθετικού τετραγώνου ως προς το ποσοστό των μπλε pixels για 6 αντιπροσωπευτικά τετραπεπτίδια (Κεφάλαιο 3, Ενότητα 3.4) μαζί με τη γραφική απεικόνιση (ένθετο) των RMSD πινάκων. Η χρησιμότητα της παραμέτρου του ποσοστού των μπλε pixels διαφαίνεται καθαρά από την ενδιαφέρουσα μορφή των κατανομών αυτών. Βλέπουμε ότι τα μπλε συμπαγή cluster των πινάκων RMSD αντικατοπτρίζονται στην κορυφή των κατανομών (μαύρα βέλη στην εικόνα) και μάλιστα η πυκνότητα των σημείων είναι ανάλογη του μεγέθους του cluster.

Αναζητήσαμε λοιπόν κάποια στατιστικά μέτρα των κατανομών αυτών τα οποία να είναι ικανά να κατατάσσουν τα πεπτίδια σωστά ως προς τη δημιουργία σταθερής δομής. Αρχικά υπολογίσαμε τη μέση τιμή, τη διάμεσο τιμή και την επικρατούσα τιμή. Προκειμένου να βρούμε το βέλτιστο συνδυασμό των παραμέτρων αυτών υπολογίσαμε τους γραμμικούς συντελεστές συσχέτισης της

κατάταξης των πεπτιδίων χρησιμοποιώντας κάθε μία παράμετρο αλλά και όλα τα μεταξύ τους πιθανά γινόμενα (Πίνακας 2.2).



Εικόνα 2.12 Γραφικές παραστάσεις του ποσοστού των μπλε pixels (άξονας ψ) ως προς την διάσταση του εκάστοτε υποθετικού τετραγώνου (άξονας χ). Εδώ βλέπουμε 6 αντιπροσωπευτικά πεπτίδια τα οποία παραθέτονται από αριστερά προς τα δεξιά και από πάνω προς τα κάτω με βάση τη βαθμολογία του πίνακα RMSD. Η γραφική απεικόνιση του πίνακα RMSD που αντιστοιχεί σε κάθε γραφική παράσταση φαίνεται στο ένθετο πάνω δεξιά, με κοινή χρωματική κλίμακα.

Οι περισσότεροι συντελεστές είναι πολύ υψηλοί (~0.9) και οι κατατάξεις των πεπτιδίων διαφέρουν μόνο κατά μετατοπίσεις της τάξης των 2-4 θέσεων. Για να επιλέξουμε τον πιο αντιπροσωπευτικό συνδυασμό υπολογίσαμε ένα δενδρόγραμμα με τους γραμμικούς συντελεστές συσχέτισης στη θέση των αποστάσεων (Εικόνα 2.13). Οι πιο απομακρυσμένοι κλάδοι του δενδρογράμματος αντιπροσωπεύουν τους συνδυασμούς των παραμέτρων που οδηγούν στις περισσότερο διαφορετικές κατατάξεις των πεπτιδίων. Έτσι ο συνδυασμός επιλέγεται ή απορρίπτεται εξετάζοντας τα πεπτίδια τα οποία διαφέρουν σημαντικά στη θέση κατάταξης. Με



βάση τη διαδικασία αυτή καταλήξαμε ότι το γινόμενο της διαμέσου τιμής με την επικρατούσα τιμή (Median x Mode) περιγράφει καλύτερα την κατάταξη των πεπτιδίων βάσει της βαθμολογίας των πινάκων RMSD.

-	BxAxMxD	BxAxM	BxMxD	AxMxD	BxA	BxM	BxD	AxM	AxD	MxD
BxAxMxD	1	0.978	0.987	0.945	0.934	0.940	0.927	0.893	0.850	0.869
BxAxM		1	0.973	0.930	0.978	0.984	0.921	0.942	0.839	0.862
BxMxD			1	0.979	0.955	0.960	0.973	0.928	0.915	0.931
AxMxD				1	0.929	0.933	0.979	0.943	0.968	0.980
BxA					1	0.997	0.943	0.971	0.876	0.893
BxM						1	0.941	0.974	0.873	0.894
BxD							1	0.929	0.970	0.975
AxM								1	0.908	0.926
AxD									1	0.996
MxD										1

-	B	A	M	D
B	1	0.937	0.934	0.788
A		1	0.988	0.837
M			1	0.827
D				1

Πίνακας 2.2 Γραμμικοί συντελεστές συσχέτισης της κατάταξης των πεπτιδίων με βάση κάθε ένα από τα στατιστικά μέτρα και όλων των πιθανών γινόμενων. Με **B** (total blues) υποδηλώνεται το πλήθος των μπλε pixels σε ολόκληρο τον πίνακα RMSD, με **A** (average) υποδηλώνεται η μέση τιμή της κατανομής, με **M** (median) υποδηλώνεται η διάμεσος τιμή και με **D** (mode) υποδηλώνεται η επικρατούσα τιμή. Οι μικρότερες τιμές των συντελεστών συσχέτισης υποδεικνύονται με κόκκινο χρώμα.

Για να καταδείξουμε την αποτελεσματικότητα του αλγόριθμου των “επεκτεινομένων παραθύρων” να αξιολογούν τους πίνακες RMSD και να κατατάσσουν σωστά τα πεπτίδια ως προς τη δημιουργία σταθερής δομής παραθέτουμε στην Εικόνα 2.13 τα αποτελέσματα από την εφαρμογή

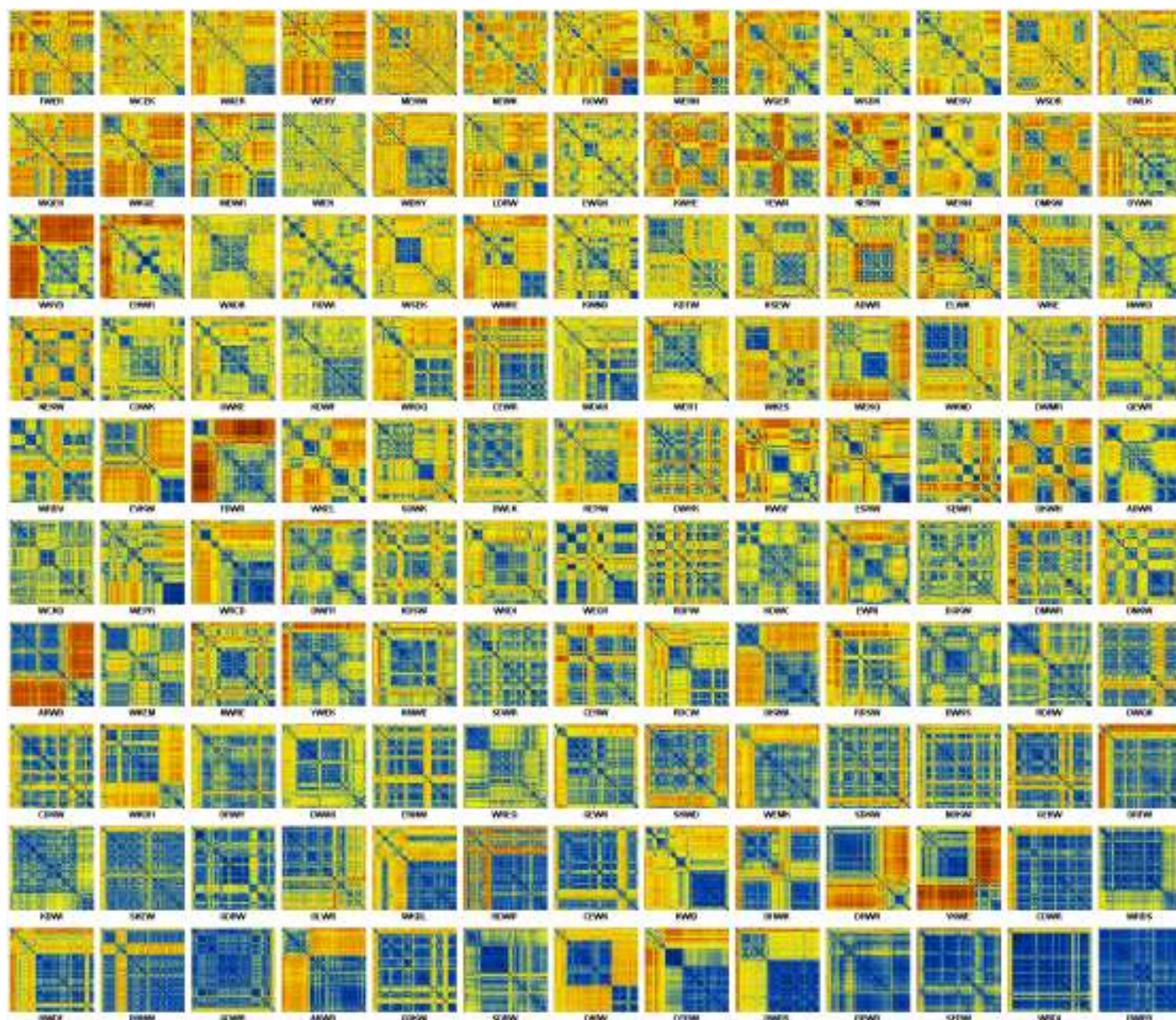
του αλγόριθμου σε ένα σύνολο 130 τετραπεπτιδίων (Κεφάλαιο 3, Ενότητα 3.4). Ο αλγόριθμος είναι ενσωματωμένος στο Perl script με τη μορφή υπορουτίνας, ενώ ο κώδικας παραθέτεται στο Παράρτημα (#10, systematic.pl, subroutine *Expanding\_Windows*).



Εικόνα 2.13 Δενδρογράμμα των κατανομών των πεπτιδίων, χρησιμοποιώντας τους γραμμικούς συντελεστές συσχέτισης του Πίνακα 2.2.

Η παράμετρος αυτή λοιπόν κατατάσσει ικανοποιητικά τους πίνακες RMSD. Προκειμένου να εξετάσουμε κατά πόσο είναι επαρκής από μόνη της ως παράμετρος εκτίμησης της αναδιπλωσιμότητάς των πεπτιδίων εξετάσαμε μία πληθώρα επιπλέον παραμέτρων οι οποίες υπολογίζονται όλες μέσω του Perl script (Ενότητα 2.2), και συνοψίζονται (μαζί με τους συντελεστές συσχέτισης της εκάστοτε κατάταξης των πεπτιδίων με βάση κάθε παράμετρο) στον Πίνακα 2.3.

Τα στοιχεία III-VII του Πίνακα 2.3 αντιστοιχούν στις βαθμολογίες της συνάρτησης TF1 για τις τρεις ατομικές αποστάσεις, στη βαθμολογία της συνάρτησης TF2 και στη βαθμολογία της συνάρτησης TF1 για τη γυροσκοπική ακτίνα. Οι επιπλέον υπολογισμοί βασίζονται στην ανάλυση Cartesian-PCA (Ενότητα 2.4) χρησιμοποιώντας όλα τα βαριά άτομα ώστε να ληφθούν υπόψη και οι πλευρικές ομάδες.



Εικόνα 2.14 Γραφική απεικόνιση των πινάκων RMSD των 130 τετραπεπτιδίων του δεύτερου κύκλου των προσομοιώσεων (διάρκειας 30ns). Η σειρά των πεπτιδίων από αριστερά προς τα δεξιά και από πάνω προς τα κάτω ακολουθεί τη βαθμολογία τους με βάση τον αλγόριθμο των “επεκτεινομένων παραθύρων”. Η χρωματική κλίμακα είναι κοινή και κυμαίνεται από σκούρο μπλε (0Å) έως σκούρο κόκκινο (6Å).

Από την ανάλυση αυτή, δώσαμε έμφαση στην εντροπία της κατανομής των τριών κυρίαρχων principal components (στοιχείο VIII), στον αριθμό από cluster (στοιχείο IX) και στην κατοχή (σε frames) του κυρίαρχου cluster (στοιχείο X). Για την κυρίαρχη ομάδα δομών διατηρήσαμε ένα αρχείο PDB με χαρακτηριστικές δομές (στιγμιότυπα) σε υπέρθεση και υπολογίσαμε τις μέσες



ατομικές διακυμάνσεις από τη μέση δομή του cluster, για συγκεκριμένα σύνολα ατόμων, όπως για όλα τα βαριά άτομα (στοιχείο XI), για τα άτομα της πλευρικής ομάδας της τρυπτοφάνης (στοιχείο XII), για τα άτομα του πεπτιδικού σκελετού (στοιχείο XIII) και για τα άτομα των υπόλοιπων πλευρικών ομάδων (στοιχείο XIV). Οι παράμετροι αυτές είναι ενδεικτικές της κινητικότητας διαφορετικών τμημάτων της δομής. Το εναπομένον στοιχείο του πίνακα είναι η βαθμολογία του πίνακα RMSD, από την εφαρμογή του αλγόριθμου των “επεκτεινομένων παραθύρων”.

-	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV
II	1	0.244	0.098	0.290	0.193	0.053	-0.460	-0.420	0.360	-0.519	-0.380	-0.316	-0.557
III		1	0.332	0.021	0.887	0.161	0.142	0.026	-0.206	-0.219	-0.186	-0.346	-0.137
IV			1	0.073	0.369	0.064	-0.129	-0.003	0.091	-0.061	0.054	-0.202	-0.054
V				1	-0.035	0.332	-0.356	-0.440	0.270	-0.244	-0.202	-0.123	-0.308
VI					1	0.088	0.196	0.095	-0.266	-0.240	-0.204	-0.406	-0.111
VII						1	0.094	0.021	-0.122	-0.197	-0.073	-0.245	-0.253
VIII							1	0.523	-0.781	0.115	0.028	-0.144	0.228
IX								1	-0.619	0.129	0.174	0.020	0.146
X									1	0.155	0.144	0.238	0.050
XI										1	0.802	0.746	0.906
XII											1	0.495	0.559
XIII												1	0.689
XIV													1

Πίνακας 2.3 Παράμετροι που εξετάστηκαν ως δυνητικοί εκτιμητές της αναδιπλωσιμότητας των πεπτιδίων της παρούσας μελέτης και οι γραμμικοί συντελεστές συσχέτισης μεταξύ των κατανομών των πεπτιδίων από την εφαρμογή κάθε μεμονωμένης παραμέτρου. Με κόκκινο χρώμα υποδεικνύεται η αρνητική συσχέτιση και με γκρι η θετική. Η μέγιστη και ελάχιστη συσχέτιση που βρέθηκε μεταξύ δύο παραμέτρων υπογραμμίζεται με κίτρινο χρώμα, ενώ με πορτοκαλί υποδεικνύονται οι παράμετροι που συμμετέχουν στην συνάρτηση TF3. Κάθε στήλη του πίνακα αντιστοιχεί στις στήλες του αρχείου των αποτελεσμάτων (με τον ίδιο αύξοντα αριθμό), όπως περιγράφεται στην Ενότητα 2.2 (σελ. 29-30).

Βλέπουμε λοιπόν με βάση τον Πίνακα 2.3 ότι η βαθμολογία του πίνακα RMSD (στοιχείο II) δεν εμφανίζει υψηλό συντελεστή συσχέτισης με τις υπόλοιπες παραμέτρους ενώ η πληροφοριακή αξία της παραμέτρου είναι υψηλή. Την μεγαλύτερη συσχέτιση (σημειώνεται με κίτρινη

υπογράμμιση στον Πίνακα 2.3) στη βαθμολογική κατάταξη των πεπτιδίων, την παρατηρούμε μεταξύ των βαθμολογιών που βασίζονται στις ατομικές αποστάσεις (στοιχεία III και VI) και μεταξύ των βαθμολογιών οι οποίες βασίζονται στις ατομικές διακυμάνσεις μεταξύ διαφορετικών ομάδων ατόμων (στοιχεία XI-XIV). Η υψηλότερη αρνητική συσχέτιση παρουσιάζεται μεταξύ του αριθμού από frames της κυρίαρχης ομάδας δομών και της εντροπίας της κατανομής των τριών κυρίαρχων principal components (στοιχεία VIII και X). Ένα δευτερεύον συμπέρασμα που προκύπτει είναι η μεγαλύτερη επίδραση που ασκεί η απόσταση μεταξύ ατόμων Ca 1-4 στην συνάρτηση TF2, ενώ οι υπόλοιπες ατομικές αποστάσεις χαρακτηρίζονται από περισσότερο ήπιες μεταβολές και έχουν μικρότερη επίδραση, όπως έχουμε προαναφέρει.

Επειδή οι συναρτήσεις που βασίζονται στις ατομικές αποστάσεις εμφανίζουν τις αδυναμίες που αναδείξαμε και αναλύονται εκτενώς στα κεφάλαια που ακολουθούν, στραφήκαμε προς το το συνδυασμό παραμέτρων με την πιο υψηλή πληροφοριακή αξία. Η παράμετρος II που αφορά τη βαθμολογία του πίνακα RMSD φέρει πληροφορία για τη σταθεροποίηση της διαμόρφωσης τόσο του πεπτιδικού σκελετού, όσο και των πλευρικών ομάδων. Η παράμετρος XII είναι ενδεικτική της κινητικότητας της πλευρικής ομάδας της τρυπτοφάνης, η οποία συνιστά και το πιο ογκώδες αμινοξύ στα πεπτίδια αυτά, η σταθεροποίηση δε της οποίας είναι κριτικής σημασίας για τη λήψη φάσματος CD ώστε να γίνει σύγκριση των προσομοιώσεων με πειραματικά δεδομένα (Κεφάλαιο 3, Ενότητα 3.1). Έτσι, καταλήξαμε στην ακόλουθη συνάρτηση για την εκτίμηση της δημιουργίας σταθερής δομής :

$$S_i = \frac{\text{Cluster}}{\text{Trp}_{cl} \text{RMSF}} \quad (\text{TF3})$$

όπου:

$S_i$  = η βαθμολογία του πεπτιδίου, ενδεικτική της σταθερότητας της δομής

**Cluster** = η βαθμολογία του πίνακα RMSD (αλγόριθμος των “επεκτεινομένων παραθύρων”)

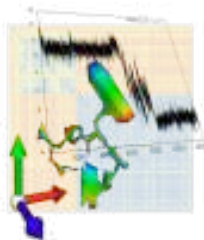
$\text{Trp}_{cl}\text{RMSF}$  = μέση τετραγωνική ρίζα των ατομικών διακυμάνσεων (rmsf) για όλα τα άτομα της πλευρικής ομάδας της τρυπτοφάνης για την κυρίαρχη ομάδα δομών (cluster), όπως αυτή υπολογίζεται από ανάλυση PCA στο Καρτεσιανό σύστημα.

*“ The only difference between a bug and a feature is the documentation. ”*

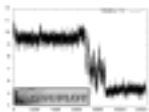
<http://www.gdargaud.net/Humor/Quotes/Programming.html>

*“Life would be so much easier if we only had the source code.”*

<http://www.gdargaud.net/Humor/QuotesProgramming.html>

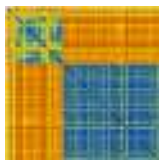


## 2.4 Μέθοδοι ανάλυσης των τροχιακών



**Εξέλιξη ατομικών αποστάσεων**

Όλοι οι υπολογισμοί της εξέλιξης στο χρόνο διαφόρων παραμέτρων, όπως οι ατομικές αποστάσεις και η γυροσκοπική ακτίνα πραγματοποιήθηκαν με το πρόγραμμα CARMA (Version 1.1, Glykos 2006). Η γραφική τους απεικόνιση έγινε με το πρόγραμμα Gnuplot (Version 4.4, patch level 3, Williams & Kelly, 1986-1993, 1998, 2004, 2007-2010) και τη συγγραφή μικρών gnuplot-scripts για τη συστηματική και μαζική παραγωγή των γραφημάτων. Η συγκριτική αντιπαράθεση των γραφικών παραστάσεων με οπτικά μέσα διεξήχθη μέσω browser (Firefox).



**Πίνακες RMSD μεταξύ διαδοχικών δομών του τροχιακού**

Οι πίνακες RMSD είναι τετράγωνοι συμμετρικοί πίνακες μεγέθους συνήθως μέχρι 5000x5000. Στον οριζόντιο και κάθετο άξονα αντιστοιχούν τα στιγμιότυπα (frames) του τροχιακού (trajectory), με συγκεκριμένο κάθε φορά βήμα, ώστε η διάσταση του πίνακα να παραμένει διαχειρίσιμη, από άποψη υπολογιστικής μνήμης. Στο εσωτερικό του πίνακα, σε κάθε θέση έχουμε την υπολογιζόμενη τιμή RMSD μεταξύ των εκάστοτε δομών μετά από υπέρθεση (least-

square fitting) των αντίστοιχων δομών (συνήθως χρησιμοποιώντας όλα τα βαριά άτομα). Το πρόγραμμα που χρησιμοποιούμε για τον υπολογισμό του πίνακα αλλά και τη γραφική του απεικόνιση είναι το CARMA (Glykos 2006) ενώ για την υπέρθεση χρησιμοποιείται (από το CARMA) ο αλγόριθμος του Kabsch (Kabsch 1976, 1994). Οι τετράγωνοι αυτοί πίνακες RMSD μπορούν να χρησιμοποιηθούν για την εξαγωγή ομάδων δομών (cluster analysis) (Gordon et al., 1992).



### Ανάλυση PCA

Η ανάλυση PCA ([principal component analysis](#)) (Pearson 1901) χρησιμοποιείται ευρέως στις προσομοιώσεις μοριακής δυναμικής βιολογικών μακρομορίων για την ανάδειξη καθολικών συσχετιζόμενων κινήσεων (Mayer et al., 2003, Lange et al., 2005). Πρόκειται για μία μαθηματική τεχνική ανάλυσης πολυδιάστατων δεδομένων, όπου, εν ολίγοις, ορίζεται ένα νέο σύστημα συντεταγμένων για τα δεδομένα, με την ιδιαιτερότητα ότι η συνδιακύμανση μεταξύ δύο οιονδίποτε συνιστώσων είναι μηδέν (uncorrelated). Οι νέες συνιστώσες κατατάσσονται με βάση τη διακύμανση (variance) των δεδομένων ως προς τη συνιστώσα. Η ελάττωση των διαστάσεων επιτυγχάνεται δίνοντας έμφαση στις συντεταγμένες με τις μεγαλύτερες διακυμάνσεις και αμελώντας αυτές με τις μικρές. Αρχικά υπολογίζεται ένας πίνακας συνδιακύμανσης  $C$  (covariance matrix), ο οποίος μετασχηματίζεται σε διαγώνιο πίνακα μέσω ενός ορθοκανονικού πίνακα μετασχηματισμού  $R$  (orthonormal transformation matrix). Οι στήλες του πίνακα  $R$  είναι οι λεγόμενες κυρίαρχες συνιστώσες (principal components) ή eigenvectors. Κάθε μία από τις κυρίαρχες συνιστώσες (eigenvectors) συνοδεύεται από μία χαρακτηριστική τιμή, eigenvalue, που ισούται με την διακύμανση (variance explained) από την προβολή των δεδομένων στην διεύθυνση του eigenvector. Έτσι, η σύνθετη κίνηση που πραγματοποιεί το μόριο προβάλλεται σε ένα μικρό άθροισμα απλών (ανεξάρτητων) κινήσεων, που αντιπροσωπεύονται από τους principal components (Ichiye et al., 1991, Garcia 1992, Amadei et al., 1993, Balsera et al., 1996).

Αυτός ο τύπος PCA ονομάζεται και Cartesian-PCA γιατί χρησιμοποιεί Καρτεσιανές συντεταγμένες. Στην περίπτωση συστημάτων που υπάρχουν κινήσεις μεγάλων διακυμάνσεων, έχειδειχθεί ότι δε μπορεί να διαχωρισθεί σωστά η εσωτερική (ενδιαφέρουσα) κίνηση που

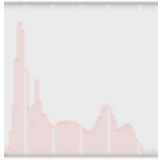
πραγματοποιεί το μόριο από την ολική (ασίμαντη) κίνηση που περιλαμβάνει τις μεταθέσεις και τις περιστροφές, γιατί δεν υπάρχει μία και μόνο ξεκάθαρη διαμόρφωση που να μπορεί να χρησιμοποιηθεί ως αναφορά (reference structure) κατά την υπέρθεση (least-square fitting) για την αφαίρεση των μεταθέσεων/περιστροφών (Hünenberger et al., 1995, Koslover et al., 2007, Altis et al., 2007, Altis 2008). Για τροχιακά αναδίπλωσης πεπτιδίων, ενδείκνυται ένας άλλος τύπος ανάλυσης PCA που βασίζεται στο μετασχηματισμό εσωτερικών συντεταγμένων, των διέδρων γωνιών  $\phi/\psi$ , εξ ου και η ονομασία, Dihedral-PCA (Mu et al., 2005, Maisuradze et al., 2007, Altis et al., 2007, Altis et al., 2008). Η ανάλυση Dihedral-PCA αναδεικνύει το πραγματικό ενεργειακό τοπίο των συστημάτων που παρουσιάζουν μεγάλες διακυμάνσεις (όπως είναι η αναδίπλωση από την εκτεταμένη διαμόρφωση), καθώς διακρίνει διαμορφώσεις παρόμοιας ενεργειακής στάθμης (Mu et al., 2005). Αυτό δίνει μία εμφάνιση με περισσότερες κορυφές (rugged and peaky) σε σχέση με το πιο λείο (smooth) ενεργειακό τοπίο της ανάλυσης Cartesian-PCA (Altis et al., 2007, Altis et al., 2008).

Στα πλαίσια της δικής μας ανάλυσης, πραγματοποιούμε κάθε φορά και τους δύο τύπους αναλύσεων και εξετάζουμε συγκριτικά τα αποτελέσματα και τις ομάδες δομών που προκύπτουν (cluster analysis) καθώς η ανάλυση Dihedral-PCA (στη μορφή που εφαρμόζεται), από τη μία πλευρά δε λαμβάνει υπόψη τις πλευρικές ομάδες, αλλά από την άλλη πλευρά, δεν εξαρτάται από τη δομή αναφοράς που έχουμε επιλέξει για την αφαίρεση των μεταθέσεων/περιστροφών.



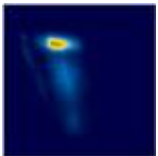
**Word-clouds**

Τα [word-clouds](#) είναι μία μορφή οπτικής απεικόνισης δεδομένων σε μορφή κειμένου, όπου το μέγεθος της γραμματοσειράς κάθε λέξης είναι ανάλογο της συχνότητας εμφάνισης της. Το πρόγραμμα που χρησιμοποιήσαμε για τη δημιουργία των word-clouds είναι το [Wordle](#) (Feinberg J.) σε συνδυασμό με ένα μικρό perl-script για τη δημιουργία κειμένου με τις αλληλουχίες των πεπτιδίων, όπου κάθε αλληλουχία εμφανίζεται με συχνότητα ίση με τη βαθμολογία του πεπτιδίου από την εκάστοτε συνάρτηση εκτίμησης της αναδιπλωσιμότητας.



### Γραφικές παραστάσεις και ιστογράμματα κατανομών

Μέρος των γραφικών παραστάσεων και όλα τα ιστογράμματα κατανομών έγιναν (και υπολογίστηκαν) με το πρόγραμμα [Xmgr](#) (μετέπειτα [Grace](#), Turner P.J. 1996 & Stambulchik E. 1996-1998).



### Γραφική απεικόνιση δισδιάστατων δεδομένων (πινάκων) και πλεγμάτων (grids)

Ο υπολογισμός όλων των δεδομένων σε μορφή [grid](#) έγινε με τα in-house προγράμματα [grid](#) και [make density](#), ελεύθερα διαθέσιμα από τις αντίστοιχες ιστοσελίδες του υπολογιστικού κέντρου [Norma](#). Η γραφική τους απεικόνιση έγινε με το πρόγραμμα CARMA (Glykos 2006).



### Δενδρογράμματα και cluster analysis

Η επεξεργασία τετράγωνων πινάκων και η ομαδοποίηση των δεδομένων (cluster analysis, Gordon et al., 1992) καθώς και η δημιουργία των δενδρογραμμάτων έγιναν με το πρόγραμμα [R](#) (R Development Core Team, 2004), ενώ οποιαδήποτε περαιτέρω επεξεργασία τους έγινε σε επίπεδο [postscript](#).



### Διαγράμματα Venn

Τα διαγράμματα [Venn](#) δημιουργήθηκαν με το πρόγραμμα [SmartDraw](#) στη δοκιμαστική του έκδοση (demo), ενώ η περαιτέρω επεξεργασία τους και προσθήκη των αλληλουχιών έγινε με το πρόγραμμα [Gimp](#).



### Δομές

Για τις αρχικές διαμορφώσεις των πεπτιδίων χρησιμοποιήθηκε το πρόγραμμα [Ribosome](#) (Shrinivasan R.). Όλα τα αρχεία PDB, κατά την ανάλυση των τροχιακών, παρήχθησαν με το πρόγραμμα CARMA (Glykos 2006) και η γραφική τους απεικόνιση έγινε με το πρόγραμμα [VMD](#) (Humphrey et al., 1996).



### Γραφικά

Η επεξεργασία όλων των εικόνων αλλά και η δημιουργία τους, εκτός αν αναφέρεται άλλο πρόγραμμα, έγινε με τα (free, open-source) προγράμματα [VMD](#) (Humphrey et al., 1996), [Rasmol](#) (Bernstein, 1999), [Raster3D](#) (Merritt et al., 1997), [Gimp](#), [ImageMagic](#).



### Adaptive tempering

Για τη διεξαγωγή των προσομοιώσεων με τη μέθοδο adaptive tempering (Ενότητα 4.8) ακολουθήσαμε τα πρωτόκολλα που βρίσκονται στο Παράρτημα (#14, NAMD script, heat.namd και #15, NAMD script, equi.namd) με την προσαρμογή για χρήση του AMBER99SB-ILDN force field όπως περιγράφεται στο Κεφάλαιο 4, Ενότητα 4.5. Η βασική διαφορά της μεθόδου αυτής σε σχέση με τις κλασικές προσομοιώσεις έγκειται στη δυναμική μεταβολή της θερμοκρασίας της προσομοίωσης βάσει της υπολογιζόμενης δυναμικής ενέργειας. Η αναπροσαρμογή της θερμοκρασίας γίνεται μέσω του θερμοστάτη (Langevin thermostat) και όχι μέσω των ταχυτήτων (velocity rescaling).





**Norma**

Computing Cluster

Όλες οι προσομοιώσεις πραγματοποιήθηκαν σε μία συστοιχία υπολογιστών τύπου Beowulf (Norma), βασιζόμενοι στη διανομή Caos NSA GNU/Linux του Τμήματος Μοριακής Βιολογίας και Γενετικής του Δημοκριτείου Πανεπιστημίου Θράκης στην Αλεξανδρούπολη, νομού Έβρου. Η Norma αποτελείται (στην παρούσα χρονική στιγμή) από τετραπύρηνους επεξεργαστές (40 πυρήνες CPU και 6GPGPUs κατανεμημένοι σε 10 κόμβους), με 46Gbytes φυσικής μνήμης, που διασυνδέονται με 1800-24G Gigabit ethernet switch. Το πρόγραμμα που χρησιμοποιήθηκε για όλες τις προσομοιώσεις είναι το NAMD (Kale et al., 1999) στις εκάστοτε αναβαθμισμένες του εκδόσεις (v.2.6, v.2.7, v.2.8). Λόγω του μικρού μεγέθους του συστήματος των πεπτιδίων, οι προσομοιώσεις πραγματοποιήθηκαν είτε σε ένα πυρήνα, είτε σε ένα (τετραπύρηννο) κόμβο, όπου και έδιναν τη μεγαλύτερη απόδοση (benchmarks), ~30ns/μέρα και ~100ns/μέρα αντίστοιχα για ένα σύστημα περίπου 900 ατόμων. Το σύνολο των προσομοιώσεων που παρουσιάζεται στην παρούσα διατριβή αθροίζεται σε 272.46ms για τον οποίο χρειάστηκαν 461 μέρες αθροιστικού φυσικού χρόνου.

*“ Unix is the answer, but only if you phrase the question very carefully. ”*

<http://www.gdargaud.net/Humor/QuotesProgramming.html>

*“ #define QUESTION ((bb) || !(bb)) ” —Shakespeare ”*







*"If we knew what it was we were doing,  
it would not be called research, would it?"*

*Albert Einstein*



### 3.1 Επιλογή τετραπεπτιδικών αλληλουχιών

**Στο** κεφάλαιο αυτό περιγράφουμε την περιπλάνησή μας στον κόσμο των τετραπεπτιδίων προς αναζήτηση “αναδιπλούμενων” αλληλουχιών. Η ερώτηση που μας γοήτευσε ήταν:

*Υπάρχει άραγε ένα τετραπεπτίδιο που να υιοθετεί μία σταθερή δομή σε υδατικό διάλυμα;*

Και αν ναι,

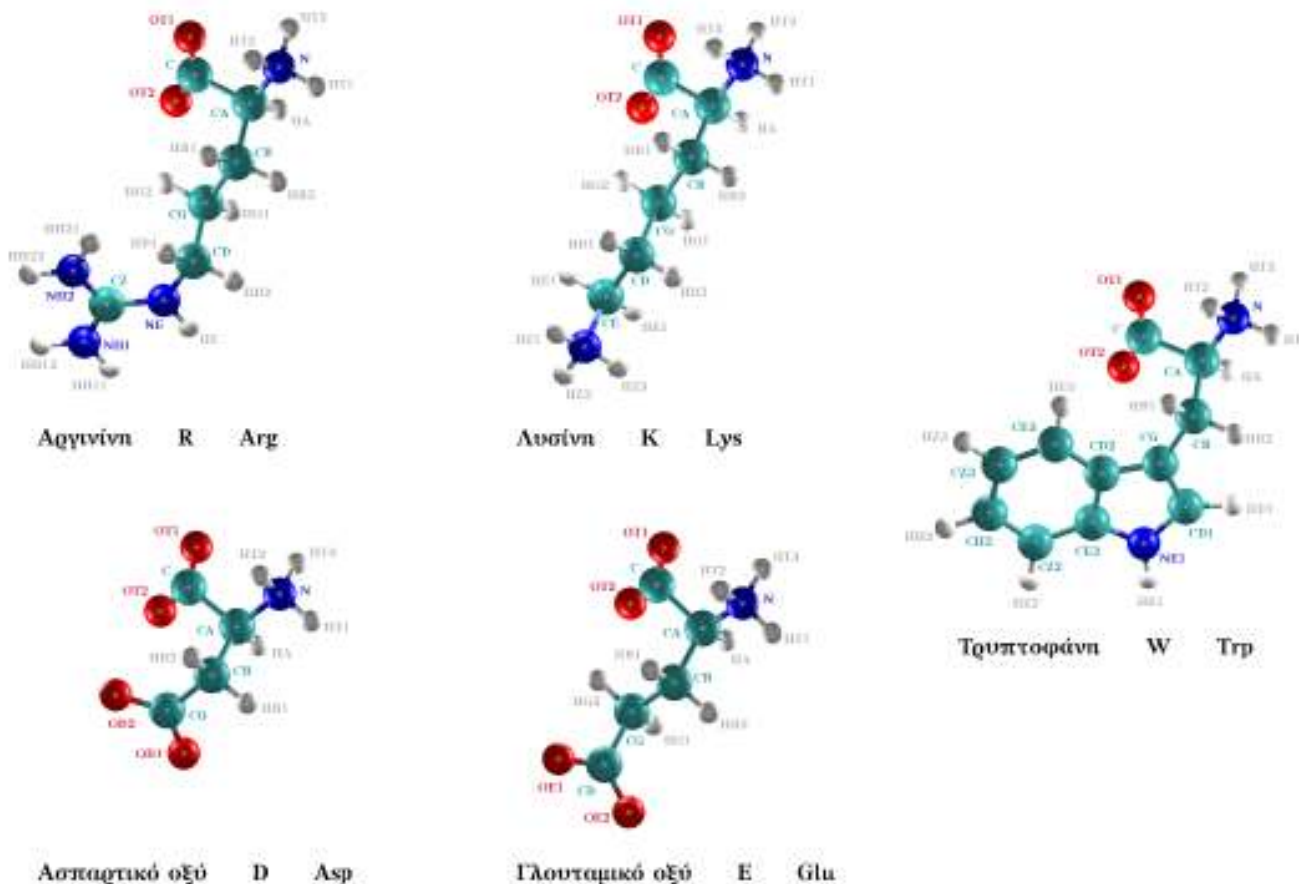
*Μπορούμε να το ταυτοποιήσουμε χρησιμοποιώντας τις προσομοιώσεις μοριακής δυναμικής;*

Ο αριθμός των πιθανών συνδυασμών αλληλουχιών για τα τετραπεπτίδια είναι 160.000. Λόγω των περιορισμών μας σε υπολογιστικό και φυσικό χρόνο, η αναδίπλωση ενός τέτοιου αριθμού πεπτιδίων είναι αδύνατο να μελετηθεί διεξοδικά μέσω προσομοιώσεων μοριακής δυναμικής. Κρίναμε λοιπόν απαραίτητη την επιβολή κάποιων περιορισμών ώστε να μειωθεί ο τελικός αριθμός των υπό μελέτη πεπτιδίων. Στον Πίνακα 3.1, παραθέτουμε διάφορους συνδυασμούς από περιορισμούς και πως αυτοί μεταβάλλουν τον αριθμό των πιθανών πεπτιδικών αλληλουχιών μήκους τεσσάρων καταλοίπων.

Ο πιο καθοριστικός περιορισμός για την μετέπειτα αναδιπλωσιμότητα είναι η επιβολή της παρουσίας ενός καταλοίπου τρυπτοφάνης (Εικόνα 3.1). Η εισαγωγή ενός ή περισσοτέρων καταλοίπων τρυπτοφάνης, φαίνεται πως προσδίδει σε μικρά πεπτίδια δομικά χαρακτηριστικά παρόμοια με αυτά των πρωτεϊνών, με χαρακτηριστικό παράδειγμα τις αλληλουχίες Trpzip (Cochran et al., 2001). Παρότι πρόκειται για ένα ογκώδες αμινοξύ για ενσωμάτωση σε τόσο



μικρές αλληλουχίες, η επιλογή αυτή έχει διττή σημασία.



Εικόνα 3.1 Τρισδιάστατο μοντέλο (cpk), πλήρες όνομα και κώδικας των τριών και του ενός γράμματος των αμινοξέων των οποίων επιβάλλαμε την παρουσία κατά τον σχεδιασμό των τετραπεπτιδικών αλληλουχιών. Τα άτομα άνθρακα απεικονίζονται με γαλάζιο χρώμα, τα άτομα οξυγόνου με κόκκινο, τα άτομα αζώτου με μπλε και τα άτομα υδρογόνου με γκρι χρώμα. Δίπλα σε κάθε άτομο αναφέρεται το όνομα όπως αυτό χρησιμοποιείται στο αρχείο παραμέτρων των force fields.

Ο ινδολικός δακτύλιος μπορεί να αναπτύξει τόσο ηλεκτροστατικές αλληλεπιδράσεις όσο και να πακεταριστεί υδρόφοβα (π-stacking interactions), συμβάλλοντας σημαντικά στη δημιουργία σταθερής δομής (Mahalakshmi et al., 2006, Eidenschink et al., 2009). Επιπλέον, ο αρωματικός δακτύλιος επιτρέπει τη λήψη φάσματος κυκλικού διχροϊσμού (CD spectra). Φάσματα στα 190-250nm (far-UV) χρησιμοποιούνται για τη ανίχνευση σταθερών στοιχείων δευτεροταγούς δομής, ενώ τα φάσματα στα 250-350nm (near-UV) επιτρέπουν την ανίχνευση ισχυρά πακεταρισμένης δομής γύρω από τον αρωματικό δακτύλιο (Greenfield et al., 2006). Εξελίξεις στις πειραματικές τεχνολογίες (Roder et al., 1999), όπως της φασματοσκοπίας φθορισμού, μας επιτρέπουν να

παρακολουθήσουμε πλέον διαδικασίες που λαμβάνουν χώρα σε χρόνο της τάξης των microsecond (Plaxco et al., 1996) και προσφάτως των nanosecond και picosecond (Fierz et al., 2007, Kubelka et al., 2008), επιτρέποντας τη σύγκριση των πειραματικών δεδομένων με τις προσομοιώσεις μοριακής δυναμικής (Feige et al., 2008).

Ο δεύτερος καθοριστικός περιορισμός είναι η επιβολή της παρουσίας φορτισμένων καταλοίπων η οποία συμβάλλει σημαντικά στην αύξηση της διαλυτότητας, ενώ προσφέρει τη δυνατότητα ανάπτυξης σταθεροποιητικών αλληλεπιδράσεων τόσο με άλλες πλευρικές ομάδες όσο και με τον πεπτιδικό σκελετό (Glättli et al., 2005, Wei et al., 2005). Για τον ίδιο λόγο τα άκρα των πεπτιδίων επιλέχθηκαν να παραμείνουν σε ελεύθερη μορφή (uncapped). Η παρουσία των φορτίων αναμένεται να διαδραματίσει ισχυρό ρόλο και στη σταθεροποίηση του καταλοίπου της τρυπτοφάνης.

Όπως φαίνεται στον Πίνακα 3.1, η παρουσία μίας τρυπτοφάνης και δύο φορτισμένων καταλοίπων αντιθέτου φορτίου οδηγεί στο μικρότερο πλήθος αλληλουχιών. Τα 1.440 τετραπεπτίδια μελετήθηκαν ως προς την αναδίπλωσή τους με προσομοιώσεις μοριακής δυναμικής με απώτερο σκοπό την ανάπτυξη μίας μεθοδολογίας για την ταυτοποίηση δυνητικά αναδιπλούμενων πεπτιδίων.

Αριθμός Τετραπεπτιδίων	Παράμετροι				
	ΟΛΑ	1 Trp	ΟΛΑ ΑΑ ΔΙΑΦΟΡΕΤΙΚΑ	2 ΘΕΤΙΚΑ ΦΟΡΤΙΣΜΕΝΑ	1 ΘΕΤΙΚΑ - 1 ΑΡΝΗΤΙΚΑ ΦΟΡΤΙΣΜΕΝΟ
160.000	X				
116.280			X		
27.436		X			
24.576				X	
23.256		X	X		
12.288					X
2.880		X		X	
2.160		X	X	X	
1.440		X			X

Πίνακας 3.1 Αριθμός πιθανών τετραπεπτιδικών αλληλουχιών και περιοριστικές παράμετροι στην επιλογή αμινοξικών καταλοίπων. Η παρουσία τρυπτοφάνης και φορτισμένων καταλοίπων αναμένεται να ενισχύσει τη δημιουργία αλληλεπιδράσεων και να αυξήσει τη διαλυτότητα των πεπτιδίων.

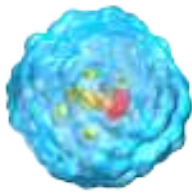
*"Give me a firm place to stand and I will move the earth."*

*Archimedes*



*"Nothing in this world is to be feared...  
only understood."*

*Marie Curie*



## 3.2 Σχεδιασμός, αριθμός και διάρκεια προσομοιώσεων

**Όταν** πραγματοποιούμε προσομοιώσεις αναδίπλωσης ενός σημαντικού αριθμού πεπτιδίων, ο χρόνος της προσομοίωσης συνιστά μία σημαντική πρόκληση. Ένα εύλογο λοιπόν ερώτημα είναι:

*πόσος είναι ο ελάχιστος χρόνος προσομοίωσης που απαιτείται για να παρατηρήσουμε ένα γεγονός αναδίπλωσης ενός πεπτιδίου μήκους τεσσάρων καταλοίπων;*

Το οποίο δίνει τροφή σε μία δεύτερη, ίσως πιο πολύπλοκη ερώτηση:

*πώς μπορούμε να ανιχνεύσουμε ένα τέτοιο γεγονός;*

Τόσο θεωρητικές όσο και πειραματικές μελέτες δείχνουν ότι το όριο ταχύτητας της πρωτεϊνικής αναδίπλωσης υπολογίζεται σε  $N/100\mu\text{s}$ , όπου  $N$  είναι ο αριθμός των καταλοίπων (Kubelka et al., 2004). Το όριο αυτό διαμορφώνεται από έναν ανταγωνισμό ανάμεσα στην εντροπία της διαμόρφωσης, το σχηματισμό δεσμών υδρογόνου, το σχηματισμό υδρόφοβου πυρήνα, τη δημιουργία ηλεκτροστατικών αλληλεπιδράσεων και την ενέργεια διαλυτοποίησης αλλά κυρίως τη διάχυση (McCammon, 1996). Μεγάλο τμήμα της γνώσης μας γύρω από το λεγόμενο όριο ταχύτητας (speed-limit) της αναδίπλωσης προέρχεται από τη μελέτη της κινητικής σχηματισμού απλών στοιχείων δευτεροταγούς δομής (Eaton et al., 1998, Bieri et al., 1999). Η δημιουργία δομών θηλιάς (end-to-end contact, loop-closure) λαμβάνει χώρα στην κλίμακα των 10ns (Lapidus et al., 2000, Portman, 2003), της  $\alpha$ -έλικας στα 200ns (Williams et al. 1996, Garcia et al. 2002), ενώ των

β-φύλλων και μίνι-πρωτεϊνών στα 1-10 $\mu$ s (Munoz et al., 1997), με βάση το ελάχιστο μήκος αλληλουχίας που μπορεί να δημιουργήσει τις αντίστοιχες δομές.

Τα πεπτίδια της παρούσας μελέτης εμπίπτουν στην πρώτη κατηγορία καθώς το μήκος τους επαρκεί για το σχηματισμό δομών θηλιάς και διαφόρων τύπων στροφών αλλά ακόμα και για τη δημιουργία μίας στροφής  $\alpha$ -έλικας.

Για την πειραματική μελέτη της κινητικής της δημιουργίας δομών θηλιάς χρησιμοποιούνται φασματοσκοπικές τεχνικές οι οποίες ανιχνεύουν τη δημιουργία δεσμών van der Waals και περιλαμβάνουν triplet-triplet energy transfer (Bieri et al., 1999, Krieger et al., 2003), fluorescence quenching (Lapidus et al., 2000, 2001, Buscaglia et al., 2006) και προσφάτως time-resolved FRET (Möglich et al., 2006). Αυτές οι δομές θηλιάς που αφορούν την επαφή των άκρων (end-to-end) ονομάζονται τύπου I και εμφανίζουν γρηγορότερη κινητική σε σχέση με τις δομές θηλιάς που συναντάμε σε ανώτερες δομές πρωτεϊνών και περιλαμβάνουν τη δημιουργία επαφής με εσωτερικές θέσεις στη δομή, τύπου II (end-to-interior) και τύπου III (interior-to-interior) (Fierz et al., 2007).

Υπάρχουν αναφορές όπου υπάρχει συμφωνία στον προβλεπόμενο πειραματικό χρόνο με τον αντίστοιχο που υπολογίζεται από τις προσομοιώσεις μοριακής δυναμικής (Yeh et al., 2002), αλλά αυτό φαίνεται να συμβαίνει κατά περίπτωση και να εξαρτάται από το μήκος της πεπτιδικής αλληλουχίας, την ίδια την αλληλουχία και το force field (Feige et al., 2008). Μία ενδεδειγμένη πειραματική και υπολογιστική μελέτη σε πολυμερή γλυκίνης-σερίνης (polyGS) διαφόρων μηκών έδειξε ότι ο σχηματισμός της δομής θηλιάς λαμβάνει χώρα σε 20-100ns για αλληλουχίες με περισσότερα από 10 κατάλοιπα, ακολουθώντας εκθετική αύξηση σε σχέση με το μήκος. Η σχέση αυτή δεν ικανοποιείται για μικρότερου μήκους πεπτίδια, λόγω της στερεοχημικής παρεμπόδισης από τη χρωστική που χρησιμοποιήθηκε στα πειράματα αυτά, δίνοντας μέσο χρόνο σχηματισμού τα 17ns και 22ns για μήκη τριών και πέντε καταλοίπων αντίστοιχα (Daidone et al., 2010). Μάλιστα το μέγιστο όριο ταχύτητας για τη δημιουργία ενός γεγονότος αναδίπλωσης στα τριπεπτίδια αυτά ορίστηκε σε 20ns (Bieri et al., 1999). Η δημιουργία αλληλεπίδρασης (contact-formation) μεταξύ καταλοίπων  $i$ ,  $i+4$  γίνεται στην κλίμακα των 12-20ns για την πλειοψηφία των αμινοξικών αλληλουχιών, ενώ η ενσωμάτωση εύκαμπτων αμινοξέων όπως η γλυκίνη μειώνει το χρόνο στα 8ns και η ενσωμάτωση άκαμπτων αμινοξέων όπως η προλίνη (σε *trans* διαμόρφωση) αυξάνει το χρόνο σε 50ns (Plaxco et al., 1998, Krieger et al., 2003, 2004).

Ακόμα, υπολογιστικές μελέτες αναδίπλωσης τετραπεπτιδίων με μοτίβο αλληλουχίας XXXP δείχνουν ότι γεγονότα αναδίπλωσης συμβαίνουν πολύ συχνά, με μέσο χρόνο 1.4ns στους 300K (Fuchs et al., 2006). Τετραπεπτίδια με μοτίβο αλληλουχίας τύπου APGD και APGN χαρακτηρίζονται επίσης από γρήγορες εναλλαγές μεταξύ αναδιπλωμένης και μη-αναδιπλωμένης δομής κάθε 1-2ns (Bashford et al., 1997), επιδεικνύοντας μία κινητική αναδίπλωσης δύο σταδίων. Η κινητική αναδίπλωσης δύο σταδίων χαρακτηρίζεται από 2 καταστάσεις (two-state kinetics), την αναδιπλωμένη (folded) και τη μη-αναδιπλωμένη (unfolded), οι οποίες σχετίζονται μέσω ενός συνόλου ενδιάμεσων (transition state ensemble, TSE) και επιδεικνύει χαρακτηριστική σιγμοειδή καμπύλη συναρτήσε κάποιου αποδιατακτικού παράγοντα (όπως θερμοκρασία, ουρία) (Barrick, 2009). Σε τέτοιες περιπτώσεις, όπου υπάρχει μόνο ένα ενεργειακό φράγμα (free energy barrier) που διαχωρίζει τις δύο καταστάσεις, η κατανομή του χρόνου αναδίπλωσης (της χρονικής στιγμής που γίνεται πέρασμα από τη μία κατάσταση στην άλλη, crossing time) έχει εκθετική μορφή και υπακούει στην εξίσωση  $P(t) = k \exp(-kt)$ , όπου  $k$  είναι ο ρυθμός αναδίπλωσης (Larson et al., 2003). Αυτό συνεπάγεται ότι πρέπει κανείς να προχωρήσει μία προσομοίωση αναδίπλωσης τουλάχιστον για χρόνο  $1/k$ , δηλαδή για χρόνο περισσότερο από το μέσο ρυθμό αναδίπλωσης, προκειμένου να έχει μία σεβαστή πιθανότητα να παρατηρήσει ένα τυχαίο γεγονός αναδίπλωσης. Δεδομένου ότι η αναδίπλωση είναι στοχαστική διαδικασία, υπάρχει μία επίσης σεβαστή πιθανότητα να μην παρατηρηθεί γεγονός αναδίπλωσης στο χρόνο αυτό εάν γίνει μία και μόνο προσομοίωση. Έτσι, εναλλακτικά μπορεί κανείς αντί να παρατείνει το χρόνο της προσομοίωσης να πραγματοποιήσει ένα σύνολο από παράλληλες προσομοιώσεις και να αναζητήσει το αντίγραφο εκείνο το οποίο θα περάσει το ενεργειακό φράγμα (Worth et al., 1998, Caves et al., 1998, Ferrara et al., 2000, Snow et al., 2002, Fersht 2002, Larson et al., 2003, Paci et al., 2003, Monticelli et al., 2008, Ensign et al., 2009).

Επιστρέφοντας στη δική μας περίπτωση και δεδομένου ότι θέλουμε να πραγματοποιήσουμε προσομοιώσεις σε ένα μεγάλο αριθμό πεπτιδίων, ποιά είναι η χρυσή τομή ανάμεσα στο χρόνο της προσομοίωσης και στον αριθμό αντιγράφων προκειμένου να μεγιστοποιήσουμε την πιθανότητα να βρούμε έναν αναδιπλωτή; Στην απόφαση μας πρέπει να προσμετρήσουμε τον αριθμό αλληλουχιών που θέλουμε να ερευνήσουμε (1.440 τετραπεπτίδια), τη διαθέσιμη υπολογιστική ισχύ, το φυσικό χρόνο πραγματοποίησης κάθε προσομοίωσης, και φυσικά το γεγονός ότι περιοριζόμαστε από το χρόνο περάτωσης της παρούσας διατριβής.

Χρησιμοποιήσαμε ένα πεπτίδιο ως μοντέλο, για να πραγματοποιήσουμε μία σειρά από

δοκιμαστικές προσομοιώσεις που ποικίλουν τόσο στο μήκος όσο και στον αριθμό επαναλήψεων. Ως πεπτίδιο-μοντέλο χρησιμοποιήσαμε το RWTDQ για το οποίο υπήρχαν ήδη διαθέσιμα (στο παρόν εργαστήριο) αποτελέσματα από μεγάλου μήκους προσομοιώσεις. Για το πεπτίδιο αυτό γνωρίζουμε ότι αναδιπλώνεται, βάσει των προσομοιώσεων, σε μία καλά καθορισμένη δομή, ενώ η αλληλουχία του πληροί τους περιορισμούς της αλληλουχίας που θέσαμε στα τετραπεπτίδια, καθότι συναντάμε τόσο την τρυπτοφάνη (W) όσο και τα φορτισμένα κατάλοιπα αργινίνη (R) και ασπαρτικό οξύ (D). Στον Πίνακα 3.2 παραθέτουμε συνοπτικά τις δοκιμαστικές προσομοιώσεις που πραγματοποιήσαμε. Οι χρόνοι των γεγονότων αναδίπλωσης που προκύπτουν από αυτές τις προσομοιώσεις είναι σε συμφωνία με τη βιβλιογραφία που προαναφέραμε.

Αριθμός Επαναλήψεων	Μήκος Τροχιακού (ns)	Αριθμός Γεγονότων Αναδίπλωσης	Χρόνος του Γεγονότος Αναδίπλωσης (ns)
1	130	1	110
1	110	1	10
12	40	6	12, 15, 17, 17, 20, 30
40	20	13	6, 8, 8, 10, 10, 12, 12, 14, 19, 20, 20, 20, 20,
20	12	2	7, 9
40	10	9	1.5, 2, 3, 5, 7, 8, 9, 10, 10
40	7	6	1.5, 4, 4, 5, 7, 7

Πίνακας 3.2 Συνοπτικός πίνακας των δοκιμαστικών προσομοιώσεων στο πεπτίδιο-μοντέλο RWTDQ.

Ο πιο περιοριστικός παράγοντας στο σχεδιασμό μίας τέτοιας κλίμακας υπολογισμών είναι η υπολογιστική ισχύς. Όλοι οι υπολογισμοί πραγματοποιήθηκαν σε μία συστοιχία υπολογιστών τύπου Beowulf, <http://norma.mbg.duth.gr/>, η οποία βρίσκεται στο Τμήμα Μοριακής Βιολογίας και Γενετικής. Οι δοκιμαστικές προσομοιώσεις χρησιμοποιήθηκαν για τη συγκριτική αξιολόγηση (benchmark) της απόδοσης του πρωτοκόλλου της προσομοίωσης, το οποίο με τις τιμές των παραμέτρων όπως αναφέρονται στο Παράρτημα (#13, #14, #15) μας δίνει μία μέση απόδοση των 30ns/πεπτίδιο/μέρα/πυρήνα ή 100ns/πεπτίδιο/μέρα/κόμβο (βλέπε Ενότητα 2.4).

Με βάση λοιπόν όλα τα προηγούμενα δεδομένα εκτιμήσαμε ότι η πραγματοποίηση 4 επαναλήψεων των 5ns (συνολικός χρόνος 20ns) για κάθε τετραπεπτιδική αλληλουχία μας δίνει ικανοποιητικές πιθανότητες για να παρατηρήσουμε τουλάχιστον ένα γεγονός αναδίπλωσης. Να

σημειωθεί, ότι ακόμα και με αυτές τις παραδοχές, χρειάστηκαν 41 μέρες φυσικού χρόνου για αυτές τις προσομοιώσεις των τετραπεπτιδίων (συνολικά 28.8μs υπολογιστικού χρόνου) απασχολώντας το 50-75% των υπολογιστών της συστοιχίας. Το πρωτόκολλο της προσομοίωσης είναι πανομοιότυπο με αυτό που βρίσκεται στο Παράρτημα (#13, NAMD script, all.namd) με μικρές διαφορές στις ακόλουθες παραμέτρους:

- dcdFreq -> 200
  - switchDist -> 8
  - cutoff -> 9
  - pairlistdist -> 10
- langevinPistonPeriod -> 500
  - langevinPistonDecay -> 200
  - run -> 2500000

Στη συνέχεια, και όπως αναλύεται στις επόμενες ενότητες, αναπτύξαμε συναρτήσεις οι οποίες ανιχνεύουν τα γεγονότα αναδίπλωσης και αξιολογούν τη δυνητική αναδιπλωσιμότητα ενός τετραπεπτιδίου. Πεπτίδια με υψηλή βαθμολογία επιλέγονται για ένα καινούργιο κύκλο προσομοιώσεων μεγαλύτερης διάρκειας. Η διαδικασία αυτή επαναλαμβάνεται μέχρις ότου να υποδειχθεί ένας μικρός αριθμός υποψήφιων πεπτιδίων με σταθερή αναδίπλωση (Πίνακας 3.3). Συνολικά, ο υπολογιστικός χρόνος των τετραπεπτιδίων αθροίζεται σε 47.1μs (37.680 core-hours) για τον οποίο χρειάστηκαν περίπου 75 μέρες (αθροιστικού) φυσικού χρόνου. Στον Πίνακα 3.3 παραθέτουμε το σύνολο των προσομοιώσεων που πραγματοποιήθηκαν στο σετ των 1.440 τετραπεπτιδίων.

Αριθμός Τετραπεπτιδίων	Χρόνος Προσομοίωσης (ns)	Αριθμός Επαναλήψεων	Αθροιστικός Υπολογιστικός Χρόνος (ns)
1.440	4	5	28.800
130	30	1	3.900
36	100	1	3.600
4	300	4*	4.800
2	1000	3 <sup>#</sup>	6.000

Πίνακας 3.3 Συγκεντρωτικός πίνακας των προσομοιώσεων που πραγματοποιήσαμε στο σύνολο των 1.440 τετραπεπτιδίων. Οι επαναλήψεις που σημειώνονται με \* αφορούν διαφορετικές θερμοκρασίες, ενώ οι επαναλήψεις που σημειώνονται με <sup>#</sup> αφορούν διαφορετικά force fields.

Όλες οι προσομοιώσεις πραγματοποιούνται με το πρόγραμμα NAMD (Kale et al., 1999) σε συνθήκες περιοδικής οριοθέτησης, με αναλυτική παρουσία του διαλύτη (explicit solvent) και πλήρη υπολογισμό των ηλεκτροστατικών αλληλεπιδράσεων με τη μέθοδο Particle Mesh Ewald (full PME electrostatics) (Darden et al., 1993). Στο Κεφάλαιο 2, Ενότητα 2.1, αναλύονται οι λεπτομέρειες της διεξαγωγής των προσομοιώσεων.

Στην πρωταρχική φάση των προσομοιώσεων, μέχρι και το στάδιο των 36 τετραπεπτιδίων, χρησιμοποιείται το force field CHARMM22 (MacKerell et al., 1998), ίσως το πιο διαδεδομένο και αξιόπιστο force field για προσομοιώσεις πρωτεϊνών, κατά την εποχή που έλαβαν χώρα οι προσομοιώσεις αυτές (Lindorff-Larsen et al., 2012). Κατά τη διάρκεια της έρευνάς μας ωστόσο, η σύγχρονη βιβλιογραφία στράφηκε προς άλλα γνωστά force fields και τη μεταξύ τους σύγκριση και απόδοση, ιδιαίτερα για την περίπτωση της αναδίπλωσης μικρών πεπτιδίων (Aliiev et al., 2010, Best et al., 2008, Lange et al., 2010). Έτσι, όταν ο αριθμός των πεπτιδίων κατέστη επιτρεπτός (<4), στραφήκαμε και εμείς στη μελέτη της επίδρασης της επιλογής της θερμοκρασίας και/ή του force field στο αποτέλεσμα μίας προσομοίωσης αναδίπλωσης (Κεφάλαιο 3, Ενότητα 3.6 Μελέτη της αναδίπλωσης των RWPD, DTRW, RPWD, EVKW σε τέσσερις θερμοκρασίες, και Ενότητα 3.7 Μελέτη της αναδίπλωσης των RWPD, DTRW με τρία force fields).

Η αναζήτηση ενός σταθερά αναδιπλούμενου τετραπεπτιδίου με τα χαρακτηριστικά αλληλουχίας τα οποία θέσαμε διήρκεσε σχεδόν ένα χρόνο, εκ του οποίου μόλις το ~20% αντιστοιχεί σε καθαρό υπολογιστικό χρόνο αφιερωμένο σε προσομοιώσεις. Ο υπόλοιπος χρόνος αντιστοιχεί αφενός στην ανάπτυξη, εφαρμογή και βελτιστοποίηση των συναρτήσεων εκτίμησης της αναδιπλωσιμότητας προς ανάδειξη μιας “μειοψηφίας” δυνητικά αναδιπλούμενων αλληλουχιών και αφετέρου στην ανάλυση της δυναμικής των υποψήφιων αυτών τετραπεπτιδίων που έδωσαν τις υψηλότερες βαθμολογίες, με προσομοιώσεις μεγάλης διάρκειας.

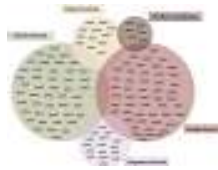
Το εναρκτήριο λάκτισμα λοιπόν για την παρούσα εργασία δόθηκε από τις προσομοιώσεις των 1.440 τετραπεπτιδίων. Στις επόμενες ενότητες του κεφαλαίου αναλύουμε τα αποτελέσματα των προσομοιώσεων αυτών και πως καταλήξαμε στην ανάδειξη των υποψήφιων “δυνητικά αναδιπλούμενων” τετραπεπτιδίων.

*"If it ain't broke, don't fix it."*

*Bert Lance*

*"The great tragedy of Science — the slaying  
of a beautiful hypothesis by an ugly fact."*

*Thomas Henry Huxley*



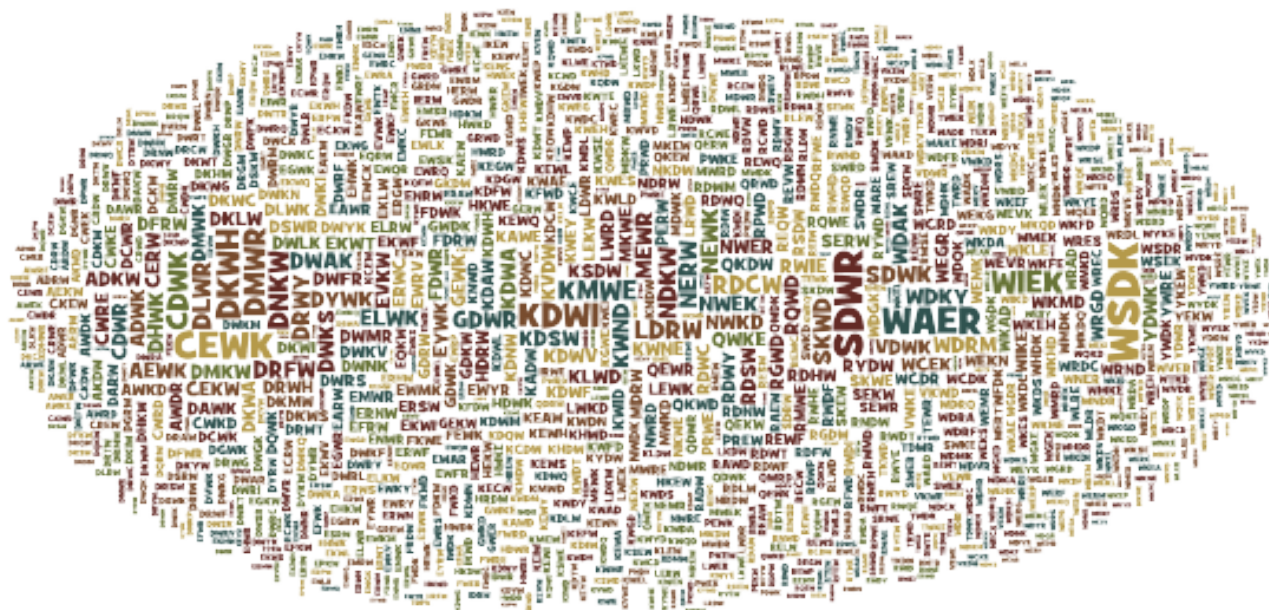
### 3.3 Επιλογή 130 υποψήφιων δυσνητικά αναδιπλούμενων τετραπεπτιδίων

**Μετά** το πέρας των 5.760 προσομοιώσεων αναδίπλωσης των τετραπεπτιδίων (4 επαναλήψεις \* 1.440 αλληλουχίες) έχουμε στη διάθεση μας τις βαθμολογίες τους που αξιολογούν τη δημιουργία ενός γεγονότος αναδίπλωσης (Κεφάλαιο 2, Ενότητα 2.3). Στο στάδιο αυτό λοιπόν, εφαρμόστηκαν οι συναρτήσεις TF1 (σελ. 40) και TF2 (σελ. 41), για τις οποίες η παράμετρος που εξετάζεται είναι η απόσταση μεταξύ ζευγών ατόμων Ca. Το πλήθος των γραφικών παραστάσεων της εξέλιξης στο χρόνο των τριών αποστάσεων είναι 17.280, γεγονός που καθιστά αναγκαία την εξέταση και την ταξινόμησή τους με συστηματικό τρόπο.

Από την εφαρμογή των παραπάνω συναρτήσεων προκύπτουν 2 σύνολα βαθμολογιών, η βαθμολογία που αφορά την απόσταση μεταξύ των ατόμων Ca 1-4, ενδεικτική της δημιουργίας δομής θηλιάς μεταξύ του N-τελικού και του C-τελικού άκρου και θα αναφέρεται με τον όρο 1-4Dist και η βαθμολογία που αφορά και τις τρεις αποστάσεις μεταξύ των ατόμων Ca και τον μεταξύ τους συγχρονισμό, ενδεικτική των συντονισμένων μεταβολών της διαμόρφωσης του πεπτιδικού σκελετού και θα αναφέρεται με τον όρο AllDist. Από τη στιγμή που έχουμε 4 επαναλήψεις για κάθε αλληλουχία, θα πρέπει να εξετάσουμε και κατά πόσο διαφοροποιούνται τα αποτελέσματα από το αντίγραφο της προσομοίωσης που επιλέγουμε να προσμετρήσουμε. Έτσι προκύπτουν 3 επιπλέον σύνολα βαθμολογιών. Στο πρώτο σύνολο ( $n = 5.760$ ) χρησιμοποιούμε και τα 4 αντίγραφα (επαναλήψεις) ως ανεξάρτητες προσομοιώσεις και θα αναφέρεται με τον όρο AllRuns. Στο επόμενο σύνολο ( $n = 1.440$ ) χρησιμοποιούμε την καλύτερη

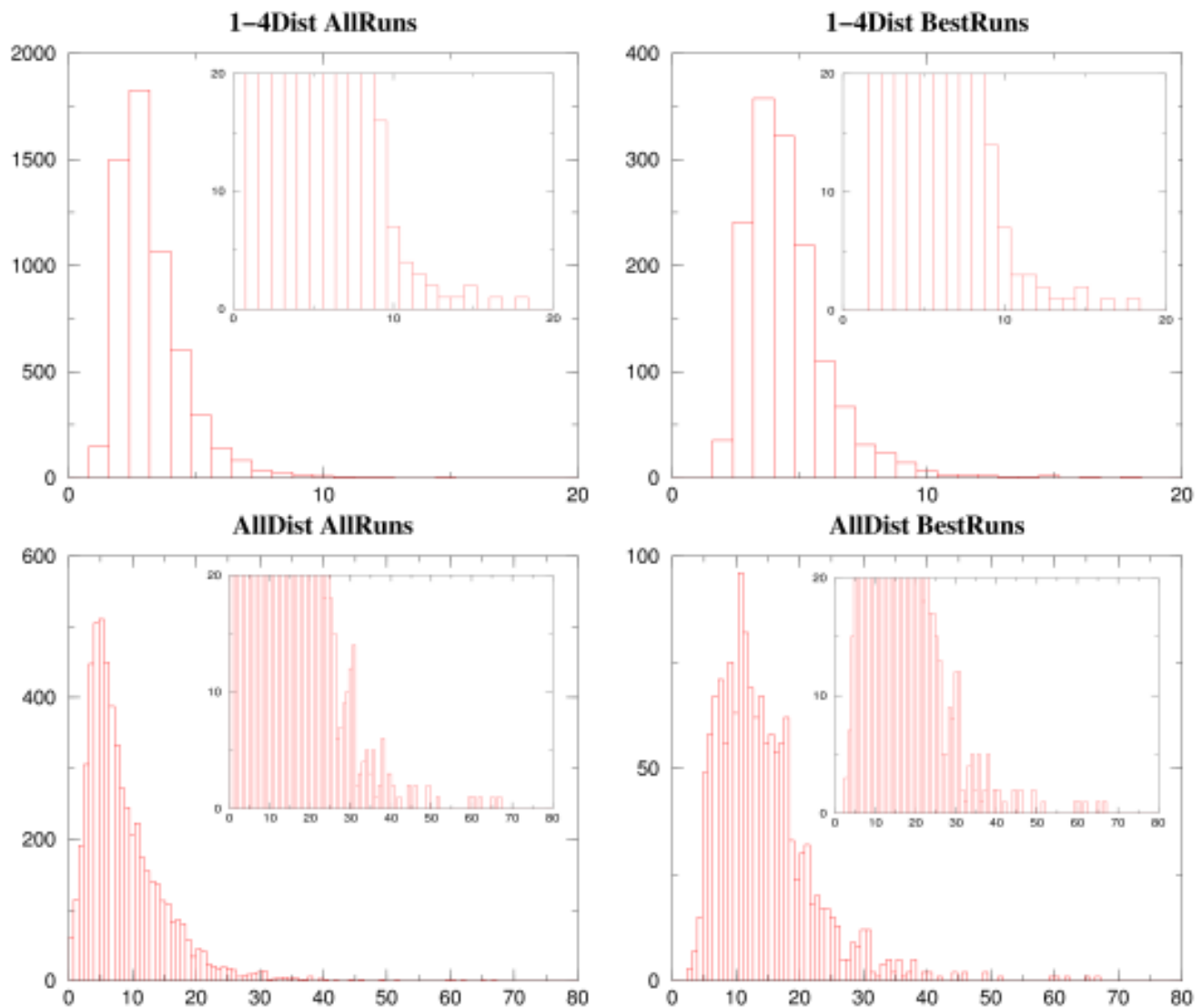


βαθμολογία από τα τέσσερα αντίγραφα της προσομοίωσης ως βαθμολογία της αλληλουχίας. Στο τρίτο σύνολο ( $n=1.440$ ) χρησιμοποιούμε το μέσο όρο της βαθμολογίας από τα τέσσερα αντίγραφα της προσομοίωσης ως βαθμολογία της αλληλουχίας. Τα προγράμματα που χρησιμοποιήθηκαν για τον υπολογισμό της υψηλότερης (καλύτερης) βαθμολογίας και της μέσης βαθμολογίας περιλαμβάνονται στο Παράρτημα (#16, `pickBestRun.pl` και #17, `pickAverRun.pl` αντίστοιχα). Τα σύνολα `AverRuns` και `BestRuns` έδωσαν σχεδόν πανομοιότυπα αποτελέσματα στις μετέπειτα αναλύσεις (cluster analysis, βαθμολογία αμινοξέων) με γραμμικό συντελεστή συσχέτισης των βαθμολογιών των πεπτιδίων 0.87 ( $>0.94$  για τις βαθμολογίες των αμινοξέων για κάθε θέση του τετραπεπτιδίου, βλέπε Κεφάλαιο 5). Τα αποτελέσματα που παρουσιάζονται στη συνέχεια της ενότητας επικεντρώνονται σε 4 σύνολα βαθμολογιών: `1-4DistAllRuns`, `1-4DistBestRuns`, `AllDistAllRuns`, `AllDistBestRuns`. Στην Εικόνα 3.2 βλέπουμε ένα word-cloud των 1440 τετραπεπτιδίων, όπου το μέγεθος της λέξης είναι ανάλογο της βαθμολογίας του συνόλου `1-4DistBestRuns`.



Εικόνα 3.2 Word-cloud των 1440 τετραπεπτιδίων, όπου το μέγεθος κάθε αλληλουχίας είναι ενδεικτικό της βαθμολογίας που έλαβε, βάσει του συνόλου `1-4DistBestRuns`.

Έχοντας στη διάθεσή μας μία βαθμολογία εκτίμησης της αναδιπλωσιμότητας, πως μπορούμε να διαχωρίσουμε τους αναδιπλωτές από τους μη-αναδιπλωτές με ένα συστηματικό και μη αυθαίρετο τρόπο;



Εικόνα 3.3 Ιστογράμματα κατανομής των βαθμολογιών με βάση τις συναρτήσεις TF1 και TF2 (1-4Dist και AllDist αντίστοιχα) και χρησιμοποιώντας όλες τις προσομοιώσεις (AllRuns) ή την προσομοίωση με την καλύτερη βαθμολογία (BestRuns). Στο ένθετο βλέπουμε σε μεγέθυνση το δεξί τμήμα της κατανομής.

Για να ορίσουμε, με ένα μη αυθαίρετο τρόπο, ένα κατώφλι βαθμολογίας υπολογίσαμε ένα ιστόγραμμα της κατανομής των βαθμολογιών για τα 4 προαναφερθέντα σύνολα. Στην Εικόνα 3.3 βλέπουμε πως σε όλες τις περιπτώσεις οι κατανομές δεν είναι κανονικές, αλλά αριστερά ασύμμετρες, με τις μεγάλες συχνότητες (μη αναδιπλωτές με χαμηλή βαθμολογία) να συγκεντρώνονται στο αριστερό άκρο της κατανομής. Ο χαρακτηρισμός αυτός επιβεβαιώνεται από το γεγονός ότι σε όλες τις κατανομές ο μέσος όρος είναι υψηλότερος από τη διάμεσο τιμή

και η επικρατούσα τιμή (με τη μεγαλύτερη συχνότητα εμφάνισης) είναι χαμηλότερη, τόσο από το μέσο όρο, όσο και από τη διάμεσο τιμή (Πίνακας 3.4).

	Μέσος Όρος	Διάμεσος Τιμή	Επικρατούσα Τιμή	Μικρότερη Τιμή	Μεγαλύτερη Τιμή
1-4Dist AllRuns	3.2	2.9	2.8	0.8	17.7
1-4Dist BestRuns	4.5	4.1	3.6	1.8	17.7
AllDist AllRuns	8.7	6.9	5.2	-2.4	66.5
AllDist BestRuns	14.4	12.8	10.8	2.6	66.5

Πίνακας 3.4 Συνοπτικός πίνακας στατιστικών μέτρων χαρακτηριστικών των κατανομών των βαθμολογιών της Εικόνας 3.3.

Το πλέον ενδιαφέρον τμήμα των κατανομών αυτών είναι το ασυνεχές τμήμα της κατανομής που εμφανίζεται με μικρή συχνότητα στα δεξιά (ένθετο στην Εικόνα 3.3). Αυτή η ουρά των ακραίων τιμών (outlier-tail) αντιστοιχεί στο μικρό σύνολο των τετραπεπτιδίων που έδωσαν ιδιαίτερα υψηλές βαθμολογίες με βάση τις συναρτήσεις TF1 και TF2. Για να πραγματοποιήσουμε την ταυτοποίηση των πεπτιδίων αυτών με ένα συστηματικό τρόπο προχωρήσαμε σε μεθόδους cluster analysis.

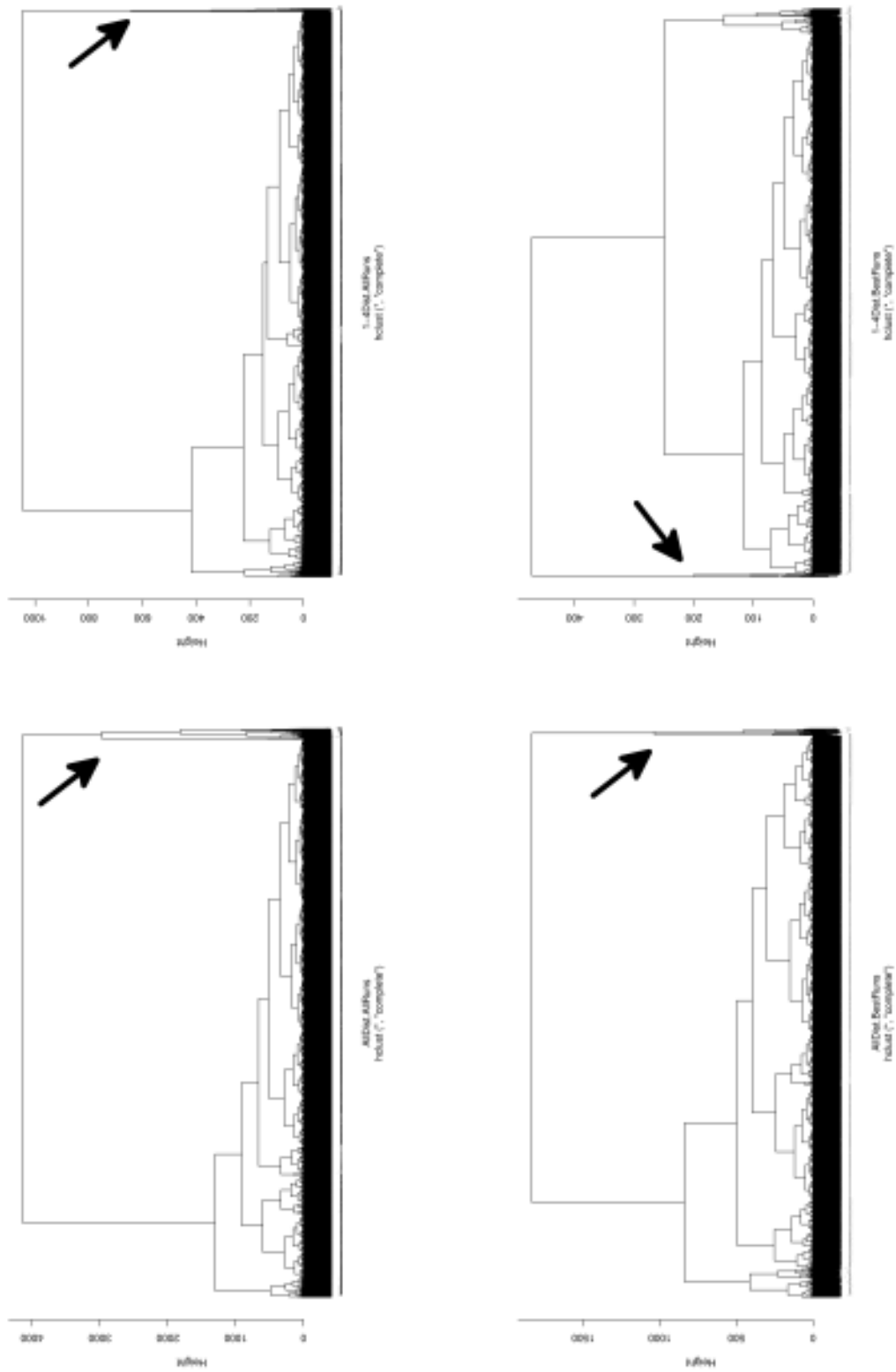
Για να κάνουμε cluster analysis στις βαθμολογίες ετοιμάσαμε ένα τετράγωνο συμμετρικό πίνακα διαστάσεων  $n \times n$ , όπου  $n=5760$  στα σύνολα AllRuns και  $n=1440$  στα σύνολα BestRuns. Τα πεπτίδια (σε αριθμητική σειρά βάσει της βαθμολογίας) αντιστοιχούν στις δύο διαστάσεις του πίνακα, ενώ το εσωτερικό του πίνακα συμπληρώνεται με τις απόλυτες τιμές των διαφορών των βαθμολογιών μεταξύ όλων των συνδυασμών ζευγών πεπτιδίων (έτσι η διαγώνιος είναι μηδενική). Το πρόγραμμα που χρησιμοποιήσαμε για να δημιουργήσουμε τους πίνακες παραθέεται στο ΠΑΡΑΡΤΗΜΑ (#18, dist\_matrix.pl) ενώ στην Εικόνα 3.4 βλέπουμε μία γραφική απεικόνιση των πινάκων. Ο τετράγωνος αυτός πίνακας χρησιμοποιείται απευθείας από τον αλγόριθμο hclust() του προγράμματος R (R Development Core Team, 2004) για τη δημιουργία δενδρογράμματος, όπου οι διαφορές στις βαθμολογίες αντιμετωπίζονται ως αποστάσεις. Ο αλγόριθμος hclust() είναι μία ιεραρχική μέθοδος (hierarchical clustering) (Shenkin et al., 1994), η οποία χρησιμοποιεί δενδρογράμματα για την οπτικοποίηση των επιπέδων ιεραρχίας που προκύπτουν με βάση μία αντικειμενική συνάρτηση. Αυτή η αντικειμενική συνάρτηση αποτελεί ένα μέτρο της ανομοιομορφίας μέσα στο cluster, την οποία προσπαθεί ο αλγόριθμος να ελαχιστοποιήσει.



Εικόνα 3.4 Γραφική απεικόνιση των τετράγωνων συμμετρικών πινάκων που χρησιμοποιούνται για cluster analysis και τη δημιουργία των δενδρογραμμάτων. Η αρχή των αξόνων είναι στην πάνω αριστερά γωνία. Η χρωματική κλίμακα κυμαίνεται από σκούρο μπλε για μηδενικές διαφορές στη βαθμολογία (όπως πάνω στη διαγώνιο) μέχρι κόκκινο για μεγάλες διαφορές (17 και 70 για τις βαθμολογίες 1-4Dist και AllDist αντίστοιχα).

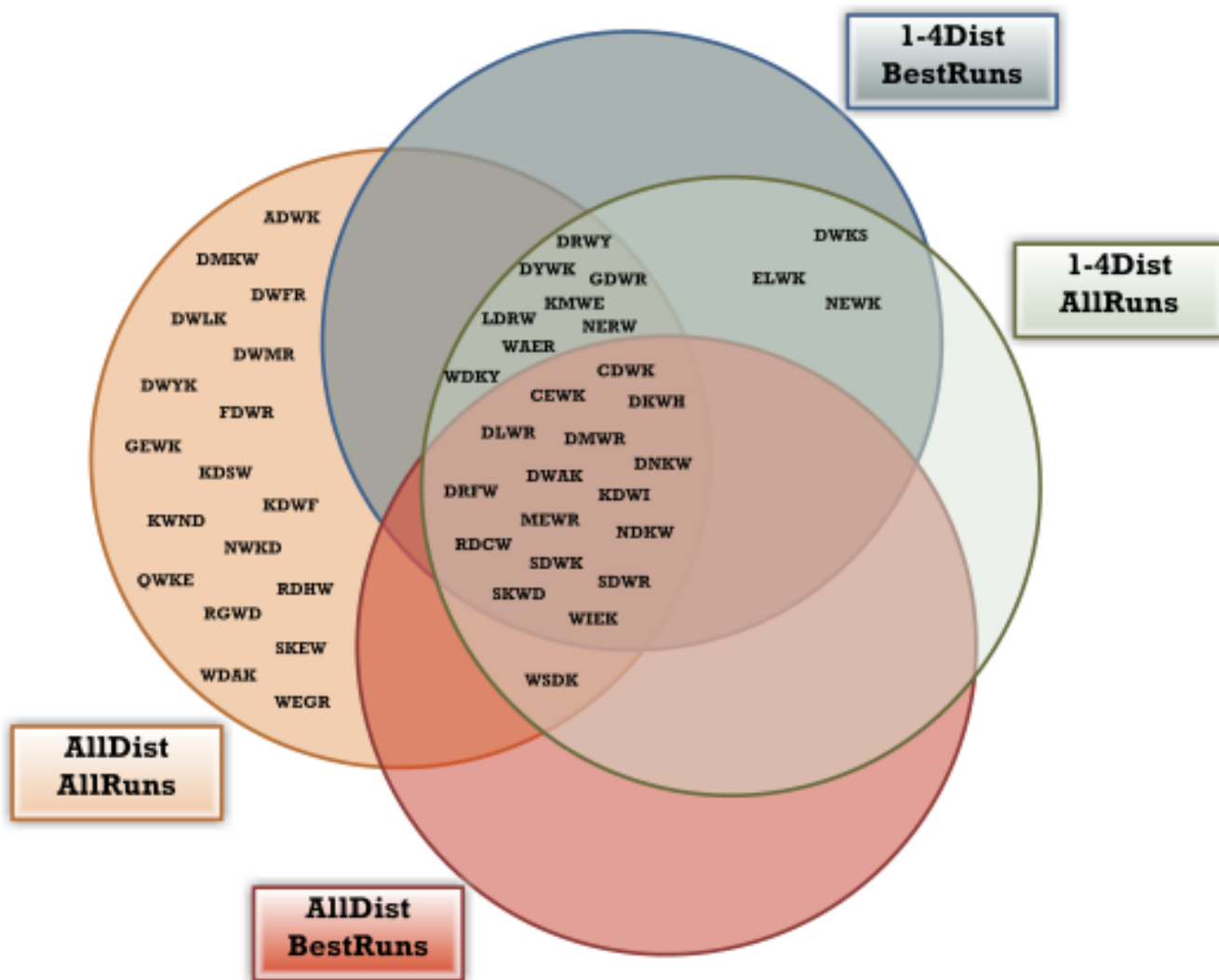
Στη δική μας περίπτωση, επιλέγουμε ο αλγόριθμος να χρησιμοποιήσει τη μέθοδο της ολοκληρωμένης σύνδεσης (complete linkage method) για να βρει παρόμοια clusters, όπου η απόσταση μεταξύ δύο cluster εξ' ισούται με τη μεγαλύτερη απόσταση οποιουδήποτε μέλους του ενός cluster από οποιοδήποτε μέλος του άλλου cluster (Jain et al., 1999).

Στην Εικόνα 3.5 βλέπουμε τα παραγόμενα (με την προηγούμενη διαδικασία) δενδρογράμματα για τα τέσσερα σύνολα βαθμολογιών. Το μικρό cluster των αναδιπλούμενων πεπτιδίων αυτή τη φορά διαχωρίζεται με ευκρίνεια στην άκρη του δενδρογράμματος (μαύρο βέλος). Το cluster αυτό φαίνεται και στην Εικόνα 3.4 με κίτρινες και κόκκινες αποχρώσεις, που συνδέουν το μικρό αριθμό υψηλόβαθμων πεπτιδίων με την πληθώρα των υπόλοιπων χαμηλόβαθμων (και συνεπώς η διαφορά στη βαθμολογία είναι μεγάλη).



Εικόνα 3.5 Cluster analysis των βαθμολογιών των πεπτιδίων όπου οι απόλυτες διαφορές των βαθμολογιών αντιμετωπίζονται ως αποστάσεις για την κατασκευή δενδρογραμμάτων.

Ο αριθμός των αναδιπλούμενων πεπτιδίων που ανήκουν στο μικρό αυτό cluster είναι διαφορετικός σε κάθε σύνολο, όπως και το κατώφλι που χρησιμοποιείται για το διαχωρισμό σε clusters. Ωστόσο, ένας σημαντικός αριθμός πεπτιδίων είναι κοινός στα τέσσερα σύνολα. Τα τετραπεπτίδια που προκύπτουν με τη μέθοδο cluster analysis είναι συνολικά 46 και φαίνονται στο διάγραμμα Venn της Εικόνας 3.6. Η λίστα αυτή των πεπτιδίων, που προκύπτει υπολογιστικά χαρακτηρίζεται ως *cluster-based* (Εικόνα 3.7).



Εικόνα 3.6 Διάγραμμα Venn των αναδιπλούμενων πεπτιδίων (cluster-based) για κάθε ένα από τα τέσσερα σύνολα βαθμολογιών.

Αφού καταλήξαμε σε μία λίστα 46 δυνητικά αναδιπλούμενων τετραπεπτιδίων, το επόμενο “λογικό” βήμα είναι να εξετάσουμε την ορθότητα της λίστας. Τα πεπτίδια αυτά προέκυψαν με συστηματικό τρόπο, βάσει τις βαθμολογίας που έλαβαν από τις συναρτήσεις TF1 και TF2 (cluster analysis based on a score-function).

*Ποιά είναι όμως η διακριτική ικανότητα των συναρτήσεων που αναπτύξαμε;*

Ή για να το διατυπώσουμε καλύτερα,

*Πληρούν οι συναρτήσεις το σκοπό της δημιουργίας τους, να ανιχνεύουν γεγονότα αναδίπλωσης;*

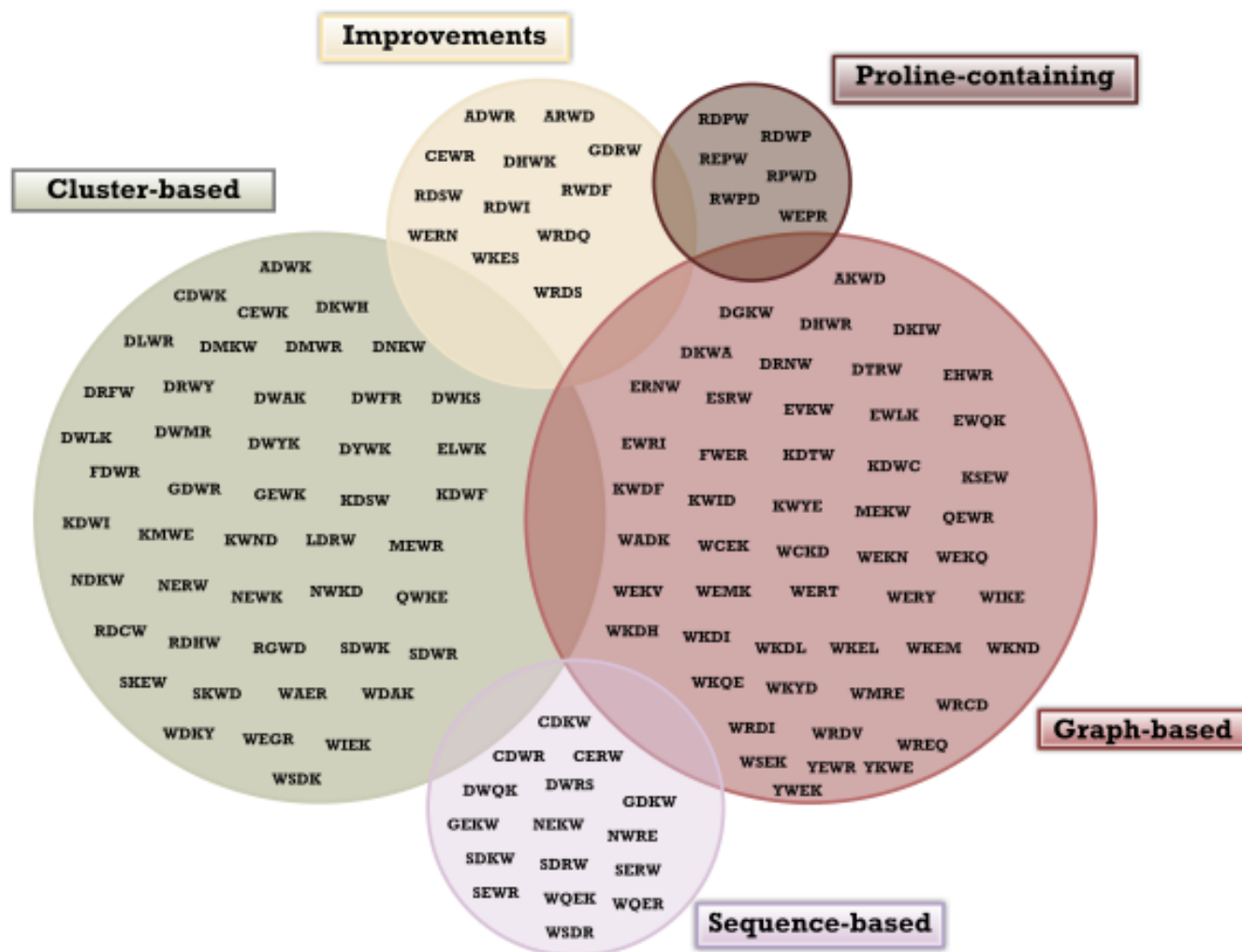
Ο πιο ασφαλής τρόπος για να απαντήσουμε στο παραπάνω ερώτημα είναι με οπτική εξέταση των γραφημάτων της απόστασης. Το στάδιο αυτό, αν και επίπονο δεδομένου ότι ο αριθμός των γραφημάτων που πρέπει να εξεταστεί για 1.440 αλληλουχίες είναι 17.280, είναι και απαραίτητο, για να εξεταστεί κατά πόσο έχουν αποκλειστεί από τη λίστα πεπτίδια που δε θα έπρεπε, αποκαλύπτοντας έτσι τις αδυναμίες και τις ατέλειες των, βασιζόμενων σε ατομικές αποστάσεις, συναρτήσεων που αναπτύξαμε.

Έτσι δημιουργήθηκε μία δεύτερη λίστα που αποτελείται από 50 τετραπεπτίδια, το βασικό χαρακτηριστικό των οποίων είναι ότι το γεγονός αναδίπλωσης συμβαίνει προς το τέλος της προσομοίωσης των 5ns (late-folders) και το παρατηρούμε τουλάχιστον σε ένα από τα τέσσερα αντίγραφα της προσομοίωσης. Πεπτίδια με ασταθή συμπεριφορά και πολλαπλά γεγονότα αναδίπλωσης σε τουλάχιστον ένα από τα αντίγραφα της προσομοίωσης αποκλείστηκαν. Η λίστα αυτή των πεπτιδίων, που προκύπτει από την οπτική εξέταση των γραφικών παραστάσεων των ατομικών αποστάσεων χαρακτηρίζεται ως *graph-based* (Εικόνα 3.7).

Στο σημείο αυτό θεωρήσαμε πρόπον να προσθέσουμε στη λίστα μας 6 πεπτίδια τα οποία περιέχουν το μινοξύ προλίνη και τα οποία γνωρίζουμε ότι δεν ευνοούνται από τη συνάρτηση βάσει του σχεδιασμού της. Ο λόγος είναι ότι λόγω των ασυνήθιστων φ/ψ γωνιών που υιοθετεί το συγκεκριμένο μινοξύ, δεν μπορούμε να παρατηρήσουμε τη δομή θηλιάς (loop-closure) σε τόσο μικρού μήκους αλυσίδες όπως στις υπόλοιπες περιπτώσεις τετραπεπτιδίων. Στα τετραπεπτίδια που περιέχουν προλίνη, η δημιουργία σταθερής δομής διαφαίνεται από τη σταθεροποίηση της τιμής των ατομικών αποστάσεων γύρω από μία μέση τιμή (μικρές διακυμάνσεις) η οποία μπορεί



να είναι ακόμα και 7-10Å. Το γεγονός αυτό καθιστά δύσκολη τη διάκριση των πεπτιδίων αυτών από τα πεπτίδια τα οποία παρέμειναν σε εκτεταμένη διαμόρφωση μέσω των συναρτήσεων TF1 και TF2. Η λίστα αυτή των πεπτιδίων, που προκύπτει επίσης από την οπτική εξέταση των γραφικών παραστάσεων των ατομικών αποστάσεων (96 πεπτίδια, 288 γραφικές παραστάσεις) χαρακτηρίζεται ως *proline-containing* (Εικόνα 3.7).



Εικόνα 3.7 Διάγραμμα Venn των 130 τετραπεπτιδίων που επιλέχθηκαν από τον πρώτο κύκλο προσομοιώσεων, για επιπλέον μελέτη με προσομοιώσεις μεγαλύτερης διάρκειας:

*cluster-based*: πεπτίδια που προέκυψαν συστηματικά μέσω cluster analysis των βαθμολογιών των συναρτήσεων TF1 και TF2

*improvements*: πεπτίδια που προέκυψαν από την εξέταση των γραφικών παραστάσεων προσομοιώσεων μεγαλύτερης διάρκειας

*proline-containing*: πεπτίδια που περιέχουν προλίνη και προέκυψαν από την εξέταση των γραφικών παραστάσεων

*graph-based*: πεπτίδια που προέκυψαν από την εξέταση των γραφικών παραστάσεων και δεν επιλέχθηκαν με το συστηματικό τρόπο

*sequence-based*: πεπτίδια που προέκυψαν από τη βαθμολογία συσχέτισης αλληλουχίας-αναδιπλωσιμότητας (Κεφάλαιο 5)

Η ενασχόληση μας με συστηματικό τρόπο με ένα σημαντικό αριθμό αλληλουχιών μας οδήγησε σε ένα αναπόφευκτο ερώτημα της βιβλιογραφίας:

*Υπάρχει κάποια σχέση ανάμεσα στην αλληλουχία ενός πεπτιδίου και την αναδιπλωσιμότητα του (sequence-structure relationships) ;*

Έτσι, κατά τη διάρκεια των προσομοιώσεων πραγματοποιήσαμε ένα συσχετισμό ανάμεσα στην αναδιπλωσιμότητα και την αλληλουχία, για την περίπτωση των πεπτιδίων της έρευνάς μας (Κεφάλαιο 5). Η ανάλυση αυτή βασίζεται σε μία συγκεντρωτική βαθμολογία που υπολογίζεται για κάθε αμινοξύ και η οποία εξαρτάται από τη θέση του στην πεπτιδική αλληλουχία και τη βαθμολογία που έλαβε η αλληλουχία από τις συναρτήσεις TF1 και TF2 (ΠΑΡΑΡΤΗΜΑ, #22-#24). Στην ανάλυση αυτή βασιστήκαμε για να δημιουργήσουμε μία λίστα πεπτιδίων των οποίων η αλληλουχία προέκυψε βάσει της βαθμολογίας για κάθε αμινοξύ και κάθε θέση, με την ακόλουθη ιδέα: Εάν έχουμε ένα τετραπεπτίδιο το οποίο περιέχει μία τρυπτοφάνη και δύο φορτισμένα κατάλοιπα, ποιά είναι η βέλτιστη (βάσει της βαθμολογίας) θέση για την τρυπτοφάνη; Εάν για παράδειγμα η τρυπτοφάνη βρίσκεται στην πρώτη θέση, ποια είναι τα φορτισμένα κατάλοιπα που προτιμώνται, και σε ποιές θέσεις; Στην εναπομείνουσα θέση, ποια αμινοξέα προτιμώνται; Με τον όρο προτίμηση εννοούμε ότι το αμινοξύ για τη συγκεκριμένη θέση έλαβε βαθμολογία πάνω από ένα συγκεκριμένο κατώφλι (2σ της κατανομής). Έτσι δημιουργήθηκε μία λίστα 62 πεπτιδίων εκ των οποίων τα 22 συμπεριλαμβάνονται (όπως αναμέναμε) ήδη στις δύο προηγούμενες λίστες, cluster-based και graph-based. Από τα εναπομείναντα 40 πεπτίδια, επιλέξαμε, βάσει των γραφικών παραστάσεων των ατομικών αποστάσεων, 16 πεπτίδια. Η λίστα αυτή των πεπτιδίων, χαρακτηρίζεται ως *sequence-based* (Εικόνα 3.7).

Τέλος, για λόγους οικονομίας χρόνου, προχωρήσαμε σε προσομοιώσεις μεγαλύτερης διάρκειας των τετραπεπτιδίων που έδιναν ιδιαίτερα υψηλή βαθμολογία με βάση τις συναρτήσεις TF1 και TF2, προτού ολοκληρωθούν οι 5760 προσομοιώσεις. Έτσι προέκυψε ένα σύνολο 12 επιπλέον πεπτιδίων. Η λίστα αυτή των πεπτιδίων, χαρακτηρίζεται ως *improvements* (Εικόνα 3.7).

Τα 130 αυτά τετραπεπτίδια αποτέλεσαν ένα καινούργιο σύνολο πεπτιδίων για τα οποία πραγματοποιήσαμε προσομοιώσεις αναδίπλωσης διάρκειας 30ns, όπως αναλύεται στην επόμενη ενότητα (Ενότητα 3.4).

*"We haven't the money,  
so we've got to think."*

*Ernest Rutherford*

*"... one of the main causes of the fall of the Roman Empire was that, lacking zero, they had no way to indicate successful termination of their C programs..."*

*Robert Firth*



### 3.4 Επιλογή 36 υποψήφιων δυνητικά αναδιπλούμενων τετραπεπτιδίων

Τα 130 τετραπεπτίδια που επιλέχθηκαν με τον τρόπο που αναλύθηκε στην προηγούμενη ενότητα, μελετήθηκαν με προσομοιώσεις αναδίπλωσης διάρκειας 30ns. Η επιλογή της διάρκειας αυτής έγινε με γνώμονα την απόδοση (benchmark) του πρωτοκόλλου της προσομοίωσης, που είναι 30ns/πυρήνα/μέρα. Έτσι, το υπολογιστικό τμήμα των προσομοιώσεων αυτών μπορεί να ολοκληρωθεί σε μόλις 5 μέρες, έχοντας στη διάθεσή μας ολόκληρη τη συστοιχία των υπολογιστών. Από την άλλη, η επιμήκυνση του χρόνου της προσομοίωσης από 5ns σε 30ns θα μας επιτρέψει να εξετάσουμε με μεγαλύτερη βεβαιότητα τη δημιουργία αναδιπλωμένης δομής, δεδομένου ότι τα γεγονότα αναδίπλωσης λαμβάνουν χώρα στην κλίμακα των 10ns (Ενότητα 3.2). Το πρωτόκολλο της προσομοίωσης είναι πανομοιότυπο με αυτό που βρίσκεται στο Παράρτημα (#13, NAMD script, all.namd) με μοναδική διαφορά τον τελικό αριθμό βημάτων (run -> 15.000.000 steps).

Η ανάλυση των αποτελεσμάτων του πρώτου κύκλου των προσομοιώσεων (Ενότητα 3.3 Επιλογή 130 υποψήφιων δυνητικά αναδιπλούμενων πεπτιδίων) κατέδειξε την ανάγκη για ανάπτυξη μίας νέας συνάρτησης η οποία δε θα φέρει τα μειονεκτήματα των συναρτήσεων TF1 και TF2 και θα μας επιτρέψει την επιλογή των “αναδιπλούμενων” πεπτιδίων, με συστηματικό τρόπο, χωρίς την επίπονη εξέταση των γραφικών παραστάσεων και την πολύπλοκη επιλογή με προσωπικά κριτήρια τα οποία φέρουν το χαρακτηριστικό της αυθαιρεσίας. Ο δεύτερος αυτός κύκλος των προσομοιώσεων πραγματοποιήθηκε και πάλι μέσω της οργανωμένης δέσμης ενεργειών, όπως

περιγράφεται στο Κεφάλαιο 2, Ενότητα 2.2. Το perl script χρησιμοποιήθηκε αυτούσιο όπως περιλαμβάνεται στο Παράρτημα (#10, systematic.pl). Στο στάδιο αυτό λοιπόν, εφαρμόσαμε εκτός από τις συναρτήσεις των ατομικών αποστάσεων TF1 και TF2, μία επιπλέον συνάρτηση TF3, η οποία βασίζεται στις ατομικές διακυμάνσεις (rmsf) και σε τετράγωνους συμμετρικούς πίνακες RMSD μεταξύ διαδοχικών δομών του τροχιακού (Ενότητα 2.3 Συναρτήσεις εκτίμησης της αναδιπλωσιμότητας, υπορουτίνα *Expand\_Windows()*).

Η νέα λοιπόν συνάρτηση (TF3) ταξινομεί τα πεπτίδια με ιδιαίτερα ικανοποιητικό τρόπο, όπως προκύπτει από την οπτική εξέταση όλων των διαθέσιμων αποτελεσμάτων (910 αρχεία) για τα 130 πεπτίδια του παρόντος κύκλου των προσομοιώσεων. Στην Εικόνα 3.8 βλέπουμε την κατάταξη των πεπτιδίων αυτών μετά την εφαρμογή της συνάρτησης TF3. Η διακριτική ικανότητα της συνάρτησης αυτής διαφαίνεται (και επιβεβαιώνεται) και στις μετέπειτα αναλύσεις που πραγματοποιήσαμε (Ενότητα 3.5 Επιλογή 4 υποψήφιων δυνητικά αναδιπλούμενων τετραπεπτιδίων, Ενότητα 4.3 Επιλογή 480 υποψήφιων δυνητικά αναδιπλούμενων πενταπεπτιδίων).

Δεδομένης της τρέχουσας κατάστασης (κατά την αντίστοιχη περίοδο) της συστοιχίας των υπολογιστών και της απόδοσης του πρωτοκόλλου της προσομοίωσης επιλέξαμε τα 36 τετραπεπτίδια που έδωσαν την καλύτερη βαθμολογία για να μελετηθούν με προσομοιώσεις αναδίπλωσης διάρκειας 100ns. Οι προσομοιώσεις αυτές ολοκληρώθηκαν σε ~10 μέρες φυσικού χρόνου χρησιμοποιώντας 12 πυρήνες. Με βάση την Εικόνα 3.9, βλέπουμε ότι το πλήθος των 36 τετραπεπτιδίων είναι ικανοποιητικό, καθώς τα πεπτίδια με μικρότερη βαθμολογική κατάταξη δε φαίνεται να σχηματίζουν κάποια σταθερή δομή με βάση τους πίνακες RMSD και συνεπώς δε χρήζουν περαιτέρω μελέτης.

Ωστόσο, προκειμένου να στηρίξουμε και να επιβεβαιώσουμε την ορθότητα της παραπάνω θέσης, ανατρέξαμε πίσω στα αποτελέσματα και αναζητήσαμε άλλους τρόπους κατάταξης των πινάκων RMSD, πέραν του αλγόριθμου των "επεκτεινομένων παραθύρων" (Ενότητα 2.3 Συναρτήσεις εκτίμησης της αναδιπλωσιμότητας, υπορουτίνα *Expand\_Windows()*). Παρατηρώντας τις κατανομές της Εικόνας 2.5 που προκύπτουν από την εφαρμογή του αλγόριθμου και την ακόλουθη κατάταξη των πινάκων RSMD της Εικόνας 3.9, διαπιστώσαμε ότι η κατάταξη είναι μεν ικανοποιητική, αλλά με το εξής "μειονέκτημα": υπάρχουν πεπτίδια, όπως τα EVKW, ESRW, REPW, WEPR, WRCD, τα οποία παρουσιάζουν ένα ιδιαίτερα συμπαγές cluster δομών (έντονο σκούρο μπλε τετράγωνο) προς το τέλος του χρόνου της προσομοίωσης των 30ns, και επομένως

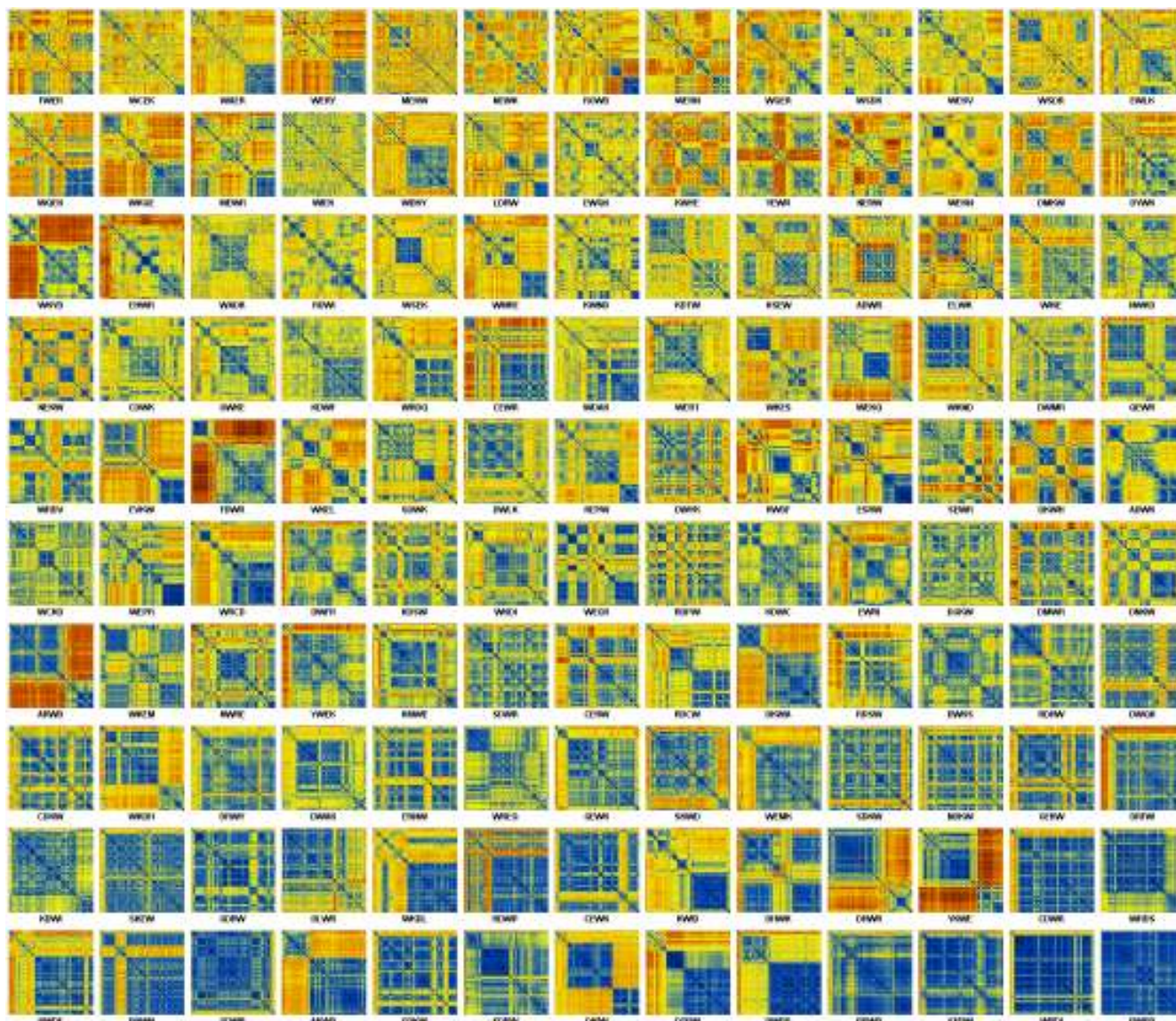
DWNR	inf
DWQK	inf
EWRI	inf
HEKN	inf
KDSW	inf
RGND	inf
WMRE	inf
WCEK	127
WEMK	177
WEKV	177
WERN	191
WAER	225
WEKY	264
YWER	286
WQER	295
EWQK	332
KWYE	346
WQKE	350
WSDK	352
WIEK	370
YENR	375
NERN	383
WERN	409
WKYD	418
WSDR	438
ELWK	448
EHLK	452
WQEK	471
EHRW	505
DMKN	506
DTWK	509
RDWI	563
MEWR	581
WDKY	622
WIKK	629
WADK	641
SDWK	661
WEQK	686
LDRN	712
FDWR	725
NKND	761
WRDV	854
NEKN	860
QWKE	883
WCKD	884
WSEK	898
CDWK	912
WKEI	916
WRDK	971
EDTN	975
EVKN	1035
KDWF	1047
SEWR	1086
YKND	1086
WKES	1093
WDAK	1204
WKDI	1234
WRCD	1257
WERT	1281
KSEW	1293
ARND	1376
ADWR	1392
CEWK	1451
WKND	1480
QENR	1497
KWYE	1546
DMWK	1566
DGKN	1594
REFN	1688
WENR	1710
DTWK	1810
DNLK	1871
DNKS	1881
KDSW	1916
DMFR	1988
DMKL	1993
WEPR	1994
KDWC	2093
ADWK	2162
WKEK	2196
RNDF	2204
DKWB	2364
ESRW	2437
RDPW	2492
WRDM	2565
RDCN	2595
NWRE	2651
DNKN	2712
KDSW	2833
YWEK	2927
CERN	3380
SDWR	3527
DRWY	3562
ERWN	3691
-----	
WREQ	3825
SKWD	4428
CEKW	4607
DNAK	4656
SDKN	5355
GEWK	5509
WEMK	5914
NEKN	6181
WKDL	6743
SKEN	6871
GEKN	6884
KEMK	7255
DRFW	7428
DLWR	8419
GDRW	8824
WRDS	9315
RNDP	9346
DBWR	9447
DWRS	10178
YWKD	10752
KWID	11250
CEWK	11833
GDWR	12333
DRNW	12893
DKIM	13014
KNDF	13135
DBWK	13235
CDWR	15570
AKWD	17368
DTEN	17449
SDRN	19310
GDKN	20438
SERN	23129
KPWO	25769
RNPD	28378
WRDI	30827



Εικόνα 3.8 *Αριστερά*: Κατάταξη των 130 τετραπεπτιδίων με βάση τη βαθμολογία της συνάρτησης TF3. Η διαβάθμιση του χρώματος από κίτρινο σε πορτοκαλί έως κόκκινο ακολουθεί την άνοδο στη βαθμολογία. Η διακεκομμένη γραμμή διαχωρίζει τα 36 τετραπεπτιδία που επιλέχθηκαν για τον επόμενο κύκλο προσομοιώσεων.

*Δεξιά*: Απεικόνιση της ίδιας κατάταξης με word-cloud. Οι πεπτιδικές αλληλουχίες κατατάσσονται με αλφαβητική σειρά ενώ το μέγεθος της γραμματοσειράς είναι ανάλογο της βαθμολογίας της συνάρτησης TF3.



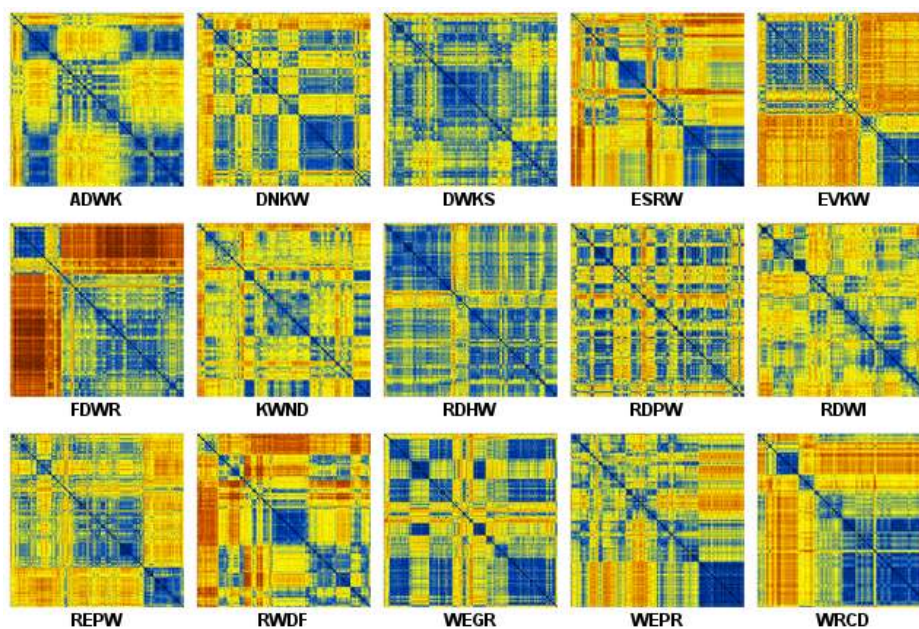


Εικόνα 3.9 Γραφική απεικόνιση των διδιάστατων πινάκων RMSD, χρησιμοποιώντας όλα τα βαριά άτομα για τον υπολογισμό. Η χρωματική κλίμακα είναι η ίδια και κυμαίνεται από σκούρο μπλε (0Å) μέχρι κόκκινο (μέγιστο rmsd 6.43Å). Η σειρά των γραφημάτων από αριστερά προς τα δεξιά (13) και από πάνω προς τα κάτω (10), όπως γίνεται η ανάγνωση κειμένου, ακολουθεί τη βαθμολογική κατάταξη με βάση τη βαθμολογία του πίνακα RMSD (αλγόριθμος των “επεκτεινομένων παραθύρων”).

μικρό σε διάρκεια ώστε να βαθμολογηθεί υψηλά από τον αλγόριθμο, το οποίο όμως δεν έχει ξαναεμφανιστεί κατά τη διάρκεια της αυτής προσομοίωσης. Τα cluster αυτά αντικατοπτρίζονται και στις κατανομές της Εικόνας 2.5 (υποδεικνύονται με μαύρα βέλη). Η πυκνότητα της κατανομής με ποσοστό μπλε pixels μεγαλύτερο από 80-90% (άξονας ψ) είναι υψηλή σε πεπτίδια

με αυτή τη συμπεριφορά. Αναπτύξαμε έτσι ένα πρόγραμμα (Παράρτημα, #19 high\_Blue.RMSDmatrix.pl) το οποίο εφαρμόζεται και πάλι στις δισδιάστατες αυτές κατανομές, αλλά αυτή τη φορά αντί να υπολογίσει διάμεσο και επικρατούσα τιμή, μετράει τον αριθμό των σημείων της κατανομής για τιμή  $\psi > 0.9$ . Το κατώφλι της τιμής 0.9 επιλέχθηκε αυθαίρετα μετά από οπτική εξέταση της κατανομής των πινάκων RMSD δοκιμάζοντας διαφορετικά cut-offs (0.80, 0.85, 0.90). Από την επεξεργασία αυτή προέκυψε μία καινούργια κατάταξη για τα 130 τετραπεπτίδια, οδηγώντας σε διαφορετική λίστα για τα 36 υποψήφια δυνητικά αναδιπλούμενα τετραπεπτίδια. Σύγκριση της με την αντίστοιχη της Εικόνας 3.8 ανέδειξε ότι το 61% των πεπτιδίων (από τα top36) είναι κοινό, και μόλις 15 πεπτίδια είναι διαφορετικά.

Έτσι προχωρήσαμε σε προσομοιώσεις μεγαλύτερης διάρκειας (100ns) των 36 τετραπεπτιδίων που αναδείχθηκαν από την εφαρμογή της συνάρτησης TF3 (Εικόνα 3.8) αλλά και των 15 αυτών τετραπεπτιδίων. Τα 15 αυτά τετραπεπτίδια (Εικόνα 3.10) αποτελούν ένα test data set, καθώς οι προσομοιώσεις μεγαλύτερης διάρκειας θα μας δείξουν την ορθότητα ή μη της πορείας που ακολουθήσαμε και τη διακριτική ικανότητα του αλγόριθμου των "επεκτεινομένων παραθύρων" που αναπτύξαμε.



Εικόνα 3.10 Γραφική απεικόνιση (σε αλφαβητική σειρά) των δισδιάστατων πινάκων RMSD των προσομοιώσεων διάρκειας 30ns, των 15 τετραπεπτιδίων που ορίσαμε ως test data-set.

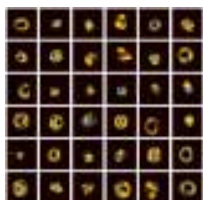
*"It's not what you look that matters, it's what you see."*

*Henry D. Thoreau*



*"Experiments should be reproducible:  
they should all fail in the same way."*

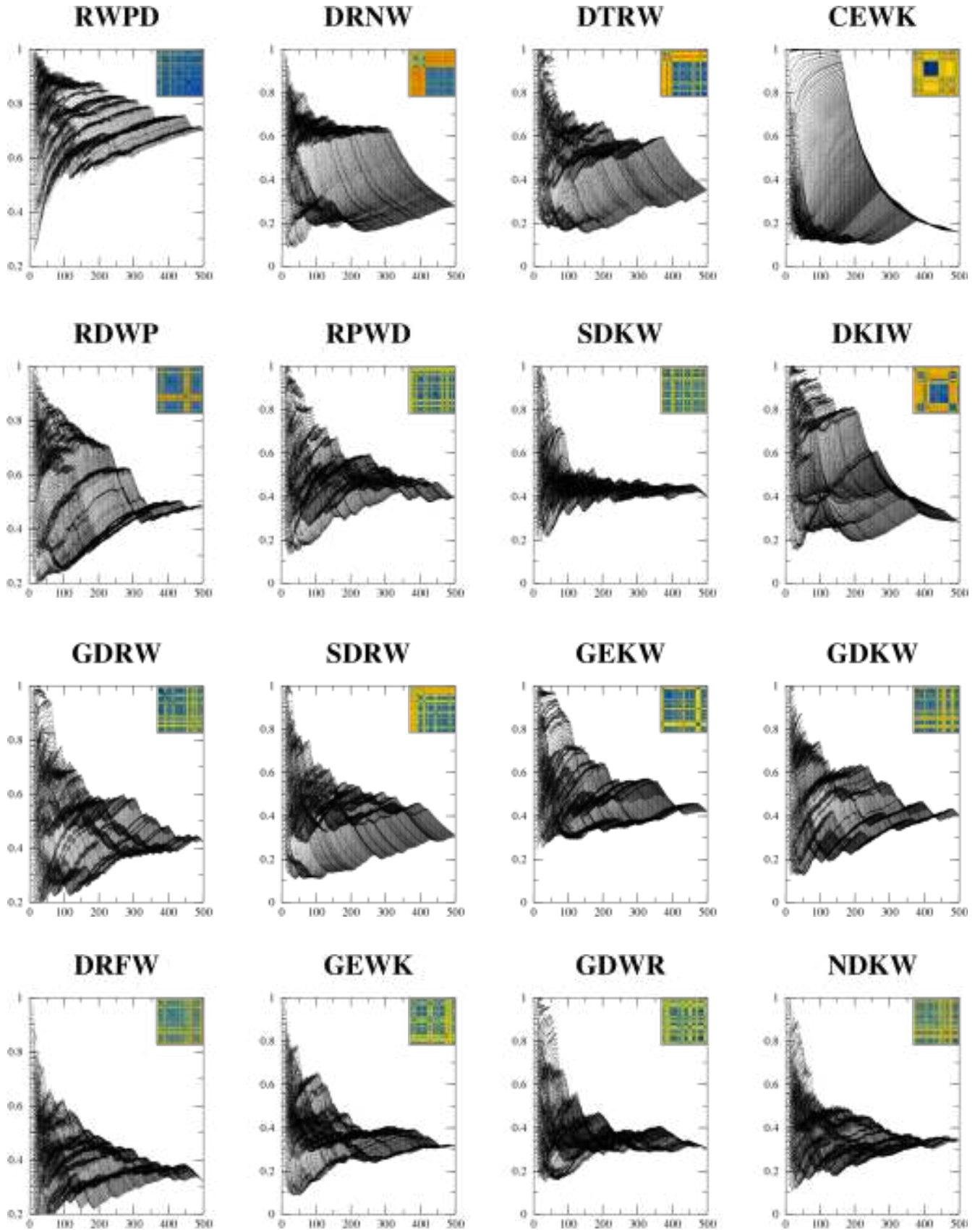
<http://www.gdargaud.net/Humor/QuotesScience.html>

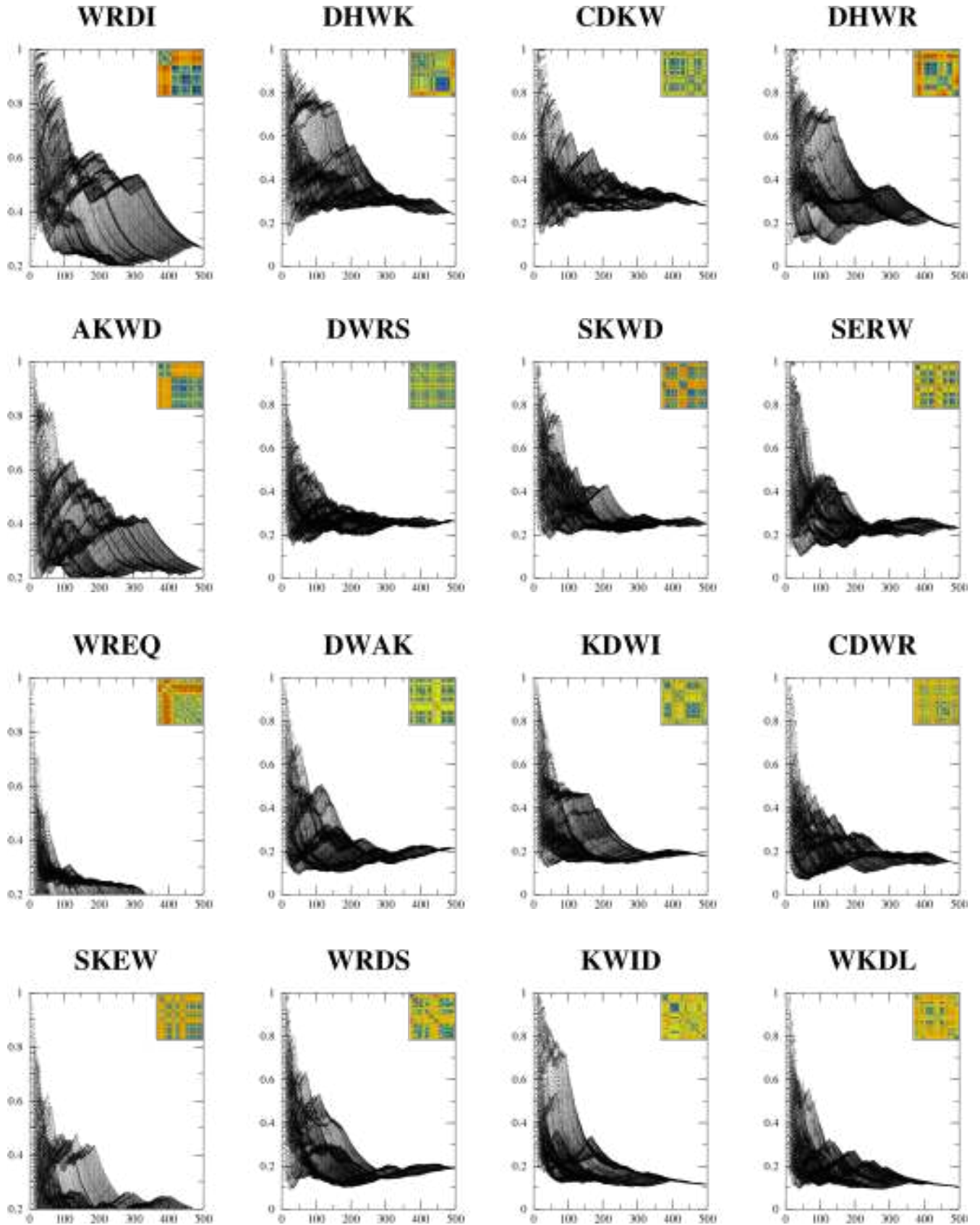


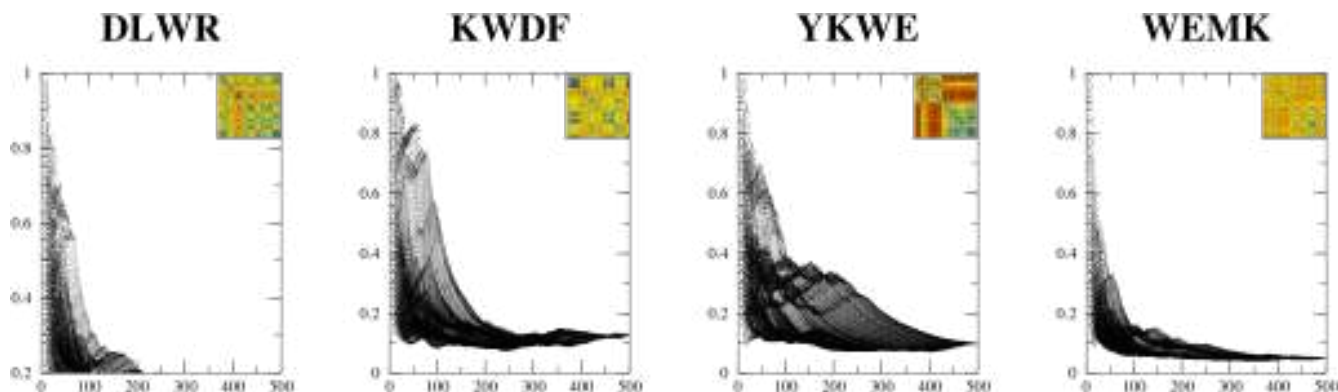
### 3.5 Επιλογή 4 υποψήφιων δυναμικά αναδιπλούμενων τετραπεπτιδίων

Τα 36 τετραπεπτίδια που επιλέχθηκαν με τον τρόπο που αναλύθηκε στην προηγούμενη ενότητα μαζί με τα 15 τετραπεπτίδια που ορίσαμε ως test data-set, μελετήθηκαν με προσομοιώσεις αναδίπλωσης διάρκειας 100ns. Η επιλογή της διάρκειας αυτής έγινε με γνώμονα την απόδοση (benchmark) του πρωτοκόλλου της προσομοίωσης, που είναι ~100ns/κόμβο/μέρα και το γεγονός ότι τα γεγονότα αναδίπλωσης σε τέτοιου μήκους πεπτίδια λαμβάνουν χώρα (με βάση πειραματικές μελέτες triplet-triplet energy transfer) στην κλίμακα των 10-20ns και η δημιουργία δομής θηλιάς στα 50-100ns (Bieri et al., 1999). Έτσι, το υπολογιστικό τμήμα των προσομοιώσεων αυτών μπορεί να ολοκληρωθεί σε ~10 μέρες, έχοντας στη διάθεσή μας ολόκληρη τη συστοιχία των υπολογιστών. Η επιμήκυνση του χρόνου της προσομοίωσης από 30ns σε 100ns θα μας επιτρέψει να εξετάσουμε όχι μόνο τη δημιουργία αναδιπλωμένης δομής, αλλά και τη δυναμική σταθερότητά της (Ενότητα 3.2). Το πρωτόκολλο της προσομοίωσης είναι πανομοιότυπο με αυτό που βρίσκεται στο Παράρτημα (#13, NAMD script, all.namd) με μοναδική διαφορά τον τελικό αριθμό βημάτων (run -> 50.000.000 steps).

Η επιμήκυνση του χρόνου της προσομοίωσης σε 100ns, μας επιτρέπει να προχωρήσουμε από τη συστηματική εκτίμηση της αναδιπλωσιμότητάς τους με συναρτήσεις, σε μια περισσότερη εις βάθος ανάλυση της δυναμικής των πεπτιδίων αυτών.







Εικόνα 3.11 Γραφικές παραστάσεις των κατανομών που προκύπτουν από τον αλγόριθμο των "επεκτεινομένων παραθύρων", του ποσοστού των μπλε pixels (άξονας  $\psi$ ) ως προς την διάσταση του εκάστοτε υποθετικού τετραγώνου (άξονας  $\chi$ ). Οι πίνακες RMSD έχουν διάσταση 500x500, καθώς έχουν προκύψει με βήμα 250 από τροχιακά των 100ns (125.000 frames). Η σειρά των 36 γραφημάτων από αριστερά προς τα δεξιά και από πάνω προς τα κάτω ακολουθεί τη (φθίνουσα) βαθμολογία τους με βάση τον αλγόριθμο των "επεκτεινομένων παραθύρων". Η γραφική απεικόνιση (σε κοινή χρωματική κλίμακα από σκούρο μπλε (0Å) ως σκούρο κόκκινο (6.23Å)) του πίνακα RMSD φαίνεται στο ένθετο κάθε γραφήματος.

Οι αναλύσεις που πραγματοποιήθηκαν (Ενότητα 2.4) αφορούν την εξέλιξη στο χρόνο των ατομικών αποστάσεων και της γυρεοσκοπικής ακτίνας, ανάλυση PCA τόσο στο Καρτεσιανό χώρο όσο και στο χώρο των διέδρων ( $\phi, \psi$ ) γωνιών (Cartesian-PCA, Dihedral-PCA) και ομαδοποίηση δομών (cluster analysis). Ακολουθεί μία συγκριτική παράθεση των αποτελεσμάτων για το σύνολο των 36 τετραπεπτιδίων.

Για την ομαδοποίηση των δομών του τροχιακού (cluster analysis), ακολουθήσαμε δύο προσεγγίσεις. Σε πρώτη φάση η ομαδοποίηση γίνεται με βάση το rmsd μεταξύ όλων των πιθανών δομών του τροχιακού. Προκύπτει έτσι ο δισδιάστατος πίνακας RMSD στον οποίο δίνουμε τη βαθμολογία μέσω του αλγόριθμου των "επεκτεινομένων παραθύρων" και η οποία αποτελεί μέρος της συνάρτησης TF3. Στην Εικόνα 3.11 παραθέτουμε τις κατανομές των 36 τετραπεπτιδίων που προκύπτουν από την εφαρμογή του αλγόριθμου.

Η δημιουργία ενός καλοσηματισμένου και συμπαγούς μπλε τετραγώνου (επί της διαγωνίου) απεικονίζεται στην κορυφή των κατανομών. Μάλιστα, όσο υψηλότερα απεικονίζονται στην κατανομή (όσο υψηλότερη η τιμή στον άξονα  $\psi$ ) τόσο πιο συμπαγές είναι το cluster, δηλαδή το σύνολο δομών που περιλαμβάνει σχετίζονται μεταξύ τους με χαμηλές τιμές rmsd. Επίσης, όσο μεγαλύτερο το εύρος της πυκνότητας των σημείων (ως προς τον άξονα  $\chi$ ) τόσο μεγαλύτερο το cluster, δηλαδή η κατοχή του σε χρόνο προσομοίωσης. Η επανεμφάνιση των clusters στη

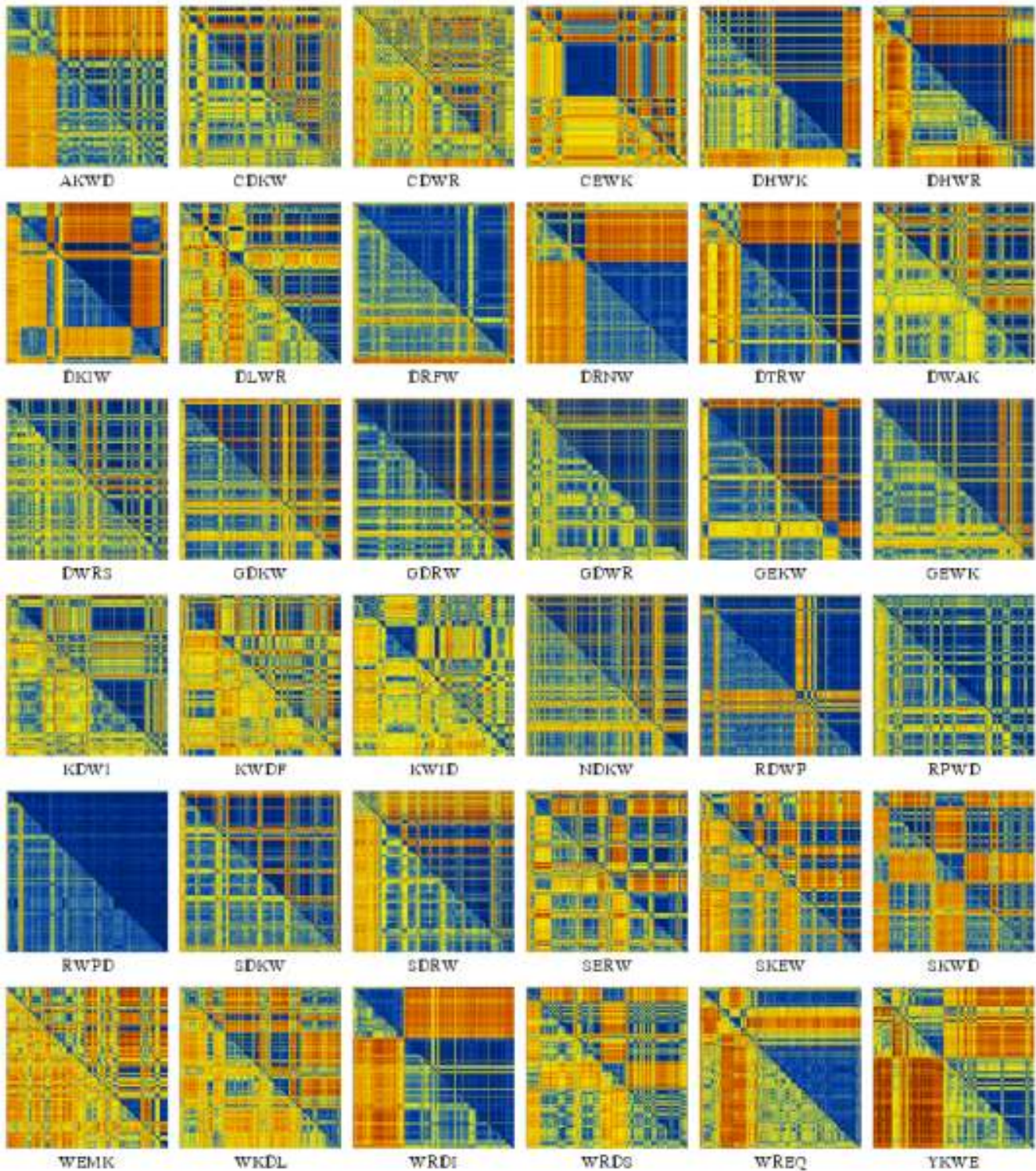
διάρκεια του τροχιακού (cross-vectors στους πίνακες RMSD) οι οποίες αντικατοπτρίζονται από τις μπλε γραμμές εκτός της διαγωνίου στη γραφική απεικόνιση των πινάκων RMSD, φαίνονται στις κατανομές ως λογαριθμικές καμπύλες με υψηλή πυκνότητα σημείων. Επίσης, όσο μεγαλώνει η διασπορά του cluster, η πυκνότητα των σημείων τείνει να είναι υψηλότερη σε χαμηλότερες τιμές του άξονα  $\psi$ . Οι κατανομές αυτές λοιπόν, μεταφέρουν τα στοιχεία τα οποία θέλουμε να εκμαιεύσουμε από τους πίνακες RMSD. Ωστόσο πρέπει να έχουμε υπόψιν μας ότι μετασηματίζοντας ένα διδιάστατο πίνακα διάστασης 500x500, στις κατανομές αυτές και ακολούθως σε δύο στατιστικά μέτρα (αλγόριθμος των “επεκτεινομένων παραθύρων”), προκειμένου να μπορεί να γίνει η εξέταση και η επεξεργασία τους με συστηματικό τρόπο, αναπόφευκτα χάνουμε πληροφορία.

Μία προσεκτική εξέταση των γραφημάτων των Εικόνων 3.11 και 3.12 μας επιτρέπει να συμπεράνουμε ότι οι πίνακες RMSD έχουν υψηλή πληροφοριακή αξία αλλά και ισχυρή διακριτική ικανότητα όσον αφορά την αναδιπλωσιμότητά μικρών πεπτιδίων. Έτσι το RWPD δικαίως παίρνει την υψηλότερη βαθμολογία καθώς εμφανίζει ένα αρκετά συμπαγές cluster που καταλαμβάνει το σύνολο της προσομοίωσης των 100ns και μάλιστα εμφανίζει τη μικρότερη μέγιστη τιμή RMSD στον πίνακα (4.56Å για όλα τα βαριά άτομα και 1.73Å για τα άτομα του σκελετού) σε σχέση με τα υπόλοιπα 36 τετραπεπτίδια. Επίσης το πεπτίδιο CEWK βαθμολογείται υψηλότερα από πεπτίδια όπως τα RDWP, RPWD και SDKW καθώς το cluster είναι μεν μικρότερο σε διάρκεια χρόνου αλλά περισσότερο συμπαγές, περιλαμβάνοντας πολύ κοντινές δομές σε αντίθεση με τα cluster των δύο άλλων πεπτιδίων, που εμφανίζουν μεγαλύτερη διασπορά.

Η παρουσία ενός cluster μεγάλης έκτασης αλλά και μεγάλης διασποράς είναι χαρακτηριστική της πληθώρας των πεπτιδίων (22 από τα 36) και οφείλεται στην (κινητική) αστάθεια αυτών και την παρουσία πολλαπλών γεγονότων αναδίπλωσης/αποδιάταξης. Μία μειονότητα πεπτιδίων φαίνεται να παρουσιάζει δύο διακριτές διαμορφώσεις που σχετίζονται με απότομη μετάβαση (DKIW, WRDI, AKWD, DRNW, SKWD) εκ των οποίων μόνο ένα (SKWD) φαίνεται να περνάει και από τις δύο διαμορφώσεις από δύο φορές. Σε κάθε περίπτωση, είναι εμφανές από την Εικόνα 3.12, ότι στην πλειοψηφία των πεπτιδίων αυτών (με ελάχιστες εξαιρέσεις, όπως το WREQ), η κινητική συμπεριφορά τους δεν διαφοροποιείται δραματικά εάν αντί για τα άτομα του πεπτιδικού σκελετού προσμετρήσουμε όλα τα βαριά άτομα. Το γεγονός αυτό υποδεικνύει ότι η σταθεροποίηση της δομής σε τέτοιου μήκους πεπτίδια δεν ακολουθεί τους κλασσικούς



## crossDCD RMSD matrices



Εικόνα 3.12 Συνοπτική γραφική αναπαράσταση των πινάκων RMSD των 36 τετραπεπτιδίων (σε αλφαβητική σειρά, για λόγους σύγκρισης). Πάνω από τη διαγώνιο, ο υπολογισμός του πίνακα RMSD έγινε

χρησιμοποιώντας τα άτομα του πεπτιδικού σκελετού (σε κοινή χρωματική κλίμακα από σκούρο μπλε (0Å) ως σκούρο κόκκινο (3.20Å)). Κάτω από τη διαγώνιο, ο υπολογισμός του πίνακα RMSD έγινε χρησιμοποιώντας όλα τα βαριά άτομα (σε κοινή χρωματική κλίμακα από σκούρο μπλε (0Å) ως σκούρο κόκκινο (6.23Å)).

κανόνες της δευτεροταγούς δομής των μεγαλύτερων πεπτιδίων, αλλά προκύπτει από αλληλεπιδράσεις μεταξύ όλων των ατόμων τόσο του σκελετού όσο και των πλευρικών ομάδων. Μόλις το 1/3 των τετραπεπτιδίων, από τα 36 που μελετήσαμε (η πρώτη σελίδα της Εικόνας 3.11) φαίνεται να δημιουργεί σταθερή δομή, εκ των οποίων μόνο 5 ή 6 πεπτιδία φαίνεται να τη διατηρούν για χρόνο αρκετό ώστε να χρήζουν περαιτέρω μελέτης.

Η ομαδοποίηση των δομών του τροχιακού (cluster analysis) μπορεί να προκύψει και μέσω της ανάλυσης PCA κατά την οποία εξάγουμε τους σημαντικούς βαθμούς ελευθερίας της κίνησης του πεπτιδίου (Ενότητα 2.4).

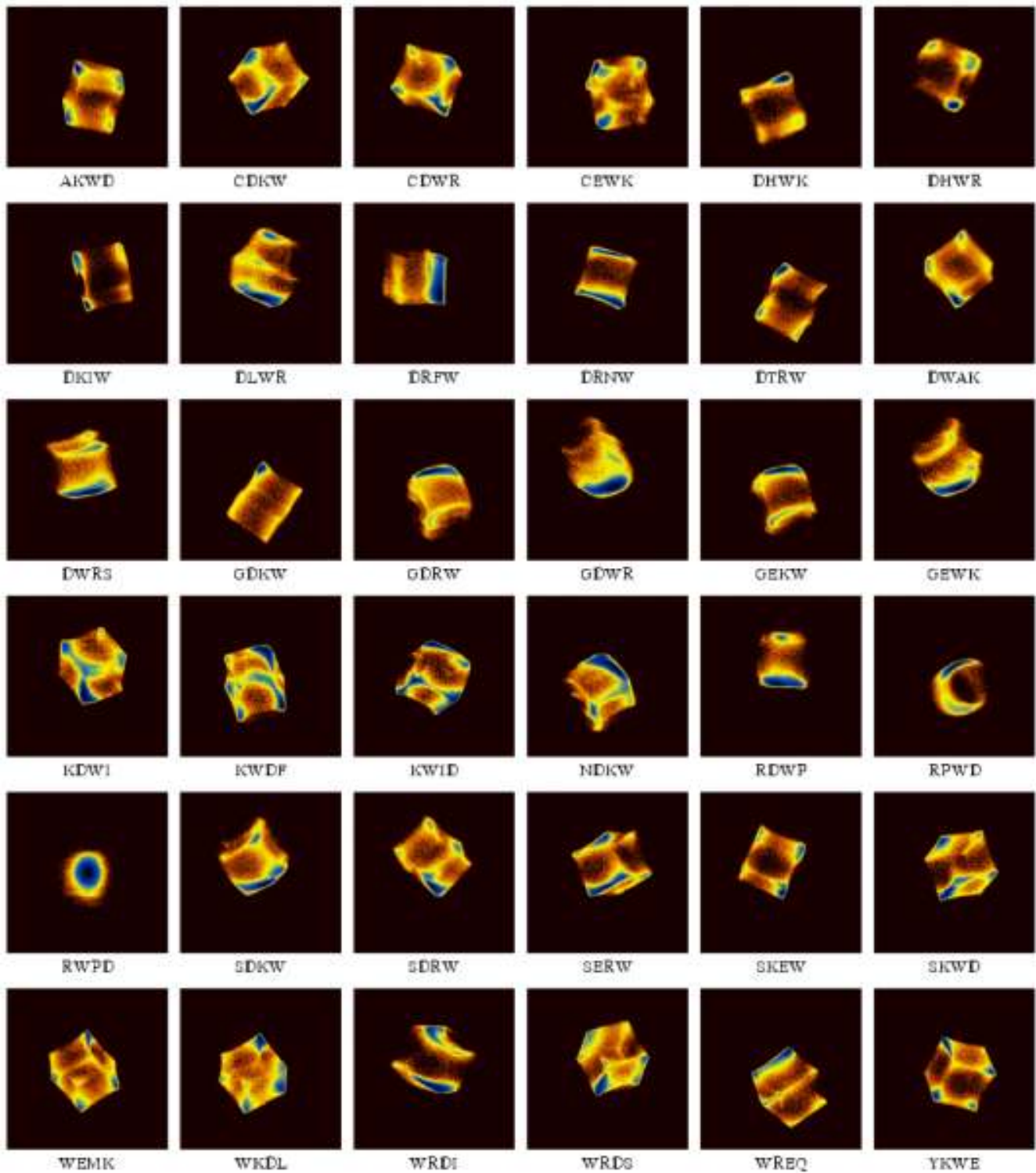
Στην περίπτωση των τετραπεπτιδίων πραγματοποιήσαμε ανάλυση δύο τύπων PCA. Στον ένα τύπο PCA η ανάλυση γίνεται στο χώρο των δίεδρων γωνιών  $\phi, \psi$  (Dihedral-PCA) και κατά συνέπεια δίνεται έμφαση στην κίνηση του πεπτιδικού σκελετού (Mu et al., 2005, Altis et al., 2007, Altis et al., 2008). Στο δεύτερο τύπο PCA (Cartesian-PCA), που γίνεται στο Καρτεσιανό σύστημα συντεταγμένων, λάβαμε υπόψιν όλα τα βαριά άτομα κατά τον υπολογισμό, ώστε να συμπεριλάβουμε και την κίνηση των πλευρικών ομάδων (Ichiye et al., 1991, Amadei et al., 1993).

Στις Εικόνες 3.13 και 3.14, βλέπουμε τα ενεργειακά τοπία που προκύπτουν από την προβολή των τροχιακών στο επίπεδο που ορίζεται από τους πρώτους δύο eigenvectors με τα μεγαλύτερα eigenvalues. Από τη μεταξύ τους σύγκριση βλέπουμε ότι τα αποτελέσματα είναι σε συμφωνία τόσο μεταξύ τους όσο και με τους πίνακες RMSD (Εικόνα 3.12). Η δημιουργία ενός καλοσηματισμένου cluster (μονή μπλε κορυφή), όπως προκύπτει από το Cartesian-PCA, των πεπτιδίων DRFW, GDKW, GDRW, GDWR, RDWP, RPWD, RWPD ισχύει και στην περίπτωση του Dihedral-PCA. Πεπτιδία όπως τα DHWK, GEWK φαίνεται να δίνουν 1 cluster με βάση το Dihedral-PCA αλλά στο Cartesian-PCA εμφανίζουν πολλαπλές κορυφές στο ενεργειακό τους τοπίο. Η διαφοροποίηση αυτή που οφείλεται στην κινητικότητα των πλευρικών ομάδων διαφαίνεται καθαρά και στους πίνακες RMSD της Εικόνας 3.12.

Εφόσον έχουμε ένα τρόπο να βαθμολογήσουμε τους πίνακες RMSD, αναζητήσαμε ένα μέτρο το οποίο θα μπορούσε να χρησιμοποιηθεί για να βαθμολογήσουμε τα ενεργειακά αυτά τοπία που προκύπτουν από την ανάλυση PCA. Στο σημείο αυτό είναι κριτικής σημασίας να έχουμε επιτύχει sufficient sampling, προκειμένου ένα τέτοιο μέτρο να μπορεί να χρησιμοποιηθεί με



## DeltaG Plots - Dihedral PCA



Εικόνα 3.13 Συνοπτική παρουσίαση των ενεργειακών τοπίων (DeltaG, energy landscapes) των προβολών των διακυμάνσεων στο επίπεδο των δύο κυρίαρχων συνιστωσών (principal components, PCs) από την

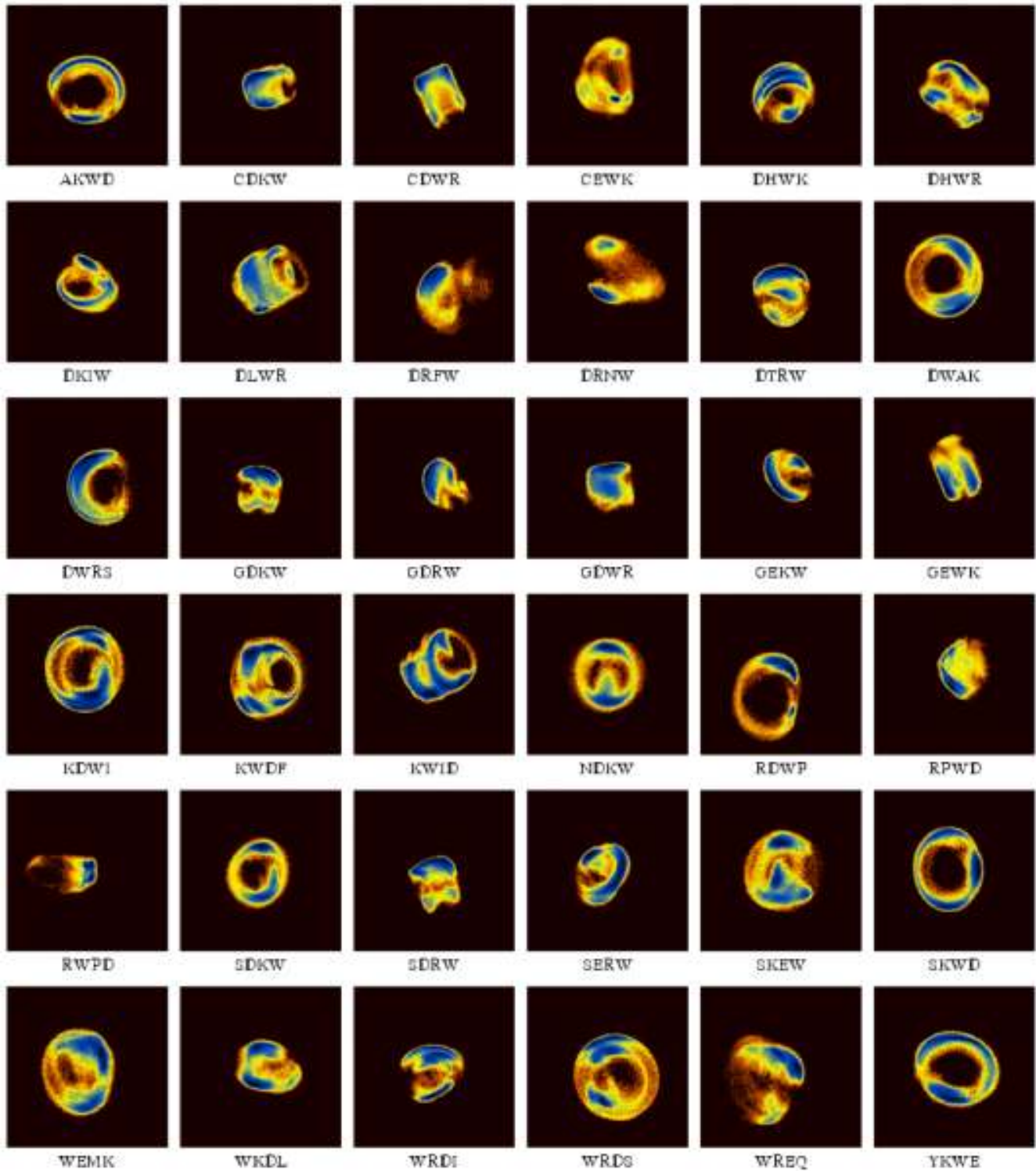
ανάλυση Dihedral-PCA, χρησιμοποιώντας 6 δίεδρες γωνίες. Όλα τα γραφήματα βρίσκονται στην ίδια κλίμακα, που κυμαίνεται από  $-3.50\text{kcal/mol}$  έως  $+3.50\text{kcal/mol}$  και στους δύο άξονες, με την αρχή των αξόνων στην πάνω αριστερά γωνία, και τον πρώτο principal component στον κάθετο άξονα.

ορθότητα (Smith et al., 2002, Matthes et al., 2009, Maisuradze et al., 2007).

Αυτό μπορεί να εξακριβωθεί με διάφορους τρόπους, όπως για παράδειγμα να πραγματοποιήσουμε ανάλυση PCA στο πρώτο και δεύτερο (μη επικαλυπτόμενο) μισό του τροχιακού και να υπολογίσουμε την επικάλυψη (overlap) των eigenvectors (Hess, 2002). Επειδή η τυχαία διάχυση μπορεί να παραληφθεί και ως συσχετιζόμενη κίνηση κατά την ανάλυση PCA (Hess 2000), ο υπολογισμός του cosine content (τιμές 0-1) ενός principal component, είναι καλός δείκτης του μη επαρκούς sampling (το cosine content τείνει στη μονάδα) (Hess, 2002, Maisuradze et al., 2007). Έτσι για πεπτίδια με μοναδική διαμόρφωση (two-state) παίρνουμε τιμές για το cosine content κοντά στο μηδέν, που δείχνει ότι ο χρόνος της προσομοίωσης των 100ns είναι επαρκής. Για πεπτίδια με περισσότερες διακριτές διαμορφώσεις, που περιλαμβάνονται στο δικό μας σύνολο των 36 τετραπεπτιδίων, η τιμή του cosine content είναι υψηλή (0.5-0.7) όταν αφορά ολόκληρο το τροχιακό, αλλά τείνει να πλησιάσει και πάλι το μηδέν εάν ο υπολογισμός γίνει μόνο στο τμήμα του τροχιακού που ανήκει σε ένα cluster.

Με την παραδοχή λοιπόν ότι έχουμε επιτύχει sufficient sampling, αναζητήσαμε ένα μέτρο που να αποδίδει την πληροφορία των ενεργειακών τοπίων των πεπτιδίων, για να μπορεί δυνητικά να χρησιμοποιηθεί με συστηματικό τρόπο, χωρίς οπτική εξέταση των γραφημάτων. Υπολογίσαμε την εντροπία κατά Shannon (Shannon, 1948) της κατανομής των τριών κυρίαρχων principal components (Li et al., 2007), οι οποίοι αντιπροσωπεύουν την πλειοψηφία των κύριων κινήσεων του πεπτιδίου (Ichiye et al., 1991, Amadei et al., 1993, Garcia 1992). Στην Εικόνα 3.15 βλέπουμε την εντροπία της κατανομής των τριών principal components, όπως προέκυψαν τόσο από Cartesian-PCA όσο και από Dihedral-PCA, καθώς και τη μεταξύ τους σύγκριση. Η συμφωνία των τιμών της εντροπίας με τα διαγράμματα των πινάκων RMSD, συμβάλλουν περαιτέρω στον ισχυρισμό ότι τα 100ns ήταν επαρκής χρόνος για να μελετήσουμε τη δημιουργία δομής και τη σταθερότητα αυτής σε πεπτίδια μήκους τεσσάρων καταλοίπων. Στην περίπτωση που η εντροπία αφορά όλα τα βαριά άτομα, την ελάχιστη τιμή την συναντάμε για το RWPD, ακολουθούμενο από τα GDRW, GDKW, DKIW. Οι υψηλότερες τιμές εντροπίας για τα υψηλά βαθμολογούμενα πεπτίδια της Εικόνας 3.11, όπως για παράδειγμα τα DRNW, DTRW, CEWK, δικαιολογείται από το γεγονός ότι στα πεπτίδια αυτά το ενεργειακό τους τοπίο δείχνει δύο

## DeltaG Plots - Cartesian PCA



Εικόνα 3.14 Συνοπτική παρουσίαση των ενεργειακών τοπίων (DeltaG, energy landscapes) των προβολών των διακυμάνσεων στο επίπεδο των δύο κυρίαρχων συνιστωσών (principal components, PCs) από την

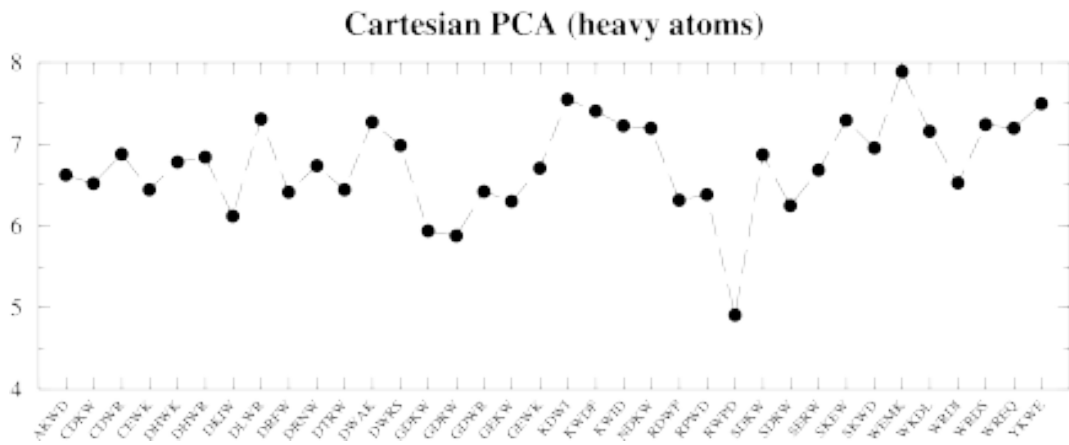
ανάλυση Cartesian-PCA, χρησιμοποιώντας 47 βαριά άτομα. Όλα τα γραφήματα βρίσκονται στην ίδια κλίμακα, που κυμαίνεται από -47.50kcal/mol έως +47.50kcal/mol και στους δύο άξονες, με την αρχή των αξόνων στην πάνω αριστερά γωνία, και τον πρώτο principal component στον κάθετο άξονα.

ξεκάθαρες κορυφές με αποτέλεσμα η κατανομή να απέχει από την κανονική.

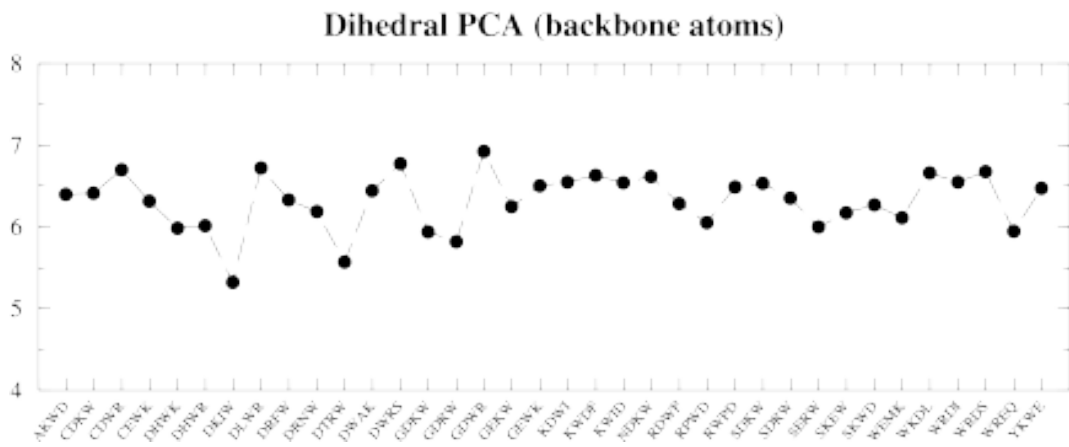
Βλέπουμε λοιπόν ότι το μέτρο αυτό δε μπορεί να χρησιμοποιηθεί με συστηματικό τρόπο για την περίπτωση πεπτιδίων με περισσότερες από μία διακριτές διαμορφώσεις, ειδικά στην περίπτωση που δεν γνωρίζουμε *a priori* την αναδιπλωμένη δομή (native state) του πεπτιδίου. Η εντροπία της κατανομής των τριών principal components από Dihedral-PCA, επίσης δεν συνιστά κατάλληλο μέτρο, καθώς περιέχει πληροφορία μόνο για τον πεπτιδικό σκελετό. Για την πλειοψηφία των πεπτιδίων οι τιμές της εντροπίας είναι κοινές μεταξύ Cartesian-PCA και Dihedral-PCA (όπως και οι πίνακες RMSD χρησιμοποιώντας όλα τα βαριά άτομα και μόνο τα άτομα του πεπτιδικού σκελετού). Υπάρχουν όμως και περιπτώσεις πεπτιδίων όπου υπάρχει σημαντική διαφορά στις δύο τιμές εντροπίας, όπως στα DHWK, DHWR, DKIW, DRNW, DTRW τα οποία αφορούν περιπτώσεις με δύο κορυφές στο ενεργειακό τους τοπίο και δύο cluster δομών στα γραφήματα των πινάκων RMSD. Στα πεπτίδια αυτά βλέπουμε όμως διαφορετική κινητικότητα των ατόμων του πεπτιδικού σκελετού και των ατόμων των πλευρικών ομάδων (Εικόνα 3.12), εξ ου και η διαφορά στην εντροπία μεταξύ Cartesian-PCA και Dihedral-PCA (Εικόνα 3.15). Η διαφορά αυτή, σε πεπτίδια όπως το WREQ είναι τόσο έντονη ώστε να είναι έκδηλη η ακαταλληλότητα της εντροπίας με βάση την ανάλυση Dihedral-PCA καθώς το πεπτίδιο αυτό έρχεται 5ο σε κατάταξη με βάση την εντροπία του Dihedral-PCA, αλλά 26ο σε κατάταξη με βάση την εντροπία του Cartesian-PCA. Στο σημείο αυτό θα πρέπει να τονίσουμε ότι τα ενεργειακά τοπία των πεπτιδίων όπως προκύπτουν από τα δύο είδη ανάλυσης PCA και παρουσιάζονται στις Εικόνες 3.13 και 3.14 αφορούν τους δύο κυρίαρχους principal components και συνεπώς είναι δύο διαστάσεων, ενώ η εντροπία (Εικόνα 3.15) υπολογίζεται για την κατανομή στις τρεις διαστάσεις που ορίζονται από τους τρεις κυρίαρχους principal components. Η απώλεια πληροφορίας λόγω της μείωσης κατά μία διάσταση θα πρέπει να ληφθεί υπόψιν κατά την κριτική εξέτασή τους.

Όπως έχουμε προαναφέρει, η εξέλιξη στο χρόνο των ατομικών αποστάσεων μεταξύ των ακραίων ατόμων Ca, είναι ενδεικτική της δημιουργίας δομής θηλιάς, και διαδραματίζει κεντρικό ρόλο στις συναρτήσεις εκτίμησης της αναδιπλωσιμότητάς TF1 και TF2 που αναπτύξαμε.

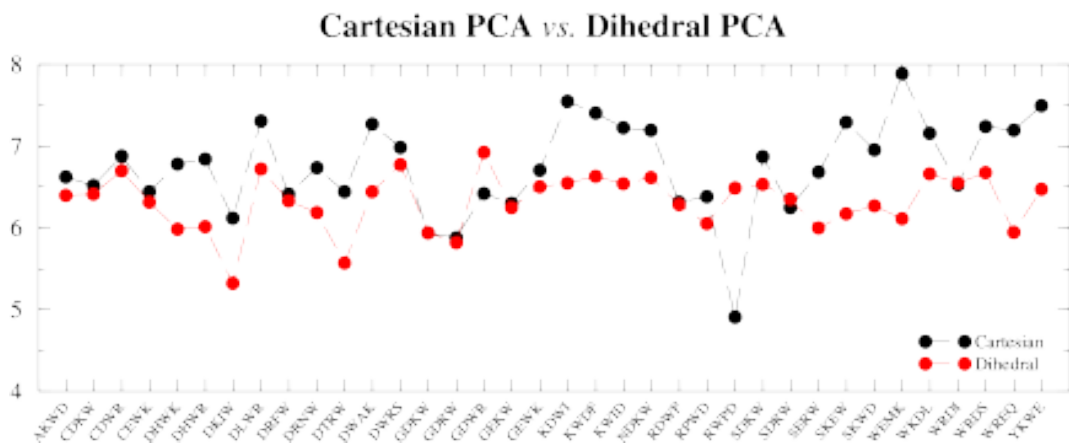
### Configurational Entropy of PC distribution



### Configurational Entropy of PC distribution

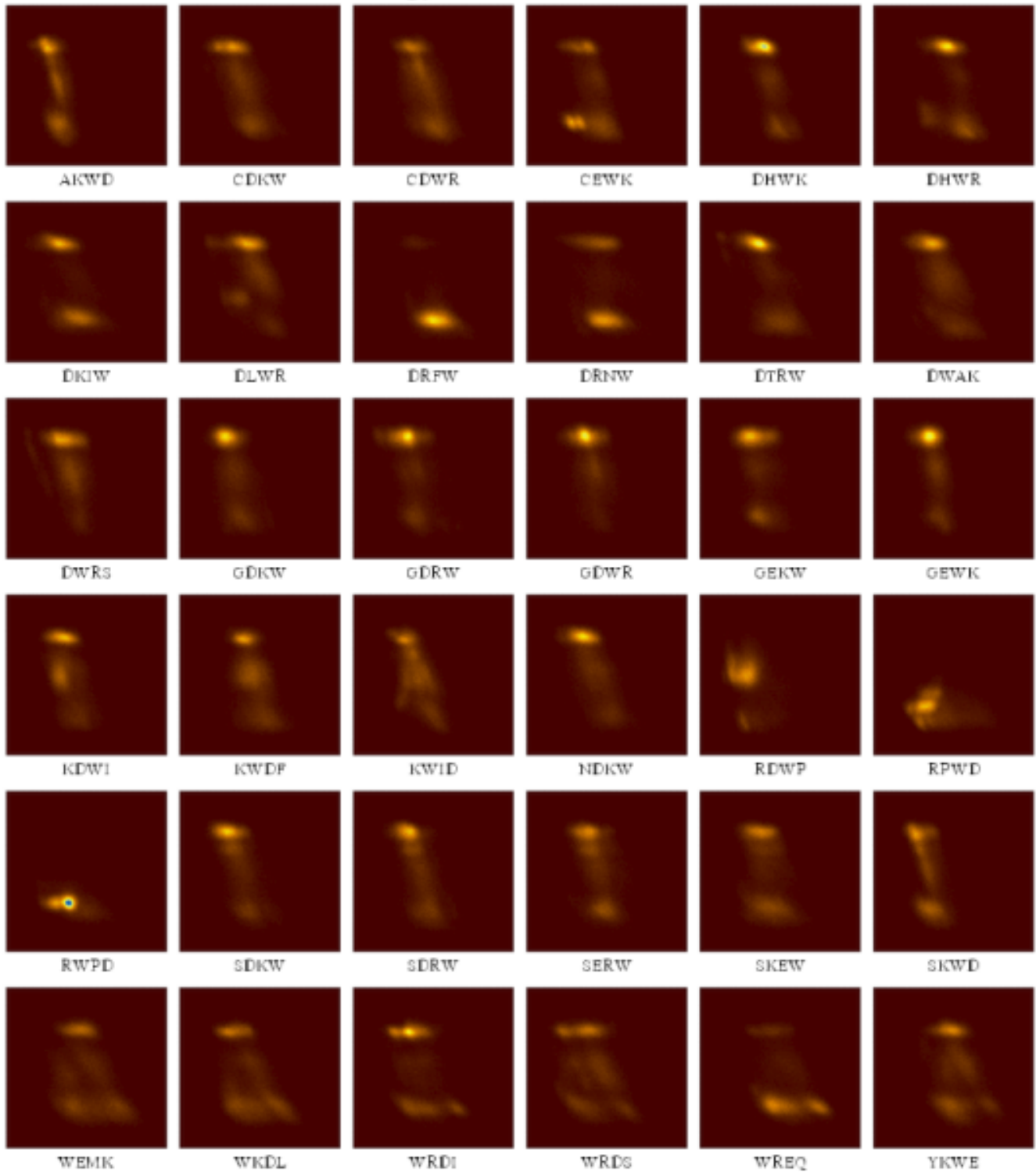


### Configurational Entropy of PC distribution



Εικόνα 3.15 Εντροπία κατά Shannon της κατανομής των τριών κυρίαρχων principal components.

Radius of gyration VS N-C distance



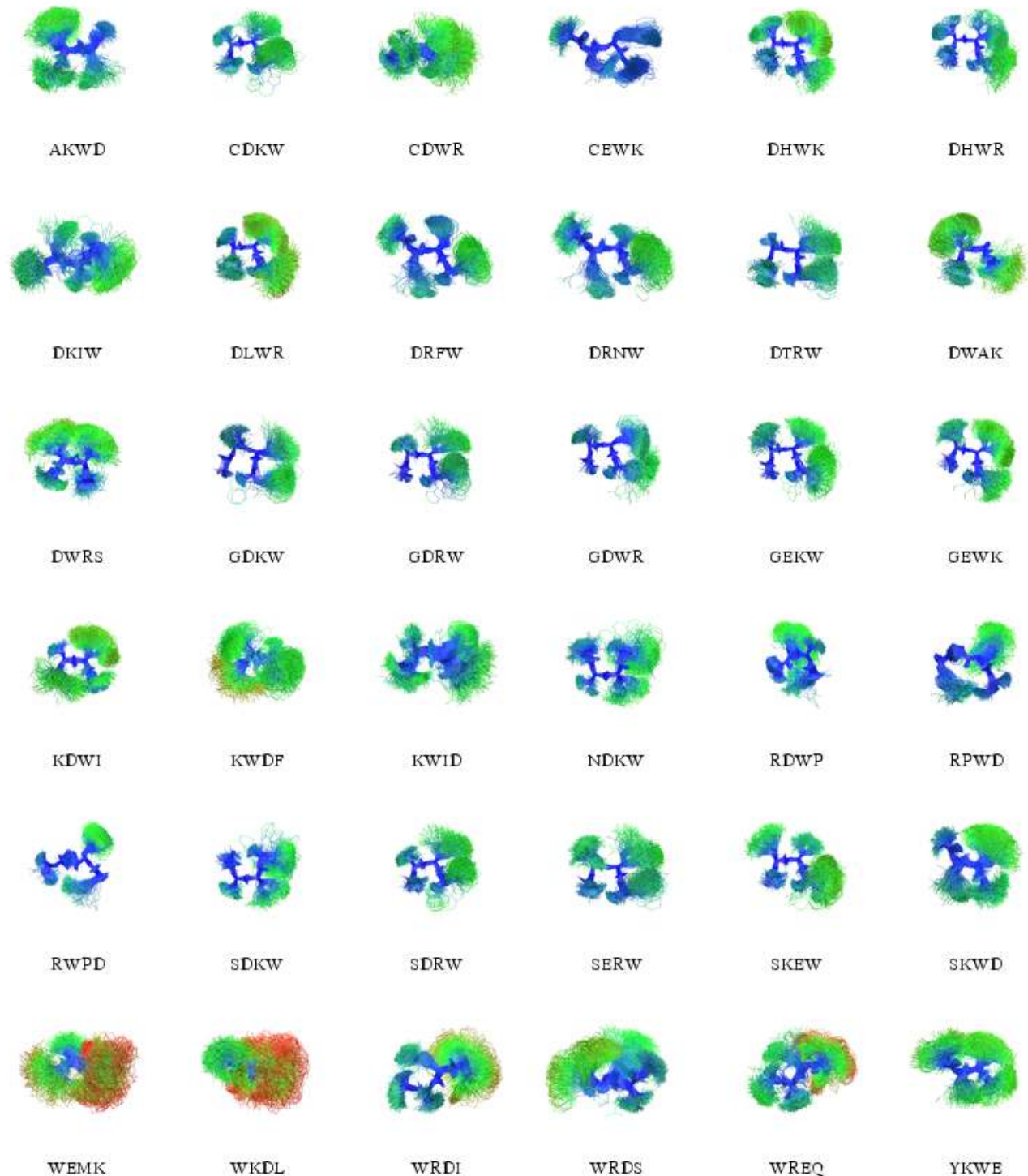
Εικόνα 3.16 Συνοπτική παρουσίαση των δισδιάστατων κατανομών της εξέλιξης της γυρεοσκοπικής ακτίνας ως προς την εξέλιξη της απόστασης μεταξύ του N- τελικού και του C- τελικού άκρου, για τα 36



τετραπεπτίδια (σε αλφαβητική σειρά). Για όλες τις κατανομές υπολογίστηκε πλέγμα (grid) με κοινές διαστάσεις (3.69-6.59 για τον άξονα  $\chi$  και 3.43-11.58 για τον άξονα  $\psi$ ) και η χρωματική κλίμακα διατηρήθηκε παντού από σκούρο κόκκινο (0) σε κίτρινο έως σκούρο μπλε (191). Η αρχή των αξόνων είναι στην πάνω αριστερά γωνία, με την γυρεοσκοπική ακτίνα στον οριζόντιο άξονα.

Μία ακόμα παράμετρος με ιδιαίτερο ενδιαφέρον είναι η εξέλιξη στο χρόνο της γυρεοσκοπικής ακτίνας (radius of gyration). Η τιμή  $R_g$  (heavy atoms, mass-weighted) αποτελεί μία ένδειξη για το πόσο συμπαγές είναι το σχήμα της δομής που υιοθετεί το μόριο, ενώ η μελέτη της εξέλιξής της συνιστά μία ένδειξη για τη σταθερότητα και τη διατήρηση της δομής (Zagrovic et al., 2005). Στην Εικόνα 3.16 βλέπουμε την εξέλιξη της γυρεοσκοπικής ακτίνας συναρτήσει της εξέλιξης της απόστασης των ακραίων ατόμων Ca για τα 36 τετραπεπτίδια. Βλέπουμε ότι οι κατανομές αυτές αναπαριστούν επιτυχώς τη δημιουργία ομάδων δομών (clusters) όπως στα πεπτίδια DHWK, DRFW, DTRW, GDKW, GDRW, GDWR, GEWK, RWPD αλλά και τη μετάβαση σε αυτές από την εκτεταμένη διαμόρφωση μέσω της διαδρομής της διάχυσης των σημείων. Πεπτίδια ασταθή όπως τα KWID, WEMK, WKDL, WRDS εμφανίζουν μεγαλύτερη διασπορά. Χαρακτηριστική είναι και η εμφάνιση των κατανομών των πεπτιδίων που περιέχουν προλίνη και τα οποία πραγματοποιούν πολλαπλά γεγονότα αναδίπλωσης/αποδιάταξης, όπως τα RDWP και RPWD. Παρότι κομψές οι κατανομές αυτές, δεν μας προσφέρουν κάποια καινούργια πληροφορία σε σχέση με τα όσα είδαμε από τις μέχρι τώρα αναλύσεις. Επιπλέον, από τις κατανομές αυτές είναι αμφίβολο κατά πόσο μπορεί να εξαχθεί ένα μέτρο που να μπορεί να χρησιμοποιηθεί με συστηματικό τρόπο για την εκτίμηση της αναδίπλωσης των πεπτιδίων και την κατάταξή τους ως προς αυτήν. Διατηρούν όμως την αξία τους ως συμπληρωματική ανάλυση για σύγκριση και επιβεβαίωση με τα υπόλοιπα αποτελέσματα.

Τέλος και για λόγους πληρότητας, στην Εικόνα 3.17 παραθέτουμε αντιπροσωπευτικές δομές (σε υπέρθεση) του κυρίαρχου cluster, όπως προέκυψε από την ανάλυση Cartesian-PCA και χρησιμοποιώντας όλα τα βαριά άτομα. Η δημιουργία σταθερών δομών (CEWK, DRFW, DTRW, RDWP, RPWD, RWPD) διαφαίνεται από τη διαβάθμιση του χρώματος που ακολουθεί τις ατομικές διακυμάνσεις (RMSFs) σε σχέση με την (υπολογιζόμενη) μέση δομή του cluster. Στην πλειοψηφία των πεπτιδικών δομών παρατηρούμε κοινά "μοτίβα αναδίπλωσης". Ο πεπτιδικός σκελετός υιοθετεί δομή β-στροφής λόγω ηλεκτροστατικών αλληλεπιδράσεων μεταξύ του N-τελικού και του C-τελικού άκρου και η τρυπτοφάνη σταθεροποιείται μέσω υδρόφοβου πακεταρίσματος με κάποιο φορτισμένο αμινοξύ (αργινίνη ή λυσίνη) ή με το δακτύλιο της



Εικόνα 3.17 Συνοπτική παρουσίαση αντιπροσωπευτικών δομών (σε υπέρθεση) του κυρίαρχου cluster, όπως προέκυψε από την ανάλυση Cartesian-PCA, για τα 36 τετραπεπίδια. Για την υπέρθεση των δομών χρησιμοποιήθηκαν μόνο τα άτομα του πεπτιδικού σκελετού. Οι δομές έχουν χρωματιστεί με βάση τις

ατομικές διακυμάνσεις (RMSFs) που υπολογίστηκαν από τη μέση δομή του cluster, και έχει διατηρηθεί κοινή χρωματική κλίμακα από μπλε (0.16Å) έως κόκκινο (6.30Å), ενδεικτική της κινητικότητας κάθε ατόμου. Τα πρωτόνια έχουν αφαιρεθεί για λόγους ευκρίνειας και απεικονίζονται μόνο τα βαριά άτομα.

προλίνης. Σε κάποιες περιπτώσεις, ο πεπτιδικός σκελετός διατηρεί περισσότερο εκτεταμένη διαμόρφωση και έτσι η δομή σταθεροποιείται με αλληλεπιδράσεις μεταξύ του σκελετού και των πλευρικών ομάδων, όπως στα πεπτίδια DKIW, WRDI, WRDS, WREQ.

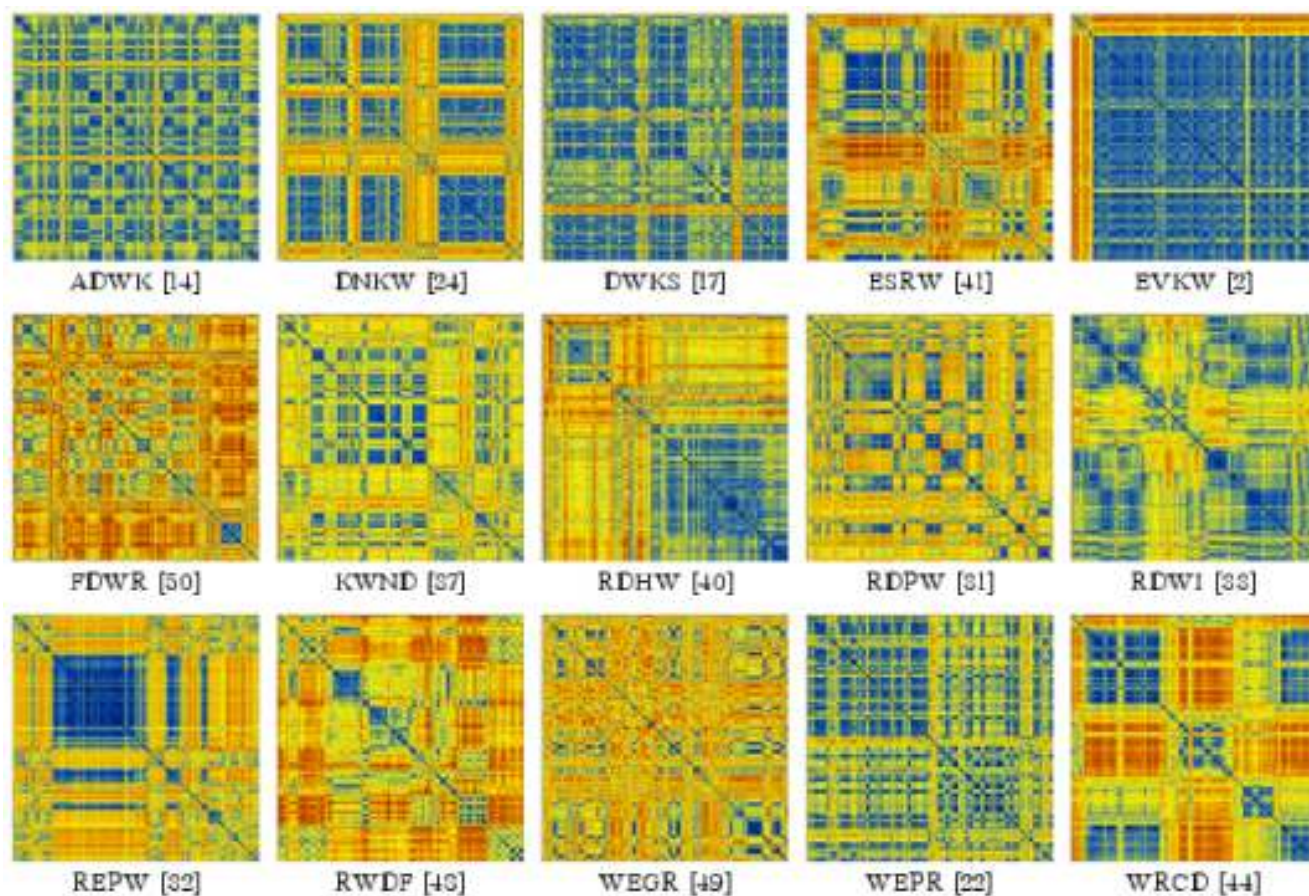
Η ανάλυση που παρουσιάστηκε στην ενότητα αυτή μας οδήγησε στην επιλογή των 4 υποψήφιων αναδιπλωμένων τετραπεπτιδίων. Από τη συστηματική μελέτη με τις συναρτήσεις εκτίμησης της αναδιπλωσιμότητάς παίρνουμε την κατάταξη των 36 τετραπεπτιδίων της Εικόνας 3.18, μετά την εφαρμογή της συνάρτησης TF3.



Εικόνα 3.18 *Αριστερά*: Κατάταξη των 36 τετραπεπτιδίων με βάση τη βαθμολογία της συνάρτησης TF3. Η διαβάθμιση του χρώματος από κίτρινο σε πορτοκαλί έως κόκκινο ακολουθεί την άνοδο στη βαθμολογία.

*Δεξιά*: Απεικόνιση της ίδιας κατάταξης με word-cloud. Οι πεπτιδικές αλληλουχίες κατατάσσονται με αλφαβητική σειρά ενώ το μέγεθος της γραμματοσειράς είναι ανάλογο της βαθμολογίας της συνάρτησης TF3.

Έτσι, το επόμενο βήμα θα ήταν να επιλέξουμε τα τέσσερα πεπτίδια με την υψηλότερη βαθμολογία με βάση τη συνάρτηση TF3, δηλαδή τα RWPД, CEWK, DTRW, DRNW. Τα ίδια τετραπεπτίδια, με σειρά RWPД, DRNW, DTRW, CEWK, έδωσαν τις τέσσερις υψηλότερες βαθμολογίες για τους πίνακες RMSD με βάση τον αλγόριθμο των "επεκτεινομένων παραθύρων". Ωστόσο, έχουμε και ένα σύνολο από 15 τετραπεπτίδια τα οποία χαρακτηρίσαμε ως test data set, για να εξετάσουμε τη διακριτική ικανότητα του αλγόριθμου αλλά και την ορθότητα της επιλογής μας των 36 τετραπεπτιδίων. Στην Εικόνα 3.19 βλέπουμε την γραφική απεικόνιση των πινάκων RMSD για τα 15 τετραπεπτίδια από τις προσομοιώσεις διάρκειας 100ns (σύγκριση με τους αντίστοιχους πίνακες της Εικόνας 3.10 που υπολογίστηκαν από προσομοιώσεις διάρκειας 30ns).



Εικόνα 3.19 Γραφική απεικόνιση (σε αλφαβητική σειρά) των δισδιάστατων πινάκων RMSD των προσομοιώσεων διάρκειας 100ns, των 15 τετραπεπτιδίων που ορίσαμε ως test data-set. Οι αριθμοί στις αγκύλες [], υποδηλώνουν την κατάταξη των πινάκων RMSD με βάση τη βαθμολογία του αλγόριθμου των "επεκτεινομένων παραθύρων" στο σύνολο των 51 τετραπεπτιδίων.

Για την πλειοψηφία των πεπτιδίων η επέκταση του χρόνου της προσομοίωσης από 30ns σε 100ns δεν είχε σημαντική διαφοροποίηση στην εικόνα των αποτελεσμάτων. Μάλιστα, η βαθμολογική τους κατάταξη στο σύνολο των 51 τετραπεπτιδίων είναι αρκετά χαμηλή, με μοναδική εξαίρεση το πεπτίδιο EVKW.

Για τον επόμενο λοιπόν κύκλο των προσομοιώσεων επιλέξαμε να μελετήσουμε τα πεπτίδια RWPD και DTRW που έδωσαν την πρώτη και τρίτη υψηλότερη βαθμολογία με βάση τη συνάρτηση TF3 στο σύνολο των 36 τετραπεπτιδίων, αντίστοιχα (Εικόνα 3.18), και το πεπτίδιο EVKW που έδωσε την υψηλότερη βαθμολογία στο σύνολο των 15 πεπτιδίων του test data set. Λόγω της ιδιαίτερης επίδρασης της παρουσίας της προλίνης στη δομή των πεπτιδίων αυτών επιλέξαμε, στη θέση του πεπτιδίου CEWK (με τη δεύτερη υψηλότερη βαθμολογία, Εικόνα 3.18), το πεπτίδιο RPWD με τη δεύτερη καλύτερη βαθμολογία μεταξύ των πεπτιδίων που περιέχουν προλίνη. Ο λόγος που δεν προχωρήσαμε με το πεπτίδιο CEWK διαφαίνεται στην Εικόνα 3.11 αλλά και στο σύνολο της ανάλυσης που πραγματοποιήσαμε (Εικόνες 3.12 – 3.17). Το πεπτίδιο αυτό εμφανίζει ένα ιδιαίτερα συμπαγές cluster από πολύ κοντινές μεταξύ τους δομές με μικρές τιμές ατομικών διακυμάνσεων (μέση τιμή rmsf για τα άτομα της πλευρικής ομάδας της τρυπτοφάνης είναι 0.84Å και 0.69Å για όλα τα βαριά άτομα), αλλά η κατοχή του σε χρόνο προσομοίωσης είναι μόλις 30%. Επιπλέον, βλέπουμε ότι το πεπτίδιο περνάει από τις διαμορφώσεις αυτές αρκετές φορές στη διάρκεια του τροχιακού, τόσο πριν τη δημιουργία του μεγάλου cluster όσο και μετά. Από την άλλη, το πεπτίδιο DRNW σχηματίζει cluster με κατοχή 56% αλλά οι μέσες ατομικές διακυμάνσεις είναι υψηλότερες (μέση τιμή rmsf για τα άτομα της πλευρικής ομάδας της τρυπτοφάνης είναι 1.53Å και 1.29Å για όλα τα βαριά άτομα).

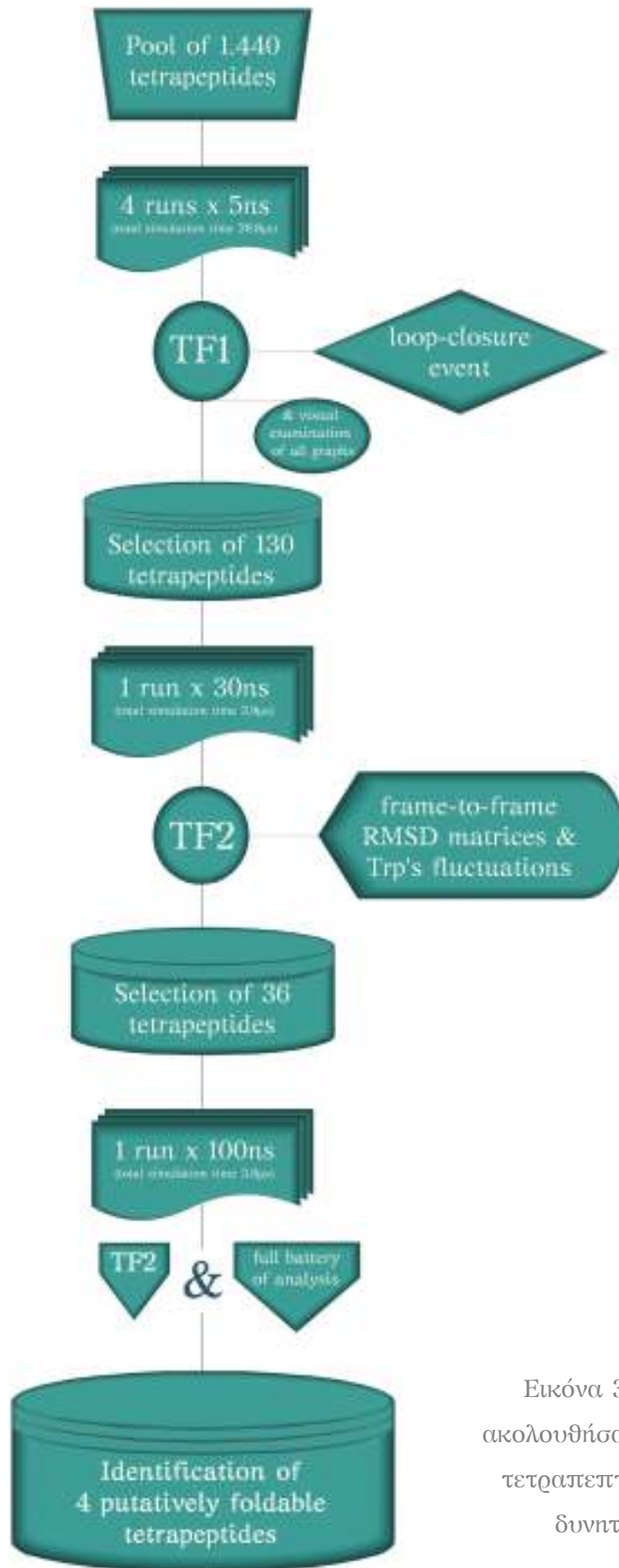
Οι παρατηρήσεις αυτές μας οδηγούν σε μία βασική διαπίστωση. Καθώς αυξάνουμε το χρόνο της προσομοίωσης και αποκτούμε καλύτερη γνώση του συνόλου των διαμορφώσεων από τα οποία περνάει το μόριο κατά την αναδίπλωσή του από την εκτεταμένη διαμόρφωση, δεν μπορούμε πλέον να στηριζόμαστε σε μία συστηματική προσέγγιση εκτίμησης της αναδιπλωσιμότητας. Έτσι όταν φτάσουμε σε ένα μικρό αριθμό υποψήφιων πεπτιδίων, προχωράμε σε μία συγκριτική και κυρίως κριτική μελέτη ενός συνόλου παραμέτρων κατά την ανάλυση της δυναμικής των πεπτιδίων.

Πριν προχωρήσουμε σε μία εις βάθος ανάλυση της δυναμικής των τεσσάρων αυτών τετραπεπτιδίων, παραθέεται στην Εικόνα 3.20 ένα διάγραμμα ροής της πορείας που ακολουθήσαμε μέχρι τώρα.



*" There are two possible outcomes:  
if the result confirms the hypothesis,  
then you've made a measurement.  
If the result is contrary to the hypothesis,  
then you've made a discovery. "*

*Enrico Fermi*

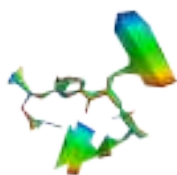


Εικόνα 3.20 Διάγραμμα ροής της πορείας που ακολουθήσαμε, ξεκινώντας από το σύνολο των 1.440 τετραπεπτιδίων για να καταλήξουμε στα τέσσερα δυνητικά αναδιπλούμενα τετραπεπτίδια.



*"If you aren't confused by quantum mechanics,  
you haven't really understood it."*

*Neil's Bohr*



### 3.6 Μελέτη της αναδίπλωσης των RWPD, DTRW, RPWD, EVKW σε τέσσερις θερμοκρασίες

Την τελευταία δεκαετία έχει δειχθεί επανειλημμένως ότι οι προσομοιώσεις μοριακής δυναμικής με τα εμπειρικά δυναμικά πεδία (force fields) της παρούσας γενιάς και με αναλυτική παρουσία του διαλύτη, είναι ένα αποτελεσματικό εργαλείο για τη μελέτη της αναδίπλωσης μικρών πεπτιδίων από την εκτεταμένη διαμόρφωση, αλλά και της αποδιάταξής τους ξεκινώντας από την πειραματικά προσδιορισμένη δομή (Daura et al., 1998, Daura et al., 1999, vanGunsteren et al., 2001, Rao et al., 2003, Rao et al., 2007). Η επιλογή της θερμοκρασίας για τη διεξαγωγή της προσομοίωσης φαίνεται να διαδραματίζει καίριο ρόλο στο αποτέλεσμα (Aliev et al., 2010) αλλά και στη σύγκριση των προσομοιώσεων με πειραματικά δεδομένα (Daura et al., 1999, Daura et al., 2001, Snow et al., 2002, Rousseau et al., 2004).

Χρησιμοποιώντας ένα εύρος θερμοκρασιών κατά τις προσομοιώσεις αναδίπλωσης των πεπτιδίων μπορούμε να εξετάσουμε τη σχετική αναλογία μεταξύ των αναδιπλωμένων και των μη αναδιπλωμένων διαμορφώσεων (Daura et al., 1999, Schäfer et al., 2001). Από τις κατανομές αυτές μπορεί κανείς να υπολογίσει τη πιθανότητα να βρεθεί το πεπτίδιο σε μία διαμόρφωση και κατ' επέκταση να υπολογίσει ελεύθερη ενέργεια αναδίπλωσης (free energy of folding) (Boned et al., 2008).

Τα τέσσερα τετραπεπτίδια που επιλέχθηκαν με τον τρόπο που αναλύθηκε στην προηγούμενη ενότητα, μελετήθηκαν με προσομοιώσεις αναδίπλωσης διάρκειας 300ns. Η επιμίκυση του χρόνου της προσομοίωσης από 100ns σε 300ns θα μας επιτρέψει να εξετάσουμε πολλαπλά

γεγονότα αναδίπλωσης/αποδιάταξης, δεδομένου ότι τα γεγονότα αναδίπλωσης λαμβάνουν χώρα στην κλίμακα των 10ns (Ενότητα 3.2) και οι μεταβάσεις των μικρών πεπτιδίων από την αναδιπλωμένη στη μη αναδιπλωμένη διαμόρφωση είναι πολύ γρήγορες, της τάξης των 0.05ns (Daura et al., 1998). Επιπλέον, σε πειράματα αποδιάταξης ενός ελικοειδούς πεπτιδίου 21 καταλοίπων (Fs peptide), τα γεγονότα αποδιάταξης συμβαίνουν σε χρόνο που κυμαίνεται από 20ns – 160ns(±60ns) ανάλογα με την θερμοκρασία (290K-300K) (Williams et al., 1996). Σε συμφωνία με την απόδοση (benchmark) του πρωτοκόλλου της προσομοίωσης, που είναι 100ns/κόμβο/μέρα, το υπολογιστικό τμήμα των προσομοιώσεων αυτών ολοκληρώθηκε σε ~4 μέρες, έχοντας στη διάθεσή μας 4 κόμβους της συστοιχίας των υπολογιστών. Το πρωτόκολλο της προσομοίωσης είναι πανομοιότυπο με αυτό που βρίσκεται στο Παράρτημα (#14, NAMD script, heat.namd και #15, NAMD script, equi.namd) με μοναδική διαφορά την τελική θερμοκρασία της προσομοίωσης (και φυσικά τον απαιτούμενο αριθμό κύκλων για τη σταδιακή άνοδο της θερμοκρασίας με βήμα  $\Delta T=20K$ ) για τη διεξαγωγή των προσομοιώσεων σε θερμοκρασίες 10°C (283K), 25°C (298K), 47°C (320K) και 67°C (340K):

🌀 langevinTemp -> 283 / 298 / 320 / 340

🌀 langevinPistonTemp -> 283 / 298 / 320 / 340

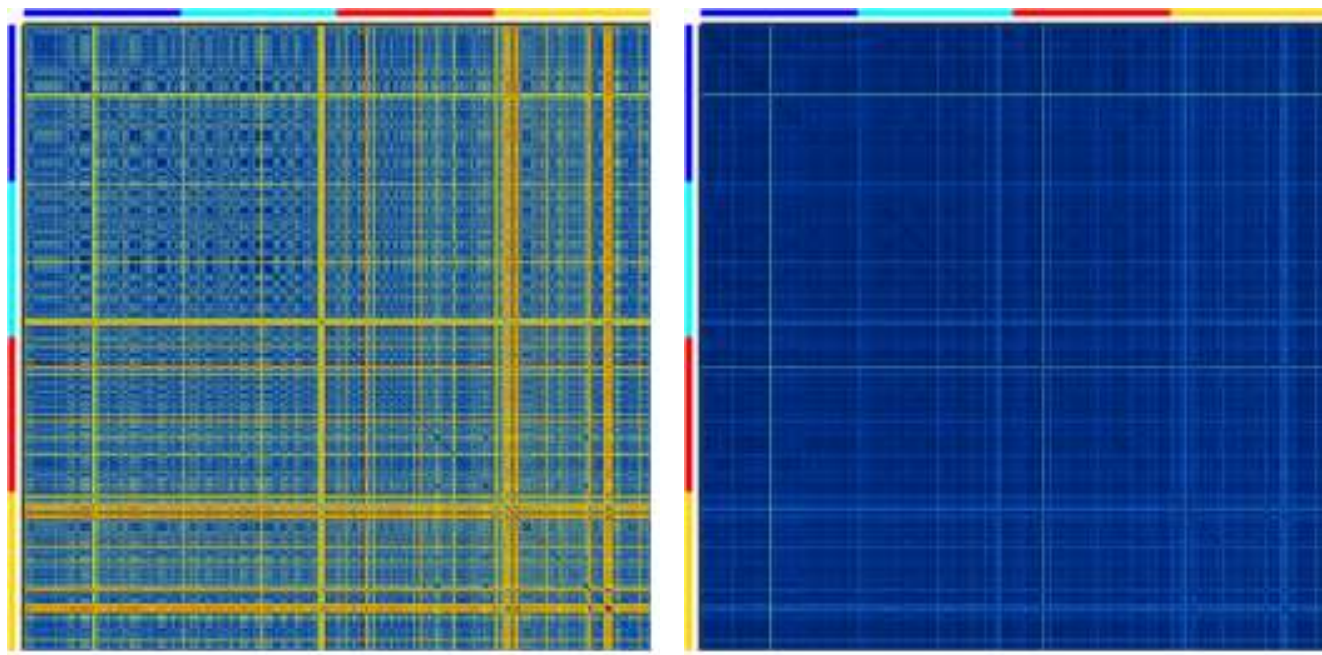
Η μελέτη της αναδίπλωσης των πεπτιδίων αυτών σε ένα εύρος θερμοκρασιών αφενός θα μας επιτρέψει να ερευνήσουμε την επίδραση της θερμοκρασίας στην αναδίπλωση και τη σταθερότητα αυτών, αφετέρου θα έχουμε τη δυνατότητα να συγκρίνουμε τα αποτελέσματα από διαφορετικές θερμοκρασίες με πειραματικά δεδομένα για να εξετάσουμε κατά πόσο τα force fields μπορούν να προβλέψουν με ακρίβεια τη δυναμική των πεπτιδίων σε ένα εύρος θερμοκρασίας.



### Σύγκριση των προβλεπόμενων δομών των τροχιακών με πίνακες RMSD

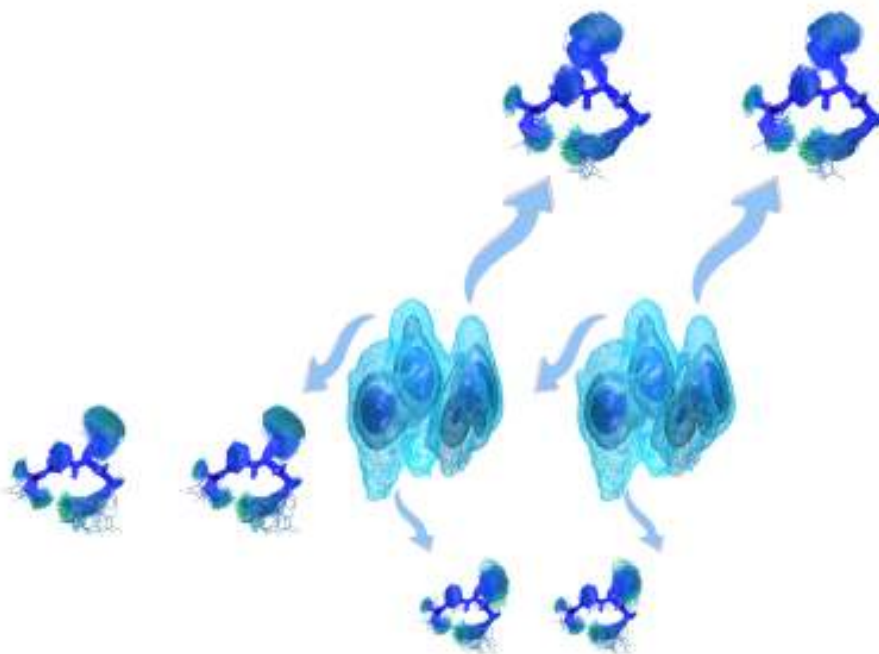
Για την άμεση σύγκριση των τροχιακών των τεσσάρων διαφορετικών θερμοκρασιών, ενώσαμε (τεχνητά) τα τέσσερα ανεξάρτητα τροχιακά και υπολογίσαμε ένα πίνακα RMSD μεταξύ διαδοχικών δομών. Έτσι προκύπτει ένας ενιαίος πίνακας που δείχνει όχι μόνο τη δημιουργία ομάδων δομών (clusters) αλλά και την ομοιότητά τους, σε επίπεδο RMSD, για το σύνολο των

προσομοιώσεων σε κάθε πεπτίδιο που μελετήσαμε. Στην Εικόνα 3.21 βλέπουμε τη γραφική απεικόνιση του ενιαίου πίνακα RMSD τόσο λαμβάνοντας υπόψιν όλα τα βαριά άτομα όσο και μόνο τα άτομα του πεπτιδικού σκελετού.



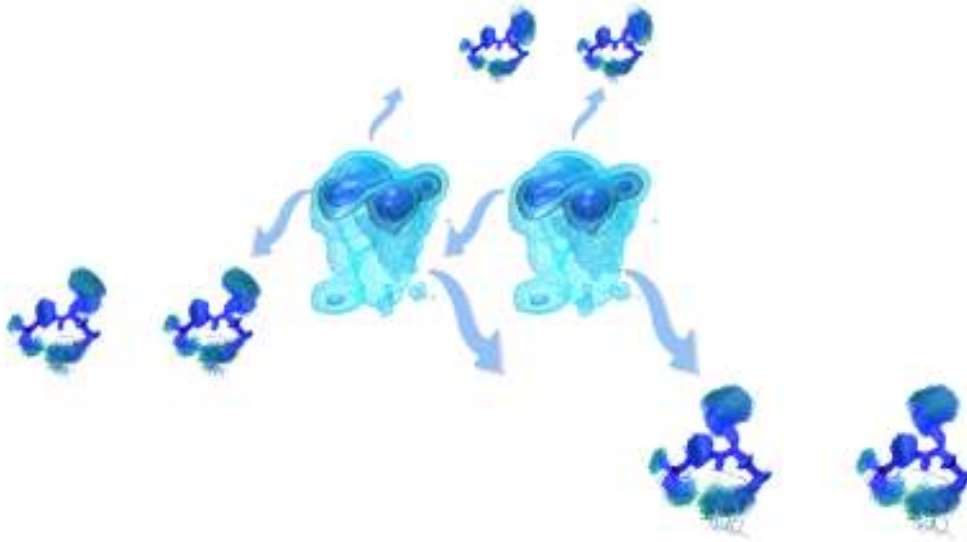
Εικόνα 3.21 Γραφική απεικόνιση του ενιαίου πίνακα RMSD μετά την ένωση των τροχιακών των 4 θερμοκρασιών του πεπτιδίου RWPD, για όλα τα βαριά άτομα (αριστερά) και για τα άτομα του πεπτιδικού σκελετού (δεξιά). Η χρωματική κλίμακα κυμαίνεται από σκούρο μπλε (0Å) έως σκούρο κόκκινο (5.8Å) για τον πίνακα αριστερά και από σκούρο μπλε (0Å) έως σκούρο κόκκινο (3.14Å) για τον πίνακα δεξιά. Οι οριζόντιες και κάθετες χρωματιστές μπάρες οριοθετούν τα τέσσερα ανεξάρτητα τροχιακά, όπου με μπλε απεικονίζεται το τροχιακό των **283K**, με γαλάζιο των **298K**, με κόκκινο των **320K** και με πορτοκαλί των **340K**. Η χρωματική κλίμακα για τα τροχιακά διατηρείται κοινή στο υπόλοιπο της ενότητας αυτής.

Στην περίπτωση του πεπτιδίου RWPD υπάρχει μία σχεδόν γραμμική αντιστρόφως ανάλογη σχέση μεταξύ σταθερότητας (κινητικής) και θερμοκρασίας όταν ο υπολογισμός συμπεριλαμβάνει όλα τα βαριά άτομα, δηλαδή και τις πλευρικές ομάδες. Ο πεπτιδικός σκελετός όμως, σταθεροποιείται πολύ γρήγορα σε μία διαμόρφωση η οποία είναι κοινή μεταξύ των τεσσάρων τροχιακών με πολύ μικρές διακυμάνσεις από τη μέση δομή (Πίνακας 3.5, Εικόνες 3.22-3.25).

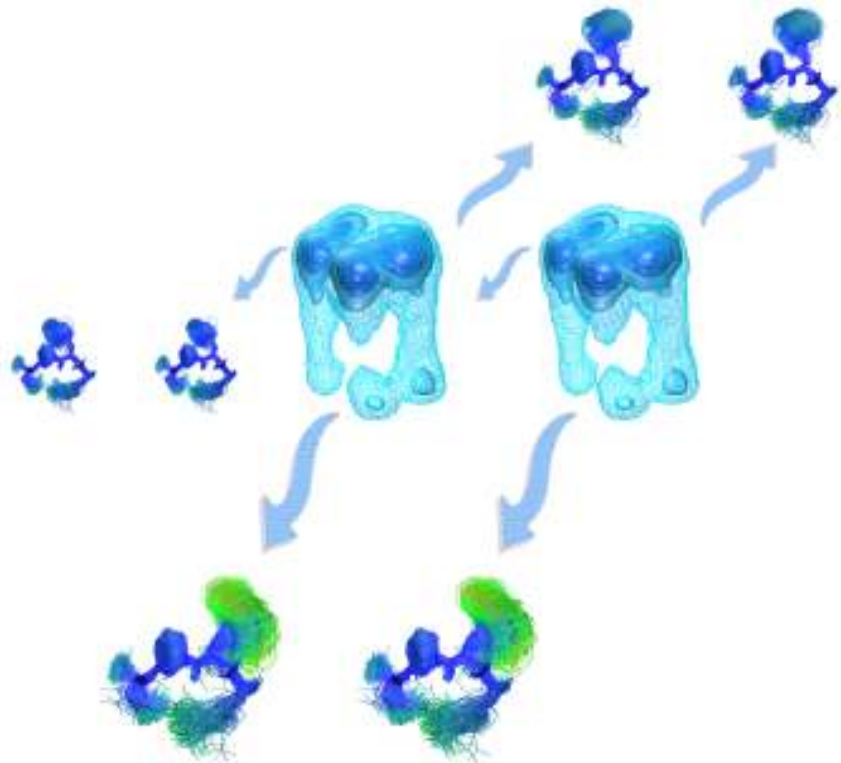


Εικόνα 3.22 Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RWPD. Στο κέντρο φαίνεται η προβολή του τροχιακού των 283K στους τρεις principal components της ανάλυσης Cartesian-PCA χρησιμοποιώντας όλα τα βαριά άτομα. Υποδεικνύονται τρία επίπεδα ισοεπιφάνειας (μέση τιμή,  $1\sigma$ ,  $6\sigma$  του χάρτη κατανομής). Κάθε ισχυρή κορυφή του ενεργειακού τοπίου αντιστοιχεί σε ένα cluster δομών (οι οποίες φαίνονται με τα βέλη), το μέγεθος αναπαράστασης των οποίων είναι ανάλογο της κατοχής του cluster σε χρόνο προσομοίωσης. Η υπέρθεση των δομών έγινε χρησιμοποιώντας τα άτομα του πεπτιδικού σκελετού για να δοθεί έμφαση στην κινητικότητα των πλευρικών ομάδων, ενώ ο χρωματισμός των ατόμων από μπλε ( $0.14\text{\AA}$ ) σε κόκκινο ( $5.07\text{\AA}$ ) έγινε με βάση τις ατομικές διακυμάνσεις (RMSFs) και διατηρείται παντού για λόγους σύγκρισης.

Η μοναδική (αλλά μικρή) διαφορά μεταξύ των αντιπροσωπευτικών δομών του κυρίαρχου cluster εντοπίζεται στη διαμόρφωση της τρυπτοφάνης και αυτό εξηγείται από την αύξηση των rmsf της πλευρικής ομάδας του καταλοίπου αυτού με την άνοδο της θερμοκρασίας (Πίνακας 3.5). Η ομοιότητα μεταξύ των δομών που προκύπτουν στις τέσσερις θερμοκρασίες διαφαίνεται και στις Εικόνες 3.22-3.25 όπου για κάθε κορυφή του ενεργειακού τοπίου υποδεικνύεται και το σύνολο δομών που απαρτίζουν το cluster. Στο τροχιακό των 283K προκύπτουν 3 cluster δομών (με RMSD cut-off 1.70 και variance-explained 0.81) από την ανάλυση Cartesian-PCA με κατοχή σε χρόνο προσομοίωσης 17.7%, 16.5% και 12.1% αντίστοιχα.

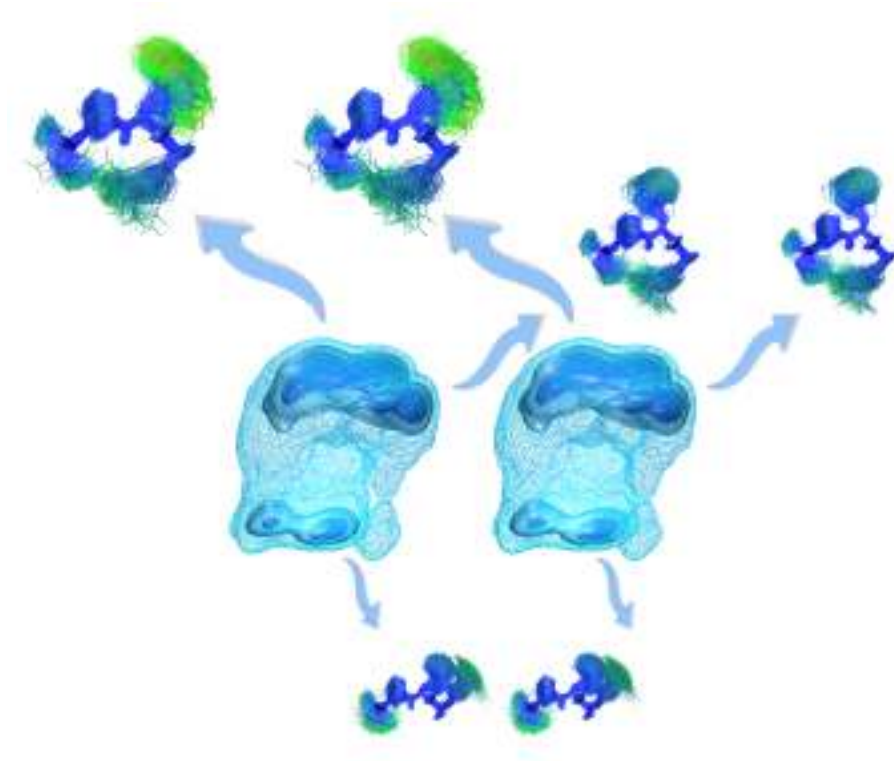


Εικόνα 3.23 Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RWPD για το τροχιακό των 298K, σε αντιστοιχία με την Εικόνα 3.22.



Εικόνα 3.24 Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RWPD για το τροχιακό των 320K, σε αντιστοιχία με την Εικόνα 3.22.





Εικόνα 3.25 Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RWPD για το τροχιακό των 340K, σε αντιστοιχία με την Εικόνα 3.22.

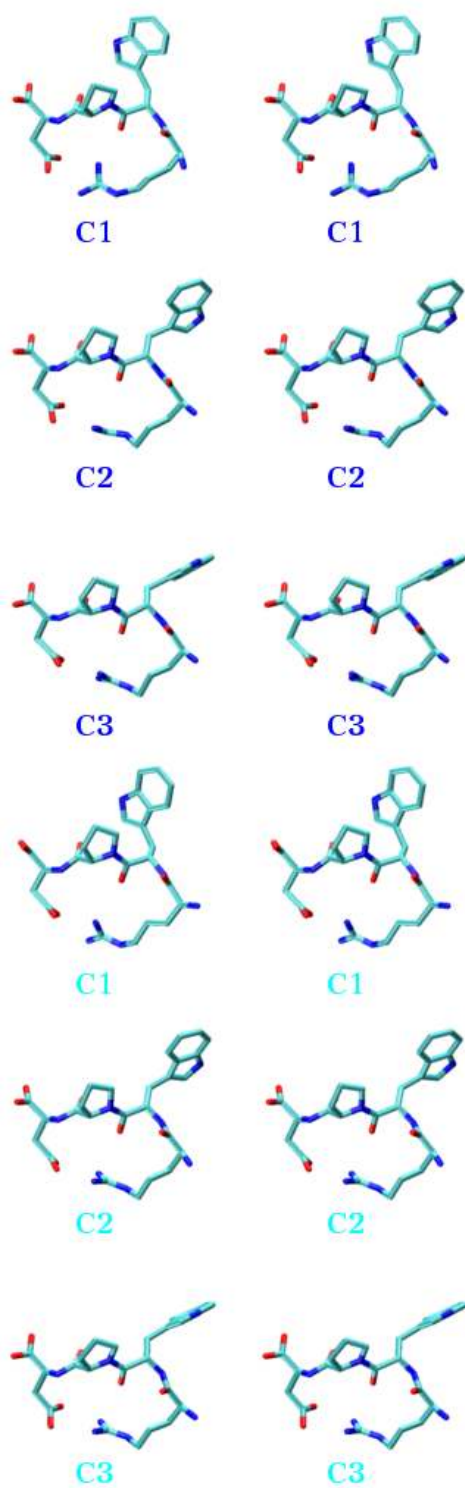
Σε σχεδόν απόλυτη συμφωνία, με ελαφρώς υψηλότερες τιμές RMSFs είναι οι δομές των 3 cluster δομών του τροχιακού των 298K (με RSMD cut-off 1.80 και variance-explained 0.88) των οποίων η κατοχή σε χρόνο προσομοίωσης είναι 30.7%, 11.8% και 12.0% αντίστοιχα. Σε υψηλότερες θερμοκρασίες των τροχιακών των 320K (με RSMD cut-off 1.00 και variance-explained 0.89) και 340K (με RSMD cut-off 1.60 και variance-explained 0.88), η κατοχή σε χρόνο προσομοίωσης διαμορφώνεται σε 53.8%, 10.2%, 1.3% και 44.4%, 9.6%, 2.9% και βλέπουμε πανομοιότυπες ομάδες δομών αλλά με αναστροφή στη σειρά τους σε σχέση με την κατοχή σε χρόνο προσομοίωσης (Εικόνα 3.26). Έτσι το πρώτο cluster των τροχιακών 283K και 298K γίνεται δεύτερο στους 320K και 340K, και το δεύτερο cluster των τροχιακών 283K και 298K γίνεται πρώτο στους 320K και 340K. Το τρίτο cluster των τροχιακών 283K και 298K είναι κοινό αλλά διαφορετικό από τα cluster που βλέπουμε στις υψηλότερες θερμοκρασίες. Στους 320K το τρίτο cluster δομών μοιάζει με το δεύτερο με τη διαφορά ότι η πλευρική ομάδα της τρυπτοφάνης έχει κάνει flip, με αποτέλεσμα στη μία περίπτωση η NH ομάδα του δακτυλίου να αλληλεπιδρά με το οξυγόνο του



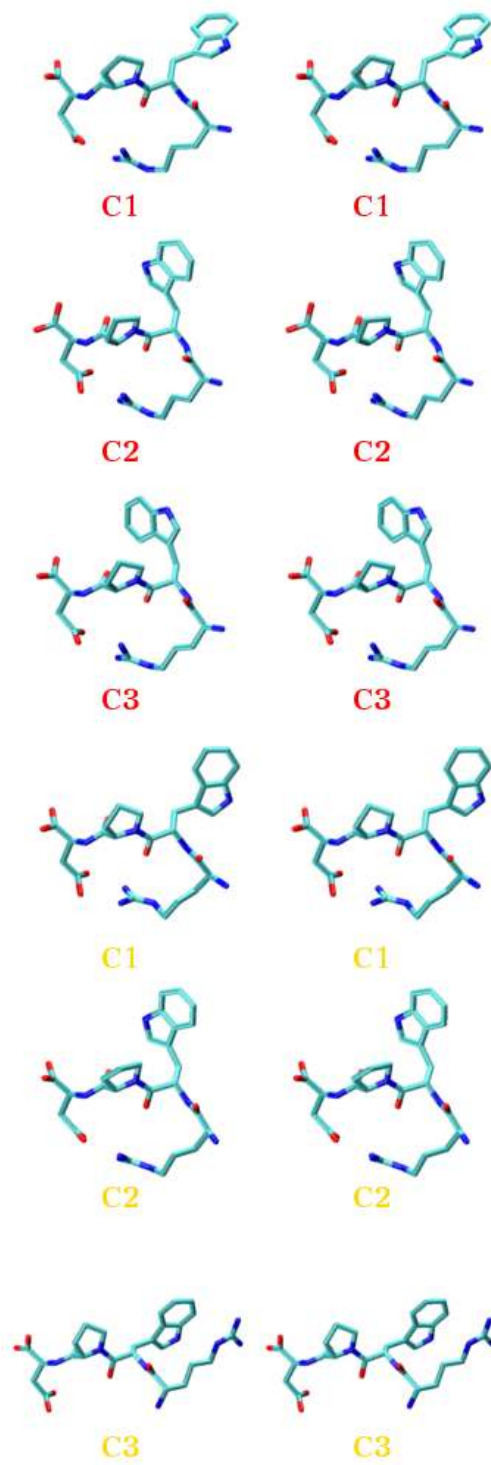
σκελετού που ανήκει στην προλίνη και στην άλλη περίπτωση να αλληλεπιδρά με το οξυγόνο του σκελετού που ανήκει στην αργινίνη. Στο τρίτο cluster δομών του τροχιακού των 340K βλέπουμε μία διαφορετική αλληλεπίδραση, με την τρυπτοφάνη να πακετάρεται απέναντι από την αργινίνη και να μην παρατηρείται η δομή θηλιάς (loop-closure). Παρατηρούμε λοιπόν, πως με την άνοδο της θερμοκρασίας, καθίσταται εφικτό για το πεπτίδιο το πέρασμα και από άλλες διαμορφώσεις του ενεργειακού τοπίου.

		Μέσο RMSF για βαριά άτομα	Μέσο RMSF για άτομα σκελετού	Μέσο RMSF για άτομα πλευρικής ομάδας Trp	Μέσο RMSF για άτομα υπόλοιπων πλευρικών ομάδων
RWPD	283K	0.59	0.25	0.67	0.87
	298K	0.63	0.27	0.74	0.93
	320K	0.98	0.31	1.97	1.06
	340K	0.98	0.32	1.93	1.08
RPWD	283K	0.91	0.28	1.80	1.01
	298K	1.35	0.49	2.62	1.43
	320K	0.96	0.30	1.84	1.10
	340K	1.07	0.31	2.37	1.05
DTRW	283K	1.06	0.30	1.09	1.54
	298K	1.09	0.32	1.14	1.57
	320K	1.07	0.33	1.14	1.50
	340K	1.05	0.34	1.13	1.44
EVKW	283K	1.03	0.31	1.19	1.32
	298K	0.97	0.30	1.06	1.30
	320K	1.11	0.36	1.20	1.51
	340K	1.18	0.37	1.30	1.57

Πίνακας 3.5 Μέση τιμή των ατομικών διακυμάνσεων (RMSFs) για τέσσερα σύνολα ατόμων για τα τέσσερα τετραπεπτίδια. Οι ατομικές διακυμάνσεις έχουν υπολογιστεί για το κυρίαρχο cluster μέσω της ανάλυσης Cartesian-PCA για όλα τα βαριά άτομα.

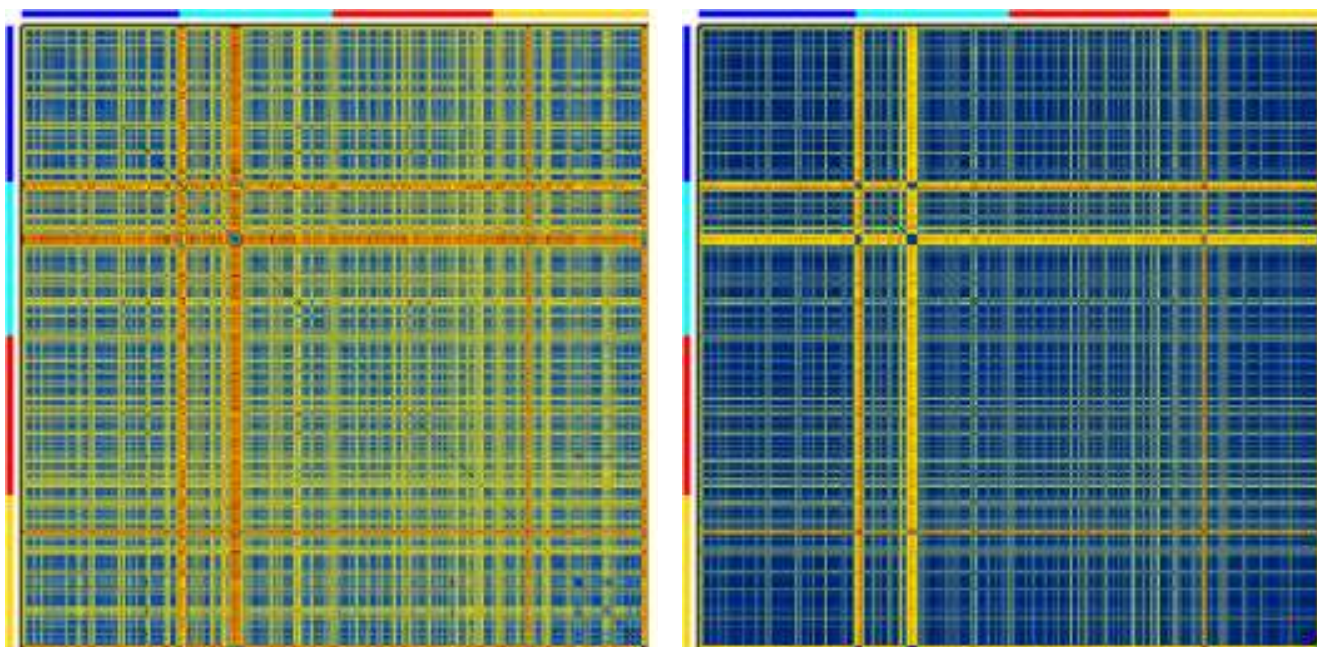


Εικόνα 3.26 Αντιπροσωπευτικές δομές (σε stereo αναπαράσταση) από τα cluster των τροχιακών των τεσσάρων θερμοκρασιών για το RYPD. Ο χρωματικός κώδικας των θερμοκρασιών είναι ίδιος με της Εικόνας 3.21.



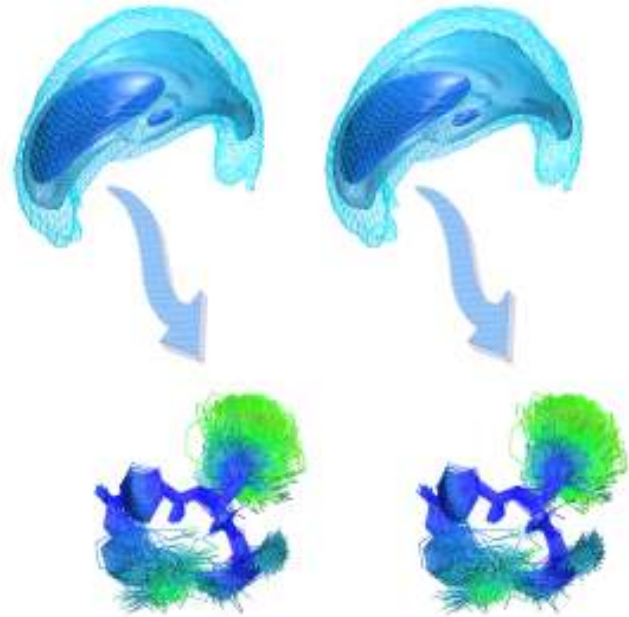
Εικόνα 3.26 (συνέχεια) Αντιπροσωπευτικές δομές (σε stereo αναπαράσταση) από τα cluster των τροχιακών των τεσσάρων θερμοκρασιών για το RWPD.

Για το πεπτίδιο RPWD (Εικόνα 3.27) βλέπουμε μεγαλύτερη αστάθεια και πολλαπλά γεγονότα αναδίπλωσης/αποδιάταξης. Με την άνοδο της θερμοκρασίας παρατηρούμε τη δημιουργία νέων cluster δομών αλλά μικρή διάρκεια. Θα μπορούσε κανείς να ισχυριστεί ότι η κατάταξη των τροχιακών με βάση τη σταθερότητα και την κινητικότητα των δομών ακολουθεί τη σειρά 283K, 320K, 340K, 298K. Δηλαδή, το τροχιακό των 298K φαίνεται να δίνει πιο ασταθείς δομές ακόμα και από το τροχιακό των 340K (Πίνακας 3.5, Εικόνες 3.28-3.31). Οι παρατηρήσεις αυτές αφορούν τόσο τη συμπεριφορά όλων των βαριών ατόμων όσο και του πεπτιδικού σκελετού.

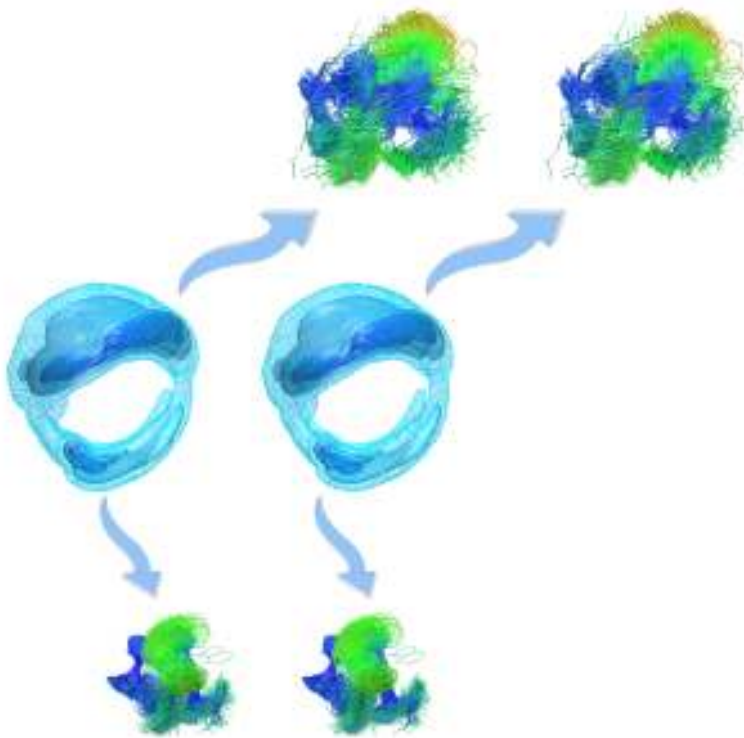


Εικόνα 3.27 Γραφική απεικόνιση του ενιαίου πίνακα RMSD μετά την ένωση των τροχιακών των 4 θερμοκρασιών του πεπτιδίου RPWD, για όλα τα βαριά άτομα (αριστερά) και για τα άτομα του πεπτιδικού σκελετού (δεξιά), σε αντιστοιχία με την Εικόνα 3.21.

Η δομή του κυρίαρχου cluster του τροχιακού των 283K (με RMSD cut-off 1.20 και variance-explained 0.92), όπως προκύπτει από την ανάλυση Cartesian-PCA για όλα τα βαριά άτομα, έχει κατοχή σε χρόνο προσομοίωσης 59.6% και μοιάζει με τις δομές των κυρίαρχων cluster των υπόλοιπων τροχιακών (Εικόνες 3.28-3.31, 3.32). Ωστόσο εμφανίζει αξιοσημείωτη ομοιότητα και με το cluster 2 των τροχιακών 283K, 298K και το cluster 1 των τροχιακών 320K, 340K του πεπτιδίου RPWD (όσον αφορά τη σχετική διεύθυνση των καταλοίπων W-P και του πεπτιδικού

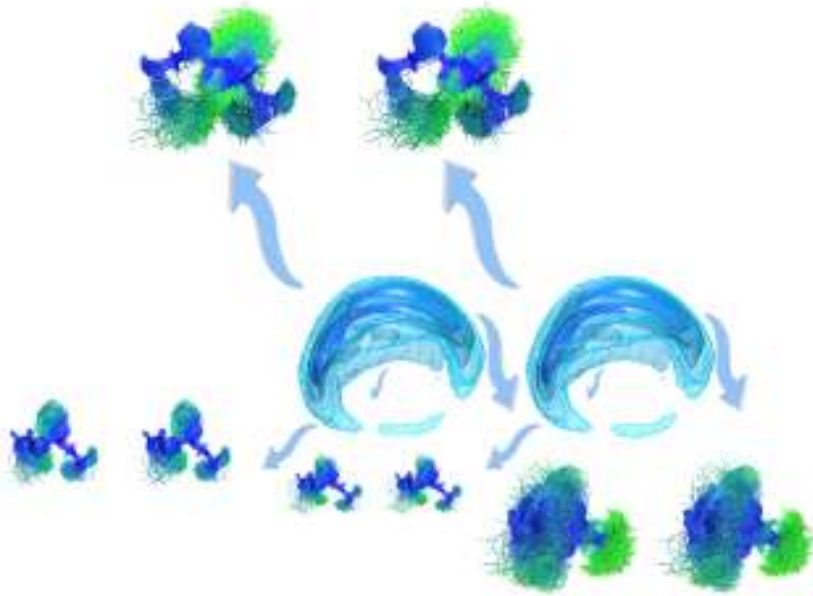


Εικόνα 3.28 Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RPWD για το τροχιακό των 283K, σε αντιστοιχία με την Εικόνα 3.22.

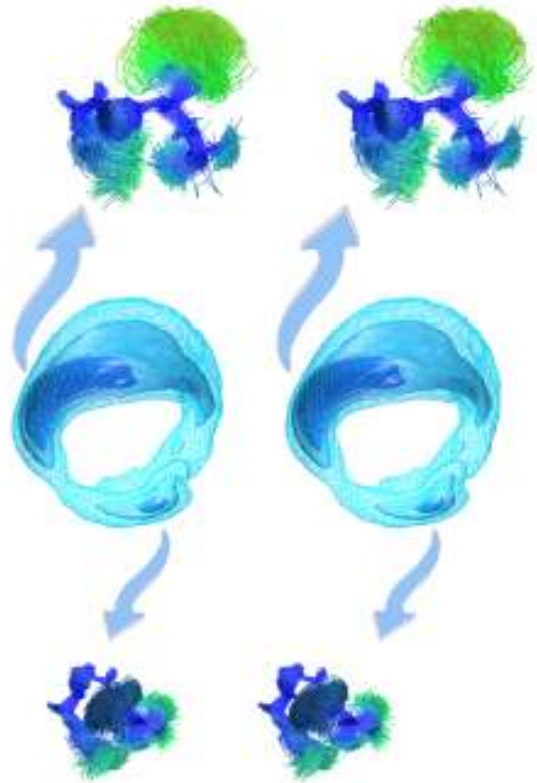


Εικόνα 3.29 Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RPWD για το τροχιακό των 298K, σε αντιστοιχία με την Εικόνα 3.22.



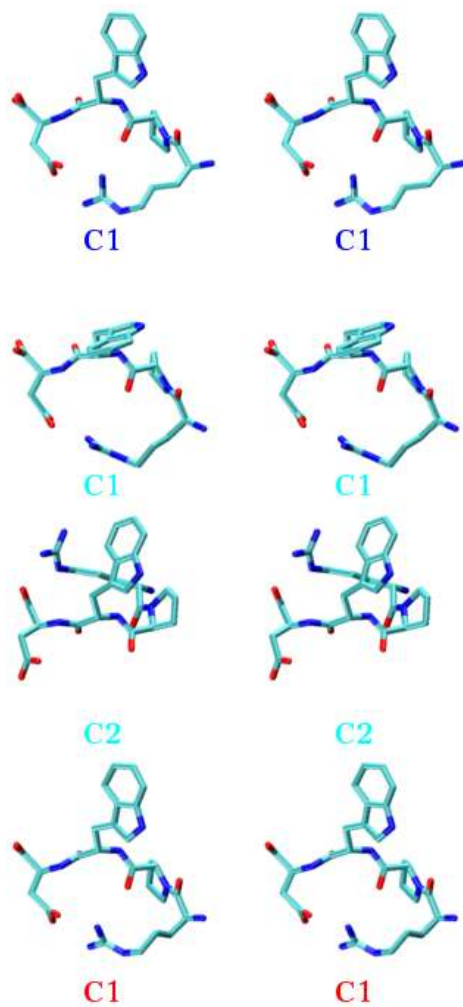


Εικόνα 3.30 Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RPWD για το τροχιακό των 320K, σε αντιστοιχία με την Εικόνα 3.22.

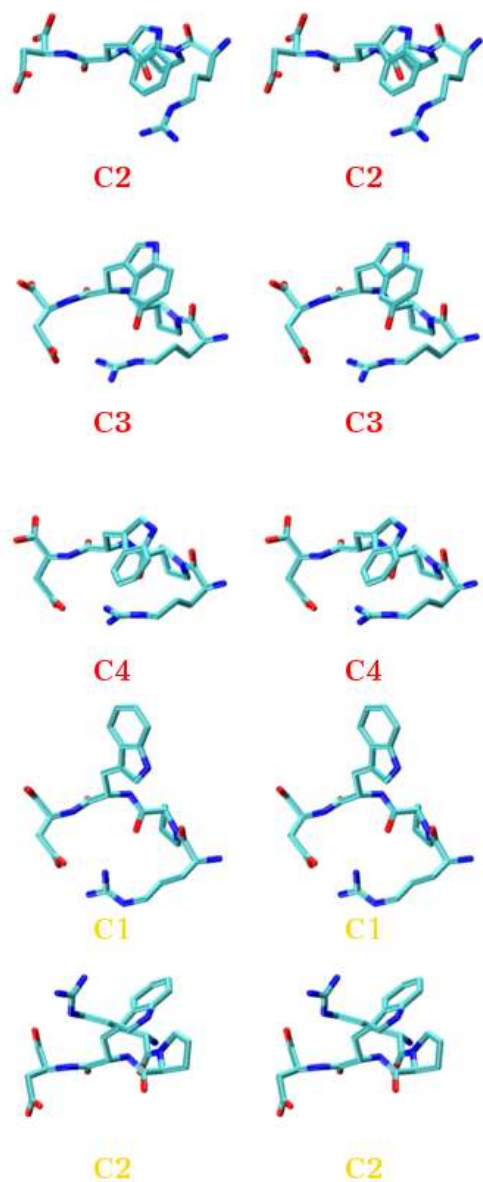


Εικόνα 3.31 Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RPWD για το τροχιακό των 340K, σε αντιστοιχία με την Εικόνα 3.22.





Εικόνα 3.32 Αντιπροσωπευτικές δομές (σε stereo αναπαράσταση) από τα cluster των τροχιακών των τεσσάρων θερμοκρασιών για το RPWD. Ο χρωματικός κώδικας των θερμοκρασιών είναι ίδιος με της Εικόνας 3.21.



Εικόνα 3.32 (συνέχεια) Αντιπροσωπευτικές δομές (σε stereo αναπαράσταση) από τα cluster των τροχιακών των τεσσάρων θερμοκρασιών για το RPWD.

σκελετού), αν θεωρήσουμε μία απλή αντιστροφή των τελικών καταλοίπων R-D (Εικόνες 3.21-3.24 *versus* Εικόνες 3.26-3.29 και Εικόνα 3.26 *versus* 3.32).

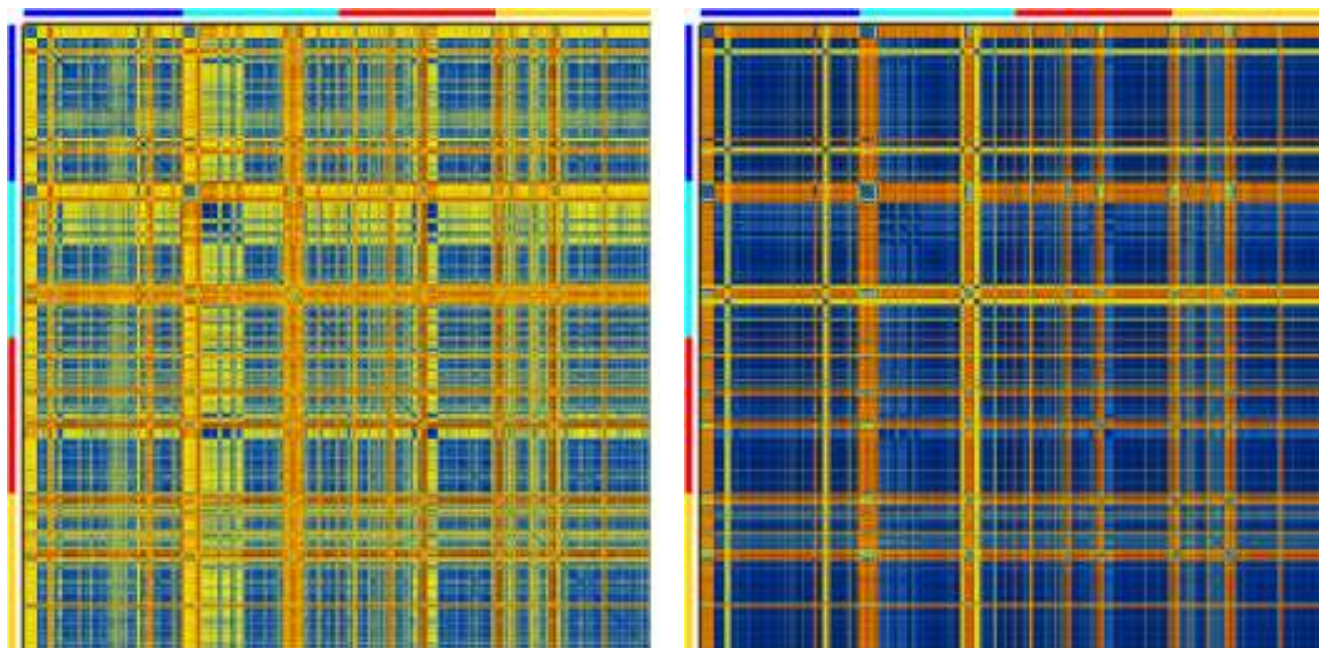
Το τροχιακό των 298K (Εικόνα 3.29) δίνει δύο cluster (με RMSD cut-off 1.00 και variance-explained 0.92) με κατοχή σε χρόνο προσομοίωσης 56.9% και 4.0% αντίστοιχα, με υψηλές ατομικές διακυμάνσεις. Στις δομές του δεύτερου cluster (Εικόνα 3.32) βλέπουμε το χαρακτηριστικό πακετάρισμα της τρυπτοφάνης με την αργινίνη.

Το τροχιακό των 320K (Εικόνα 3.30) δίνει 4 cluster δομών (με RMSD cut-off 1.90 και variance-explained 0.80) με κατοχή σε χρόνο προσομοίωσης 42.7%, 5.4%, 0.2% και 0.03% αντίστοιχα. Στο δεύτερο cluster δομών (Εικόνα 3.32) βλέπουμε και πάλι το πακετάρισμα των Trp-Arg αλλά αυτή τη φορά πίσω από τον πεπτιδικό σκελετό. Οι δομές των cluster 3 και 4 είναι πολύ κοντινές μεταξύ τους και διαφέρουν μόνο στη σχετική διεύθυνση των πλευρικών ομάδων της τρυπτοφάνης και της αργινίνης (Εικόνα 3.32).

Το τροχιακό των 340K (Εικόνα 3.31) δίνει δυο κυρίαρχα cluster (με RMSD cut-off 1.10 και variance-explained 0.92) με κατοχή σε χρόνο προσομοίωσης 56.9% και 1.2% αντίστοιχα και είναι οι δομές που συναντήσαμε και στα τροχιακά στις άλλες θερμοκρασίες (Εικόνα 3.32).

Βλέπουμε λοιπόν και πάλι συμφωνία μεταξύ των προβλεπόμενων δομών των κυρίαρχων cluster των διαφόρων τροχιακών, με το πρώτο (και κυρίαρχο) cluster να περιλαμβάνει παρόμοιες δομές και στα τέσσερα τροχιακά. Το δεύτερο (σε εμφάνιση) cluster δεν εμφανίζεται για περισσότερο από 5% του συνολικού χρόνου προσομοίωσης και είναι επίσης κοινό μεταξύ των τροχιακών και περιλαμβάνει το σύνολο δομών στις οποίες η πλευρική ομάδα της τρυπτοφάνης κλίνει προς την πλευρική ομάδα της αργινίνης, με ταυτόχρονη δημιουργία της δομής θηλιάς. Τα υπόλοιπα cluster εμφανίζουν αλληλεπίδραση μεταξύ των ιδίων καταλοίπων με μικρές διαφορές στη σχετική διεύθυνση στο χώρο, διατηρώντας χαμηλή συχνότητα εμφάνισης (Εικόνα 3.32).

Στο πεπτίδιο DTRW (Εικόνα 3.33) βλέπουμε και πάλι μία κυρίαρχη ομάδα δομών που είναι κοινή μεταξύ των τεσσάρων τροχιακών των διαφορετικών θερμοκρασιών και αρκετά cluster μικρής διάρκειας. Σε επίπεδο πεπτιδικού σκελετού, τα cluster δομών είναι ιδιαίτερος συμπαγή και περιλαμβάνουν παρόμοιες δομές, όπως φαίνεται και από τις μέσες τιμές RMSFs (Πίνακας 3.5). Εάν συνυπολογίσουμε και τις πλευρικές ομάδες, το κυρίαρχο cluster δομών εμφανίζει μεγαλύτερη διασπορά, ενώ τα cluster με μικρότερη κατοχή σε χρόνο προσομοίωσης παραμένουν συμπαγή. Η κυρίαρχη δομή που προβλέπεται από όλα τα τροχιακά είναι μία δομή θηλιάς σχήματος "π" με χαρακτηριστικό πακετάρισμα μεταξύ των καταλοίπων Trp-Arg (Εικόνα 3.38).

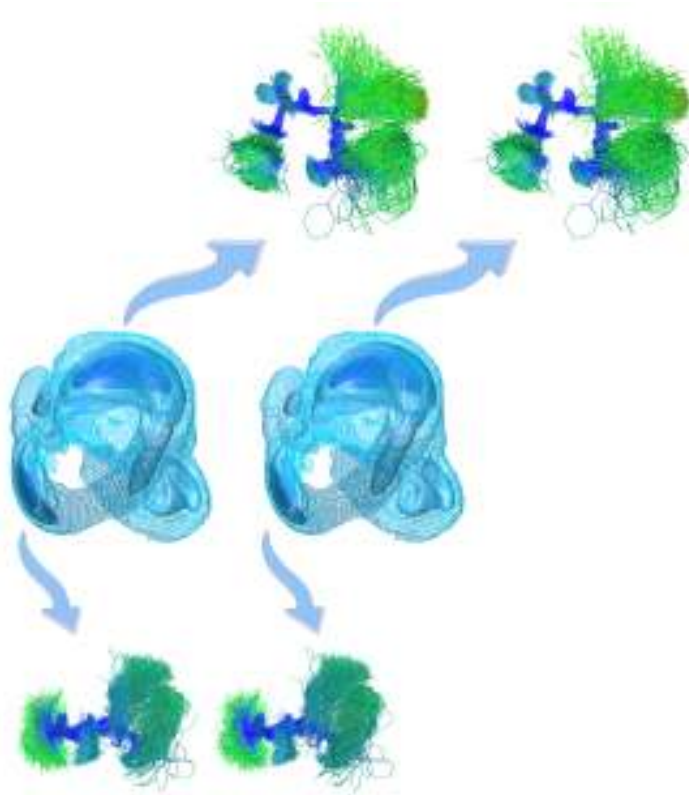


Εικόνα 3.33 Πάνω: γραφική απεικόνιση του ενιαίου πίνακα RMSD μετά την ένωση των τροχιακών των 4 θερμοκρασιών του πεπτιδίου DTRW, για όλα τα βαριά άτομα (αριστερά) και για τα άτομα του πεπτιδικού σκελετού (δεξιά), σε αντιστοιχία με την Εικόνα 3.21.

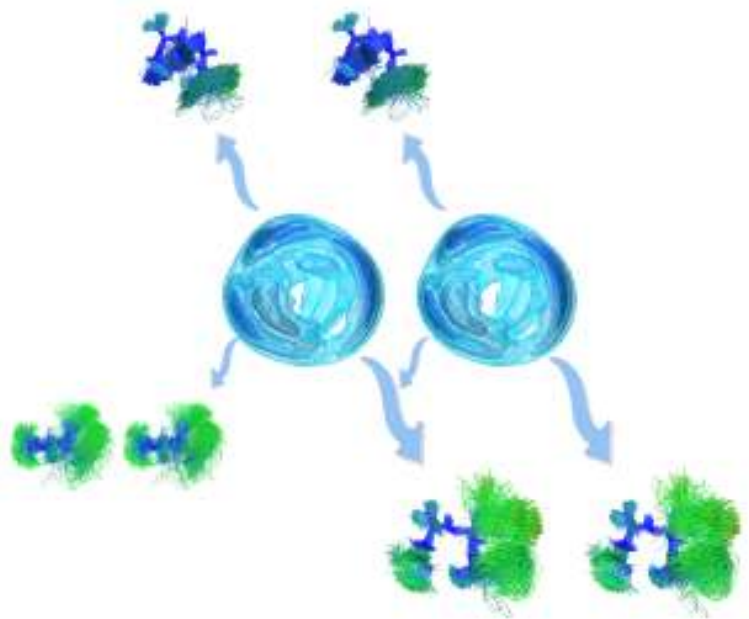
Το τροχιακό των 283K (Εικόνα 3.34) δίνει δύο cluster (με RMSD cut-off 1.00 και variance-explained 0.92) με κατοχή σε χρόνο προσομοίωσης 55.4% και 5.2% αντίστοιχα. Το δεύτερο cluster περιλαμβάνει δομές με τον πεπτιδικό σκελετό να παίρνει πιο εκτεταμένη διαμόρφωση και το χαρακτηριστικό πακετάρισμα των καταλοίπων Trp-Arg (Εικόνα 3.38).

Το τροχιακό των 298K (Εικόνα 3.35) έχει τρία cluster (με RMSD cut-off 1.00 και variance-explained 0.94) με κατοχή σε χρόνο προσομοίωσης 40.8%, 13.1% και 6.6% αντίστοιχα. Το πρώτο και τρίτο cluster είναι πανομοιότυπα με το πρώτο και δεύτερο του τροχιακού των 283K (Εικόνα 3.38). Το δεύτερο cluster περιλαμβάνει δομές με πολύ μικρές διακυμάνσεις, ο πεπτιδικός σκελετός παίρνει πάλι τη δομή θηλιάς σχήματος "π" αλλά αναπτύσσονται και αρκετές αλληλεπιδράσεις μεταξύ Asp-Arg-Trp, με την ομάδα NE της τρυπτοφάνης να αλληλεπιδρά μία με το OD του ασπαρτικού οξέος και μία με το O του πεπτιδικού σκελετού που ανήκει στην αργινίνη (Εικόνα 3.38).

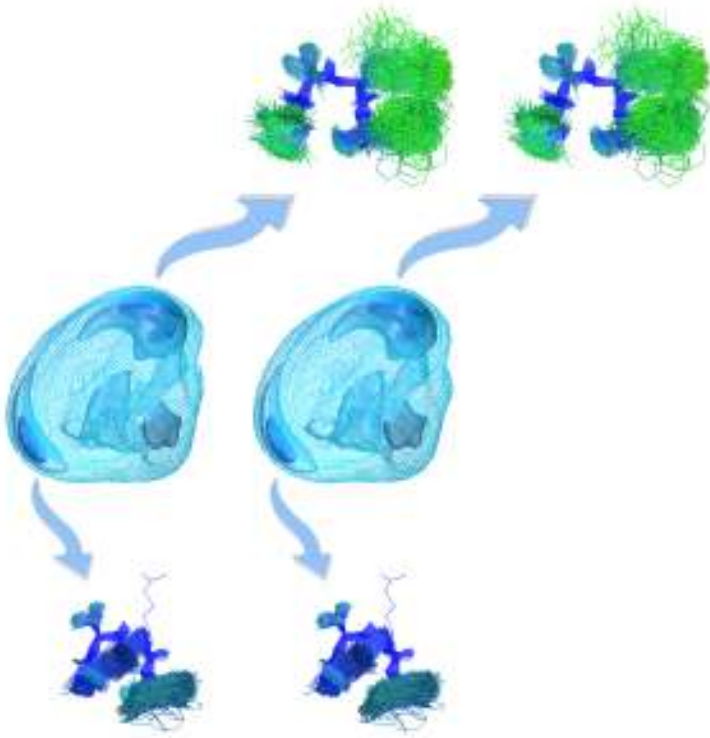
Στο τροχιακό των 320K (Εικόνα 3.36) βλέπουμε δύο cluster (με RMSD cut-off 1.00 και variance-explained 0.96) με κατοχή σε χρόνο προσομοίωσης 57.8% και 5.8% αντίστοιχα. Το πρώτο cluster



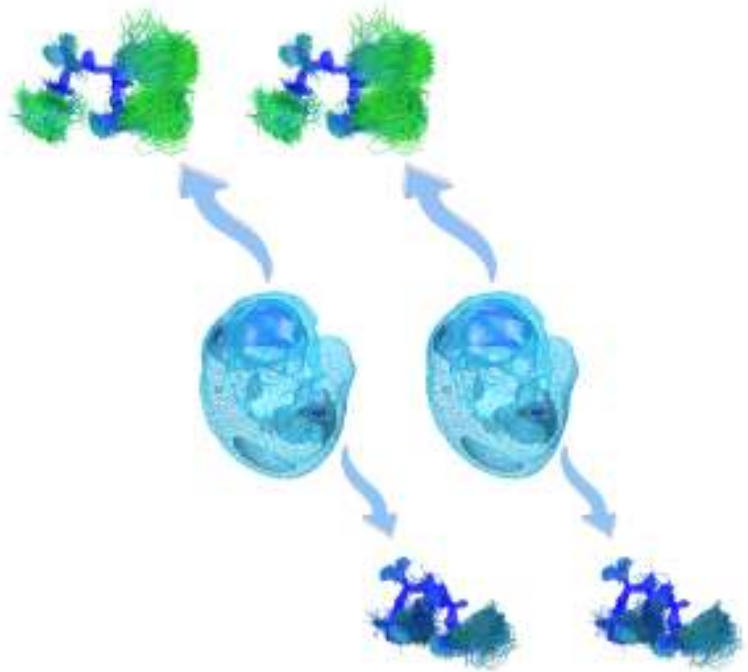
Εικόνα 3.34 Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου DRTW για το τροχιακό των 283K, σε αντιστοιχία με την Εικόνα 3.22.



Εικόνα 3.35 Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου DTRW για το τροχιακό των 298K, σε αντιστοιχία με την Εικόνα 3.22.

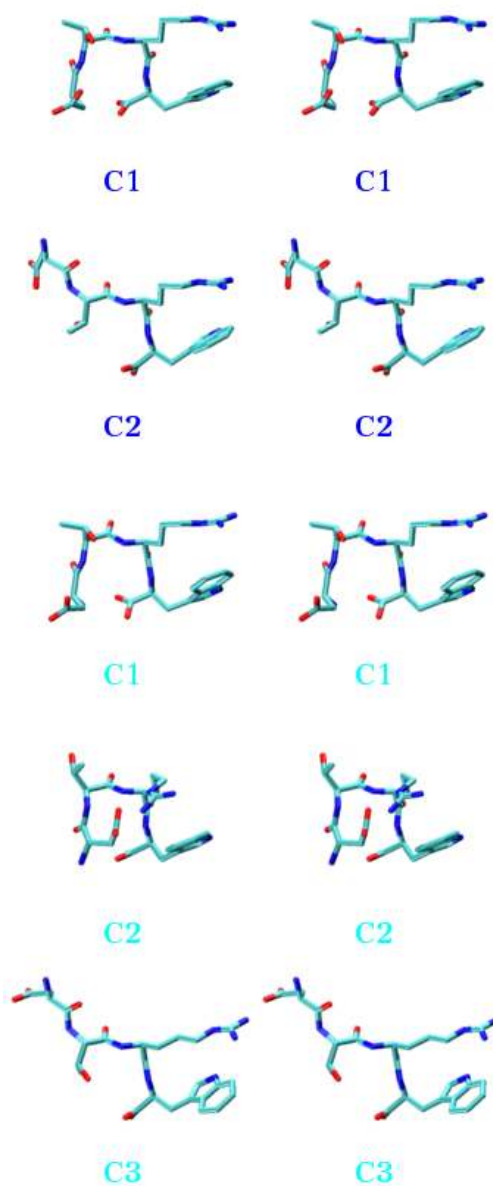


Εικόνα 3.36 Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου DTRW για το τροχιακό των 320K, σε αντιστοιχία με την Εικόνα 3.22.

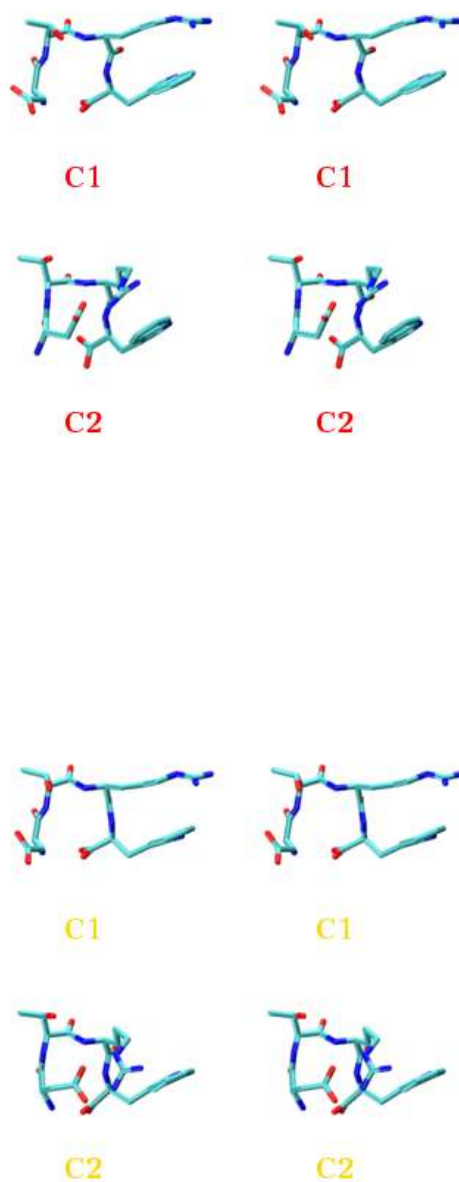


Εικόνα 3.37 Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου DTRW για το τροχιακό των 340K, σε αντιστοιχία με την Εικόνα 3.22.





Εικόνα 3.38 Αντιπροσωπευτικές δομές (σε stereo αναπαράσταση) από τα cluster των τροχιακών των τεσσάρων θερμοκρασιών για το DTRW. Ο χρωματικός κώδικας των θερμοκρασιών είναι ίδιος με της Εικόνας 3.21.

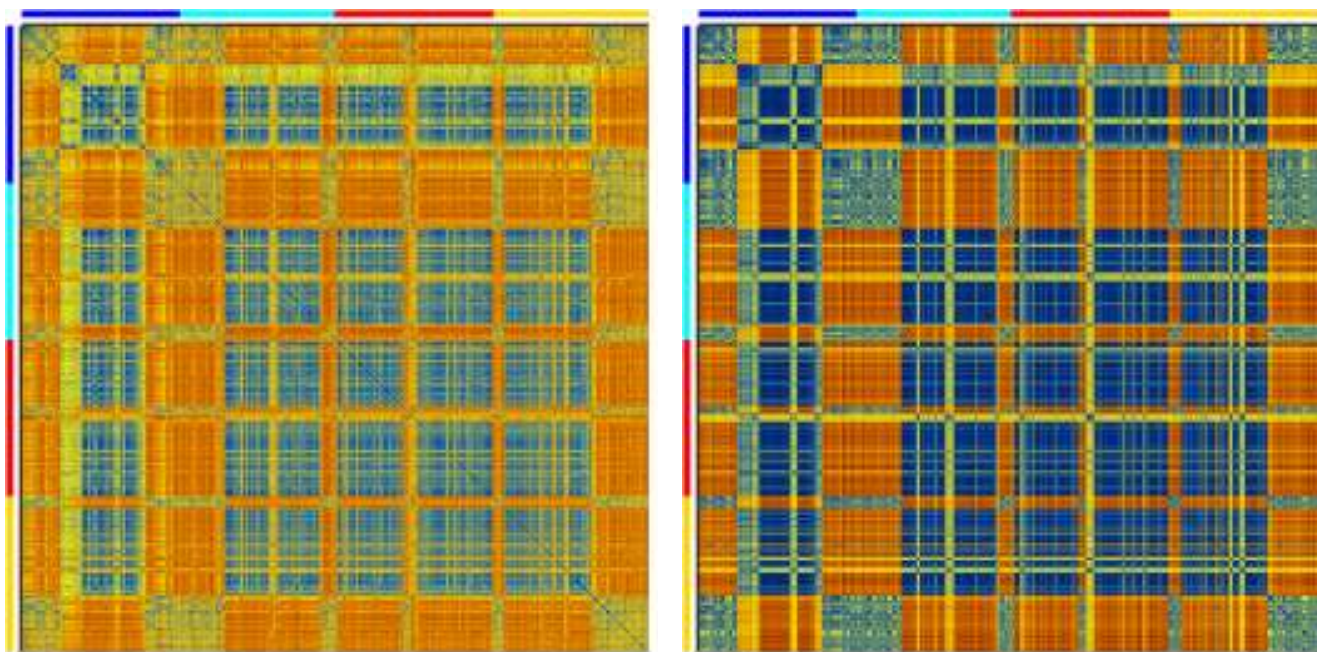


Εικόνα 3.38 (συνέχεια) Αντιπροσωπευτικές δομές (σε stereo αναπαράσταση) από τα cluster των τροχιακών των τεσσάρων θερμοκρασιών για το DTRW.

είναι πανομοιότυπο με το πρώτο cluster όλων τροχιακών και το δεύτερο cluster είναι πανομοιότυπο με το δεύτερο cluster του τροχιακού των 298K (Εικόνα 3.38).

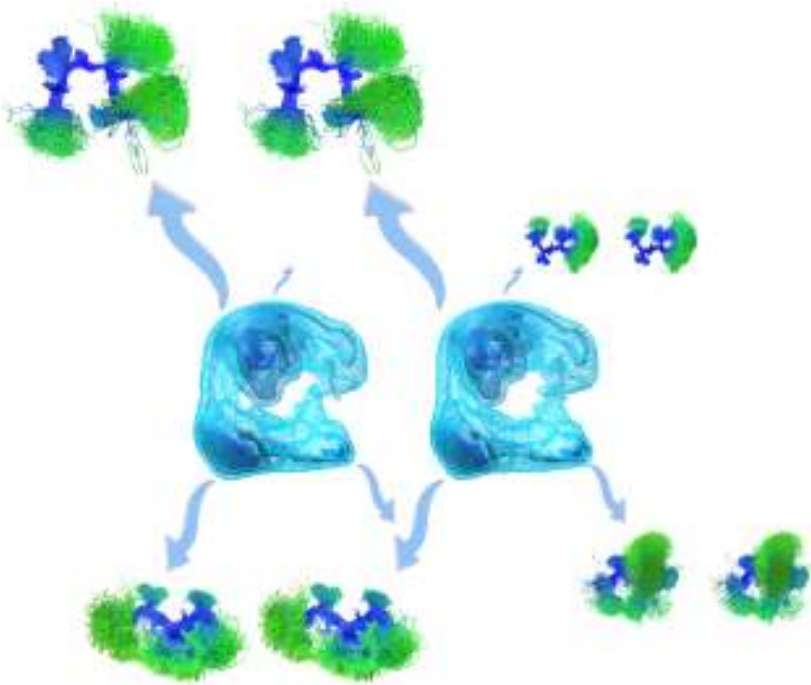
Τέλος στο τροχιακό των 340K (Εικόνα 3.37) βλέπουμε τα ίδια δύο cluster (με RMSD cut-off 1.70 και variance-explained 0.93) με το τροχιακό των 320K, με κατοχή σε χρόνο προσομοίωσης 52.5% και 1.6% αντίστοιχα (Εικόνα 3.38).

Το πεπτίδιο EVKW δείχνει τη μεγαλύτερη αστάθεια σε σχέση με τα προηγούμενα πεπτίδια (Εικόνα 3.39). Το τροχιακό με τη μεγαλύτερη σταθερότητα είναι αυτό των 320K όπως φαίνεται από τους πίνακες RMSD (Εικόνα 3.39) αλλά και τις μέσες τιμές RMSFs (Πίνακας 3.5). Το τροχιακό των 283K (Εικόνα 3.40) δίνει 4 cluster (με RMSD cut-off 1.00 και variance-explained 0.89) με κατοχή σε χρόνο προσομοίωσης 22.3%, 22.9%, 2.6% και 4.0% αντίστοιχα.

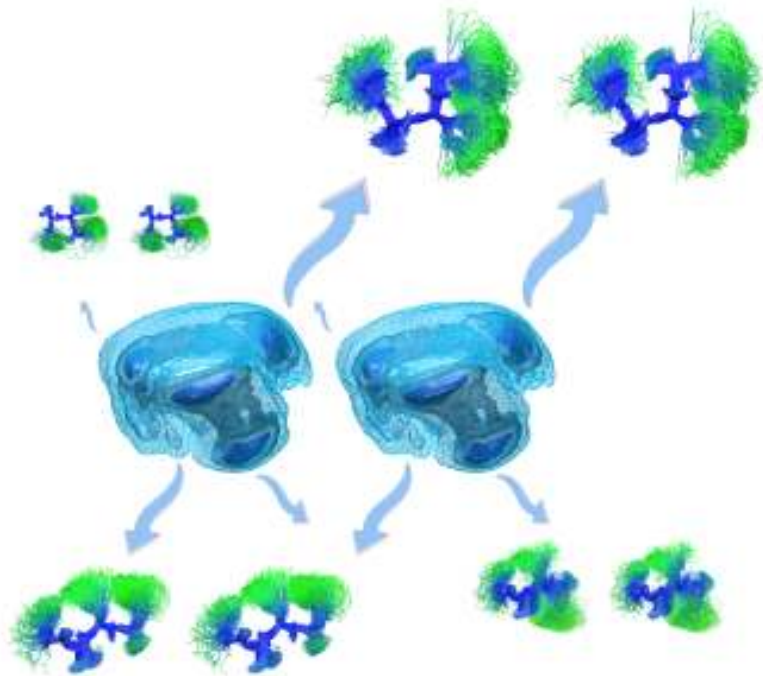


Εικόνα 3.39 Γραφική απεικόνιση του ενιαίου πίνακα RMSD μετά την ένωση των τροχιακών των 4 θερμοκρασιών του πεπτιδίου EVKW, για όλα τα βαριά άτομα (αριστερά) και για τα άτομα του πεπτιδικού σκελετού (δεξιά), σε αντιστοιχία με την Εικόνα 3.21. Ο χρωματικός κώδικας για τα τροχιακά διατηρείται κοινός.

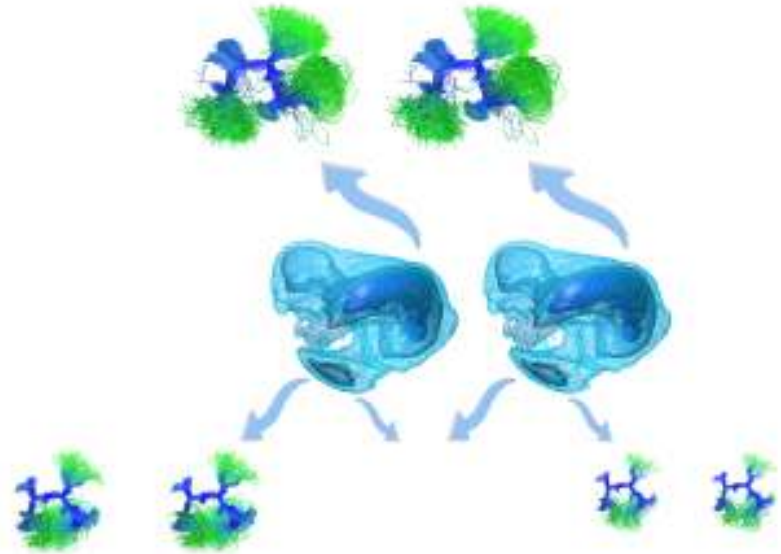
Η δομή του κυρίαρχου cluster μοιάζει πολύ με την αντίστοιχη δομή θηλιάς σχήματος "π" του πεπτιδίου DTRW (Εικόνα 3.38 *versus* Εικόνα 3.44), με ισχυρό πακετάρισμα μεταξύ των καταλοίπων Trp-Lys αυτή τη φορά.



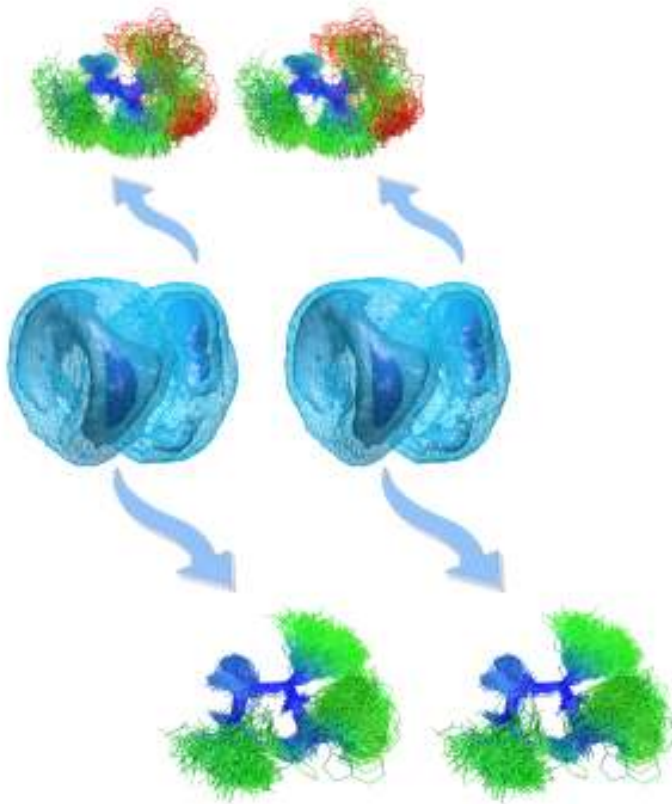
Εικόνα 3.40 Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου EVKW για το τροχιακό των 283K, σε αντιστοιχία με την Εικόνα 3.22.



Εικόνα 3.41 Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου EVKW για το τροχιακό των 298K, σε αντιστοιχία με την Εικόνα 3.22.

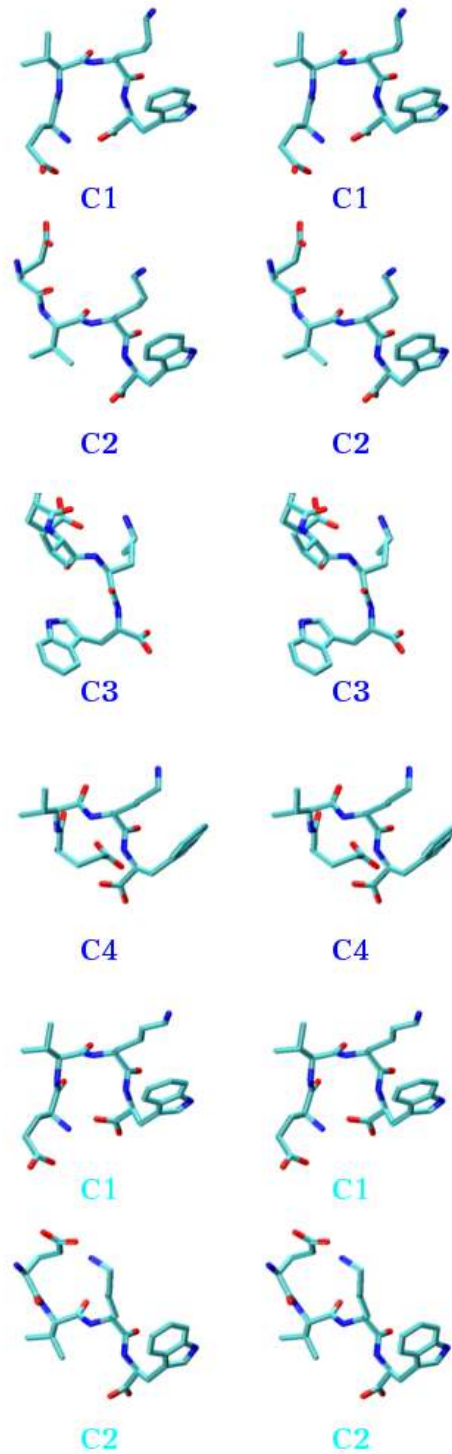


Εικόνα 3.42 Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου EVKW για το τροχιακό των 320K, σε αντιστοιχία με την Εικόνα 3.22.



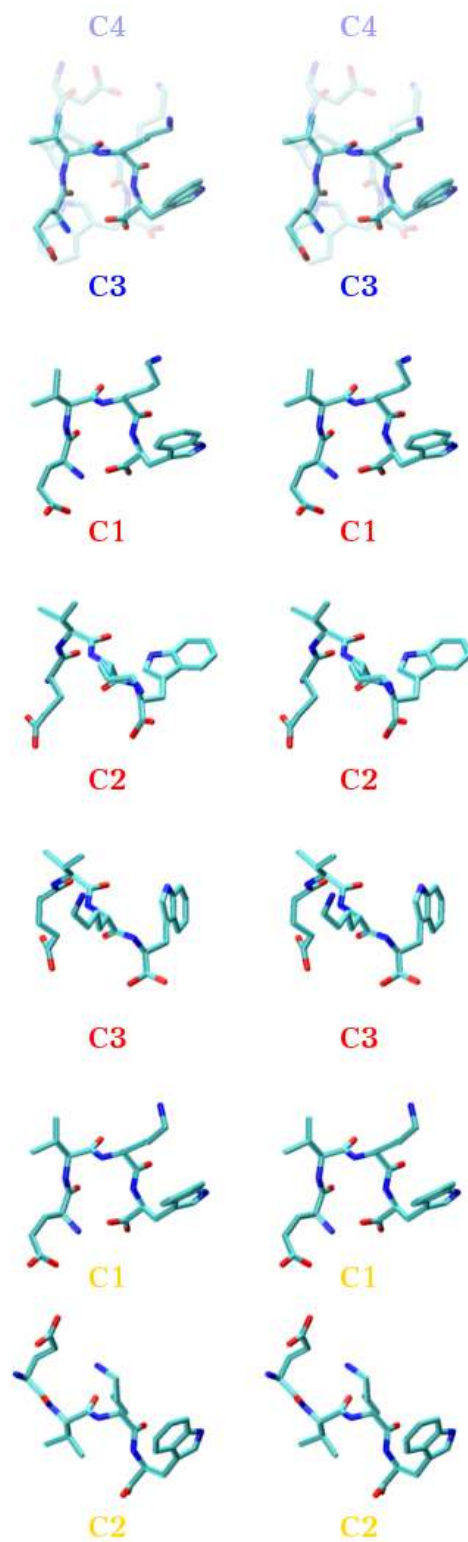
Εικόνα 3.43 Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου EVKW για το τροχιακό των 340K, σε αντιστοιχία με την Εικόνα 3.22.





Εικόνα 3.44 Αντιπροσωπευτικές δομές (σε stereo αναπαράσταση) από τα cluster των τροχιακών των τεσσάρων θερμοκρασιών για το EVKW. Ο χρωματικός κώδικας των θερμοκρασιών είναι ίδιος με της Εικόνας 3.21.





Εικόνα 3.44 (συνέχεια) Αντιπροσωπευτικές δομές (σε stereo αναπαράσταση) από τα cluster των τροχιακών των τεσσάρων θερμοκρασιών για το EVKW.

Η δομή του δεύτερου cluster (Εικόνα 3.44) περιλαμβάνει σχεδόν εκτεταμένη διαμόρφωση του πεπτιδικού σκελετού, και όλες τις πλευρικές ομάδες στη μία πλευρά του πεπτιδικού σκελετού, με τη λυσίνη στη μέση, να αλληλεπιδρά από τη μία με την τρυπτοφάνη και από την άλλη με το γλουταμικό οξύ. Στο τρίτο cluster δομών βλέπουμε τη δομή θηλιάς να σχηματίζεται μέσω της ηλεκτροστατικής αλληλεπίδρασης των πλευρικών ομάδων της αργινίνης με το γλουταμικό οξύ (και όχι μεταξύ των ελεύθερων N- και C- άκρων). Η τρυπτοφάνη πακετάρεται από την άλλη πλευρά, πάνω από τη βαλίνη ενώ η NE ομάδα της αλληλεπιδρά με το καρβονυλικό οξυγόνο της βαλίνης (Εικόνα 3.44). Το τέταρτο cluster δομών είναι πολύ κοντινό με το πρώτο, με μόνη διαφορά τη διαμόρφωση της πλευρικής ομάδας της αργινίνης.

Το τροχιακό των 298K (Εικόνα 3.41) δίνει 4 cluster (με RMSD cut-off 2.90 και variance-explained 0.76) με κατοχή σε χρόνο προσομοίωσης 16.6%, 8.1%, 2.7% και 3.2% αντίστοιχα. Οι δομές των 4 cluster είναι σε αντιστοιχία με τις δομές που συναντήσαμε στα 4 cluster του τροχιακού των 283K και με την ίδια σειρά εμφάνισης (Εικόνα 3.44).

Το τροχιακό των 320K (Εικόνα 3.42) δίνει 3 cluster (με RMSD cut-off 1.00 και variance-explained 0.86) με κατοχή σε χρόνο προσομοίωσης 55.0%, 0.02% και <0.01% αντίστοιχα. Η δομή του κυρίαρχου cluster είναι αυτή που συναντήσαμε και στις άλλες θερμοκρασίες (Εικόνα 3.44). Οι δομές των δύο άλλων cluster είναι διαφορετικές από όσες έχουμε συναντήσει μέχρι στιγμής, όπου αναπτύσσεται μία αλληλεπίδραση μεταξύ της τρυπτοφάνης με το γλουταμικό οξύ. Ωστόσο οι δομές αυτές διαρκούν πολύ λίγο μόλις 127 και 43 frames (Εικόνα 3.44, σε υπέρθεση με τη δομή του cluster 4 σε διαφάνεια) αντίστοιχα.

Το τροχιακό των 340K (Εικόνα 3.43) δίνει 2 cluster (με RMSD cut-off 1.00 και variance-explained 0.88) με κατοχή σε χρόνο προσομοίωσης 32.8% και 15.2% αντίστοιχα και δομές αυτές που συναντήσαμε στα δύο κυρίαρχα cluster και στις υπόλοιπες θερμοκρασίες αλλά με εμφανώς υψηλότερες ατομικές διακυμάνσεις (Εικόνα 3.44).

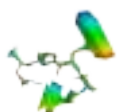


### Torsion Angles

Τα τέσσερα πεπτίδια που επιλέχθηκαν για περαιτέρω μελέτη και παρουσιάζονται στην ενότητα αυτή, παρουσιάζουν κάποιες ομοιότητες σε επίπεδο αλληλουχίας. Τα πεπτίδια RWPD και RPWD περιέχουν τα ίδια αμινοξικά κατάλοιπα και διαφέρουν μόνο στη σειρά των καταλοίπων

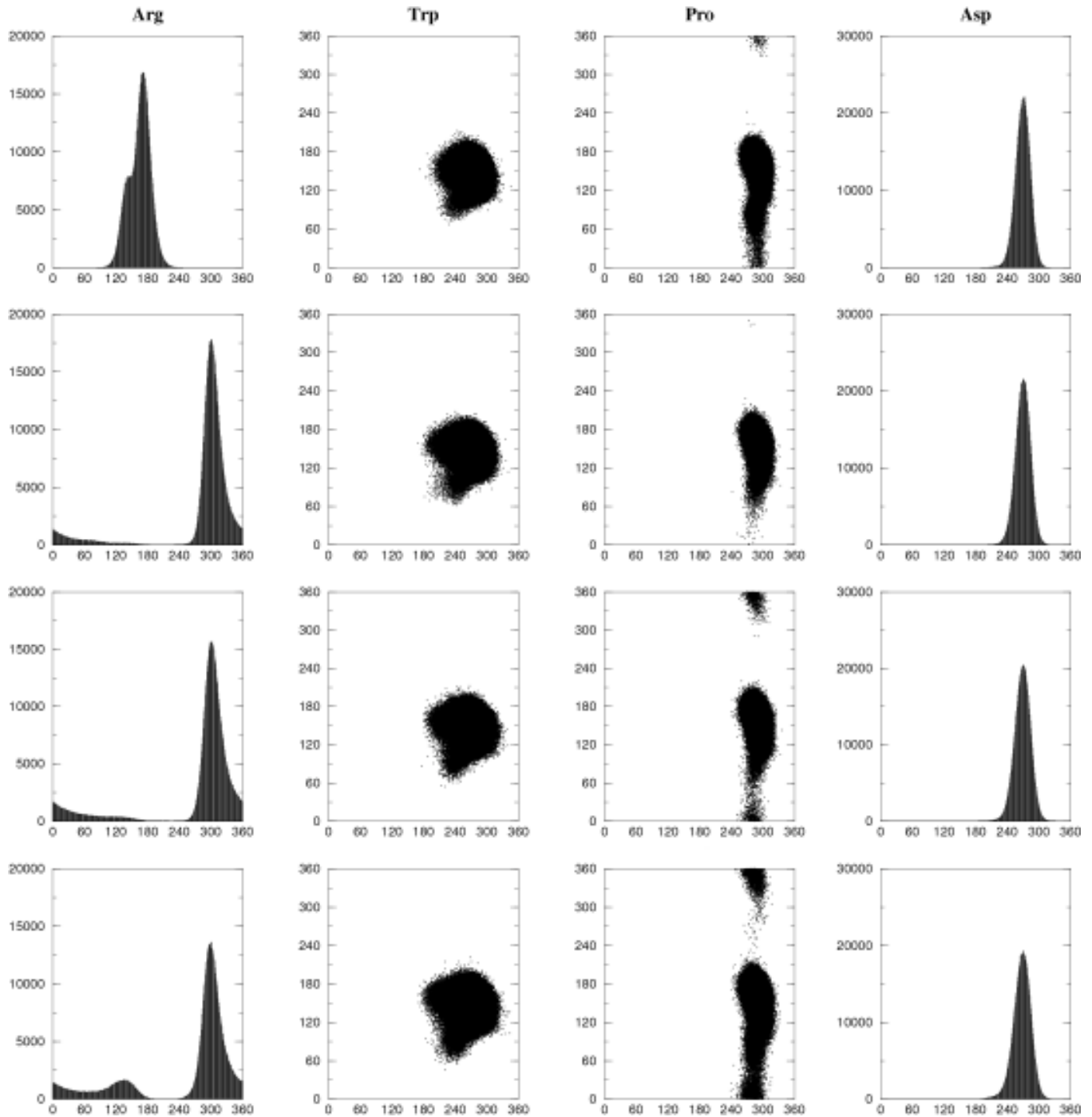
Pro-Trp. Από την άλλη, τα πεπτίδια DTRW και EVKW, δεν περιλαμβάνουν προλίνη, έχουν την Trp στην τελευταία θέση, από της οποίας προηγείται ένα θετικά φορτισμένο αμινοξύ με μεγάλη πλευρική ομάδα ενώ το αρνητικό φορτίο, την παρουσία του οποίου έχουμε επιβάλει εμείς, το συναντάμε στην πρώτη θέση. Η ομοιότητα στο επίπεδο της αλληλουχίας μεταφέρεται και στο επίπεδο της δομής, καθώς βλέπουμε τα πεπτίδια RWPD-RPWD και DTRW-EVKW να υιοθετούν συναφείς δομές, κυρίως σε επίπεδο πεπτιδικού σκελετού. Αυτό γίνεται ιδιαίτερος εμφανές εάν υπολογίσουμε την εξέλιξη στο χρόνο των διέδρων φ/ψ γωνιών για κάθε κατάλοιπο (Εικόνες 3.45-3.48). Ο υπολογισμός των τιμών των διέδρων γωνιών (torsion angles) έγινε με το πρόγραμμα CARMA (Glykos, 2006).

Από τις κατανομές αυτές φαίνεται ότι η διαμόρφωση του πεπτιδικού σκελετού είναι παρόμοια μεταξύ των πεπτιδίων RWPD-RPWD και DTRW-EVKW. Η διαφοροποίηση των κατανομών στα τροχιακά διαφορετικών θερμοκρασιών είναι σε συμφωνία με τις προηγούμενες παρατηρήσεις αναφορικά με τις διαμορφώσεις που υιοθετεί κάθε πεπτίδιο στο εύρος θερμοκρασίας που εξετάσαμε. Βλέπουμε για παράδειγμα ότι η τρυπτοφάνη έχει μόνο ένα cluster τιμών στο διάγραμμα Ramachandran στο πεπτίδιο RWPD ενώ έχει δύο στο πεπτίδιο RPWD, ενώ τη μεγαλύτερη διακύμανση την εμφανίζει η αργινίνη και στις δύο περιπτώσεις. Η ανάπτυξη παρόμοιων μοτίβων αλληλεπιδράσεων στα πεπτίδια DTRW και EVKW διαφαίνεται και από τις κατανομές των διέδρων γωνιών των δύο κεντρικών καταλοίπων αλλά και της τρυπτοφάνης.

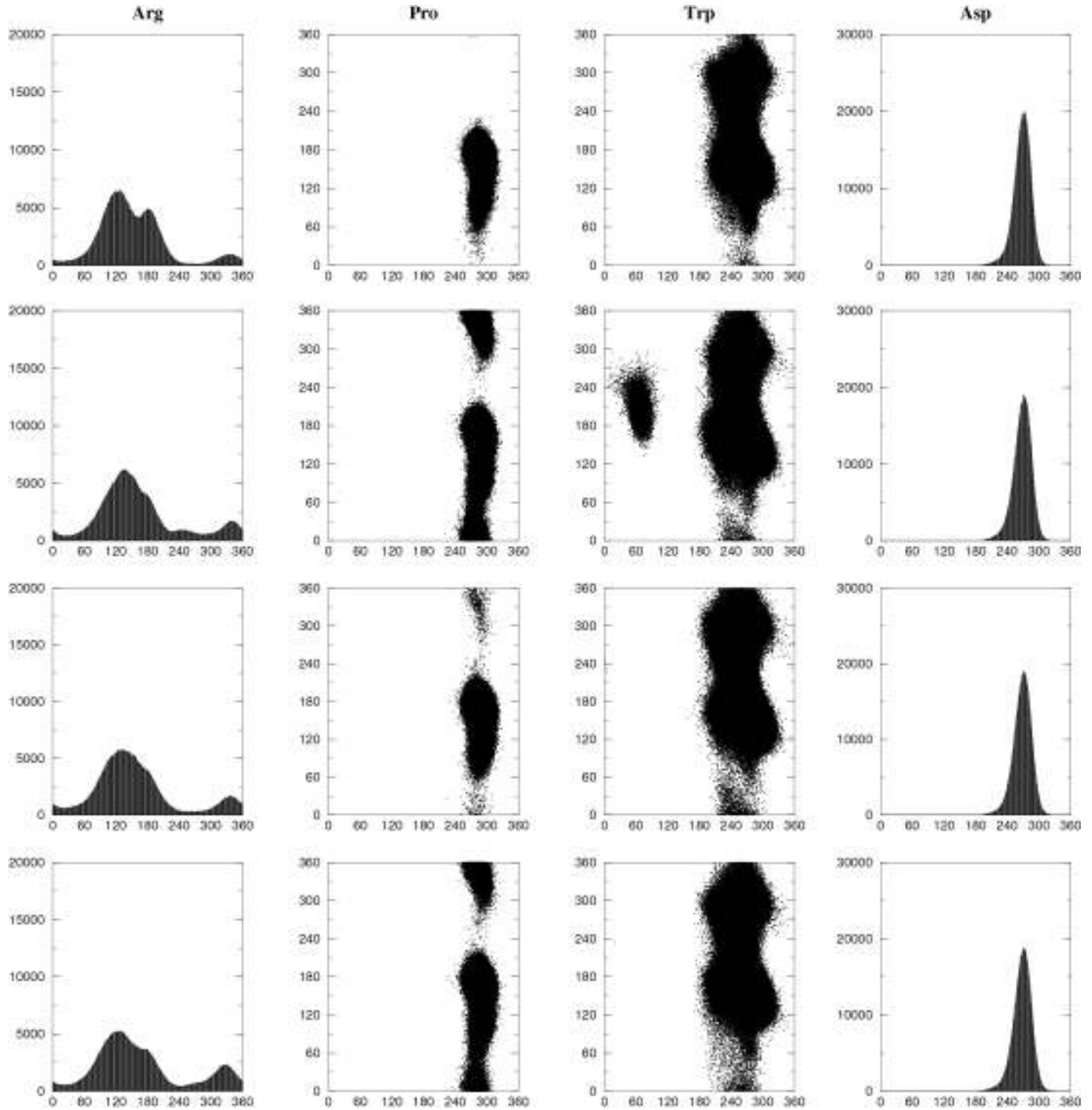


### Χαρακτηριστικές κινήσεις

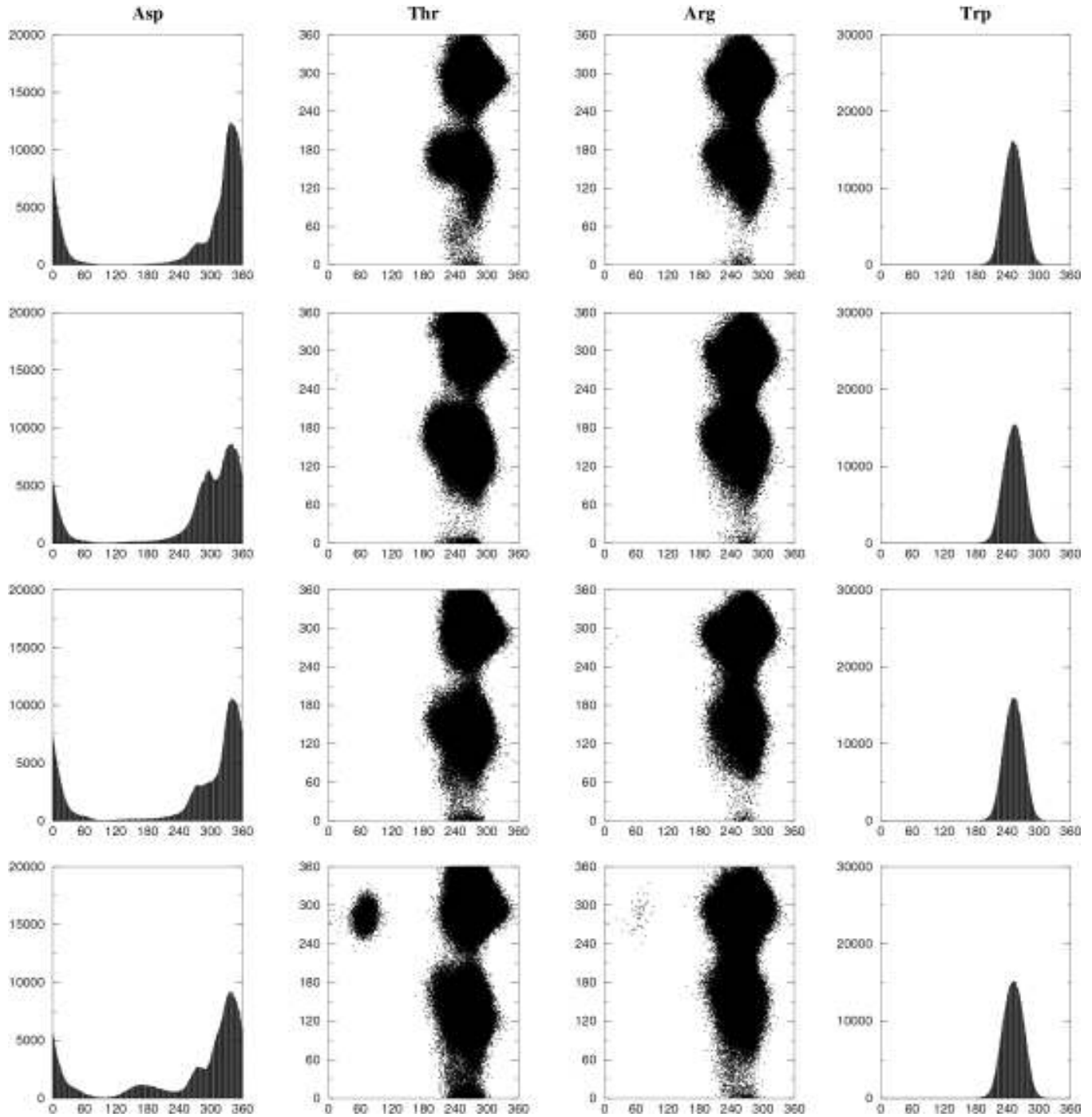
Από την ανάλυση PCA, μπορούμε να αναγάγουμε τη σύνθετη κίνηση που πραγματοποιεί το μόριο σε ένα άθροισμα απλών κινήσεων που περιγράφονται από τα ζεύγη των κυρίαρχων συνιστωσών (eigenvector-eigenvalue pairs) στα οποία έχει γίνει προβολή του τροχιακού (Ενότητα 2.4). Στην Εικόνα 3.49 βλέπουμε τη χαρακτηριστική κίνηση που πραγματοποιούν τα τέσσερα πεπτίδια, στις τέσσερις θερμοκρασίες με βάση τον πρώτο eigenvector. Το εύρος της κίνησης κάθε αμινοξέος είναι συμβατό με τις ατομικές διακυμάνσεις που έχουμε υπολογίσει (Πίνακας 3.5), με την άνοδο της θερμοκρασίας να οδηγεί σε μεγαλύτερο εύρος κίνησης, όπως φαίνεται καθαρά στο πεπτίδιο RWPD για το οποίο υπάρχει μία κυρίαρχη διαμόρφωση. Τα αμινοξέα με ογκώδεις πλευρικές ομάδες δείχνουν χαρακτηριστικά μεγαλύτερο εύρος κίνησης.



Εικόνα 3.45 Κατανομές των διεδρων γωνιών (torsion angles) για το πεπτίδιο RWPD για τα τέσσερα τροχιακά από τους 283K (πάνω) έως τους 340K (κάτω). Για τα κατάλοιπα στις θέσεις 1 και 4 υπολογίζεται μόνο η  $\psi$  και  $\phi$  διεδρη γωνία αντιστοίχως (ιστόγραμμα κατανομής), ενώ για τα κατάλοιπα στις θέσεις 2 και 4 υπολογίζονται και οι δύο διεδρες γωνίες (Ramachandran plots).

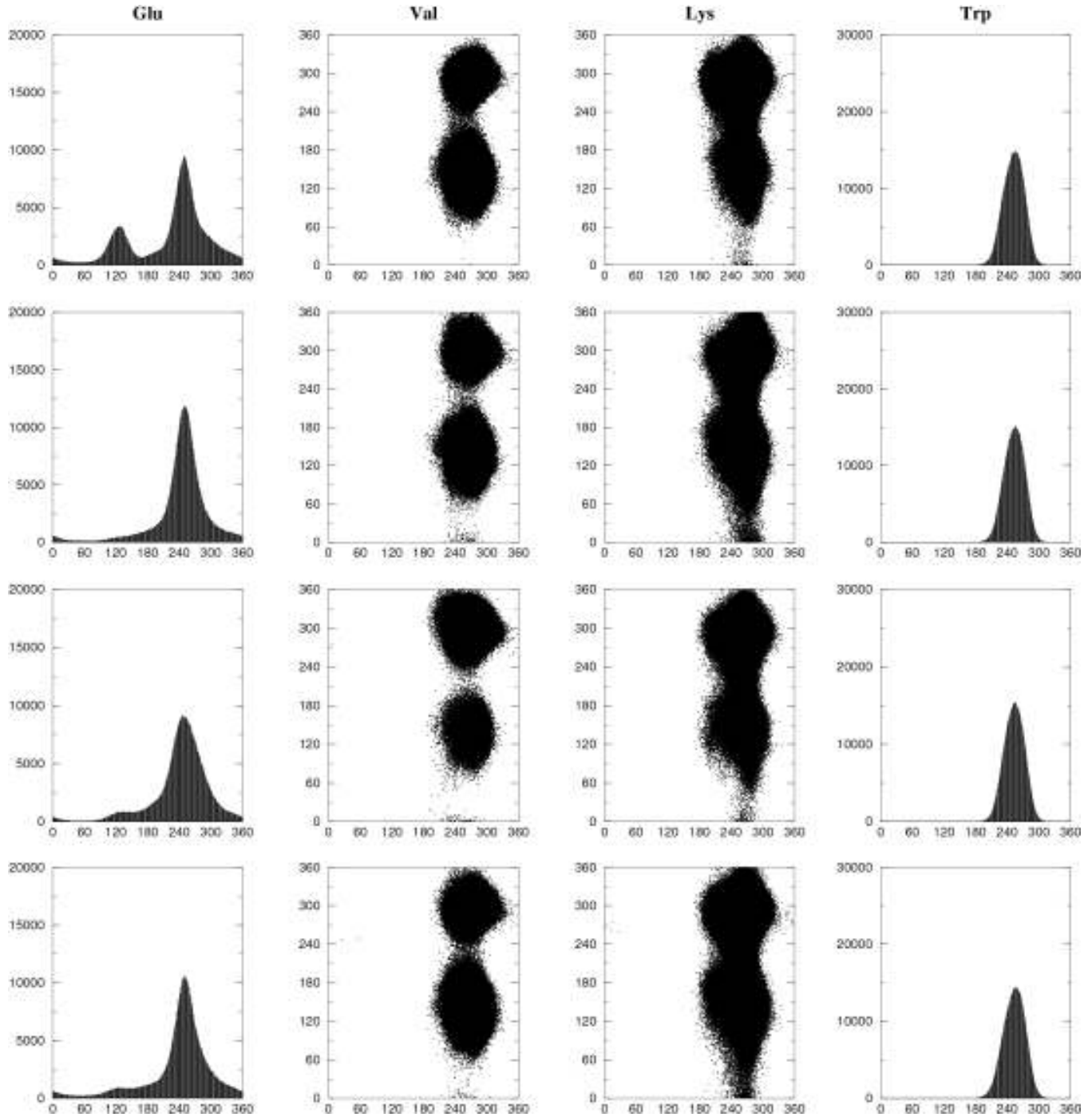


Εικόνα 3.46 Κατανομές των διεδρων γωνιών (torsion angles) για το πεπτίδιο RPWD για τα τέσσερα τροχιακά από τους 283K (πάνω) έως τους 340K (κάτω). Για τα κατάλοιπα στις θέσεις 1 και 4 υπολογίζεται μόνο  $\psi$  και  $\phi$  διεδρη γωνία αντιστοίχως (ιστόγραμμα κατανομής), ενώ για τα κατάλοιπα στις θέσεις 2 και 4 υπολογίζονται και οι δύο διεδρες γωνίες (Ramachandran plots).

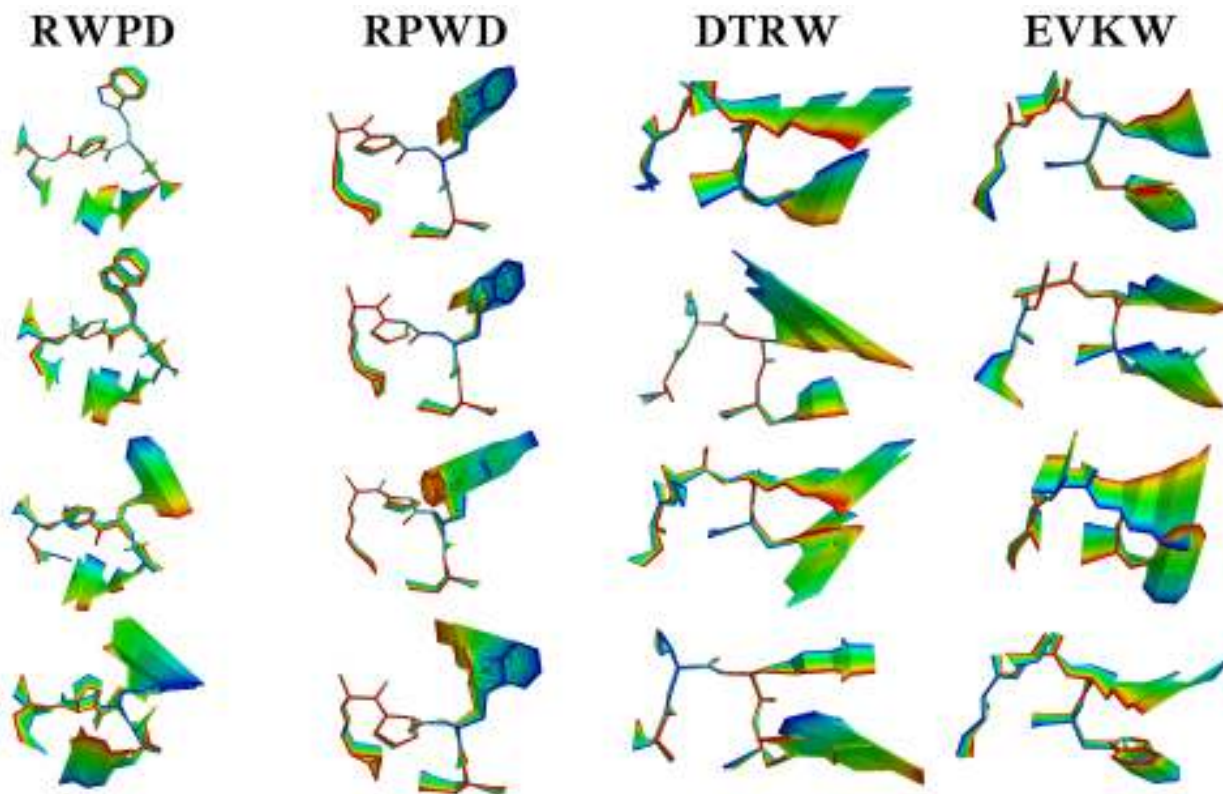


Εικόνα 3.47 Κατανομές των διεδρων γωνιών (torsion angles) για το πεπτίδιο DTRW για τα τέσσερα τροχιακά από τους 283K (πάνω) έως τους 340K (κάτω). Για τα κατάλοιπα στις θέσεις 1 και 4 υπολογίζεται μόνο η  $\psi$  και  $\phi$  διεδρη γωνία αντιστοίχως (ιστόγραμμα κατανομής), ενώ για τα κατάλοιπα στις θέσεις 2 και 4 υπολογίζονται και οι δύο διεδρες γωνίες (Ramachandran plots).





Εικόνα 3.48 Κατανομές των διεδρων γωνιών (torsion angles) για το πεπτιδίο EVKW για τα τέσσερα τροχιακά από τους 283K (πάνω) έως τους 340K (κάτω). Για τα κατάλοιπα στις θέσεις 1 και 4 υπολογίζεται μόνο η  $\psi$  και  $\phi$  διεδρη γωνία αντιστοίχως (ιστόγραμμα κατανομής), ενώ για τα κατάλοιπα στις θέσεις 2 και 4 υπολογίζονται και οι δύο διεδρες γωνίες (Ramachandran plots).

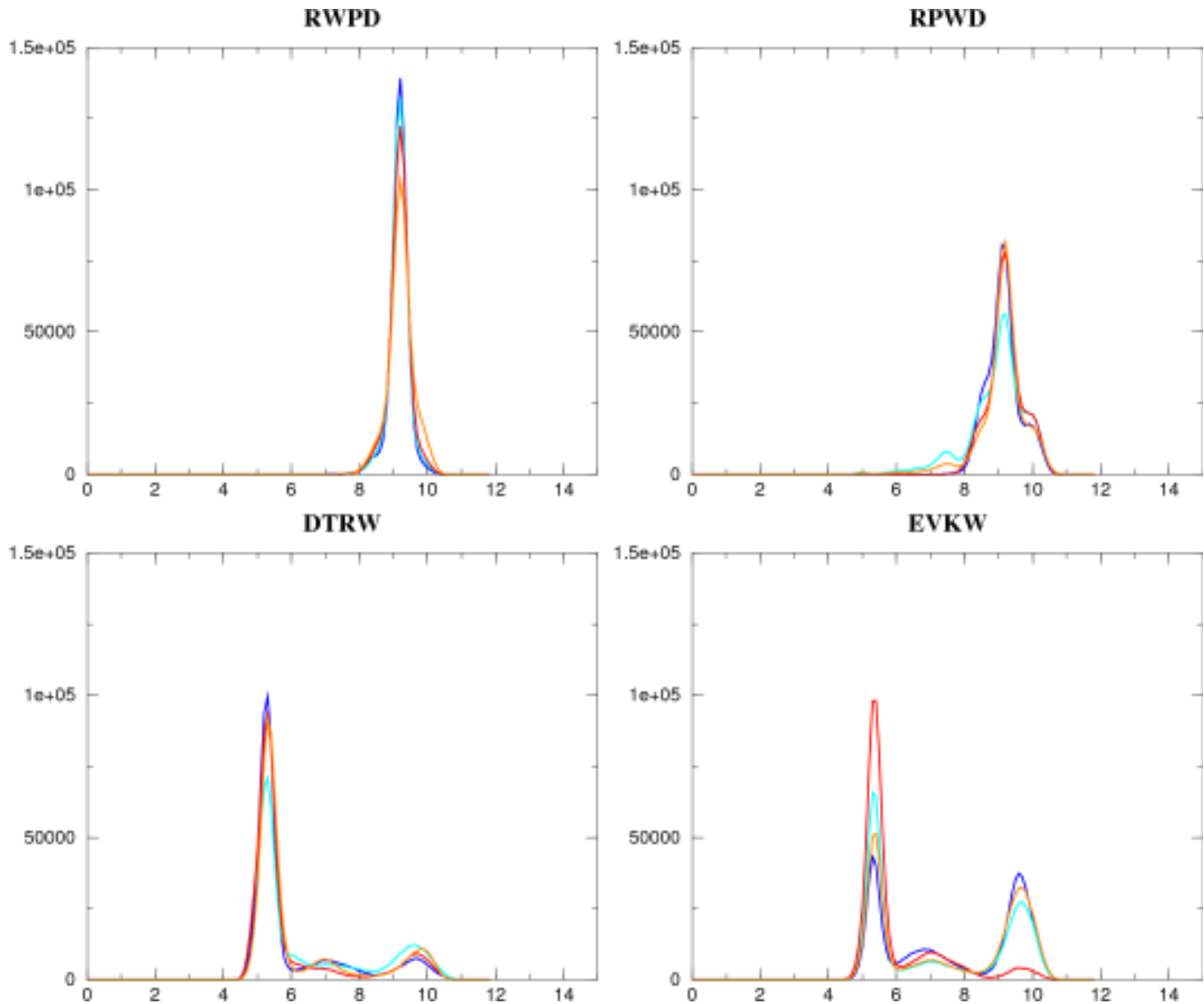


Εικόνα 3.49 Χαρακτηριστική κίνηση που περιγράφεται από τον eigenvector με την υψηλότερη τιμή eigenvalue της ανάλυσης Cartesian-PCA (heavy atoms) για τα τέσσερα πεπτίδια και τα τέσσερα τροχιακά από τους 283K (πάνω) έως τους 340K (κάτω). Η περιγραφή όλου του φάσματος της κίνησης, γίνεται μέσω της υπέρθεσης 80 δομών, με το χρωματισμό από μπλε σε κόκκινο να δείχνει το εύρος αυτής.



### Loop-closure

Η εξέλιξη της απόστασης μεταξύ των άκρων είναι μία παράμετρος που μελετήσαμε εκτενώς στις προηγούμενες ενότητες και την οποία ενσωματώσαμε στις συναρτήσεις εκτίμησης της αναδιπλωσιμότητας TF1 και TF2. Αποτελεί έναν εύκολο και άμεσο τρόπο μελέτης της δημιουργίας της δομής θηλιάς, της συχνότητας με την οποία συμβαίνει το γεγονός αλλά και της διάρκειας της στο χρόνο της προσομοίωσης.



Εικόνα 3.50 Κατανομές των διαφόρων πληθυσμών από διαμορφώσεις για τα τέσσερα πεπτίδια ως συνάρτηση της απόστασης μεταξύ των άκρων (N-C distance). Ο χρωματικός κώδικας των θερμοκρασιών είναι ίδιος με της Εικόνας 3.21.

Στην Εικόνα 3.50 βλέπουμε τις κατανομές των τιμών των αποστάσεων μεταξύ N- τελικού και C- τελικού άκρου με βάση τα Ca άτομα των καταλοίπων 1 και 4 (Κεφάλαιο 2, Ενότητα 2.3, 1-4Dist). Για το πεπτίδιο RWPD βλέπουμε έναν κύριο πληθυσμό, ο οποίος μειώνεται όσο αυξάνεται η θερμοκρασία της προσομοίωσης. Στα λιγότερο σταθερά πεπτίδια, αρχίζουμε να

βλέπουμε τη δημιουργία "ώμων" στα δεξιά της κατανομής, που ανταποκρίνονται σε πληθυσμούς με περισσότερο εκτεταμένη διαμόρφωση. Οι πληθυσμοί αυτοί όμως μπορεί να περιλαμβάνουν είτε μη-αναδιπλωμένες δομές (disordered) ή εναλλακτικές διαμορφώσεις των πεπτιδίων όπου η απόσταση των άκρων έχει υψηλή τιμή (όπως είδαμε στην Ενότητα 2.3 κατά το σχεδιασμό των συναρτήσεων εκτίμησης της αναδιπλωσιμότητας TF1 και TF2) είτε σε μείγμα αυτών. Η προφανής παρερμηνεία ως μεγαλύτερης αστάθειας κάποιων τροχιακών, όπως για παράδειγμα στους 298K των πεπτιδίων RPWD και DTRW, εξηγείται από την παρουσία περισσότερων από μία διακριτών διαμορφώσεων για τα πεπτίδια αυτά, όπως προέκυψε και από την ανάλυση Cartesian-PCA. Έτσι, το πεπτίδιο EVKW, με βάση μόνο την απόσταση μεταξύ των άκρων, φαίνεται πως έχει σημαντικό αριθμό από εκτεταμένες διαμορφώσεις, με εξαίρεση το τροχιακό στους 320K. Στην πραγματικότητα, στο τροχιακό των 320K βλέπουμε μία κύρια διαμόρφωση με κατοχή ~55% του χρόνου προσομοίωσης, ενώ στα τροχιακά των υπόλοιπων θερμοκρασιών βλέπουμε τουλάχιστον 2 κύριες διαμορφώσεις με κατοχή 15-30% του χρόνου προσομοίωσης και αρκετές παραδοικές (με κατοχή <5% του χρόνου προσομοίωσης). Οι διακυμάνσεις αυτές στους πληθυσμούς που παρατηρούνται μπορεί να οφείλονται στον περιορισμένο χρόνο της προσομοίωσης ή στην αδυναμία του συγκεκριμένου force field να αναπαραστήσει σωστά τη σχέση αυτή σε θερμοκρασίες πέραν της παραμετροποίησής του (room temperature). Σε κάθε περίπτωση, η ελάττωση της θερμοκρασίας της προσομοίωσης, μπορεί να μειώνει την κινητικότητα και επομένως να ευνοεί τη σταθεροποίηση του πεπτιδίου, αλλά ταυτόχρονα μειώνεται η πιθανότητα δημιουργίας ενός γεγονότος αναδίπλωσης, με αποτέλεσμα να χρειαζόμαστε περισσότερο χρόνο για να δούμε αναδίπλωση.

$$\frac{\Delta G_{\text{folding}}}{T} = -R \ln \frac{P_{\text{folded}}}{P_{\text{unfolded}}}$$

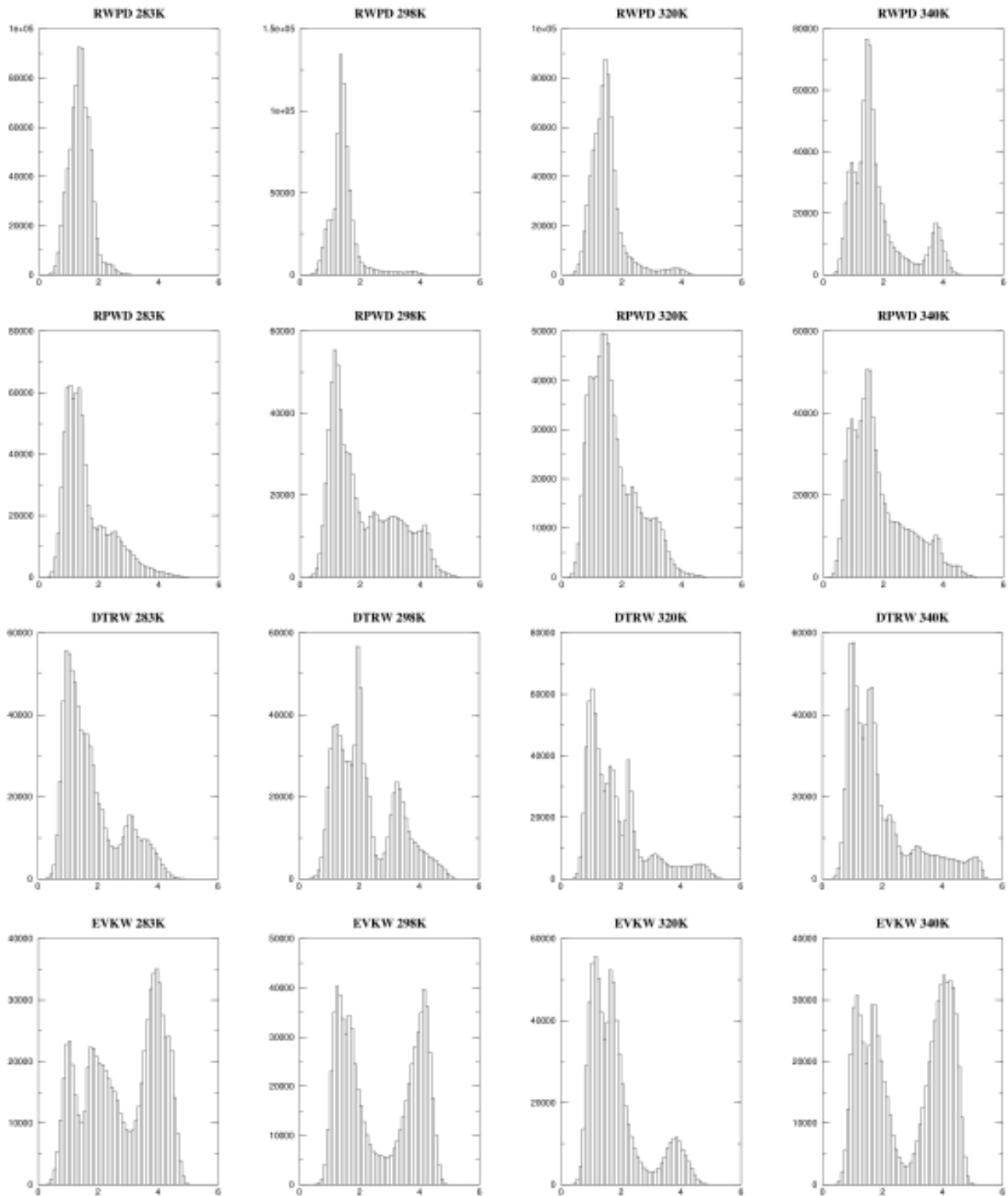
Για κάθε ένα από τα πεπτίδια και για κάθε μία από τις θερμοκρασίες, όπως δείξαμε, υπολογίσαμε μέσω cluster analysis τόσο το κυρίαρχο cluster όσο και την αντιπροσωπευτική δομή, δηλαδή το στιγμιότυπο του τροχιακού που βρίσκεται πλησιέστερα (σε επίπεδο RMSD) στην υπολογιζόμενη μέση δομή. Η δομή αυτή μπορεί να χρησιμοποιηθεί ως κριτήριο για το διαχωρισμό των δομών του τροχιακού σε δύο πληθυσμούς, F (Folded) και U (Unfolded), με την παραδοχή ότι έχουμε κινητική δύο σταδίων (two-state folders). Στη συνέχεια, οι (υπολογισμένες βάσει των προσομοιώσεων) πιθανότητες να βρεθεί το πεπτίδιο στη μία ή την άλλη κατάσταση

μπορούν να χρησιμοποιηθούν για μία αδρή εκτίμηση της ελεύθερης ενέργειας αναδίπλωσης ( $\Delta G_{\text{folding}}$ ) με βάση τον τύπο:

$$\frac{\Delta G_{FU}}{T} = -R \ln \frac{P_{\text{FOLDED}}}{P_{\text{UNFOLDED}}}$$

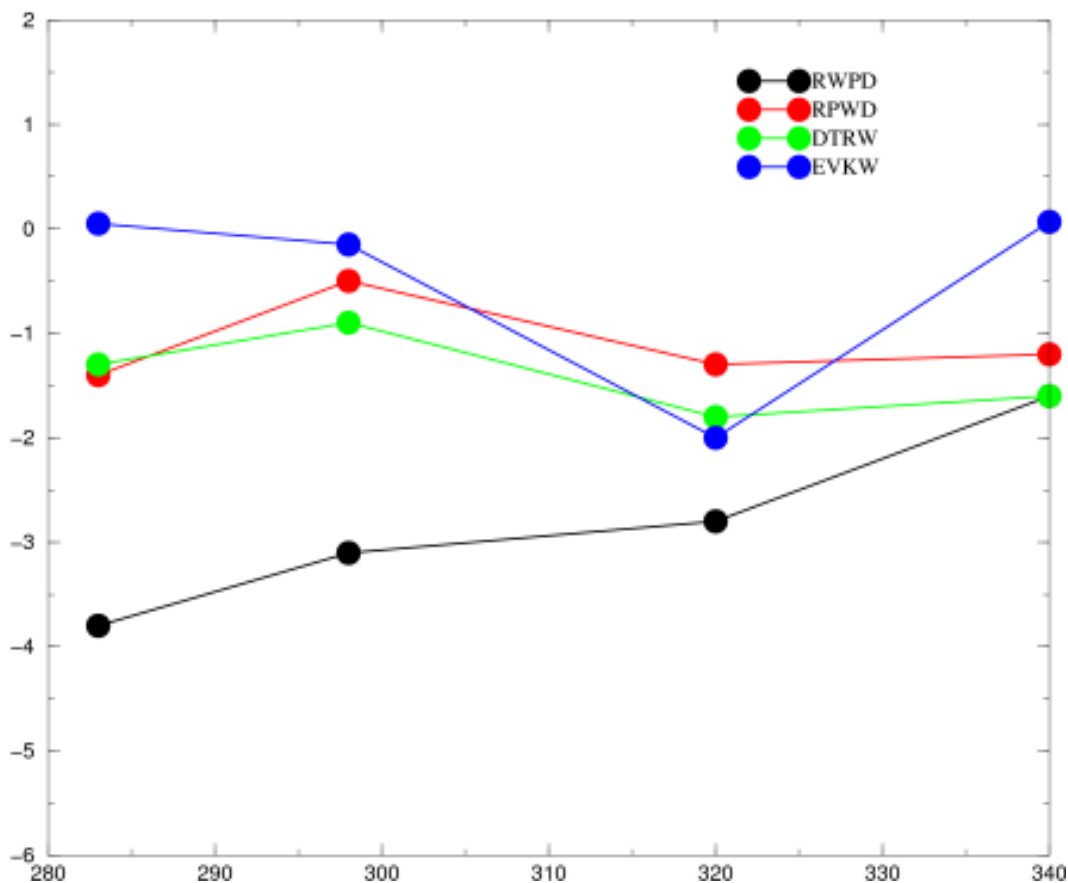
Ο προσδιορισμός της τιμής που χρησιμοποιούμε για να διαχωρίσουμε τις δομές του τροχιακού στους δύο διακριτούς πληθυσμούς προκύπτει από τα ιστογράμματα κατανομών των τιμών RMSD ολόκληρου του τροχιακού από την αντιπροσωπευτική δομή. Οι υπολογισμοί αυτοί μπορούν να γίνουν για διάφορα σύνολα ατόμων (πεπτιδικός σκελετός, βαριά άτομα). Εδώ παρουσιάζουμε τα αποτελέσματα χρησιμοποιώντας όλα τα βαριά άτομα, καθώς σε αυτά έχουμε επικεντρωθεί σε όλη την ενότητα. Ωστόσο, η επιλογή της τιμής που ορίζεται ως κατώφλι για το διαχωρισμό στους πληθυσμούς δεν είναι εύκολη στην περίπτωση πεπτιδίων με περισσότερες από μία διακριτές διαμορφώσεις. Έτσι, σύμφωνα με τις κατανομές της Εικόνας 3.51, μπορούμε να θεωρήσουμε ότι όλες οι δομές του τροχιακού με  $\text{RMSD} < 2.2\text{\AA}$  ανήκουν στον πληθυσμό F, και κατά συνέπεια όλες οι υπόλοιπες χαρακτηρίζονται ως U για την περίπτωση των πεπτιδίων RWPD και RPWD. Για τα πεπτίδια DTRW και EVKW βλέπουμε πολλαπλές κορυφές που αντιστοιχούν στα διακριτά cluster δομών που έχουμε προσδιορίσει. Έτσι, αν επιλέξουμε ως κατώφλι τα  $2.5\text{\AA}$  και  $2.8\text{\AA}$ , αντίστοιχα, περιλαμβάνουμε όλες τις διακριτές διαμορφώσεις στον πληθυσμό F ενώ αν επιλέξουμε ως κατώφλι τα  $2.2\text{\AA}$  και  $1.5\text{\AA}$  αντίστοιχα, περιλαμβάνουμε μόνο την κυρίαρχη διαμόρφωση.

Εάν χρησιμοποιήσουμε τις πιθανότητες αυτές, που υπολογίζουμε μέσω των προσομοιώσεων, στην παραπάνω συνάρτηση, θεωρώντας όλο το μείγμα διαμορφώσεων ως πληθυσμό F, προκύπτουν οι τιμές  $\Delta G_{\text{folding}}$  που παρουσιάζουμε συγκεντρωτικά στην Εικόνα 3.52. Η εικόνα αυτή περιλαμβάνει συνοπτικά όλα όσα παρατηρήσαμε στην ενότητα αυτή. Έτσι βλέπουμε τη σχεδόν γραμμικώς αντίστροφη σχέση στη σταθερότητα του πεπτιδίου RWPD (όπως εκφράζεται μέσω του  $\Delta G_{\text{folding}}$ ) με την άνοδο της θερμοκρασίας καθώς υπάρχει μία κυρίαρχη διαμόρφωση σε όλα τα τροχιακά. Η σχέση αυτή διατηρείται και για το πεπτίδιο RPWD με εξαίρεση το τροχιακό των 298K λόγω της παρουσίας του μείγματος των διαμορφώσεων, φαινόμενο το οποίο είναι περισσότερο εμφανές στο πεπτίδιο DTRW, ενώ διαφαίνεται και η αστάθεια του πεπτιδίου



Εικόνα 3.51 Ιστογράμματα κατανομής των RMSD όλων των δομών κάθε τροχιακού από την αντιπροσωπευτική δομή του κυρίαρχου cluster, όπως αυτή προσδιορίστηκε μέσω Cartesian-PCA και για όλα τα βαριά άτομα.





Εικόνα 3.52 Εκτιμώμενη ελεύθερη ενέργεια αναδίπλωσης (KJ/mol) ως συνάρτηση της θερμοκρασίας (K) διεξαγωγής της προσομοίωσης.

EVKW με εξαίρεση το τροχιακό στους 320K όπου βλέπουμε και πάλι μία κυρίαρχη διαμόρφωση.

Οι συμπεριφορές αυτές μπορούν να δικαιολογηθούν ποικιλοτρόπως, και θα μπορούσαν να αποδοθούν στην αδυναμία του συγκεκριμένου force field να παράγει σωστά αποτελέσματα σε θερμοκρασίες πέρα από την παραμετροποίησή του (Room Temperature) (Zhou, 2003) ή στο μη επαρκή χρόνο της προσομοίωσης (insufficient sampling). Ένα επίσης σημαντικό δεδομένο είναι ότι ο διαχωρισμός σε πληθυσμούς F και U δε γίνεται με βάση κάποιο ενεργειακό κριτήριο, αλλά με βάση το RMSD, δηλαδή μία παράμετρο μετρική. Από την άλλη, πρέπει να έχουμε πάντα υπόψιν ότι η αναδίπλωση δε γίνεται ερήμην του διαλύτη, και οι μεταβολές τόσο στην εντροπία του διαλύτη (Makhatadze et al., 1996, Harano et al., 2005) όσο και στην εντροπία της διαμόρφωσης (configurational entropy) των πλευρικών ομάδων (rotamer restriction) και σε

μικρότερο βαθμό στην εντροπία της ταλάντωσης του πεπτιδικού δεσμού (Doig et al., 1995), διαδραματίζουν σημαντικό ρόλο σε αυτήν την διαδικασία (και δεν έχουν ληφθεί υπόψη σε καμία ανάλυση). Για παράδειγμα, η δημιουργία στροφής στο τετραπεπτίδιο YKGQ φαίνεται να οφείλεται στην ενθαλπία στις χαμηλές θερμοκρασίες και στην εντροπία στις υψηλότερες (Kaur et al., 2012).



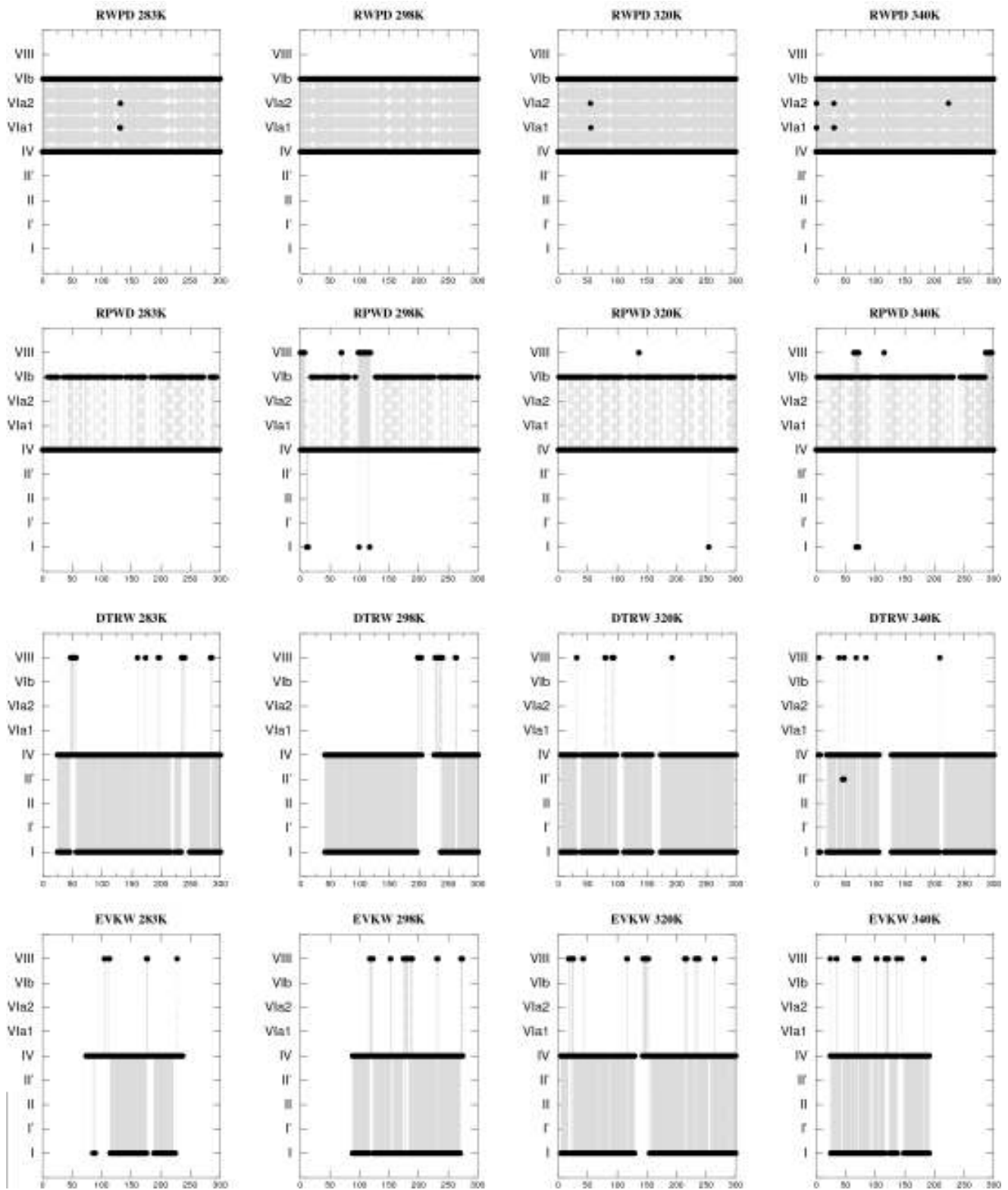
### β-στροφή

Αλληλουχίες μήκους τεσσάρων καταλοίπων δύνανται να σχηματίσουν β-στροφή, ένα δομικό μοτίβο με καίριο δομικό και λειτουργικό ρόλο σε ανώτερες πρωτεϊνικές δομές (Richardson, 1981). Το δομικό αυτό μοτίβο ταυτοποιήθηκε αρχικά με τη δημιουργία ενός δεσμού υδρογόνου μεταξύ της ομάδας CO του πεπτιδικού σκελετού του καταλοίπου  $i$  και της ομάδας NH του πεπτιδικού σκελετού του καταλοίπου  $i+3$  (Venkatachalam, 1968). Μετέπειτα όμως βρέθηκε ότι περίπου το 25% των β-στροφών υιοθετούν ανοιχτή διαμόρφωση, δηλαδή δεν παρατηρείται αυτός ο υδρογονοδεσμός (Lewis et al., 1973). Έτσι ο ορισμός μετασηματίστηκε και διευρύνθηκε ώστε να θεωρείται ότι το μοτίβο της β-στροφής συνίσταται από τέσσερα διαδοχικά κατάλοιπα με κριτήρια ότι (1) η απόσταση μεταξύ των ατόμων Ca των καταλοίπων  $i$  και  $i+3$  είναι λιγότερο από 7Å και (2) τα κατάλοιπα δεν βρίσκονται σε ελικοειδή διαμόρφωση (Richardson, 1981, Rose et al., 1985). Ένα τρίτο των καταλοίπων των σφαιρικών (globular) πρωτεϊνών ανήκουν σε αυτό το δομικό μοτίβο (Creighton, 1993). Οι β-στροφές κατηγοριοποιούνται περαιτέρω σε διάφορους τύπους με βάση έρευνες στην πρωτεϊνική βάση δεδομένων PDB (Bernstein et al., 1977). Δύο ανεξάρτητες έρευνες σε 205 (Hutchinson et al., 1994) και σε 426 πρωτεϊνικές αλυσίδες (Guruprasad et al., 2000) οδήγησαν στην ταυτοποίηση 3899 και 7153 β-στροφών αντίστοιχα, τις οποίες κατηγοριοποίησαν σε 9 τύπους, I, I', II, II', IV, VIa1, Via2, Vib, VIII, ανάλογα με τις διέδρες γωνίες του πεπτιδικού σκελετού (Lewis et al., 1973, Hutchinson et al., 1994, Wilmot et al., 1990). Οι τύποι IV και I είναι οι πιο συχνά παρατηρούμενοι, ο τύπος VIII αντιστοιχεί επακριβώς στα κριτήρια (1) και (2), ενώ για τους τύπους VIa1, Via2 και Vib είναι απαραίτητη η παρουσία προλίνης στην τρίτη θέση. Ο τύπος IV στην ουσία είναι μία ευρεία κατηγορία που περιλαμβάνει όλες τις β-στροφές που δεν ανήκουν στους υπόλοιπους τύπους (Richardson, 1981). Η ενασχόληση μας με τα τετραπεπτίδια μας οδήγησε αναπόφευκτα στη διερεύνηση ερωτημάτων

όπως: (1) οι δομές που παρατηρούμε στα τέσσερα αυτά τετραπεπτίδια είναι συμβατές με το μοτίβο της β-στροφής; Και αν ναι, (2) σε ποιόν τύπο ανήκουν; (3) Παρατηρείται διαφοροποίηση στις διάφορες θερμοκρασίες;

Για το σκοπό αυτό χρησιμοποιήσαμε το πρόγραμμα EUCB (Tsoulos et al., 2011) το οποίο ερευνά το τροχιακό για τη δημιουργία β-στροφής, βάσει των κριτηρίων που θέτει ο χρήστης για (1) την απόσταση μεταξύ των ατόμων  $Ca_i$  και  $Ca_{i+3}$ , (2) την τιμή της διέδρης γωνίας  $Ca_i-Ca_{i+1}-Ca_{i+2}-Ca_{i+3}$  και (3) την (ελάχιστη) κατοχή σε χρόνο προσομοίωσης. Στη συνέχεια οι β-στροφές αντιστοιχίζονται στους διάφορους τύπους βάσει των διέδρων  $\phi$ ,  $\psi$  γωνιών των δύο μεσαίων καταλοίπων. Για την επιλογή της απόστασης βασιστήκαμε στις κατανομές της Εικόνας 3.50, και ορίστηκε σε 9.5Å για τα πεπτίδια RWPD και RPWD και σε 6Å για τα πεπτίδια DTRW και EVKW. Η διέδρη γωνία ορίστηκε σε 90° (λόγω της παρουσίας της προλίνης στα δύο πεπτίδια). Με την προϋπόθεση ότι η β-στροφή παραμένει για τουλάχιστον 10% του χρόνου της προσομοίωσης (75.000frame ή 30ns), τα αποτελέσματα παρουσιάζονται στην Εικόνα 3.53. Βλέπουμε πως δεν υπάρχει μοναδική προτίμηση προς κάποιο τύπο, αλλά τα συμπεράσματα είναι εξίσου ενδιαφέροντα. Τα πεπτίδια RWPD και RPWD που περιέχουν προλίνη παίρνουν μόνο τους τύπους IV και Vib. Σποραδικά συναντάμε και τους τύπους VIa1, VIa2 και I, VIII στα δύο πεπτίδια αντίστοιχα. Από την άλλη, στα πεπτίδια DTRW και EVKW, συναντάμε τους τύπους I, IV και σπανίως τον τύπο VIII. Πέρα από το γεγονός ότι παρατηρούμε αλληπαλλήλες μετατροπές μεταξύ των διαφόρων τύπων β-στροφών, ο παράγοντας της θερμοκρασίας δε φαίνεται να διαδραματίζει κάποιο ρόλο ούτε στην επιλογή του τύπου της β-στροφής αλλά ούτε και στις συχνότητες εμφάνισης, τουλάχιστον για το συγκεκριμένο force field και για το συγκεκριμένο χρόνο προσομοίωσης. Οι τύποι στροφών που συναντάμε, είναι σε συμφωνία με την υπόλοιπη βιβλιογραφία, όσων αφορά τις προτιμήσεις των διαφόρων αμινοξέων και τους τύπους β-στροφών στις πρωτεΐνες (Hutchinson et al., 1994, Guruprasad et al., 2000).

Με βάση τις προτιμήσεις των αμινοξέων για τη δημιουργία β-στροφής (και των διαφόρων τύπων της) από τις κρυσταλλογραφικά προσδιορισμένες δομές αναπτύχθηκαν διάφορα μοντέλα όπως Site-Independent, 1-4/2-3 Residue-Correlation, Sequence-coupled στα οποία στηρίζονται οι διάφοροι αλγόριθμοι που κάνουν πρόβλεψη για δομή β-στροφής (Cai et al., 1999, Chou, 1997, 2000, Chou et al., 1979, 1997, Cid et al., 1982, Cohen et al., 1986, Kaur et al., 2003, McGregor et al., 1989, Shepherd et al., 1999, Wilmot et al., 1988, Zhang et al., 1997).



Εικόνα 3.53 Εξέλιξη στο χρόνο της προσομοίωσης των διαφόρων τύπων β-στροφών. Οι γκρι γραμμές που ενώνουν τα διαδοχικά σημεία είναι ενδεικτικές των μεταπτώσεων μεταξύ των τύπων στροφών.

Όλες αυτές οι αναλύσεις δεν αφορούν μεμονωμένα τετραπεπτίδια αλλά αλληλουχίες που αποτελούν μέρος μιας μεγαλύτερης πρωτεϊνικής αλυσίδας. Υπάρχουν αρκετοί διαθέσιμοι servers στο διαδίκτυο, όπως [DEBT](#) (Kountouris et al., 2010), [BTPRED](#) (Shepherd et al., 1999), [BetaTPred](#) (Kaur et al., 2002), [BetaTPred2](#) (Kaur et al., 2003), [BetaTurns](#) (Kaur et al., 2004), [MOLEBRNN](#) (Kirschner et al., 2008), [NetTurnP](#) (Petersen et al., 2010). Η πλειοψηφία αυτών χρειάζεται μήκος αλληλουχίας 15-20 κατάλοιπα κατ' ελάχιστο και οι δοκιμές μας με τα υπόλοιπα είχαν την τύχη που τους άξιζε: με βάση το BetaTPred2 βλέπουμε μοτίβο β-στροφής για τα πεπτίδια RWPD και RPWD αλλά όχι με το BetaTurns, παρόλο που το δεύτερο χρησιμοποιεί την πρόβλεψη του πρώτου και απλά κάνει αντιστοίχιση σε συγκεκριμένο τύπο β-στροφής. Οι υπόλοιποι servers είτε έχουν τεθεί εκτός λειτουργίας είτε δεν έδωσαν θετική πρόγνωση.

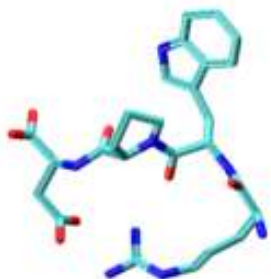
Τέλος στον Πίνακα 3.6 βλέπουμε συγκεντρωτικά τις προτιμήσεις των αμινοξέων προς β-στροφή με βάση την PDB και την μέχρι τώρα ανάλυση που πραγματοποιήσαμε εμείς.

	PDB (Hutchinson et al., 1994, Guruprasad et al., 2000)	Προσομοιώσεις σε 1.440 τετραπεπτίδια (Ενότητα 3.3)	Προσομοιώσεις σε 130 τετραπεπτίδια (Ενότητα 3.4)	Προσομοιώσεις σε 36 τετραπεπτίδια (Ενότητα 3.5)
Συνολικά	G, P, N, D, S, H, T, C, L	S	P	T, V
Πρώτη θέση	D, N, P, C	S, N, C	G, S	G, E
Δεύτερη θέση	P	A, G, M	P, T	T, P, V
Τρίτη θέση	G, N, D	A	I	P, K, R
Τέταρτη θέση	G	K, H, S	P, S	K, W, D

Πίνακας 3.6 Αμινοξέα με υψηλή συχνότητα παρατήρησης σε μοτίβα β-στροφής στην PDB και όπως προσδιορίστηκαν από τις συναρτήσεις εκτίμησης της αναδιπλωσιμότητας μέσω των προσομοιώσεων μοριακής δυναμικής που πραγματοποιήσαμε.

Η διαφοροποίηση των προτιμήσεων είναι αναμενόμενη για δύο λόγους: (1) στις κρυσταλλογραφικά προσδιορισμένες δομές, οι αλληλουχίες βρίσκονται σε ένα ευρύτερο πρωτεϊνικό περιεχόμενο και (2) έχουμε θέσει ισχυρούς περιορισμούς στην αλληλουχία κατά τον αρχικό σχεδιασμό των πεπτιδικών αλληλουχιών (Πίνακας 3.1). Παρόλα αυτά, το γενικό συμπέρασμα που βγαίνει είναι ότι τα αμινοξέα Phe και Tyr με επίσης ογκώδεις πλευρικές ομάδες δεν προτιμώνται, όπως και τα αμινοξέα Leu και Gln, ενώ υπάρχει ισχυρή προτίμηση προς Pro και Gly.

Ολοκληρώνοντας την ανάλυση των τεσσάρων αυτών τετραπεπτιδίων συναρτήσει της θερμοκρασίας με προσομοιώσεις μοριακής δυναμικής διάρκειας 300ns και με το force field CHARMM22 συνοψίζουμε τα ακόλουθα:

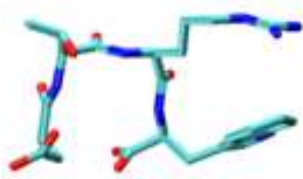


Το RWPD αναδιπλώνεται σε όλες τις θερμοκρασίες σε μία κυρίαρχη διαμόρφωση για τουλάχιστον 40-50% του χρόνου προσομοίωσης. Οι εναλλακτικές διαμορφώσεις που προκύπτουν κατά την ανάλυση Cartesian-PCA αλλά και από τους πίνακες RMSD έχουν παρόμοια διαμόρφωση για τον πεπτιδικό σκελετό και διαφέρουν μόνο στη σχετική διεύθυνση της πλευρικής ομάδας της τρυπτοφάνης. Αυτή η two-state συμπεριφορά που επιδεικνύει το συγκεκριμένο πεπτίδιο μας επιτρέπει να δούμε και τη σχεδόν γραμμική αντίστροφη σχέση μεταξύ θερμοκρασίας και σταθερότητας.

Για το RPWD βλέπουμε συχνά γεγονότα αναδίπλωσης/αποδιάταξης ενώ η άνοδος της θερμοκρασίας επιτρέπει την παρατήρηση επιπλέον διαμορφώσεων. Στο μείγμα των διαμορφώσεων διακρίνεται μία κυρίαρχη δομή (43-59% του χρόνου προσομοίωσης) ενώ οι υπόλοιπες παρατηρούνται παροδικά (<5% του χρόνου της προσομοίωσης).

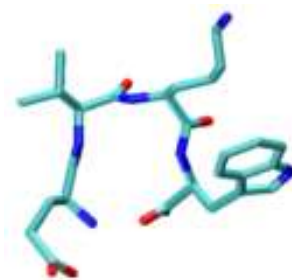


Για το DTRW βλέπουμε επίσης ένα μείγμα διαμορφώσεων με διακριτή εμφάνιση κατά τη διάρκεια της προσομοίωσης. Η κυρίαρχη ομάδα δομών είναι κοινή μεταξύ των θερμοκρασιών με χρόνο εμφάνισης 40-55% αλλά μεγαλύτερη διασπορά και υψηλότερες ατομικές διακυμάνσεις. Τα cluster με μικρότερη συχνότητα εμφάνισης (2-13%) είναι πιο συμπαγή. Η διαμόρφωση του πεπτιδικού σκελετού παραμένει σταθερή σχεδόν για το σύνολο της προσομοίωσης με εξαίρεση εναλλακτικές αλλά ιδιαίτερα παροδικές διαμορφώσεις.



Το EVKW δείχνει τη μεγαλύτερη αστάθεια σε σχέση με τα υπόλοιπα τετραπεπτίδια. Βλέπουμε αρκετές διαμορφώσεις με χρόνο εμφάνισης που κυμαίνεται στα 20-30% για τις κυρίαρχες ομάδες δομών και στα 3-4% για τις υπόλοιπες, με εξαίρεση το τροχιακό στους 320K όπου η κύρια διαμόρφωση εμφανίζεται για ~55% του χρόνου προσομοίωσης.

Έτσι, τα πεπτίδια RWPD και DTRW, που έδειξαν τη μεγαλύτερη σταθερότητα και αποκλίνουν στο μοτίβο της δομής που υιοθετούν, μελετήθηκαν περαιτέρω ως προς τη συμπεριφορά τους με διάφορα force fields.



Το EVKW δείχνει τη μεγαλύτερη αστάθεια σε σχέση με τα υπόλοιπα τετραπεπτίδια. Βλέπουμε αρκετές διαμορφώσεις με χρόνο εμφάνισης που κυμαίνεται στα 20-30% για τις κυρίαρχες ομάδες δομών και στα 3-4% για τις υπόλοιπες, με εξαίρεση το τροχιακό στους 320K όπου η κύρια διαμόρφωση εμφανίζεται για ~55% του χρόνου προσομοίωσης.

Έτσι, τα πεπτίδια RWPD και DTRW, που έδειξαν τη μεγαλύτερη σταθερότητα και αποκλίνουν στο μοτίβο της δομής που υιοθετούν, μελετήθηκαν περαιτέρω ως προς τη συμπεριφορά τους με διάφορα force fields.

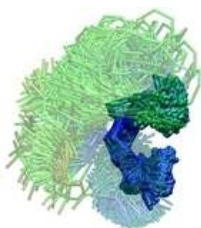
*"C is quirky, flawed, and an enormous success."*

*Dennis M. Ritchie*



*"Knowledge is invariably a matter of degree:  
you cannot put your finger upon even the  
simplest datum and say this we know."*

*T.S. Eliot*



### 3.7 Μελέτη της αναδίπλωσης των RYPD, DTRW με τρία force fields

Η μεγάλη πρόκληση των διαφόρων force fields που αναπτύχθηκαν τα τελευταία χρόνια είναι η ανάπτυξη συναρτήσεων, με εμπειρικές μεθόδους, που να περιγράφουν με ακρίβεια τη δυναμική ενέργεια του συστήματος συναρτήσει των ατομικών συντεταγμένων (MacKerell et al., 1998, MacKerell 2004 [CHARMM family], vanGunsteren et al., 1996 [GROMOS family], Jorgensen et al., 1996 [OPLS family], Cornell et al., 1995 [AMBER family]). Αυτό έχει σαν συνέπεια, η παραμετροποίησή τους να παίζει κυρίαρχο ρόλο στη μετέπειτα ακρίβεια των προβλέψεών τους και να οδηγεί σε αποκλίσεις λόγω συστηματικών προτιμήσεων κατά την περιγραφή των πρωτεϊνικών διαμορφώσεων (Yoda et al., 2004, Best et al., 2008, Matthes et al., 2009, Freddolino et al., 2009, Mittal et al., 2010, Piana et al., 2011, Lindorff-Larsen et al., 2012).

Στις πρώτες απόπειρες σύγκρισης φάνηκε να υπάρχει μία σύγκλιση των αποτελεσμάτων των διαφόρων force fields καθώς ο χρόνος προσομοίωσης την εποχή εκείνη δε μπορούσε να επεκταθεί πέρα από μερικά nanoseconds (Price et al., 2002) ενώ η επιλογή της φυσικής δομής ως αρχικής ήταν καθοριστική για το αποτέλεσμα (Rueda et al., 2007). Οι προσομοιώσεις αναδίπλωσης πεπτιδίων και μίνι-πρωτεϊνών (Ferrara et al., 2000, Snow et al., 2002, Gnanakaran et al., 2003, Ensign et al., 2007, Freddolino et al., 2008, Matthes et al., 2009, Mittal et al., 2010, Shaw et al., 2010), έγιναν προσιτές μόλις τα τελευταία χρόνια με την εξέλιξη της τεχνολογίας των υπολογιστών και των αλγόριθμων (Klepeis et al., 2009, Dror et al., 2012).

Οι διαφορετικές προτιμήσεις των διαφόρων εκδόσεων των force fields έγιναν αντικείμενο εκτεταμένης μελέτης. Συνοπτικά αναφέρουμε ότι τα force fields ff03, ff94 και ff99 της

οικογένειας AMBER δείχνουν προτίμηση προς α-ελικοειδείς διαμορφώσεις (Okur et al., 2002, Yoda et al., 2004, Hornak et al., 2006, Matthes et al., 2009), όπως και οι εκδόσεις CHARMM19, CHARMM22, CHARMM22/CMAP της οικογένειας των CHARMM force fields (Steinbach 2004). Δύο από τα δημοφιλέστερα force fields, το CHARMM27 φαίνεται να υπερ-σταθεροποιεί τις ελικοειδείς διαμορφώσεις (Best et al., 2008, Freddolino et al., 2009) ενώ αντιθέτως οι ελικοειδείς εκτιμήσεις του ff99SB είναι υποτιμημένες (Best et al., 2009). Το GROMOS96 προτιμά περισσότερο εκτεταμένες διαμορφώσεις (Yoda et al., 2004) ενώ τα AMBER ff96, CHARMM22 και OPLS, ενώ επιτυγχάνουν μεγαλύτερη ισορροπία μεταξύ των διαφόρων διαμορφώσεων, παρουσιάζουν μειωμένη συμφωνία με τα πειραματικά δεδομένα. Αυτές οι “προκαταλήψεις” που δείχνουν τα force fields προς συγκεκριμένες περιοχές του διαγράμματος Ramachandran διαφοροποιούνται ανάλογα με το μοντέλο που χρησιμοποιείται για την περιγραφή του διαλύτη (explicit or implicit solvents) (Shell et al., 2008) αλλά και το μήκος του πεπτιδίου (Gnanakaran et al., 2003, 2005). Για το λόγο αυτό συχνά η επιλογή του force field στις προσομοιώσεις αναδίπλωσης γινόταν με γνώμονα τη φυσική δομή ώστε να υπάρχει ένα εγγενές bias προς αυτήν (Freddolino et al., 2010). Η πραγματική αξία ωστόσο ενός force field φαίνεται στη σωστή αναδίπλωση όλων των δομών (α και β) και ιδιαίτερος αυτών προς τις οποίες δε δείχνει προτίμηση βάση της παραμετροποίησής του (Shaw et al., 2010, Best et al., 2010).

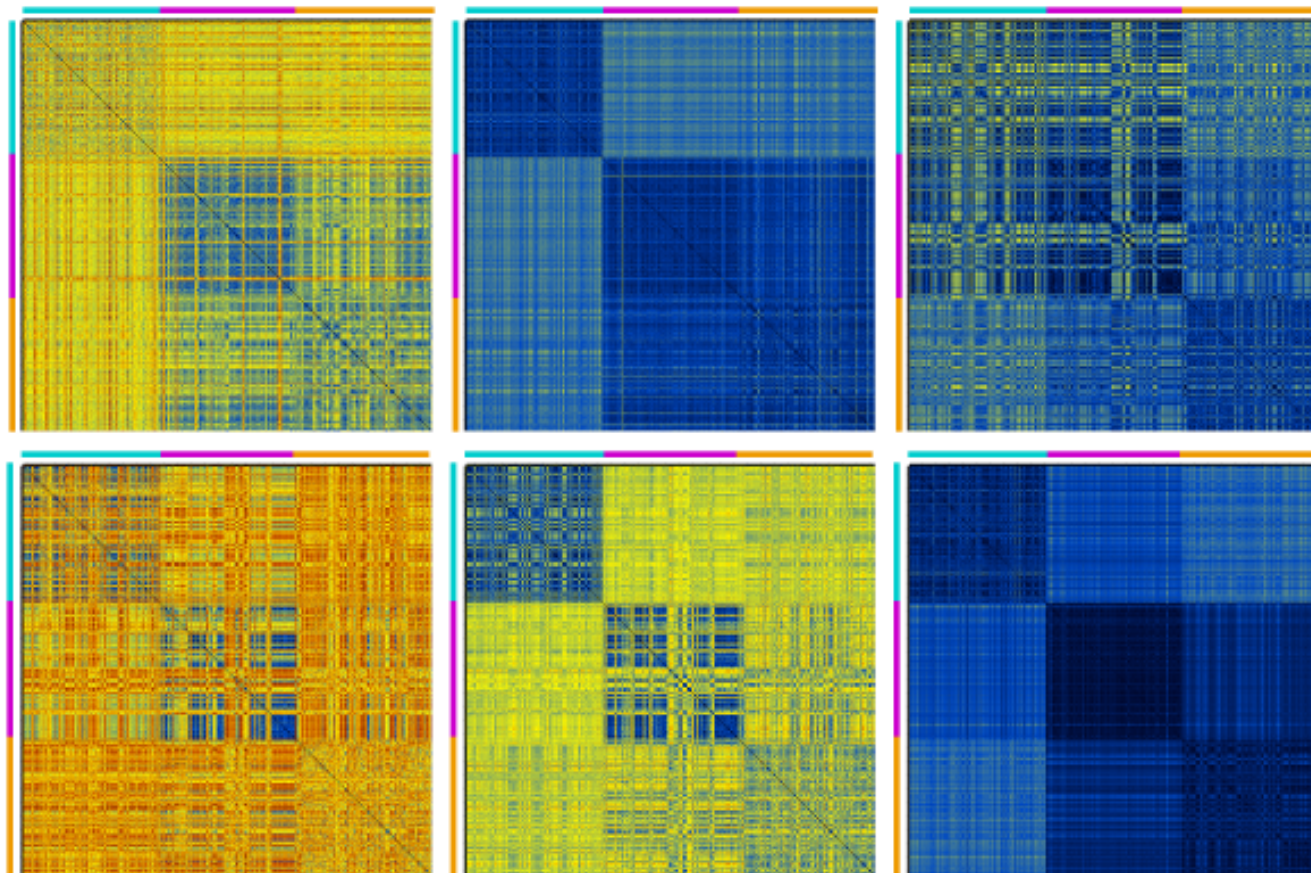
Έτσι, τον τελευταίο καιρό έχουν σημειωθεί προσπάθειες βελτίωσης των δημοφιλέστερων force fields, AMBER (Duan et al., 2003 *ff03*, Hornak et al., 2006 *ff99SB*, Best et al., 2009 *ff99SB\**, *ff03\**, Lindorff-Larsen et al., 2010 *ff99SB-ILDN*, *ff99SB\*-ILDN*), CHARMM (MacKerell et al., 2004 *CHARMM-CMAP*, Buck et al., 2006, Lindorff-Larsen et al., 2010 *CHARMM22\**, Piana et al., 2011), OPLS (Kaminski et al., 2001) και GROMOS96 (Oostenbrink et al., 2004), οι οποίες στοχεύουν συνήθως στη βελτίωση των παραμέτρων του πεπτιδικού σκελετού (torsional potential) και σπανίως των πλευρικών ομάδων (η περίπτωση ILDN) για την ορθότερη αντιπροσώπευση των εκτεταμένων και ελικοειδών διαμορφώσεων. Πάντως, για πολλές πρωτεΐνες, η ελεύθερη ενέργεια αναδίπλωσης (folding free energy) είναι αρκετά μικρή έτσι ώστε μικρές ατέλειες των force fields να ενισχύονται σε μεγάλα πρωτεϊνικά συστήματα με αποτέλεσμα η φυσική δομή να μην αποτελεί το ενεργειακό ελάχιστο του energy landscape όπως αυτό περιγράφεται από το force field (Freddolino et al., 2009, Faver et al., 2011).

Στις μέρες μας, τα force field ff99SB (Matthes et al., 2009, Lange et al., 2010) για μικρότερου μήκους πεπτίδια και τα ff99SB\*-ILDN και CHARMM22\* (Piana et al., 2011, Lindorff-Larsen et

al., 2012) για ένα μεγάλο φάσμα πρωτεϊνών, πεπτιδίων αλλά και δομών, φαίνεται πλέον να δείχνουν τη μεγαλύτερη συμφωνία με τα πειραματικά δεδομένα τόσο σε δομικό (folded structure) και κινητικό επίπεδο (folding rate) όσο και σε θερμοδυναμικό (description of free-energy surface and folding mechanism).

Οι εξελίξεις των force fields, όπως διαφαίνεται από τη βιβλιογραφία (ένα μικρό δείγμα της οποίας παρουσιάσαμε) ήταν ραγδαίες και συσσωρευμένες στα τελευταία χρόνια, ενώ οι αποτελεσματικότερες βελτιώσεις έγιναν μεταξύ 2008-2011 (Lindorff-Larsen et al., 2012, Figure 3). Στην προσπάθειά μας να μείνουμε συγχρονισμένοι πραγματοποιήσαμε προσομοιώσεις αναδίπλωσης των δύο καλύτερων τετραπεπτιδίων, RWPD και DTRW, με τρεις από τις δημοφιλέστερες οικογένειες force fields (AMBER, CHARMM, OPLS), στις καλύτερες εκδόσεις τους (ff99SB, CHARMM-CMAP, OPLS-AA) με βάση την εποχή που πραγματοποιήσαμε εμείς τις προσομοιώσεις (Guvench et al., 2008). Η διάρκεια των προσομοιώσεων ορίστηκε σε 1μs ώστε να εξασφαλίσουμε sufficient sampling και οι υπολογισμοί (συνολικά 6μs) διήρκεσαν περίπου 1 μήνα. Οι λεπτομέρειες των πρωτοκόλλων διεξαγωγής της προσομοίωσης με το εκάστοτε force field, αναλύονται εκτενώς στην Ενότητα 4.5. Στην ανάλυση που ακολουθεί θα πρέπει να έχουμε υπόψιν μας ότι γίνεται σύγκριση μεταξύ των force fields καθώς δε γνωρίζουμε την πειραματικά προσδιορισμένη δομή για τα πεπτίδια αυτά.

Ο πιο άμεσος τρόπος να συγκρίνουμε το σύνολο δομών της προσομοίωσης που προβλέπεται από το κάθε force field είναι μέσω των πινάκων RMSD (Εικόνα 3.54) που υπολογίζονται για το τεχνητό τροχιακό που έχει προκύψει μετά την ένωση των τριών ανεξάρτητων τροχιακών με τα τρία force fields (που για λόγους συντομίας θα αναφέρονται συνοπτικά ως AMBER, CHARMM, OPLS). Η ποικιλομορφία των αποτελεσμάτων είναι έκδηλη: για το πεπτίδιο RWPD το CHARMM βρίσκεται πιο κοντά στο OPLS όταν επικεντρωθούμε στη συμπεριφορά όλων των βαριών ατόμων ή μόνο του πεπτιδικού σκελετού, αλλά αν εστιάσουμε μόνο στα τέσσερα Ca άτομα τότε βλέπουμε μεγαλύτερη ομοιότητα μεταξύ AMBER και CHARMM και σε μικρότερο βαθμό με το OPLS. Σε κάθε περίπτωση, τις πιο σταθερές δομές για το RWPD τις παρατηρούμε με το CHARMM. Για το πεπτίδιο DTRW βλέπουμε μεγάλη ασυμφωνία μεταξύ AMBER, CHARMM, OPLS με το CHARMM να δίνει και πάλι τις σταθερότερες δομές ενώ σύμφωνα με το OPLS το πεπτίδιο αυτό είναι τελείως ασταθές. Αν αφαιρέσουμε τις πλευρικές ομάδες, βλέπουμε ότι ο πεπτιδικός σκελετός παίρνει σταθερή διαμόρφωση, η οποία όμως διαφέρει σημαντικά ανάμεσα στα force fields, ενώ ομοιότητες εντοπίζονται μόνο σε επίπεδο ατόμων Ca (CHARMM/OPLS).



Εικόνα 3.54 Γραφική απεικόνιση του ενιαίου πίνακα RMSD μετά την ένωση των τροχιακών των 3 force fields για τα πεπτίδια RWPD (πάνω) και DTRW (κάτω). Από αριστερά προς τα δεξιά, ο υπολογισμός έγινε με όλα τα βαριά άτομα (left, all-heavy atoms), με τα άτομα του πεπτιδικού σκελετού (middle, backbone atoms) και μόνο με τα άτομα Ca (right, Ca atoms). Οι οριζόντιες και κάθετες χρωματιστές μπάρες οριοθετούν τα τρία ανεξάρτητα τροχιακά, όπου με γαλάζιο (cyan) απεικονίζεται το τροχιακό με το **AMBER**, με μωβ (magenta) το τροχιακό με το **CHARMM** και με πορτοκαλί (orange) το τροχιακό με το **OPLS**. Η χρωματική κλίμακα κυμαίνεται από σκούρο μπλε (0Å) έως σκούρο κόκκινο (5.98Å).

Για να εξετάσουμε περαιτέρω τις παρατηρήσεις αυτές, οι οποίες στηρίζονται στο RMSD, πραγματοποιήσαμε και τους δύο τύπους PCA, Cartesian-PCA (all-heavy atoms) και Dihedral-PCA που εξετάζουν τη συμπεριφορά ολόκληρου του πεπτιδίου και του πεπτιδικού σκελετού μεμονωμένα, αντιστοίχως. Από την cluster analysis που προκύπτει, προσδιορίσαμε αντιπροσωπευτικές δομές για κάθε cluster (δηλαδή τη δομή του cluster που βρίσκεται πλησιέστερα στην υπολογιζόμενη μέση δομή). Προκειμένου να εξετάσουμε την ομοιότητα μεταξύ των αντιπροσωπευτικών δομών υπολογίσαμε RMSD μεταξύ κάθε πιθανού ζεύγους.

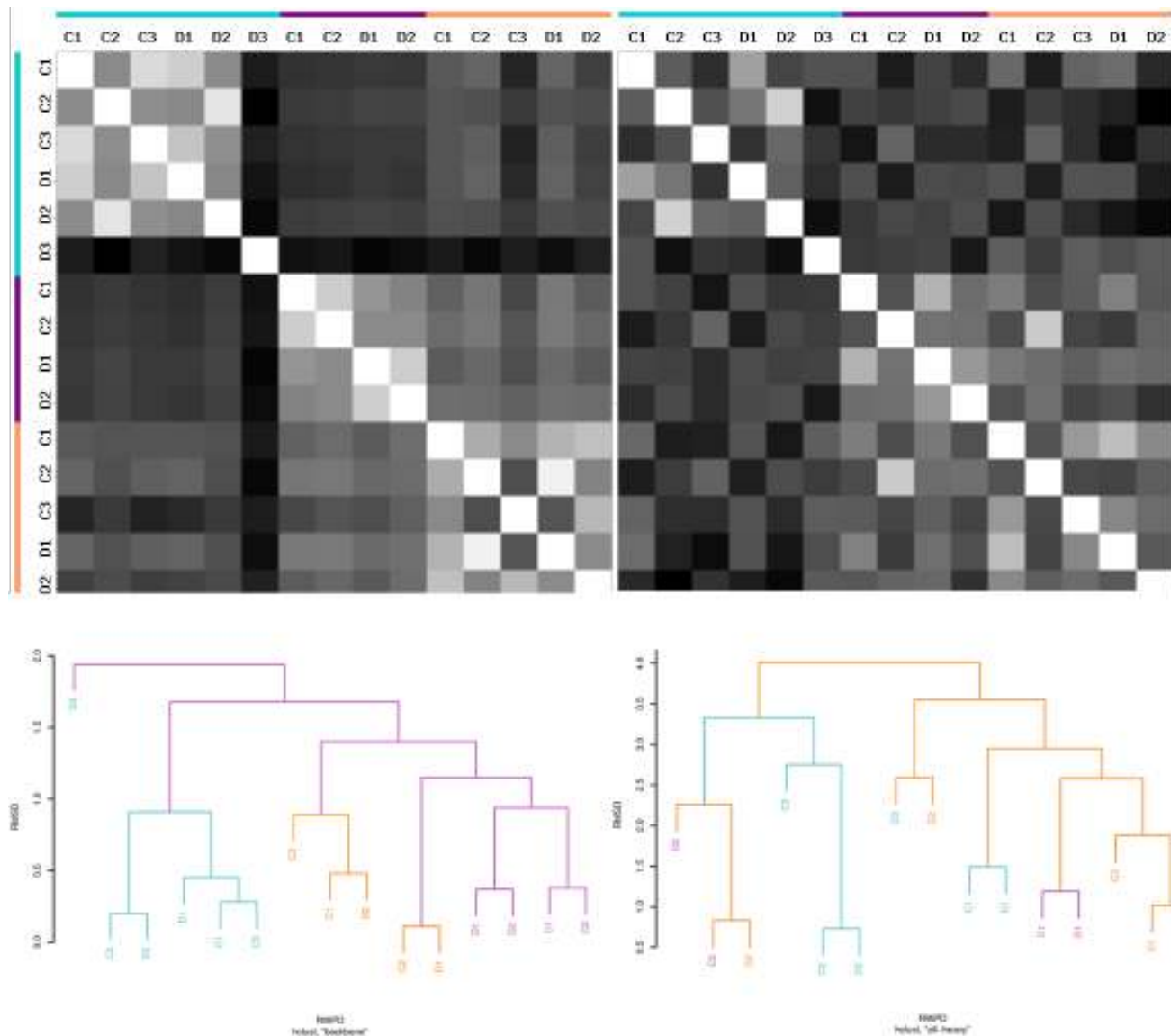
	C1	C2	C3	D1	D2	D3	C1	C2	D1	D2	C1	C2	C3	D1	D2
C1	0.0	0.89	0.28	0.37	0.88	1.73	1.56	1.54	1.51	1.52	1.26	1.17	1.66	1.17	1.45
C2		0.0	0.86	0.90	0.20	1.94	1.50	1.48	1.42	1.43	1.30	1.33	1.50	1.32	1.36
C3			0.0	0.45	0.85	1.69	1.55	1.52	1.51	1.51	1.29	1.22	1.68	1.22	1.47
D1				0.0	0.91	1.79	1.59	1.57	1.50	1.54	1.29	1.19	1.63	1.19	1.44
D2					0.0	1.88	1.48	1.46	1.42	1.45	1.32	1.35	1.51	1.33	1.38
D3						0.0	1.81	1.77	1.90	1.85	1.74	1.88	1.72	1.83	1.69
C1							0.0	0.38	0.81	0.94	1.20	1.04	1.40	1.02	1.24
C2								0.0	0.88	0.89	1.12	1.03	1.29	1.01	1.15
D1									0.0	0.37	1.25	1.15	1.35	1.13	1.26
D2										0.0	1.10	1.12	1.21	1.09	1.11
C1											0.0	0.63	0.89	0.58	0.48
C2												0.0	1.35	0.11	0.94
C3													0.0	1.29	0.54
D1														0.0	0.88
D2															0.0

Πίνακας 3.7. Τιμές RMSD μεταξύ όλων των αντιπροσωπευτικών δομών όλων των cluster όπως προκύπτουν από την ανάλυση Cartesian-PCA (με το γράμμα 'C') και Dihedral-PCA (με το γράμμα 'D') για το πεπτίδιο RWPD. Ο υπολογισμός αφορά πάντα μόνο τα άτομα του πεπτιδικού σκελετού (15 άτομα). Ο χρωματικός κώδικας για τα force fields διατηρείται κοινός.

	C1	C2	C3	D1	D2	D3	C1	C2	D1	D2	C1	C2	C3	D1	D2
C1	0.0	2.56	3.28	1.49	2.92	2.72	2.73	3.57	2.95	3.32	2.38	3.56	2.45	2.31	3.36
C2		0.0	2.75	2.17	0.73	3.76	2.99	3.14	2.98	2.84	3.55	3.06	3.28	3.49	4.01
C3			0.0	3.22	2.37	3.18	3.68	2.43	3.33	3.33	3.51	2.47	3.31	3.82	3.22
D1				0.0	2.48	3.30	2.76	3.57	2.81	2.87	2.71	3.52	2.71	2.72	3.55
D2					0.0	3.79	3.14	2.88	2.98	2.78	3.64	2.80	3.35	3.66	3.88
D3						0.0	3.11	3.04	2.96	3.62	2.53	3.06	2.55	2.78	2.59
C1							0.0	2.73	1.19	2.30	2.04	2.81	2.58	1.98	2.63
C2								0.0	2.21	2.26	2.80	0.83	2.93	3.08	2.46
D1									0.0	1.62	2.10	2.30	2.52	2.27	2.37
D2										0.0	2.73	2.25	2.94	2.75	3.24
C1											0.0	2.71	1.60	1.02	1.86
C2												0.0	2.87	2.94	2.53
C3													0.0	1.88	2.31
D1														0.0	2.61
D2															0.0

Πίνακας 3.8. Τιμές RMSD μεταξύ όλων των αντιπροσωπευτικών δομών όλων των cluster όπως προκύπτουν από την ανάλυση Cartesian-PCA (με το γράμμα 'C') και Dihedral-PCA (με το γράμμα 'D') (για λόγους πληρότητας) για το πεπτίδιο RWPD. Ο υπολογισμός αφορά πάντα όλα τα βαριά άτομα (41 άτομα). Ο χρωματικός κώδικας για τα force fields διατηρείται κοινός.





Εικόνα 3.55. Γραφική απεικόνιση των πινάκων RMSD 3.7-3.8 των αντιπροσωπευτικών δομών για το πεπτίδιο RWPD και τα αντίστοιχα δενδρογράμματα. Η χρωματική κλίμακα κυμαίνεται από λευκό (0Å) σε μαύρο (1.94Å) για τα άτομα του πεπτιδικού σκελετού (αριστερά) και 4.0Å για όλα τα βαριά άτομα (δεξιά). Ο χρωματικός κώδικας για τα force fields διατηρείται κοινός.

Τα αποτελέσματα συνοψίζονται στους Πίνακες 3.7-3.10 ενώ η ομαδοποίηση των δομών που προκύπτει απεικονίζεται στις Εικόνες 3.55 και 3.60 με τους ασπρόμαυρους πίνακες (με όψη “σκακιέρας”) και τα αντίστοιχα δενδρογράμματα. Η δημιουργία διακριτών cluster δομών για το

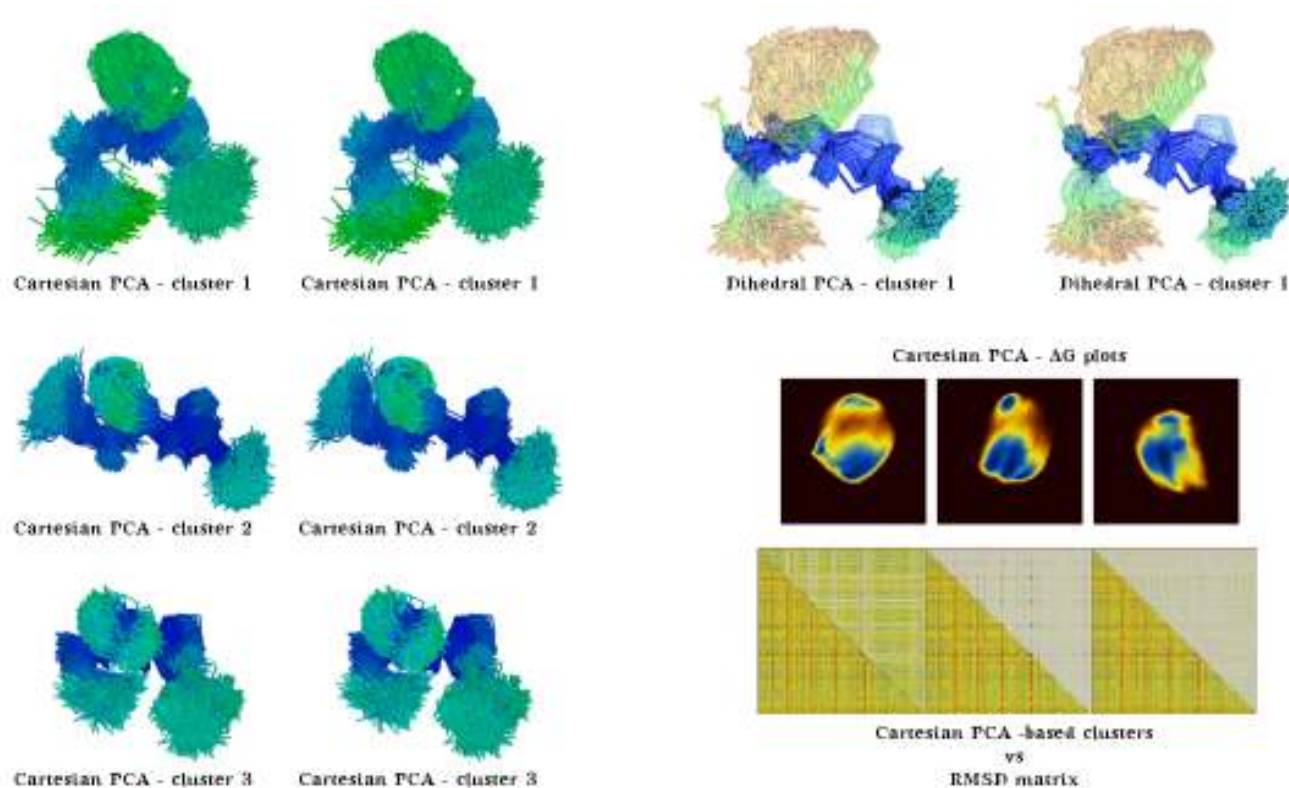


πεπτίδιο RWPD με κάθε ένα από τα force fields απεικονίζεται καθαρά στα δενδρογράμματα, ενώ οι ασπρόμαυρες αποχρώσεις της “σκακιέρας” μας δίνουν πληροφορία για το πόσο όμοιες ή ανόμοιες είναι οι δομές, σε επίπεδο RMSD. Η διαμόρφωση του πεπτιδικού σκελετού που προβλέπεται από τα CHARMM και OPLS είναι σαφώς πιο κοντινή. Η εικόνα περιπλέκεται αρκετά όταν αρχίζουμε να λαμβάνουμε υπόψιν και τις πλευρικές ομάδες. Το πιο ξεκάθαρο μήνυμα της Εικόνας 3.55 είναι ότι βλέπουμε πάντα τις δομές 'C' κοντά στις αντίστοιχες 'D' του ίδιου cluster γεγονός που δείχνει ότι η διαμόρφωση του σκελετού παίζει πρωτεύοντα δομικό ρόλο. Μία άλλη σημαντική παρατήρηση είναι πως βλέπουμε κοντά τις δομές που ανήκουν σε cluster με τον ίδιο αύξοντα αριθμό μεταξύ των force fields, δηλαδή προβλέπουν την ίδια σειρά για τα cluster, βάσει της κατοχής τους σε χρόνο προσομοίωσης.

Αυτό γίνεται εμφανές και από τις Εικόνες 3.56-3.58 που ακολουθούν, όπου παραθέτουμε συγκεντρωτικά αποτελέσματα από την ανάλυση των τροχιακών των τριών force fields. Τα αποτελέσματα αυτά καθιστούν σαφή τα ακόλουθα: (1) Το CHARMM φαίνεται να δίνει τις πιο σταθερές δομές και με τη μεγαλύτερη κατοχή σε χρόνο προσομοίωσης, ακολουθούμενο από το OPLS και μετά το AMBER. (2) Υπάρχει συμφωνία στις δομές του κυρίαρχου cluster που προβλέπονται από το CHARMM και το OPLS αλλά όχι και με το AMBER, ενώ το δεύτερο σε σειρά cluster είναι κοινό. (3) Η κύρια διαμόρφωση του πεπτιδικού σκελετού είναι παρόμοια μεταξύ των τριών force fields (Dihedral-PCA).

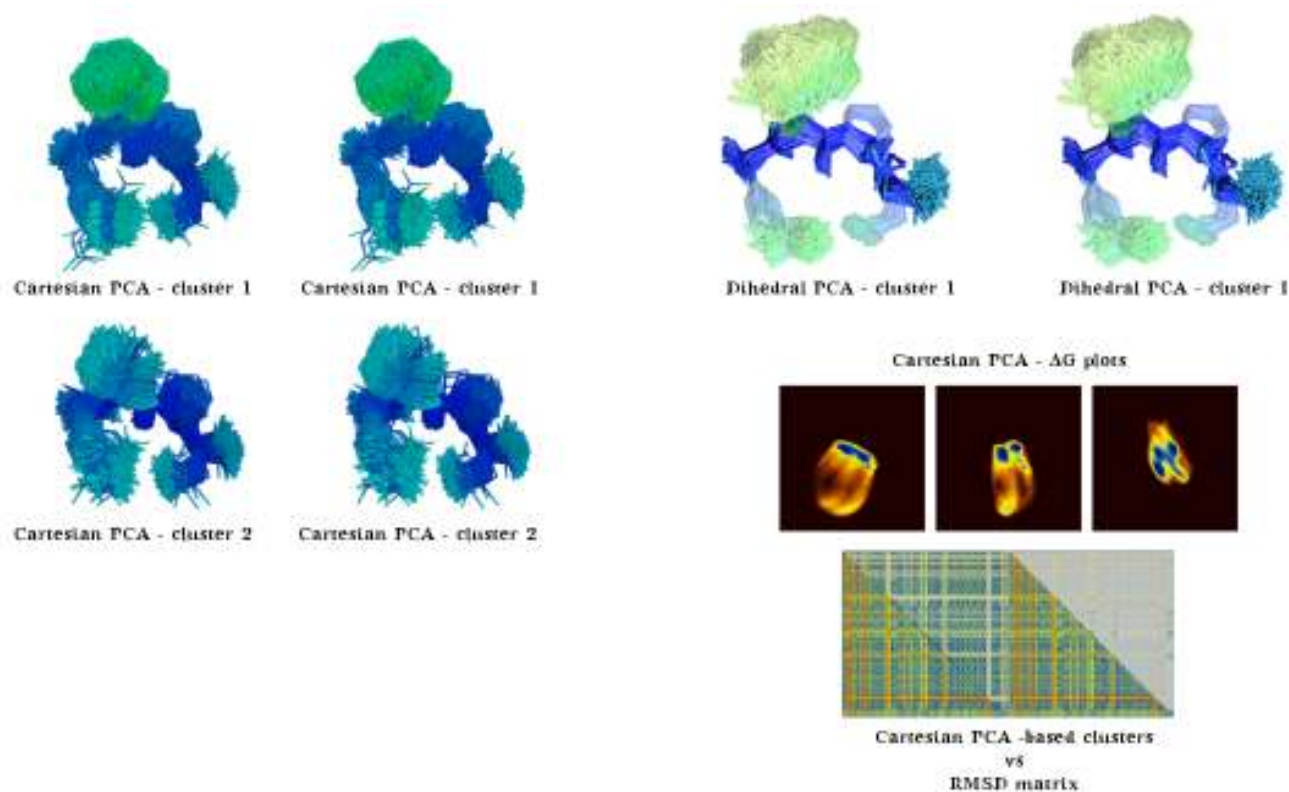
Τέλος, αν θα θέλαμε να συγκρίνουμε τις δομές των τριών αυτών force fields με τα αποτελέσματα από το CHARMM22 που αναλύσαμε στην προηγούμενη ενότητα (Ενότητα 3.6), έχοντας υπόψιν ότι ο χρόνος προσομοίωσης διαφέρει σημαντικά (1000ns/100ns), βλέπουμε ότι:

- ♦ η κυρίαρχη δομή του **AMBER** είναι πιο κοντά στις δομές του cluster 1 των 283K και 298K και κατά συνέπεια στις δομές του cluster 2 των 320K και 340K (με all-heavy RMSD ~ 1.9Å).
- ♦ η κυρίαρχη δομή του **CHARMM** είναι πιο κοντά στις δομές του cluster 2 των 283K και 298K και κατά συνέπεια στις δομές του cluster 1 των 320K και 340K (με all-heavy RMSD ~ 1.1Å).
- ♦ η κυρίαρχη δομή του **OPLS** είναι πιο κοντά στις δομές του cluster 1 των 283K και 298K και κατά συνέπεια στις δομές του cluster 2 των 320K και 340K (με all-heavy RMSD ~ 1.3Å).
- ♦ Επίσης η αντιπροσωπευτική δομή του **AMBER** έχει RMSD 1.7Å από την αντιπροσωπευτική του **CHARMM** και 1.8Å από την αντιπροσωπευτική του **OPLS** και η αντιπροσωπευτική δομή του **CHARMM** έχει RMSD 1.0Å από την αντιπροσωπευτική του **OPLS**.



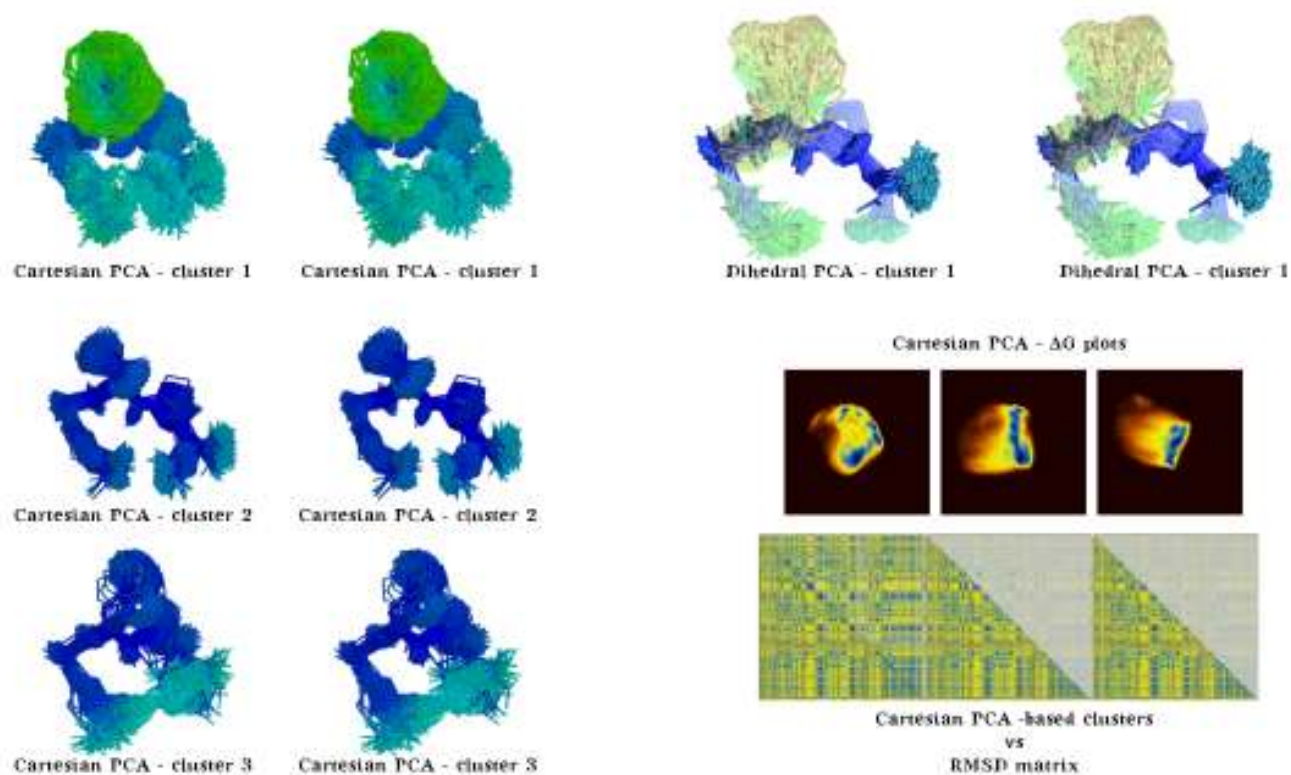
Εικόνα 3.56 *Αριστερά*: υπέρθεση (χρησιμοποιώντας όλα τα βαριά άτομα) αντιπροσωπευτικών δομών από τα κυρίαρχα cluster (40%, 6%, 5%) που προέκυψαν με βάση την ανάλυση Cartesian-PCA για το τροχιακό **AMBER** και το πεπτίδιο RWPD. *Δεξιά-πάνω*: υπέρθεση (χρησιμοποιώντας τα άτομα του πεπτιδικού σκελετού) αντιπροσωπευτικών δομών από το κυρίαρχο cluster (54%) που προέκυψε με βάση την ανάλυση Dihedral-PCA για το τροχιακό **AMBER** και το πεπτίδιο RWPD. Ο χρωματισμός των δομών έγινε με βάση τις τιμές RMSF σε κοινή κλίμακα από μπλε (0.15Å) σε κόκκινο (5.55Å), ενώ για λόγους ευκρίνειας χρησιμοποιείται διαφάνεια για τις πλευρικές ομάδες στις δομές του Dihedral-PCA cluster. *Δεξιά-κάτω*: ενεργειακά τοπία ( $\Delta G$  energy plots) της προβολής του τροχιακού στο επίπεδο των 1-2, 1-3, 2-3 κυρίαρχων συνιστωσών (principal components) της ανάλυσης Cartesian-PCA. Από κάτω βλέπουμε την προβολή των Cartesian-PCA clusters πάνω στον πίνακα RMSD. Κάτω από τη διαγώνιο διατηρείται ο πίνακας RMSD ενώ πάνω από τη διαγώνιο βλέπουμε έντονα μόνο τις δομές που ανήκουν στο εκάστοτε cluster.

Βλέπουμε δηλαδή, πως οι διαφορές μεταξύ των force fields δεν εστιάζονται στις δομές που περιγράφουν αλλά στη γενικότερη περιγραφή του ενεργειακού τοπίου του πεπτιδίου: βλέπουμε τις ίδιες δομές, αλλά για διαφορετικό χρόνο και με διαφορετική σειρά. Αυτό καθίσταται εμφανές στην Εικόνα 3.59 όπου βλέπουμε τις αντιπροσωπευτικές δομές (τα στιγμιότυπα που είναι



Εικόνα 3.57 *Αριστερά*: υπέρθεση (χρησιμοποιώντας όλα τα βαριά άτομα) αντιπροσωπευτικών δομών από τα κυρίαρχα cluster (63%, 4%) που προέκυψαν με βάση την ανάλυση Cartesian-PCA για το τροχιακό **CHARMM** και το πεπτίδιο RWPD. *Δεξιά-πάνω*: υπέρθεση (χρησιμοποιώντας τα άτομα του πεπτιδικού σκελετού) αντιπροσωπευτικών δομών από το κυρίαρχο cluster (59%) που προέκυψε με βάση την ανάλυση Dihedral-PCA για το τροχιακό **CHARMM** και το πεπτίδιο RWPD. Ο χρωματισμός των δομών έγινε με βάση τις τιμές RMSF σε κοινή κλίμακα από μπλε (0.15Å) σε κόκκινο (5.55Å), ενώ για λόγους ευκρίνειας χρησιμοποιείται διαφάνεια για τις πλευρικές ομάδες στις δομές του Dihedral-PCA cluster. *Δεξιά-κάτω*: ενεργειακά τοπία (ΔG energy plots) της προβολής του τροχιακού στο επίπεδο των 1-2, 1-3, 2-3 κυρίαρχων συνιστωσών (principal components) της ανάλυσης Cartesian-PCA. Από κάτω βλέπουμε την προβολή των Cartesian-PCA clusters πάνω στον πίνακα RMSD. Κάτω από τη διαγώνιο διατηρείται ο πίνακας RMSD ενώ πάνω από τη διαγώνιο βλέπουμε έντονα μόνο τις δομές που ανήκουν στο εκάστοτε cluster.

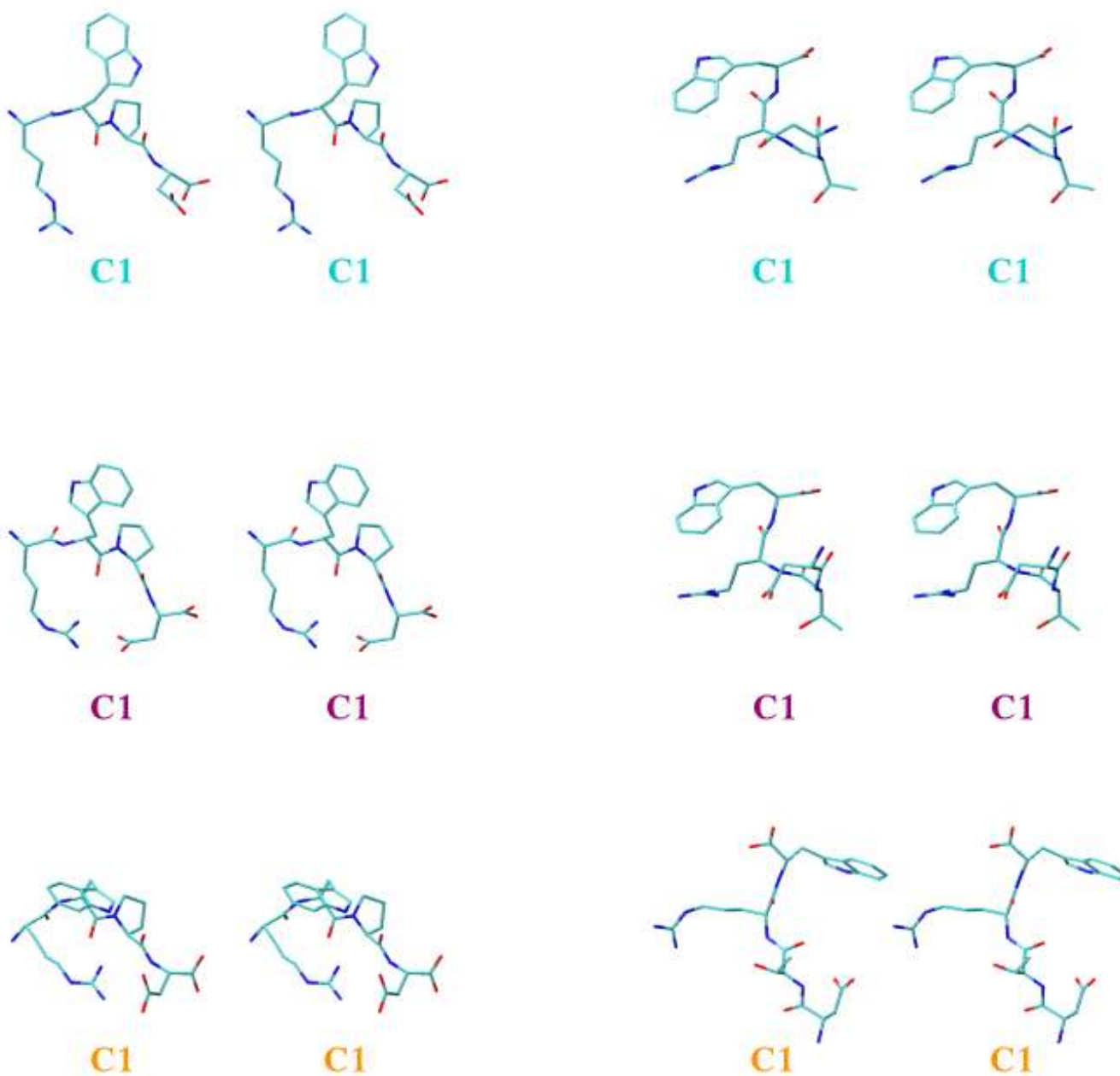
πλησιέστερα στην υπολογιζόμενη μέση δομή) για τα κυρίαρχα cluster που προβλέπονται για τα πεπτίδια από τα τρία force fields, AMBER, CHARMM, OPLS (σύγκριση με Εικόνα 3.26, σελ. 117/118).



Εικόνα 3.58 *Αριστερά*: υπέρθεση (χρησιμοποιώντας όλα τα βαριά άτομα) αντιπροσωπευτικών δομών από τα κυρίαρχα cluster (53%, 0.3%, 0.2%) που προέκυψαν με βάση την ανάλυση Cartesian-PCA για το τροχιακό **OPLS** και το πεπτίδιο RWPD. *Δεξιά-πάνω*: υπέρθεση (χρησιμοποιώντας τα άτομα του πεπτιδικού σκελετού) αντιπροσωπευτικών δομών από το κυρίαρχο cluster (32%) που προέκυψε με βάση την ανάλυση Dihedral-PCA για το τροχιακό **OPLS** και το πεπτίδιο RWPD. Ο χρωματισμός των δομών έγινε με βάση τις τιμές RMSF σε κοινή κλίμακα από μπλε ( $0.15\text{\AA}$ ) σε κόκκινο ( $5.55\text{\AA}$ ), ενώ για λόγους ευκρίνειας χρησιμοποιείται διαφάνεια για τις πλευρικές ομάδες στις δομές του Dihedral-PCA cluster. *Δεξιά-κάτω*: ενεργειακά τοπία ( $\Delta G$  energy plots) της προβολής του τροχιακού στο επίπεδο των 1-2, 1-3, 2-3 κυρίαρχων συνιστωσών (principal components) της ανάλυσης Cartesian-PCA. Από κάτω βλέπουμε την προβολή των Cartesian-PCA clusters πάνω στον πίνακα RMSD. Κάτω από τη διαγώνιο διατηρείται ο πίνακας RMSD ενώ πάνω από τη διαγώνιο βλέπουμε έντονα μόνο τις δομές που ανήκουν στο εκάστοτε cluster.

Στη συνέχεια μελετήσαμε το πεπτίδιο DTRW με τα τρία αυτά force fields. Το πεπτίδιο αυτό εμφανίζεται τελείως ασταθές με το OPLS ενώ στοιχειώδης δομή προβλέπεται μόνο από το CHARMM.





Εικόνα 3.59 Αντιπροσωπευτικές δομές (σε stereo αναπαράσταση) από τα κυρίαρχα cluster των τροχιακών των δύο τετραπεπτιδίων RWPD (αριστερά) και DTRW (δεξιά) για τα force fields AMBER (πάνω), CHARMM (μέση) και OPLS (κάτω). Ο χρωματικός κώδικας των τροχιακών διατηρείται ίδιος με της Εικόνας 3.56

Τα αποτελέσματα αυτά δε διαφοροποιούνται σημαντικά εάν αγνοήσουμε τις πλευρικές ομάδες. Η διάκριση του AMBER ως προς τα υπόλοιπα διαφαίνεται τόσο στο επίπεδο των πλευρικών ομάδων όσο και στο επίπεδο του πεπτιδικού σκελετού.

Πιο συγκεκριμένα, ακολουθήσαμε στο πεπτίδιο DTRW την ίδια πορεία με το RWPD για τη σύγκριση των αντιπροσωπευτικών δομών όπως αυτές προκύπτουν από cluster analysis μέσω Cartesian-PCA και Dihedral-PCA (Πίνακες 3.9 – 3.10 και Εικόνες 3.60 – 3.63). Εάν θεωρήσουμε μόνο τα άτομα του πεπτιδικού σκελετού, βλέπουμε πως η δομή C1 του AMBER είναι πιο κοντά στη δομή C1 του CHARMM. Από την άλλη, οι δομές C1 και C2 του OPLS είναι πιο κοντά στις δομές C2 και C3 του AMBER και C2 του CHARMM. Αυτό σημαίνει πως το κυρίαρχο (αλλά ασταθές και διάσπαρτο) cluster του OPLS ταιριάζει καλύτερα με τα μικρά cluster των άλλων 2 force fields. Οι παρατηρήσεις αυτές ανάγονται και στην περίπτωση που θεωρήσουμε όλα τα βαριά άτομα. Παρατηρούμε με συνέπεια καλύτερη συμφωνία μεταξύ AMBER και CHARMM, τουλάχιστον για το πολυπληθέστερο cluster δομών ενώ για τα μικρότερα cluster βλέπουμε συμφωνία μεταξύ του OPLS και του CHARMM, και λιγότερο του AMBER.

Για την οπτικοποίηση των παραπάνω παρατηρήσεων παραθέτουμε στις Εικόνες 3.61 – 3.63 συνοπτικά αποτελέσματα από την ανάλυση των τριών τροχιακών με τα τρία force fields. Τα συμπεράσματα συνοψίζονται στα ακόλουθα: (1) Το CHARMM δίνει τις πιο σταθερές δομές ακολουθούμενο από το AMBER και τέλος το OPLS. (2) Υπάρχει συμφωνία στις προβλεπόμενες δομές του κυρίαρχου cluster μεταξύ AMBER και CHARMM αλλά όχι του OPLS, ενώ υπάρχει μία σύγκλιση στα cluster με μικρές κατοχές σε χρόνο προσομοίωσης. (3) Η κύρια διαμόρφωση του πεπτιδικού σκελετού (σχήματος 'U') είναι παρόμοια μεταξύ και των τριών force fields.

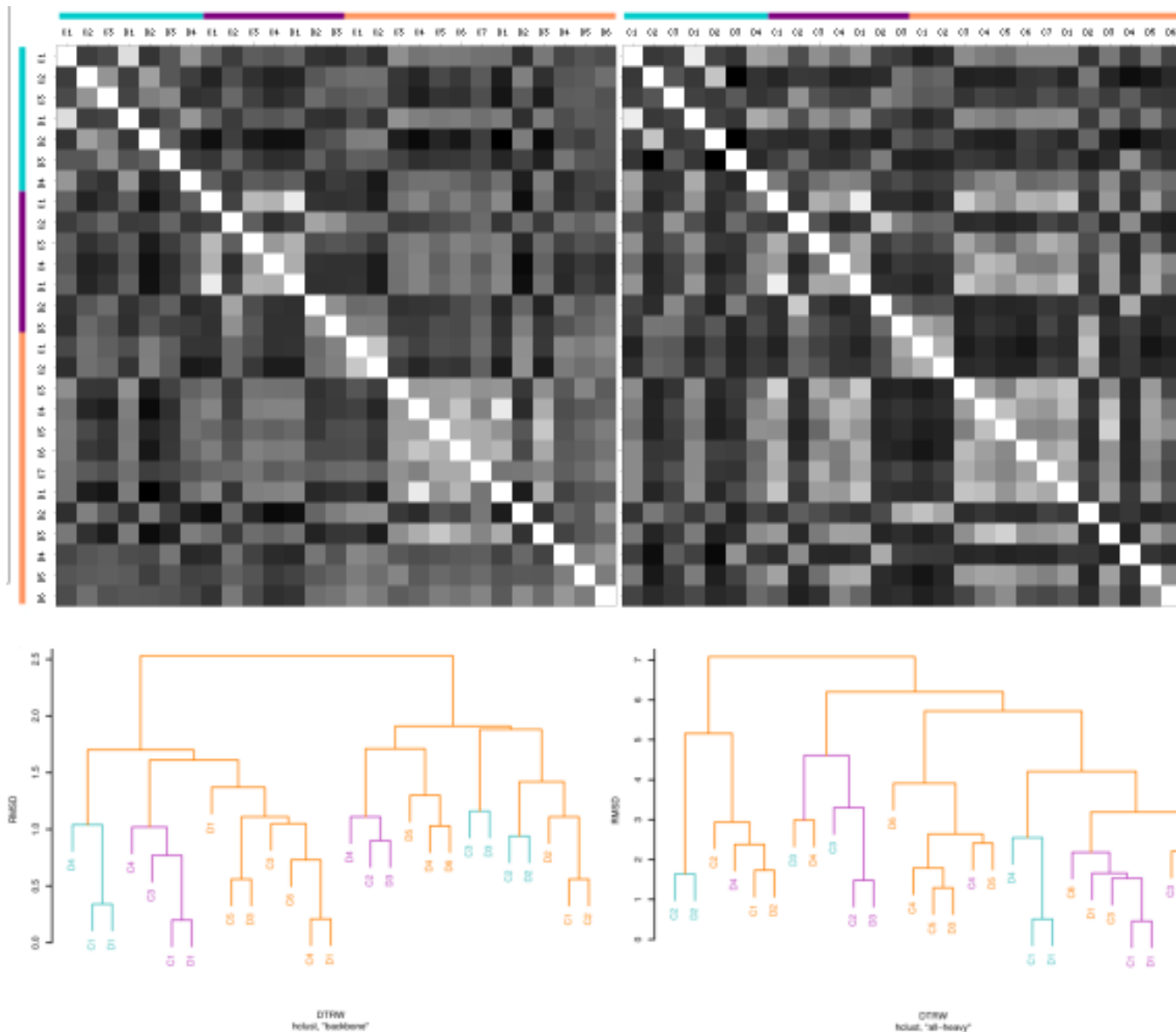
Τέλος, αν θα θέλαμε να συγκρίνουμε τις δομές των τριών αυτών force fields (Εικόνα 3.59 versus Εικόνα 3.38, σελ. 128/129) με τα αποτελέσματα από το CHARMM22 που αναλύσαμε στην προηγούμενη ενότητα (3.6), έχοντας υπόψιν ότι ο χρόνος προσομοίωσης διαφέρει σημαντικά (1000ns/100ns), βλέπουμε ότι:

♦ η κυρίαρχη δομή του **AMBER** είναι πιο κοντά στις δομές του cluster 1 των 283K (με all-heavy RMSD  $\sim 3.4\text{\AA}$ ), του cluster 2 των 298K (με all-heavy RMSD  $\sim 2.6\text{\AA}$ ) και του cluster 2 των 320K (με all-heavy RMSD  $\sim 2.6\text{\AA}$ ) και 340K (με all-heavy RMSD  $\sim 2.8\text{\AA}$ ).

♦ η κυρίαρχη δομή του **CHARMM** είναι πιο κοντά στις δομές του cluster 1 των 283K (με all-heavy RMSD  $\sim 2.4\text{\AA}$ ), του cluster 2 των 298K (με all-heavy RMSD  $\sim 0.8\text{\AA}$ ) και του cluster 2 των 320K (με all-heavy RMSD  $\sim 0.7\text{\AA}$ ) και 340K (με all-heavy RMSD  $\sim 1.2\text{\AA}$ ).



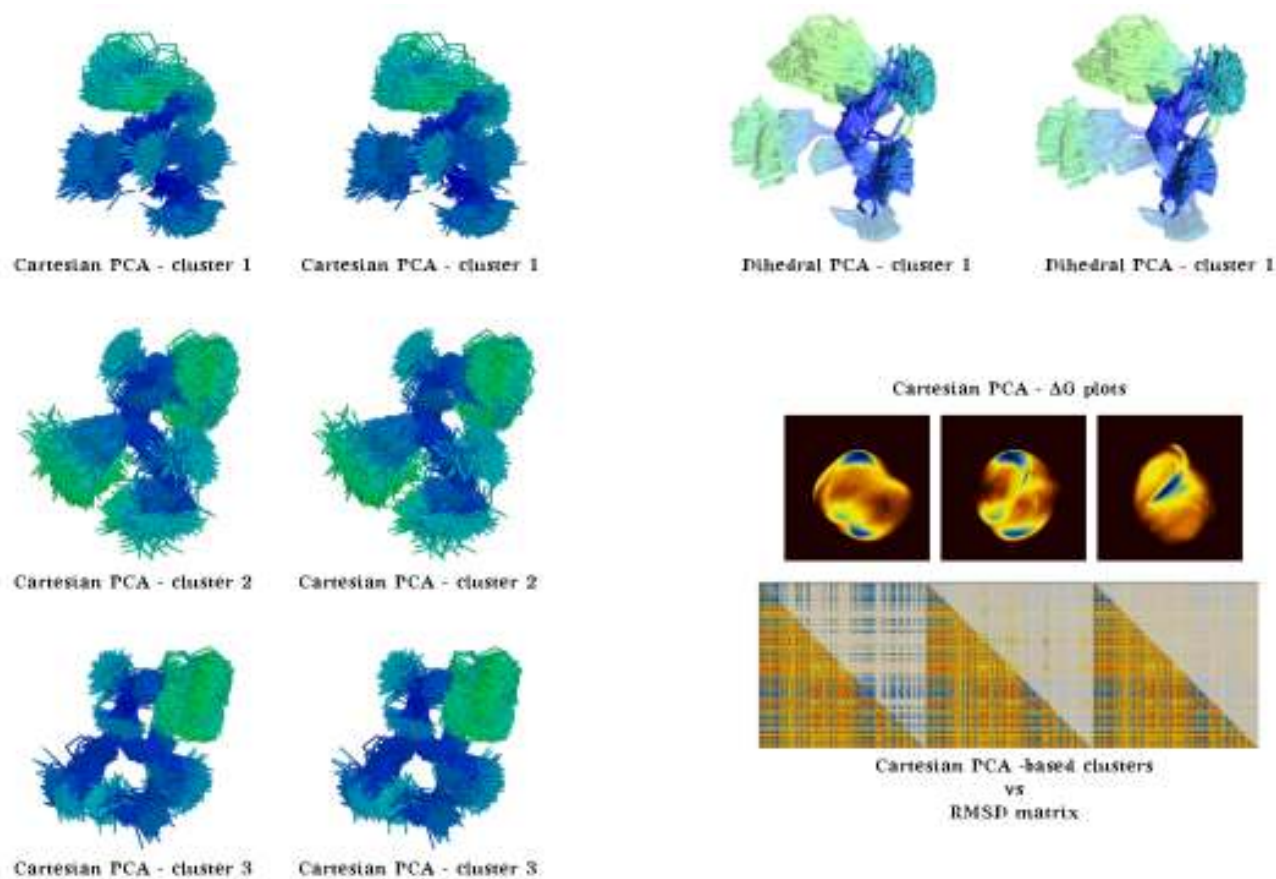




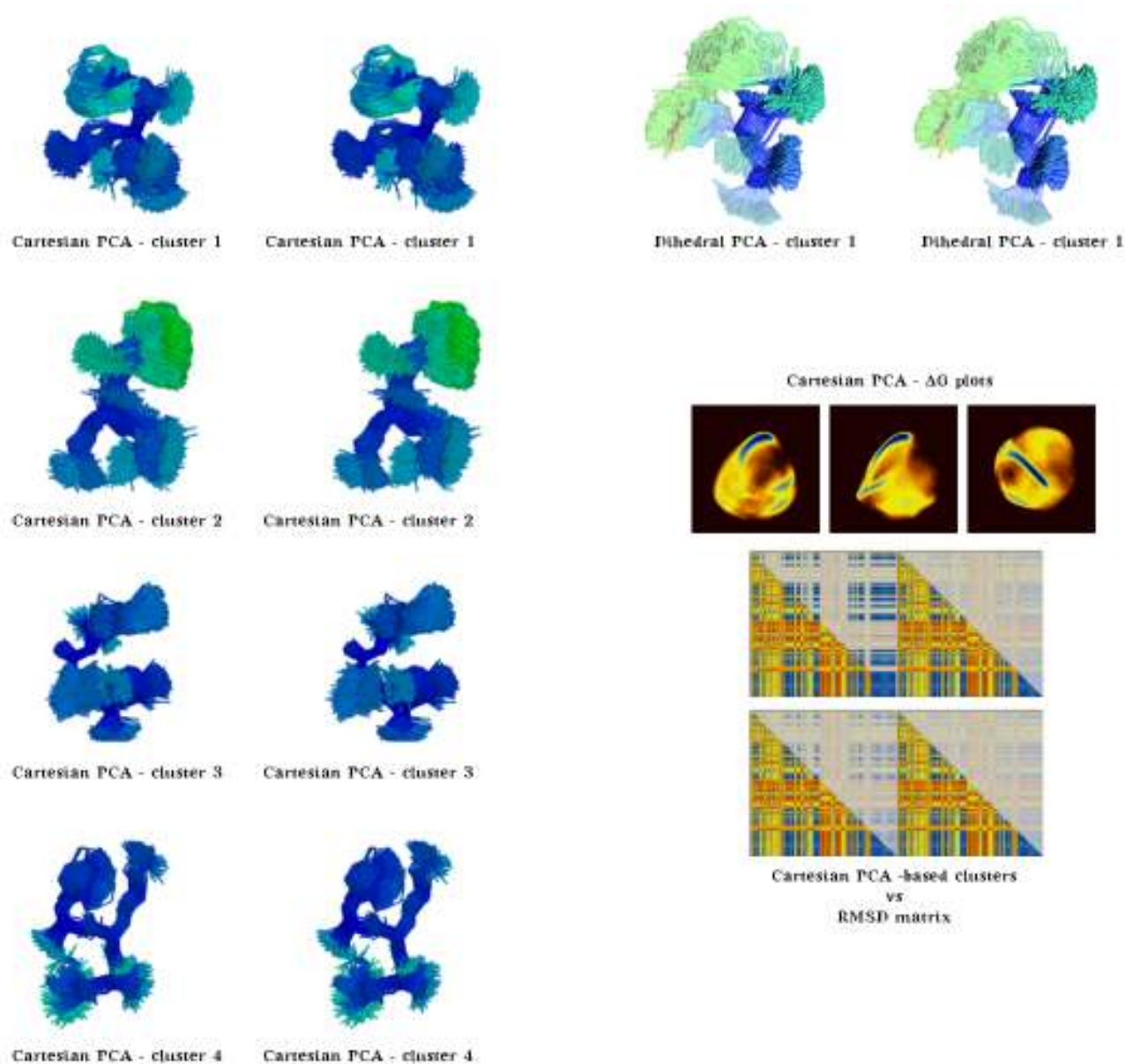
Εικόνα 3.60 Γραφική απεικόνιση των πινάκων RMSD 3.9-3.10 των αντιπροσωπευτικών δομών για το πεπτίδιο DTRW και τα αντίστοιχα δενδρογράμματα. Η χρωματική κλίμακα κυμαίνεται από λευκό (0Å) σε μαύρο (2.53Å για τα άτομα του πεπτιδικού σκελετού (αριστερά) και 7.0Å για όλα τα βαριά άτομα (δεξιά)).

Ο χρωματικός κώδικας για τα force fields διατηρείται κοινός.

♦ η κυρίαρχη δομή του **OPLS** είναι πιο κοντά στις δομές του cluster 2 των 283K (με all-heavy RMSD ~3.4Å), του cluster 3 των 298K (με all-heavy RMSD ~3.4Å) και του cluster 1 των 320K (με all-heavy RMSD ~3.6Å) και 340K (με all-heavy RMSD ~3.7Å).



Εικόνα 3.61 *Αριστερά*: υπέρθεση (χρησιμοποιώντας όλα τα βαριά άτομα) αντιπροσωπευτικών δομών από τα κυρίαρχα cluster (37%, 11%, 0.6%) που προέκυψαν με βάση την ανάλυση Cartesian-PCA για το τροχιακό **AMBER** και το πεπτίδιο DTRW. *Δεξιά-πάνω*: υπέρθεση (χρησιμοποιώντας τα άτομα του πεπτιδικού σκελετού) αντιπροσωπευτικών δομών από το κυρίαρχο cluster (44%) που προέκυψε με βάση την ανάλυση Dihedral-PCA για το τροχιακό **AMBER** και το πεπτίδιο DTRW. Ο χρωματισμός των δομών έγινε με βάση τις τιμές RMSF σε κοινή κλίμακα από μπλε (0.15Å) σε κόκκινο (5.55Å), ενώ για λόγους ευκρίνειας χρησιμοποιείται διαφάνεια για τις πλευρικές ομάδες στις δομές του Dihedral-PCA cluster. *Δεξιά-κάτω*: ενεργειακά τοπία (ΔG energy plots) της προβολής του τροχιακού στο επίπεδο των 1-2, 1-3, 2-3 κυρίαρχων συνιστωσών (principal components) της ανάλυσης Cartesian-PCA. Από κάτω βλέπουμε την προβολή των Cartesian-PCA clusters πάνω στον πίνακα RMSD. Κάτω από τη διαγώνιο διατηρείται ο πίνακας RMSD ενώ πάνω από τη διαγώνιο βλέπουμε έντονα μόνο τις δομές που ανήκουν στο εκάστοτε cluster.



Εικόνα 3.62 *Αριστερά*: υπέρθεση (χρησιμοποιώντας όλα τα βαριά άτομα) αντιπροσωπευτικών δομών από τα κυρίαρχα cluster (38%, 5%, 1%) που προέκυψαν με βάση την ανάλυση Cartesian-PCA για το τροχιακό **CHARMM** και το πεπτικό DTRW. *Δεξιά-πάνω*: υπέρθεση (χρησιμοποιώντας τα άτομα του πεπτιδικού σκελετού) αντιπροσωπευτικών δομών από το κυρίαρχο cluster (56%) που προέκυψε με βάση την ανάλυση Dihedral-PCA για το τροχιακό **CHARMM** και το πεπτικό DTRW. Ο χρωματισμός των δομών έγινε με βάση τις τιμές RMSF σε κοινή κλίμακα από μπλε (0.15Å) σε κόκκινο (5.55Å), ενώ για λόγους ευκρίνειας χρησιμοποιείται διαφάνεια για τις πλευρικές ομάδες στις δομές του Dihedral-PCA cluster. *Δεξιά-κάτω*: ενεργειακά τοπία (ΔG energy plots) της προβολής του τροχιακού στο επίπεδο των 1-2, 1-3, 2-3 κυρίαρχων

συνιστωσών (principal components) της ανάλυσης Cartesian-PCA. Από κάτω βλέπουμε την προβολή των Cartesian-PCA clusters πάνω στον πίνακα RMSD. Κάτω από τη διαγώνιο διατηρείται ο πίνακας RMSD ενώ πάνω από τη διαγώνιο βλέπουμε έντονα μόνο τις δομές που ανήκουν στο εκάστοτε cluster.

◇ Επίσης η αντιπροσωπευτική δομή του **AMBER** έχει RMSD 2.4Å από την αντιπροσωπευτική του **CHARMM** και 4.0Å από την αντιπροσωπευτική του **OPLS** και η αντιπροσωπευτική δομή του **CHARMM** έχει RMSD 4.3Å από την αντιπροσωπευτική του **OPLS**.

Στην περίπτωση του πεπτιδίου αυτού, το οποίο είναι και σημαντικά πιο ασταθές από το RWPD, βλέπουμε ότι υπάρχει μεγαλύτερη απόκλιση στην περιγραφή του ενεργειακού τοπίου. Υπάρχει σύγκλιση στην πρόβλεψη κάποιων δομών αλλά γίνεται περισσότερο ευδιάκριτη η αδυναμία των force fields να περιγράψουν τις ασταθείς (disordered) δομές.

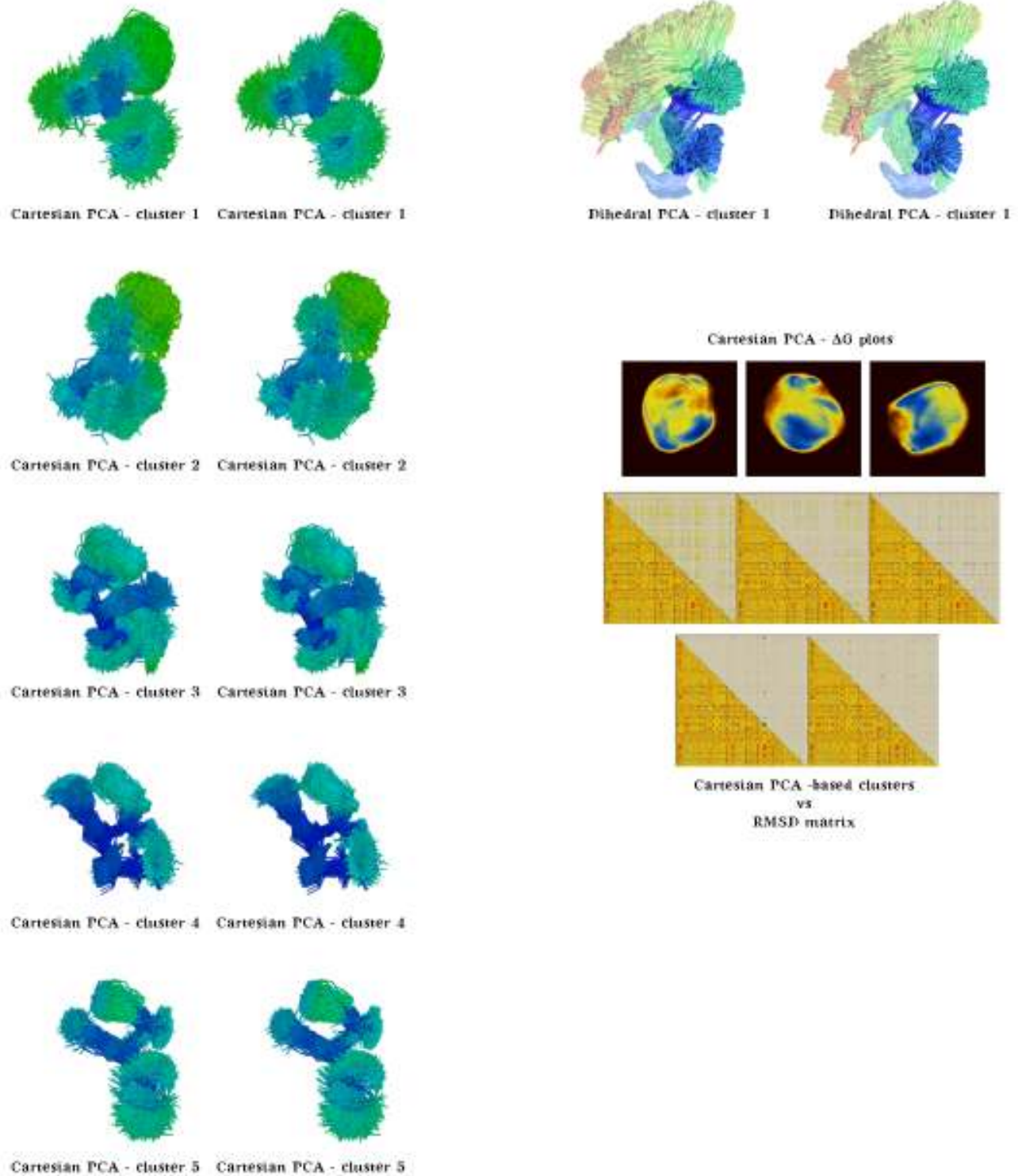


### Sufficient Sampling

Όλες οι παραπάνω παρατηρήσεις αποκτούν αξία μόνο στην περίπτωση που ο χρόνος του 1μs ήταν επαρκής για τα force fields για την περιγραφή των διαμορφώσεων των πεπτιδίων (sufficient sampling). Ο τρόπος που επιλέξαμε να το εξετάσουμε αυτό είναι μέσω της επικάλυψης (overlap) των eigenvectors (Hess, 2002) από την ανάλυση PCA στο πρώτο και δεύτερο μη επικαλυπτόμενο μισό του τροχιακού (Εικόνα 3.64).

Εναλλακτικά θα μπορούσαμε να υπολογίσουμε cosine content (Ενότητα 3.5) αλλά η εξάρτηση του από τον αριθμό των κύριων διαμορφώσεων (two-state, three-state folders) των πεπτιδίων μας αποτρέπει από τη γενικευμένη χρήση του, ειδικά στην περίπτωση που δεν γνωρίζουμε πειραματικά τη δομή (ή τις δομές) των πεπτιδίων της παρούσας μελέτης. Βλέπουμε ότι οι τιμές για όλα τα force fields συγκλίνουν τελικά σε τιμές πάνω από 0.9, με τα OPLS και CHARMM να δείχνουν πιο αργή σύγκλιση στα RWPD και DTRW, αντίστοιχα, όσον αφορά τη συμπεριφορά όλων των βαριών ατόμων (Cartesian-PCA, all-heavy atoms). Το ίδιο μοτίβο διακρίνεται και όταν ο υπολογισμός αφορά μόνο τον πεπτιδικό σκελετό, όπου βλέπουμε ότι μόλις από τον 15ο eigenvector οι τιμές αγγίζουν την μονάδα (Dihedral-PCA).





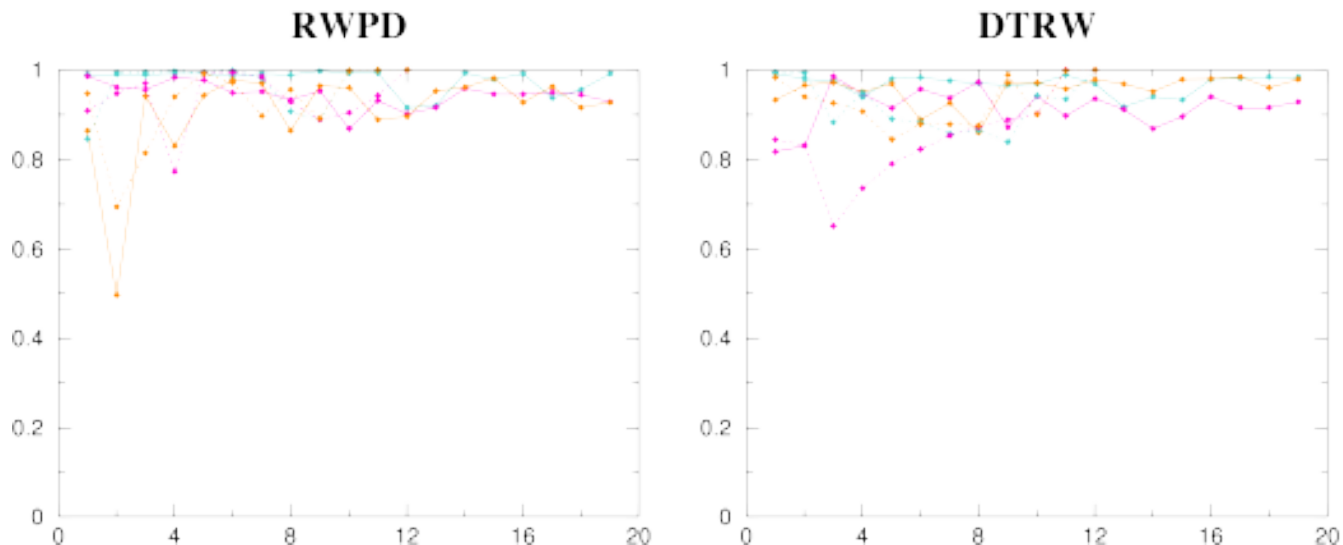
Εικόνα 3.63 Αριστερά: υπέρθεση (χρησιμοποιώντας όλα τα βαριά άτομα) αντιπροσωπευτικών δομών από τα κυρίαρχα cluster (22%, 13%, 5%, 2%, 3%) που προέκυψαν με βάση την ανάλυση Cartesian-PCA για το



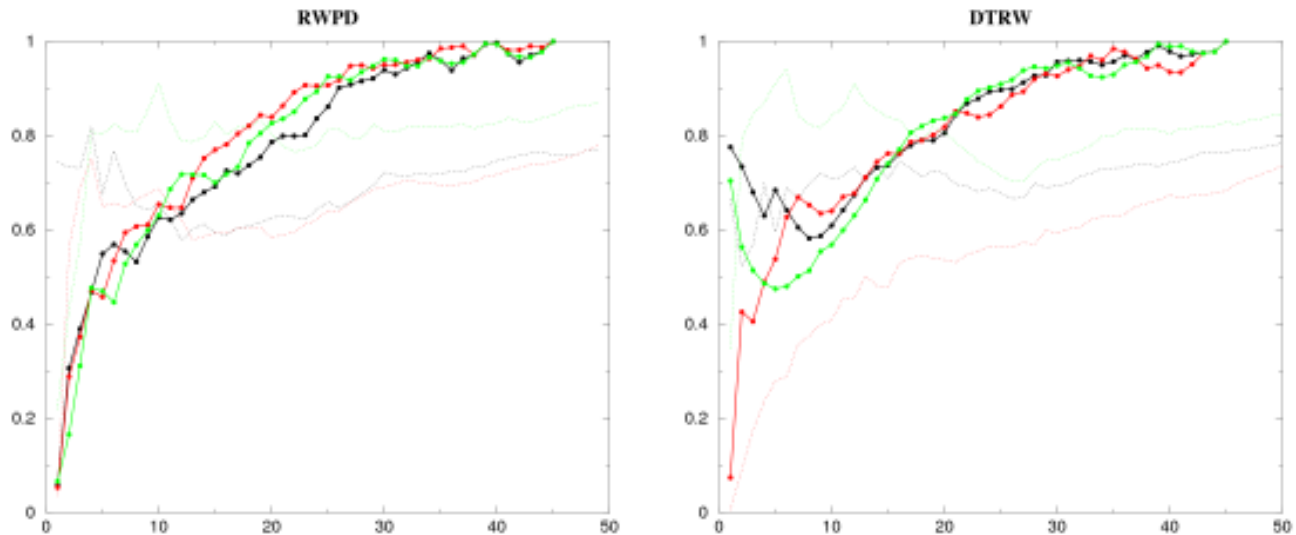
τροχιακό **OPLS** και το πεπτίδιο RWPD. Δεξιά: υπέρθεση (χρησιμοποιώντας τα άτομα του πεπτιδικού σκελετού) αντιπροσωπευτικών δομών από το κυρίαρχο cluster (29%) που προέκυψε με βάση την ανάλυση Dihedral-PCA για το τροχιακό **OPLS** και το πεπτίδιο RWPD. Ο χρωματισμός των δομών έγινε με βάση τις τιμές RMSF σε κοινή κλίμακα από μπλε (0.15Å) σε κόκκινο (5.55Å), ενώ για λόγους ευκρίνειας χρησιμοποιείται διαφάνεια για τις πλευρικές ομάδες στις δομές του Dihedral-PCA cluster. Κέντρο: ενεργειακά τοπία ( $\Delta G$  energy plots) της προβολής του τροχιακού στο επίπεδο των 1-2, 1-3, 2-3 κυρίαρχων συστατικών (principal components) της ανάλυσης Cartesian-PCA. Από κάτω βλέπουμε την προβολή των Cartesian-PCA clusters πάνω στον πίνακα RMSD. Κάτω από τη διαγώνιο διατηρείται ο πίνακας RMSD ενώ πάνω από τη διαγώνιο βλέπουμε έντονα μόνο τις δομές που ανήκουν στο εκάστοτε cluster.

Η επικάλυψη των eigenvectors μπορεί να χρησιμοποιηθεί και για να εξεταστεί η σύγκλιση μεταξύ των τροχιακών των διαφόρων force fields σε επίπεδο ομοιότητας της κίνησης που περιγράφεται από τα ζεύγη των eigenvectors-eigenvalues.

Στην Εικόνα 3.65 βλέπουμε την επικάλυψη (eigenspace overlap) των πρώτων 50 eigenvectors (με τα 50 υψηλότερα eigenvalues) από την ανάλυση Cartesian-PCA για όλους τους συνδυασμούς μεταξύ των force fields. Για το πεπτίδιο RWPD βλέπουμε σχεδόν την ίδια επικάλυψη μεταξύ CHARMM και OPLS, με το AMBER να ξεχωρίζει από αυτά τα δύο αν λάβουμε υπόψη και τις πλευρικές ομάδες (διακεκομμένες γραμμές, all-heavy atoms) ενώ η επικάλυψη είναι ίδια μεταξύ και των τριών αν θεωρήσουμε μόνο τον πεπτιδικό σκελετό (συνεχείς γραμμές, backbone atoms). Για το λιγότερο σταθερό πεπτίδιο DTRW η εικόνα είναι πιο πολύπλοκη. Υπάρχει πολύ καλύτερη σύγκλιση μεταξύ CHARMM και OPLS, ακολουθούμενη από τη σύγκλιση μεταξύ AMBER και CHARMM, με τα AMBER και OPLS να δείχνουν τη μεγαλύτερη μεταξύ τους απόκλιση, λαμβάνοντας υπόψη και τις πλευρικές ομάδες (διακεκομμένες γραμμές, all-heavy atoms). Η εικόνα αυτή αντιστρέφεται εάν εστιάσουμε στον πεπτιδικό σκελετό, όπου βλέπουμε καλύτερη συμφωνία μεταξύ AMBER και CHARMM. Η εικόνα που μας αποτυπώνει η μελέτη αυτή που στηρίζεται στην ανάλυση PCA έρχεται σε πλήρη συμφωνία με τα αποτελέσματα που παίρνουμε από τους πίνακες RMSD (Εικόνα 3.54).



Εικόνα 3.64 Eigenspace overlap μεταξύ του πρώτου και δεύτερου μη επικαλυπτόμενου μισού κάθε ανεξάρτητης προσομοίωσης με τα τρία force fields ως συνάρτηση των 20 eigenvectors με τα υψηλότερα eigenvalues από την ανάλυση Cartesian-PCA (συνεχείς γραμμές) και Dihedral-PCA (διακεκομμένες γραμμές). Ο χρωματικός κώδικας για τα force fields διατηρείται κοινός.



Εικόνα 3.65 Eigenspace overlap μεταξύ των force fields ως συνάρτηση των 50 eigenvectors με τα υψηλότερα eigenvalues από την ανάλυση Cartesian-PCA. Με μαύρο χρώμα φαίνεται η σύγκριση μεταξύ AMBER/CHARMM, με κόκκινο μεταξύ AMBER/OPLS και με πράσινο μεταξύ CHARMM/OPLS. Με συνεχή γραμμή φαίνεται ο υπολογισμός μόνο για τα άτομα του πεπτιδικού σκελετού ενώ η διακεκομμένη χρησιμοποιείται για τον υπολογισμό με όλα τα βαριά άτομα.

$$\frac{\Delta G_{folded} - \Delta G_{unfolded}}{T} = \Delta G_{folding}$$

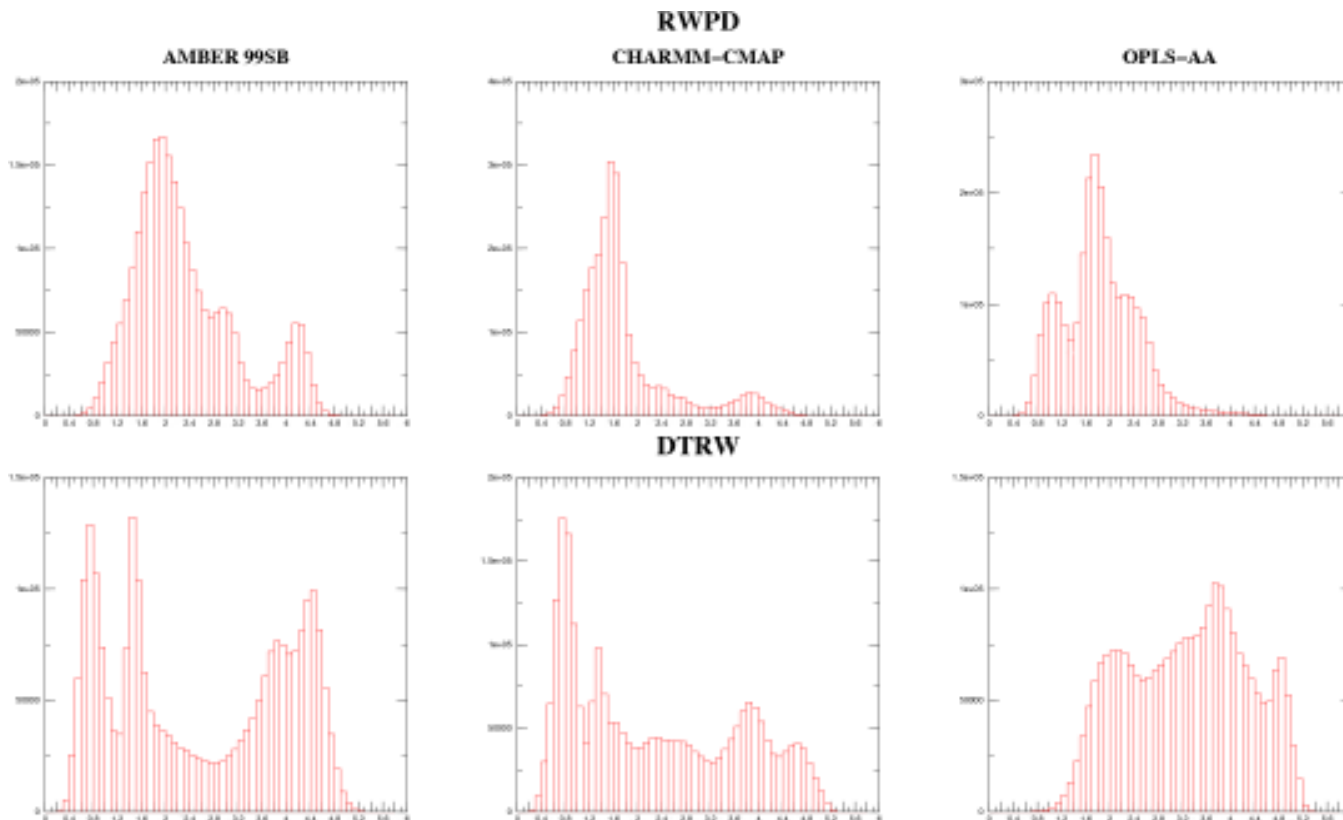
Στην ενότητα αυτή εστιαστήκαμε στη σύγκριση μεταξύ των force fields για τα δύο πεπτίδια. Ωστόσο όλα τα αποτελέσματα συνηγορούν προς τη σημαντικά μεγαλύτερη σταθερότητα του πεπτιδίου RWPD. Ο πιο άμεσος τρόπος ποσοτικοποίησης της παραπάνω παρατήρησης είναι μέσω του υπολογισμού του  $\Delta G_{folding}$  που αναλύσαμε στην Ενότητα 3.6 κατά τη μελέτη της επίδρασης της θερμοκρασίας. Επιλέγοντας το κατώφλι (2.7Å για τα AMBER, CHARMM και 2.5Å για το OPLS για το RWPD και 2.8Å για το AMBER, 3.2Å για το CHARMM και 2.6Å για το OPLS για το DTRW) με βάση τις κατανομές της Εικόνας 3.66 διαχωρίσαμε τα στιγμιότυπα κάθε τροχιακού στους πληθυσμούς F και U και υπολογίσαμε τις αντίστοιχες πιθανότητες εύρεσης του πεπτιδίου σε μία από τις δύο καταστάσεις. Η παρουσία του μείγματος των διαμορφώσεων, οι οποίες παρουσιάστηκαν κατά την cluster analysis, δίνει και πάλι την παρουσία πολλαπλών κορυφών στα ιστογράμματα κατανομής των RMSD από την κυρίαρχη δομή. Οι τιμές εδώ επιλέχθηκαν ώστε να περιληφθούν όλες οι διακριτές διαμορφώσεις.

Στη συνέχεια, από την εφαρμογή της συνάρτησης (5) παίρνουμε μία εκτίμηση της ελεύθερης ενέργειας αναδίπλωσης  $\Delta G_{folding}$ , για κάθε ένα force field (για προσομοιώσεις σε θερμοκρασία 320K). Για το πεπτίδιο RWPD υπολογίσαμε  $\Delta G_{folding}$  -1.1KJ/mol για το τροχιακό με το AMBER force field, -2.4KJ/mol για το τροχιακό με το CHARMM force field και -2.0KJ/mol για το τροχιακό με το OPLS force field. Για το πεπτίδιο DTRW υπολογίσαμε  $\Delta G_{folding}$  -0.15KJ/mol για το τροχιακό με το AMBER force field, -0.95KJ/mol για το τροχιακό με το CHARMM force field και +1.0KJ/mol για το τροχιακό με το OPLS force field.

Οι εκτιμώμενες αυτές τιμές ενέργειας αναδίπλωσης, σε συνδυασμό με τις κατανομές που παρατηρούμε στην Εικόνα 3.66 μας δείχνουν ότι (1) δεν μπορούμε να θέσουμε μία καθολική τιμή RMSD (παράμετρος απόστασης) ως κατώφλι για την εκτίμηση της αναδίπλωσης (παράμετρος ενεργειακή) μεταξύ των force fields και (2) τα πεπτίδια αυτά περνούν ένα σημαντικό κομμάτι του χρόνου της προσομοίωσης σε καταστάσεις τις οποίες θεωρούμε μη αναδιπλωμένες, οι οποίες όμως περιλαμβάνουν κάποιες παροδικές δομές. Σε γενικές γραμμές, η σειρά 'σταθερότητας' των force fields για το πεπτίδιο RWPD είναι CHARMM-OPLS-AMBER και για το πεπτίδιο DTRW είναι CHARMM-AMBER-OPLS.

Οι αδυναμίες ωστόσο των συγκεκριμένων εκδόσεων των force fields που χρησιμοποιήσαμε και που άρχισαν να διαφαίνονται στη βιβλιογραφία τα τελευταία χρόνια μας οδηγούν να

αντιμετωπίσουμε τα αποτελέσματα αυτά με σκεπτικισμό, κάνοντας επιτακτική την ανάγκη πειραματικής επιβεβαίωσης ειδικά για πεπτίδια τόσο μικρού μήκους.



Εικόνα 3.66 Κατανομή των τιμών RMSD των δομών ολόκληρου του τροχιακού από την αντιπροσωπευτική δομή του κυρίαρχου cluster όπως ορίστηκε μέσω της ανάλυσης Cartesian-PCA χρησιμοποιώντας όλα τα βαριά άτομα. Ένα τοπικό ελάχιστο της κατανομής ορίστηκε ως κατώφλι για το διαχωρισμό των δομών σε δύο πληθυσμούς, F (Folded) και U (Unfolded).

*"The true delight is in the finding out rather than in the knowing."*

*Isaac Asimov*







# Κεφάλαιο 4 ΠΕΝΤΑΠΕΠΤΙΔΙΑ



*"Being a PhD advisor is like sex:  
one mistake and you're providing  
support for a lifetime."*

*Michael Sinz*



## 4.1 Επιλογή πενταπεπτιδικών αλληλουχιών

Τα τετραπεπτίδια αποτέλεσαν για εμάς τα "ινδικά χοιρίδια" πάνω στα οποία εμπνευστήκαμε, εφαρμόσαμε και βελτιώσαμε τα προγράμματα και τις συναρτήσεις μας. Έτσι όταν στραφήκαμε προς τα μεγαλύτερου μήκους πενταπεπτίδια προς αναζήτηση "αναδιπλούμενων" αλληλουχιών, ακολουθήσαμε τη μεθοδολογία την οποία αναπτύξαμε στο σύνολο των τετραπεπτιδίων και αναλύσαμε εκτενώς στο Κεφάλαιο 3 της παρούσας διατριβής. Η ερώτηση αυτή τη φορά διαμορφώνεται:

*Υπάρχει άραγε ένα πενταπεπτίδιο που να υιοθετεί μία σταθερή δομή σε υδατικό διάλυμα;*

Και αν ναι,

*Μπορούμε να το ταυτοποιήσουμε χρησιμοποιώντας τις προσομοιώσεις μοριακής δυναμικής, με τρόπο πανομοιότυπο με αυτόν που ακολουθήσαμε στα πεπτίδια μήκους τεσσάρων καταλοίπων;*

Ο αριθμός των πιθανών συνδυασμών αλληλουχιών για τα πενταπεπτίδια είναι 3.200.000. Η επιβολή των ίδιων περιορισμών (Εικόνα 3.1, Πίνακας 3.1) με τα τετραπεπτίδια οδηγεί σε 54.000 συνδυασμούς, αριθμός ο οποίος είναι ακόμα αρκετά υψηλός για περαιτέρω μελέτη. Έτσι εξετάστηκαν κι άλλοι περιορισμοί, όπως η παρουσία της τρυπτοφάνης μόνο σε εσωτερικές θέσεις της αλληλουχίας (1 Trp εσωτερικά) και ο αποκλεισμός της προλίνης, η οποία λόγω των

ασυνήθιστων φ/ψ δίδεδρων γωνιών περιορίζει τους βαθμούς ελευθερίας του πεπτιδικού σκελετού (MacArthur et al., 1991). Ωστόσο, οι περιορισμοί αυτοί δε μειώνουν σημαντικά τον αριθμό των πεπτιδίων, όπως φαίνεται και στον Πίνακα 4.1. Η επιβολή της παρουσίας ενός επιπλέον φορτίου (3 φορτισμένα) όμως (Glättli et al., 2005, Wei et al., 2005), αφενός συμβάλλει θετικά στη αύξηση της διαλυτότητας, αφετέρου έχει σημαντική επίπτωση στον αριθμό των πεπτιδικών αλληλουχιών (από 3.200.000 σε 651.605 με 1 Trp, σε 19.200 με 1 Trp και 3 φορτισμένα). Ο τελικός συνδυασμός που επιλέχτηκε ήταν 1 κατάλοιπο τρυπτοφάνης, 3 φορτισμένα κατάλοιπα, και η απουσία επαναλήψεων αμινοξικών καταλοίπων, που οδήγησε σε ένα σύνολο 7.200 πενταπεπτιδίων. Η τελευταία παράμετρος επιλέχτηκε για την αποφυγή επαναλήψεων των φορτισμένων αμινοξέων, ενώ μείωσε περαιτέρω τον αριθμό των πεπτιδίων. Να σημειωθεί σε αυτό το σημείο ότι η απουσία επαναλήψεων ισχύει και στην περίπτωση των τετραπεπτιδίων (χωρίς την ανάγκη επιβολής επιπλέον περιορισμού λόγω του μικρότερου μήκους). Στο Παράρτημα, παραθέτονται όλα τα προγράμματα (#1 - #7) που χρησιμοποιήθηκαν για να παραχθούν οι πιθανοί συνδυασμοί αλληλουχιών μετά την επιβολή των παραπάνω περιορισμών.

Μία ενδιαφέρουσα μελέτη θα ήταν να πραγματοποιηθεί μία έρευνα στη βάση δεδομένων Protein Data Bank (Bernstein et al., 1977), για την εύρεση πενταπεπτιδίων για τα οποία δεν υπάρχει κάποια δομική πληροφορία. Ο στόχος δηλαδή είναι να ταυτοποιήσουμε πενταπεπτιδικές αλληλουχίες οι οποίες δεν συμπεριλαμβάνονται σε αλληλουχία πρωτεΐνης για την οποία έχει προσδιοριστεί και καταχωρηθεί η τρισδιάστατη δομή της. Για το σκοπό αυτό, ετοιμάσαμε ένα πρόγραμμα (Παράρτημα, #8-#9) το οποίο σε πρώτη φάση διαβάζει όλες τις αλληλουχίες (σε μορφή fasta) του αρχείου της βάσης δεδομένων της PDB (pdb archive, 05.07.2009), και στη συνέχεια παράγει όλα τα πιθανά πενταπεπτίδια. Η σύγκριση των πενταπεπτιδίων που περιλαμβάνονται στην PDB με τα 7.200 πενταπεπτίδια που επιλέξαμε οδηγεί στο διαχωρισμό των τελευταίων σε 2 κατηγορίες (PDB και NonPDB).

Όπως βλέπουμε και στον Πίνακα 4.1, 4.519 από τα 7.200 πενταπεπτίδια, δηλαδή πάνω από το 60%, δεν συμπεριλαμβάνονται σε αλληλουχία πρωτεΐνης με προσδιορισμένη δομή. Ο λόγος για την πραγματοποίηση μίας τέτοιας διερεύνησης είναι να εξετάσουμε σε ποια από τις δύο κατηγορίες ανήκουν (ή τείνουν να ανήκουν) τα δυνητικά αναδιπλούμενα πεπτίδια που τυχόν θα προσδιορίσουμε στη συνέχεια. Να σημειωθεί ότι ο εμπλουτισμός της βάσης δεδομένων από την ημερομηνία πραγματοποίησης της ανάλυσης (2009) έως σήμερα (2012) ενδέχεται να έχει αλλάξει τα ποσοστά αυτά.

Αριθμός Πενταπεπτιδίων	Παράμετροι									
	ΟΛΑ	NonPDB	PDB	1 Trp	1 Trp ΕΣΩΤΕΡΙΚΑ	2 ΦΟΡΤΙΣΜΕΝΑ	1 ΘΕΤ.- 1 ΑΡΝΗΤ. ΦΟΡΤΙΣΜΕΝΟ	3 ΦΟΡΤΙΣΜΕΝΑ	ΟΛΑ ΑΑ ΔΙΑΦΟΡΕΤΙΚΑ	No Pro
3.200.000	X									
1.672.626		X								
1.527.374			X							
491.795		X		X						
294.938		X			X					
159.810			X	X						
96.025			X		X					
74.732		X		X		X				
52.657		X		X		X			X	
44.751		X			X	X				
36.940		X		X			X			
34.600		X		X			X		X	
33.268			X	X		X				
31.503		X			X	X			X	
22.943			X	X		X			X	
22.121		X			X		X			
20.717		X			X		X		X	
20.049			X		X	X				
17.961		X			X		X		X	X
17.060			X	X			X			
15.800			X	X			X		X	
13.857			X		X	X			X	
11.813		X		X				X		
10.279			X		X		X			
9.523			X		X		X		X	
8.247			X		X		X		X	X
7.387			X	X				X		
6.934		X			X			X		
4.586			X		X			X		
<b>4.519</b>		<b>X</b>		<b>X</b>				<b>X</b>	<b>X</b>	
2.682		X			X			X	X	
<b>2.681</b>			<b>X</b>	<b>X</b>				<b>X</b>	<b>X</b>	
1.638			X		X			X	X	

Πίνακας 4.1 Αριθμός πιθανών πενταπεπτιδικών αλληλουχιών και περιοριστικές παράμετροι.

Σε περίπτωση που προσδιορίσουμε έναν “καλό αναδιπλωτή” ο οποίος ανήκει στην πρώτη κατηγορία, θα ήταν ιδιαίτερα ενδιαφέρον να συγκρίνουμε τη δομή του πεπτιδίου αυτού όταν βρίσκεται ελεύθερο στο διάλυμα και όταν βρίσκεται σε πρωτεϊνικό περιβάλλον. Η πεποίθησή μας είναι ότι περιμένουμε σημαντική διαφορά στις δύο δομές, καθώς η δομή που υιοθετεί ένα πενταπεπτίδιο εξαρτάται από το πρωτεϊνικό περιεχόμενο στο οποίο θα βρεθεί, καθώς έχουν ταυτοποιηθεί αλληλουχίες “χαμαιλέοντες” που παίρνουν είτε α- είτε β- δομές (Minor et al., 1996, Mezei, 1998). Οι Kabsch & Sander σε μία έρευνα για εύρεση πανομοιότυπων πενταπεπτιδικών αλληλουχιών σε πρωτεΐνες με πειραματικά προσδιορισμένη δομή, κατέδειξαν περιπτώσεις όπου το ίδιο πενταπεπτίδιο συμμετέχει τόσο σε δομή α-έλικας όσο και σε δομή β-φύλλου (αλλά και πενταπεπτίδια που υιοθετούν πανομοιότυπη δομή σε διαφορετικές πρωτεΐνες (Kabsch et al., 1984). Για παράδειγμα, το εξαπεπτίδιο AAGDYY-NH<sub>2</sub> (B1) της πρωτεΐνης BLIP (TEM-1 β-lactamase inhibitor protein) παίρνει δομή β-στροφής και απουσία της υπόλοιπης πρωτεΐνης, ο δε τύπος της β-στροφής ποικίλλει ανάλογα με τον διαλύτη (Gao et al., 2002). Τα 7.200 πενταπεπτίδια μελετήθηκαν ως προς την αναδίπλωσή τους χρησιμοποιώντας τις προσομοιώσεις μοριακής δυναμικής και ακολουθώντας τη μεθοδολογία που αναπτύξαμε στα τετραπεπτίδια.

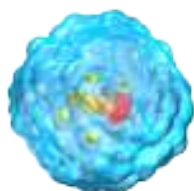
*"More than any other time in history,  
mankind faces a crossroads.  
One path leads to  
despair and utter hopelessness.  
The other, to total extinction.  
Let us pray we have the wisdom  
to choose correctly."*

*Woody Allen*



*"That's not right.  
That's not even wrong."*

*Wolfgang Pauli*



## 4.2 Σχεδιασμός, αριθμός και διάρκεια προσομοιώσεων

**ΣΤΟ** σύνολο των πενταπεπτιδίων ακολουθήσαμε την ίδια λογική με τα τετραπεπτίδια για να σχεδιάσουμε το γενικό πλάνο των προσομοιώσεων (Ενότητα 3.2). Από μία μικρή ανασκόπηση στη διεθνή βιβλιογραφία, είδαμε από προσομοιώσεις που έχουν γίνει σε πενταπεπτίδια, ξεκινώντας από εκτεταμένη διαμόρφωση, ότι τα γεγονότα αναδίπλωσης λαμβάνουν χώρα στην κλίμακα των 2-14ns.

Για παράδειγμα το πεπτίδιο AYPYD αναδιπλώθηκε στην NMR δομή (type VI reverse  $\beta$ -turn) σε 3ns (Demchuck et al., 1997), όπως και τα πεπτίδια SYPFDV, SYPYD, SYPFD που αναδιπλώνονται σε 2-4ns, λόγω του ισχυρού πακεταρίσματος του δακτυλίου της προλίνης με τα αρωματικά αμινοξέα του τμήματος YPF (Mohanty et al., 1997). Στο πεπτίδιο YP<sub>(trans)</sub>GDV τα γεγονότα αναδίπλωσης λαμβάνουν χώρα στα πρώτα 2ns λόγω των ισχυρών ηλεκτροστατικών αλληλεπιδράσεων μεταξύ N-τελικού και C-τελικού άκρου και των φορτισμένων αμινοξέων Tyr1 και Asp4 (Wu et al., 2000). Η δημιουργία της δομής θηλιάς συμβαίνει στην κλίμακα των 10ns με 14ns (από προσομοιώσεις και μετρήσεις φθορισμού, αντίστοιχα) στο πεπτίδιο CAGQW (Yeh et al., 2009). Επίσης, πειραματικές και θεωρητικές μελέτες στο ίδιο πενταπεπτίδιο αλλά και οκταπεπτίδιο του τύπου C-(AGQ)<sub>v</sub>-W έδειξαν ότι ο σχηματισμός της δομής θηλιάς γίνεται στα 5ns και στα 6-8ns αντίστοιχα (Yeh et al., 2002). Από πειραματικές μελέτες triplet-triplet energy transfer, φαίνεται ότι ο σχηματισμός δεσμικών αλληλεπιδράσεων (contact formation) λαμβάνει χώρα στα 11.6( $\pm$ 0.4)ns, 25.0( $\pm$ 1.3)ns, 57.1( $\pm$ 3.3)ns για πεπτίδια μήκους 10, 18 και 30 καταλοίπων



αντίστοιχα (Kreiger et al., 2003). Τα β-επταπεπτίδια δείχνουν γεγονότα αναδίπλωσης (σε μεθανόλη) στους 298K στην κλίμακα των 40ns, ενώ στους 340K-360K τα γεγονότα αναδίπλωσης συμβαίνουν στα 2-4ns (Daura et al., 1998, vanGunsteren et al., 2001).

Έτσι με βάση τους χρόνους αναδίπλωσης που αναφέρονται στη διεθνή βιβλιογραφία, τους χρόνους αναδίπλωσης που παρατηρήσαμε από τη μελέτη μας στα 1.440 τετραπεπτίδια και τις δοκιμαστικές προσομοιώσεις που πραγματοποιήσαμε στο πεπτίδιο-μοντέλο RWTDQ (Πίνακας 3.2) καταλήξαμε στην επιμήκυνση του χρόνου της προσομοίωσης (Naganathan et al., 2005) σε 20ns, αντί των 4 επαναλήψεων των 5ns (συνολικά 20ns) στα τετραπεπτίδια. Λόγω του μεγαλύτερου πλήθους αλληλουχιών αλλά και της επιμήκυνσης του χρόνου της προσομοίωσης, ο πρώτος κύκλος των προσομοιώσεων διήρκεσε λίγο παραπάνω από 7 μήνες (231 μέρες φυσικού χρόνου) και οδήγησε σε ένα σύνολο 144μs υπολογιστικού χρόνου. Το πρωτόκολλο της προσομοίωσης είναι αυτούσιο με αυτό που παρουσιάζεται στο Παράρτημα (#13, NAMD script, all.namd).

Αριθμός Πενταπεπτιδίων	Χρόνος Προσομοίωσης (ns)	Αριθμός Επαναλήψεων	Αθροιστικός Υπολογιστικός Χρόνος (ns)
7.200	20	1	144.000
480	100	1	48.000
32	120	4 <sup>#</sup>	15.360
8	1000	1	8.000
1	2000	1 <sup>*</sup>	2.000
1	2000	3 <sup>*</sup>	6.000
1	1000	2 <sup>§</sup>	2.000

Πίνακας 4.2 Συγκεντρωτικός πίνακας των προσομοιώσεων που πραγματοποιήσαμε στο σύνολο των 7.200 πενταπεπτιδίων. Οι επαναλήψεις που σημειώνονται με \* αφορούν διαφορετικές θερμοκρασίες, οι επαναλήψεις που σημειώνονται με <sup>#</sup> αφορούν διαφορετικά force fields και οι επαναλήψεις που σημειώνονται με <sup>§</sup> αφορούν τη μέθοδο adaptive tempering.

Στη συνέχεια, και όπως αναλύεται στις επόμενες ενότητες, εφαρμόσαμε τις δύο βασικές συναρτήσεις εκτίμησης της αναδιπλωσιμότητας TF2 και TF3 (Ενότητα 2.3) που αναπτύξαμε κατά τη μελέτη των τετραπεπτιδίων, και οι οποίες ανιχνεύουν τα γεγονότα αναδίπλωσης και

αξιολογούν τη δυναμική αναδιπλωσιμότητα ενός πεπτιδίου. Πεπτιδία με υψηλή βαθμολογία επιλέγονται για ένα καινούργιο κύκλο προσομοιώσεων μεγαλύτερης διάρκειας. Η διαδικασία αυτή επαναλαμβάνεται μέχρις ότου να υποδειχθεί ένας μικρός αριθμός υποψήφιων πενταπεπτιδίων με σταθερή αναδίπλωση.

Συνολικά, ο υπολογιστικός χρόνος των πενταπεπτιδίων αθροίζεται σε 225.36 $\mu$ s (135.216 core-hours) για τον οποίο χρειάστηκαν περίπου 386 μέρες (αθροιστικού) φυσικού χρόνου. Στον Πίνακα 4.2 παραθέτουμε το σύνολο των προσομοιώσεων που πραγματοποιήθηκαν στο σύνολο των 7.200 πενταπεπτιδίων.

Όλες οι προσομοιώσεις πραγματοποιούνται με το πρόγραμμα NAMD (Kale et al., 1999) σε συνθήκες περιοδικής οριοθέτησης, με αναλυτική παρουσία του διαλύτη (explicit solvent) και πλήρη υπολογισμό των ηλεκτροστατικών αλληλεπιδράσεων με τη μέθοδο Particle Mesh Ewald (full PME electrostatics), όπως αναλύεται στο Κεφάλαιο 2, Ενότητα 2.1.

Στις πρώτες δύο φάσεις των προσομοιώσεων των πενταπεπτιδίων (Ενότητα 4.3-7200 πενταπεπτιδία και Ενότητα 4.4-480 πενταπεπτιδία) χρησιμοποιείται το force field CHARMM22 (MacKerell et al., 1998), όπως και στους πρώτους κύκλους προσομοιώσεων των τετραπεπτιδίων. Στη συνέχεια (στο στάδιο των 32 πενταπεπτιδίων) χρησιμοποιούνται τα τέσσερα force fields, CHARMM-CMAP (MacKerell et al., 2004), OPLS-AA (Jorgensen et al., 1996, Kaminski et al., 2001), AMBER99SB (Hornak, et al., 2006, Wickstrom et al., 2009) και AMBER99SB-ILDN (Lindorff-Larsen et al., 2010). Ακολούθως τα 8 πενταπεπτιδία που επιλέχθηκαν, μελετήθηκαν με προσομοιώσεις διάρκειας 1 $\mu$ s με το force field AMBER99SB-ILDN, το οποίο φαίνεται να δίνει τα περισσότερα αξιόπιστα αποτελέσματα για μικρά πεπτιδία, σε σύγκριση με πειραματικά δεδομένα (Lindorff-Larsen et al., 2012). Δύο πεπτιδία που έδειξαν αναδίπλωση σε μία σταθερή και καλά καθορισμένη δομή μελετήθηκαν σε διαφορετικές θερμοκρασίες. Τέλος, η παρουσία της προλίνης στην αλληλουχία του πενταπεπτιδίου με τα καλύτερα αποτελέσματα, μας έδωσε την ευκαιρία να μελετήσουμε το φαινόμενο της ισομερείωσης του πεπτιδικού δεσμού (cis/trans isomerization), χρησιμοποιώντας τη μέθοδο adaptive tempering (Zhang et al., 2010).

Η αναζήτηση ενός σταθερά αναδιπλούμενου πενταπεπτιδίου με τα χαρακτηριστικά αλληλουχίας τα οποία θέσαμε διήρκεσε σχεδόν δύο χρόνια, εκ του οποίου το ~50% αντιστοιχεί σε καθαρό υπολογιστικό χρόνο αφιερωμένο σε προσομοιώσεις. Ο υπόλοιπος χρόνος αντιστοιχεί στην ανάλυση των αποτελεσμάτων, καθώς σε αυτό το σύνολο των πεπτιδίων πραγματοποιήσαμε ένα σεβαστό αριθμό (>50) από προσομοιώσεις μεγάλης διάρκειας, των οποίων η ανάλυση δεν έγινε

με συστηματικό τρόπο μέσω των συναρτήσεων.

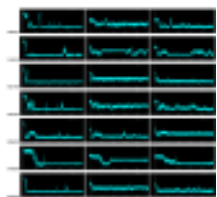
Στις επόμενες ενότητες του κεφαλαίου αυτού περιγράφουμε την περιπλάνησή μας στον κόσμο των πενταπεπτιδίων και πως καταλήξαμε στην ανάδειξη των υποψήφιων “δυσνητικά αναδιπλούμενων”.

*"A hypothesis or theory is clear, decisive, and positive, but it is believed by no one but the man who created it. Experimental findings, on the other hand, are messy, inexact things, which are believed by everyone except the man who did that work."*

*Harlow Shapley*

*"Support bacteria,  
it's the only culture some people have."*

<http://www.gdargaud.net/Humor/QuotesScience.html>



### 4.3 Επιλογή 480 υποψήφιων δυνητικά αναδιπλούμενων πενταπεπτιδίων

Η αλληλεπίδραση μας με τα τετραπεπτίδια που αναλύθηκε εκτενώς στο Κεφάλαιο 3 μας οδήγησε σε δύο συναρτήσεις εκτίμησης της αναδιπλωσιμότητας TF2 και TF3 (Ενότητα 2.3). Η συνάρτηση TF2 βασίζεται σε ατομικές αποστάσεις και εφαρμόστηκε σε τροχιακά μικρής διάρκειας (5ns) με σκοπό την ανίχνευση γεγονότων αναδίπλωσης. Η συνάρτηση TF3 βασίζεται σε πίνακες RMSD μεταξύ διαδοχικών δομών του τροχιακού και σε ατομικές διακυμάνσεις με σκοπό την ανίχνευση της δημιουργίας δομής αλλά και την εκτίμηση της σταθερότητας της. Στο δεύτερο κύκλο προσομοιώσεων του συνόλου των 130 τετραπεπτιδίων, που ο χρόνος προσομοίωσης ήταν 30ns, είδαμε ότι τα μειονεκτήματα που εντοπίσαμε στην πρώτη συνάρτηση, TF2, οδηγούσαν σε λανθασμένη εκτίμηση της αναδιπλωσιμότητας των πεπτιδίων, εν αντιθέσει με την συνάρτηση TF3. Συνεπώς, η ανάλυση των προσομοιώσεων των 7.200 πενταπεπτιδίων μπορεί να συνεχιστεί εφαρμόζοντας τη συνάρτηση TF3 και επιλέγοντας τον κατάλληλο αριθμό πενταπεπτιδίων για τον επόμενο κύκλο προσομοιώσεων.

Για λόγους οικονομίας χρόνου, καθώς η ανάλυση των αποτελεσμάτων του συνόλου των τετραπεπτιδίων και η ανάπτυξη των συναρτήσεων ήταν αρκετά χρονοβόρα (σχεδόν ένα χρόνο, εκ του οποίου ~20% αντιστοιχεί σε υπολογιστικό χρόνο στο cluster), προχωρήσαμε στις προσομοιώσεις των 7.200 πενταπεπτιδίων προτού ανακαλύψουμε τα μειονεκτήματα της συνάρτησης TF2 και προχωρήσουμε στο σχεδιασμό της συνάρτησης TF3. Αυτό είχε σαν συνέπεια την απουσία των πινάκων RMSD για την εφαρμογή της συνάρτησης TF3. Έτσι, με βάση την πορεία που ακολουθήσαμε στα τετραπεπτίδια, εφαρμόσαμε τη συνάρτηση TF2 που εξετάζει τις

τρεις ατομικές αποστάσεις και το μεταξύ τους συγχρονισμό και επιλέξαμε 480 πενταπεπτίδια με βάση τη βαθμολογική τους κατάταξη. Το πλήθος των 480 πενταπεπτιδίων επιλέχθηκε με βάση την απόδοση των 100ns/κόμβο/μέρα. Έτσι, εάν έχουμε στη διάθεσή μας όλη τη συστοιχία των υπολογιστών μπορούμε να κάνουμε προσομοιώσεις των 100ns σε 8 πεπτίδια την ημέρα, ή 240 πεπτίδια σε 1 μήνα, ή 480 πεπτίδια σε 2 μήνες. Αυτή η λίστα που προέκυψε από την απευθείας εφαρμογή της συνάρτησης TF2 θα την ονομάσουμε *top480A* (Εικόνα 4.1).



Εικόνα 4.1 Word-cloud των 7.200 πενταπεπτιδίων, όπου το μέγεθος της αλληλουχίας είναι ενδεικτικό της βαθμολογίας που έλαβε με βάση τη συνάρτηση TF2.

Η εμπειρία μας από τα τετραπεπτίδια ωστόσο, μας υπέδειξε την ανάγκη επιπλέον εξέτασης των αποτελεσμάτων για εμπλουτισμό της λίστας. Επειδή τα τροχιακά σβήνονται αυτόματα μετά την επεξεργασία τους μέσω του Perl script, ο εκ των υστέρων υπολογισμός των πινάκων RMSD είναι αδύνατος. Από την άλλη, η οπτική εξέταση όλων των διαθέσιμων αποτελεσμάτων (42.674 αρχεία) δεν είναι μόνο δύσκολη και επίπονη αλλά εμπεριέχει και το στοιχείο της υποκειμενικότητας. Οι παράμετροι που έχουμε στη διάθεσή μας (όπως και στην περίπτωση των τετραπεπτιδίων, Ενότητα 3.4) είναι οι τρεις αποστάσεις μεταξύ ατόμων Ca, η γυρεοσκοπική ακτίνα (mass-weighted, heavy atoms), η εντροπία της κατανομής των τριών principal components

από την ανάλυση Cartesian-PCA (όλα τα βαριά άτομα) και ένα αρχείο PDB με δομές, σε υπέρθεση, του κυρίαρχου cluster, εφόσον αυτό διαρκεί για τουλάχιστον 10% του χρόνου της προσομοίωσης, περιορισμό τον οποίο πληροί το 73% των πενταπεπτιδίων (5.265 από τα 7.200). Προσπαθήσαμε λοιπόν να σχεδιάσουμε μία συνάρτηση η οποία να μπορεί να αναπληρώσει την πληροφορία που εμπεριέχεται στους πίνακες RMSD.

Η πρώτη μας σκέψη ήταν να εξετάσουμε την κατανομή των βαθμολογιών για κάθε μία από τις παραπάνω παραμέτρους ώστε να προσδιορίσουμε ένα κατώφλι βαθμολογίας (Εικόνα 4.2). Ωστόσο οι κατανομές αυτές δεν έχουν τη δικόρυφη (bimodal) εμφάνιση που θα ελπίζαμε ώστε να μπορούμε να ορίσουμε το πρώτο τοπικό ελάχιστο ως κατώφλι. Αν προχωρήσουμε στη δημιουργία ενός δενδρογράμματος (Εικόνα 4.3) όπου οι διαφορές των βαθμολογιών αντιμετωπίζονται ως αποστάσεις (Ενότητα 3.3) δημιουργούμε περισσότερες ερωτήσεις παρά απαντήσεις: ο αριθμός των cluster που θα προκύψουν εξαρτάται από το ύψος του δένδρου που θα ορίσουμε. Εάν για παράδειγμα, στο δενδρογράμμα της Εικόνας 4.3 επιλέξουμε να έχουμε 3 cluster, αυτό θα ομαδοποιήσει τα 7.200 πεπτίδια σε 3 πληθυσμούς των 7.010, 186 και 4 πεπτιδίων. Συνεπώς, σύμφωνα με αυτή τη μέθοδο θα μπορούσαμε να επιλέξουμε 190 πενταπεπτίδια.

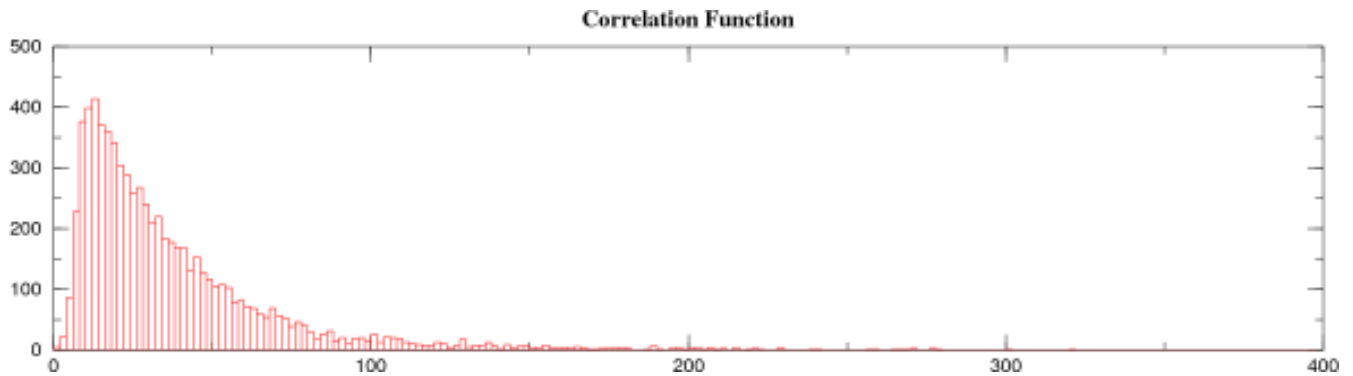
Ένας εναλλακτικός τρόπος της κατασκευής δενδρογράμματος που θα μπορούσε να εφαρμοστεί καθολικά και με τον ίδιο τρόπο σε όλες τις παραμέτρους είναι να υπολογίσουμε για κάθε κατανομή μέση τιμή και τυπική απόκλιση και να επιλέξουμε όλα τα πεπτίδια που βρίσκονται στο ακραίο δεξιό τμήμα της κατανομής, δηλαδή έχουν βαθμολογία πάνω από το μέσο όρο + 3σ (mean+3σ).

Βλέποντας συγκεντρωτικά τα πενταπεπτίδια αυτά για όλες τις παραμέτρους καταλήξαμε λίγο-πολύ στα ίδια συμπεράσματα με αυτά του Πίνακα 2.2 για τα 130 τετραπεπτίδια αναφορικά με τη σχέση μεταξύ των διαφόρων παραμέτρων. Χρησιμοποιώντας τα 130 τετραπεπτίδια, για τα όποια έχουμε τόσο τους πίνακες RMSD όσο και τις υπόλοιπες παραμέτρους, ως "test data-set" προσπαθήσαμε να βρούμε ένα συνδυασμό παραμέτρων που να μπορούν να αναπληρώσουν την πληροφορία που εμπεριέχεται στους πίνακες RMSD, η προσπάθεια όμως αυτή απέβη άκαρπη.

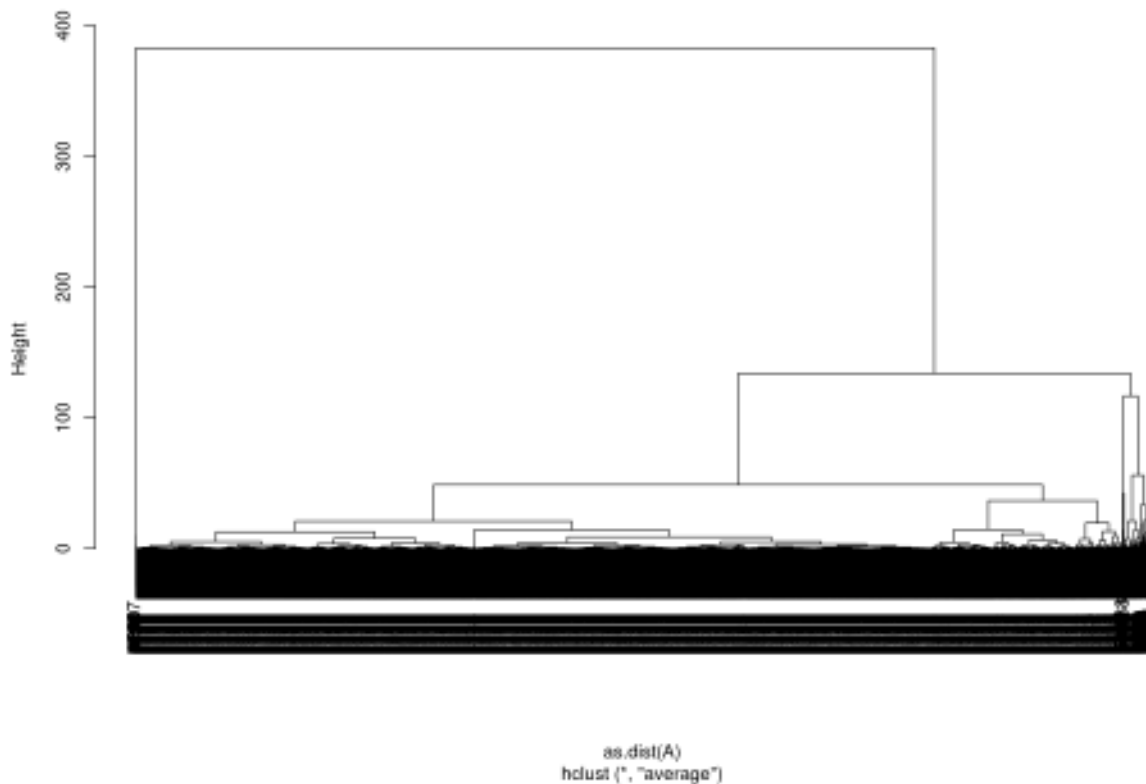
Επιστρέφοντας στην αρχική συνάρτηση των ατομικών αποστάσεων εξετάσαμε οπτικά όλες τις γραφικές παραστάσεις των ατομικών αποστάσεων (21.600 γραφικές παραστάσεις) και καταλήξαμε σε δύο συμπεράσματα: (α) μπορούμε αυξάνοντας τον αριθμό των πεπτιδίων που θα επιλεγθούν για τον επόμενο κύκλο προσομοιώσεων, να ξεπεράσουμε μερικώς την "ανεπάρκεια"



της συνάρτησης των ατομικών αποστάσεων (συνάρτηση TF2) και  $(\beta)$  η επιμήκυνση του χρόνου της προσομοίωσης σε 20ns, σε σχέση με τα 5ns της περίπτωσης του πρώτου κύκλου προσομοιώσεων των τετραπεπτιδίων, μας επιτρέπει να ελέγχουμε το μέγεθος του κυρίαρχου cluster μέσω της τιμής που θα ορίσουμε ως κατώφλι για να γίνει cluster analysis.



Εικόνα 4.2 Κατανομή της βαθμολογίας των 7.200 πενταπεπτιδίων με βάση τις τρεις αποστάσεις των ατόμων Ca (συνάρτηση TF2).



Εικόνα 4.3 Cluster analysis των βαθμολογιών των πεπτιδίων της Εικόνας 4.1, όπου οι απόλυτες διαφορές των βαθμολογιών αντιμετωπίζονται ως αποστάσεις για την κατασκευή δενδρογράμματος.

Αυξομειώνοντας το κατώφλι για την πραγματοποίηση cluster analysis, πεπτίδια τα οποία σχηματίζουν μεγάλα cluster δε θα επηρεαστούν, αλλά για τα πεπτίδια που σχηματίζουν cluster με μικρό αριθμό από frames μεν, αλλά συμπαγές (χαμηλές τιμές RMSFs), δε θα γίνει cluster analysis όσο μεγαλώνει το cut-off. Για τις τέσσερις τιμές cut-off που ορίσαμε τα αποτελέσματα διαμορφώθηκαν ως εξής:

A) Εάν το κυρίαρχο cluster διαρκεί περισσότερο από 10% (2.500 frames) του χρόνου της προσομοίωσης, τότε cluster analysis γίνεται στο 73.1% των πεπτιδίων (5.265).

B) Εάν το κυρίαρχο cluster διαρκεί περισσότερο από 20% (5.000 frames) του χρόνου της προσομοίωσης, τότε cluster analysis γίνεται στο 46.5% των πεπτιδίων (3.349).

Γ) Εάν το κυρίαρχο cluster διαρκεί περισσότερο από 30% (7.500 frames) του χρόνου της προσομοίωσης, τότε cluster analysis γίνεται στο 22.9% των πεπτιδίων (1.649).

Δ) Εάν το κυρίαρχο cluster διαρκεί περισσότερο από 40% (10.000 frames) του χρόνου της προσομοίωσης, τότε cluster analysis γίνεται στο 8.4% των πεπτιδίων (609).

Η λίστα των 480 πενταπεπτιδίων με την υψηλότερη βαθμολογία διαφοροποιείται αρκετά ανάλογα με το επιλεγμένο κατώφλι, με την τέταρτη επιλογή (κατώφλι 40%) να δίνει την καλύτερη κατάταξη των πενταπεπτιδίων (*top480B*) με βάση τις ατομικές αποστάσεις (συνάρτηση TF2). Ωστόσο και τα υπόλοιπα 129 πενταπεπτίδια από το σύνολο των 609, έχουν εξίσου ενδιαφέρον με βάση την εξέλιξη των ατομικών αποστάσεων.

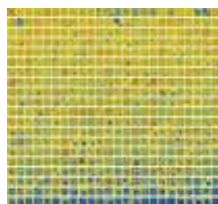
Δεδομένου ότι έχουν ήδη προηγηθεί οι προσομοιώσεις διάρκειας 2 μηνών στα 480 πενταπεπτίδια της λίστας *top480A*, δεν υπάρχει επαρκής φυσικός χρόνος για προσομοιώσεις επιπλέον 609 πεπτιδίων. Ο πλεονασμός αυτός όμως μπορεί να είναι και μη αναγκαίος. Αν συγκρίνουμε τις λίστες των πεπτιδίων, *top480A* και *top480B*, που προκύπτουν με τις δύο τιμές cut-off των περιπτώσεων A και Δ αντίστοιχα, βλέπουμε ότι τα 172 πενταπεπτίδια είναι κοινά. Επομένως απομένουν 437 πεπτίδια τα οποία θα πρέπει ενδεχομένως να εξεταστούν περαιτέρω, αφού εξεταστούν τα αποτελέσματα των πενταπεπτιδίων της λίστας *top480A*, των οποίων οι προσομοιώσεις έχουν ήδη πραγματοποιηθεί.

*"There's a common myth  
that evidence speaks for itself.  
It doesn't.  
It just sits there, on the lab table,  
incapable of speaking."*

<http://www.gdargaud.net/Humor/QuotesScience.html>

*"Love is a matter of chemistry,  
but sex is a matter of physics."*

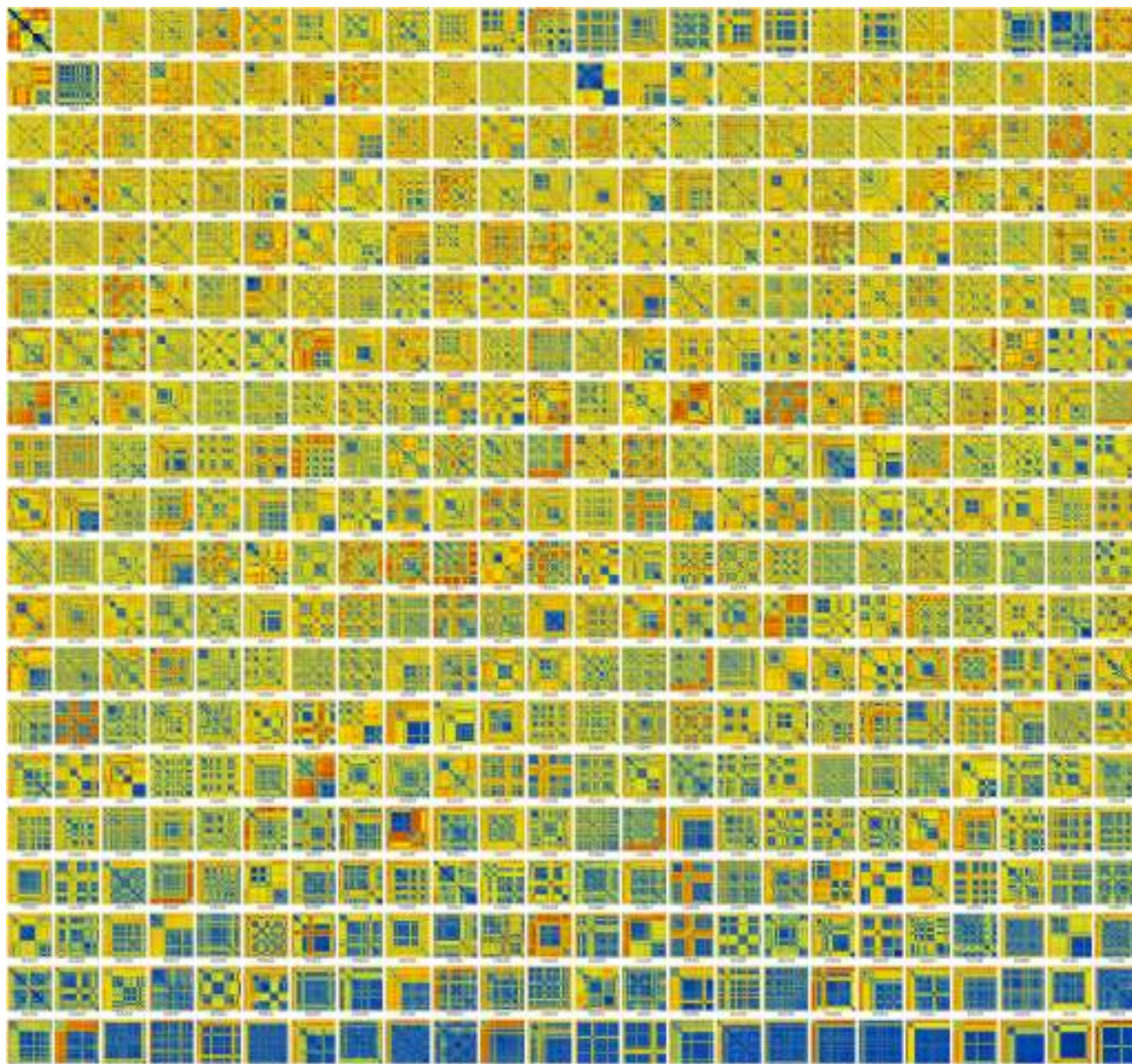
<http://www.gdargaud.net/Humor/QuotesScience.html>



## 4.4 Επιλογή 32 υποψήφιων δυνητικά αναδιπλούμενων πενταπεπτιδίων

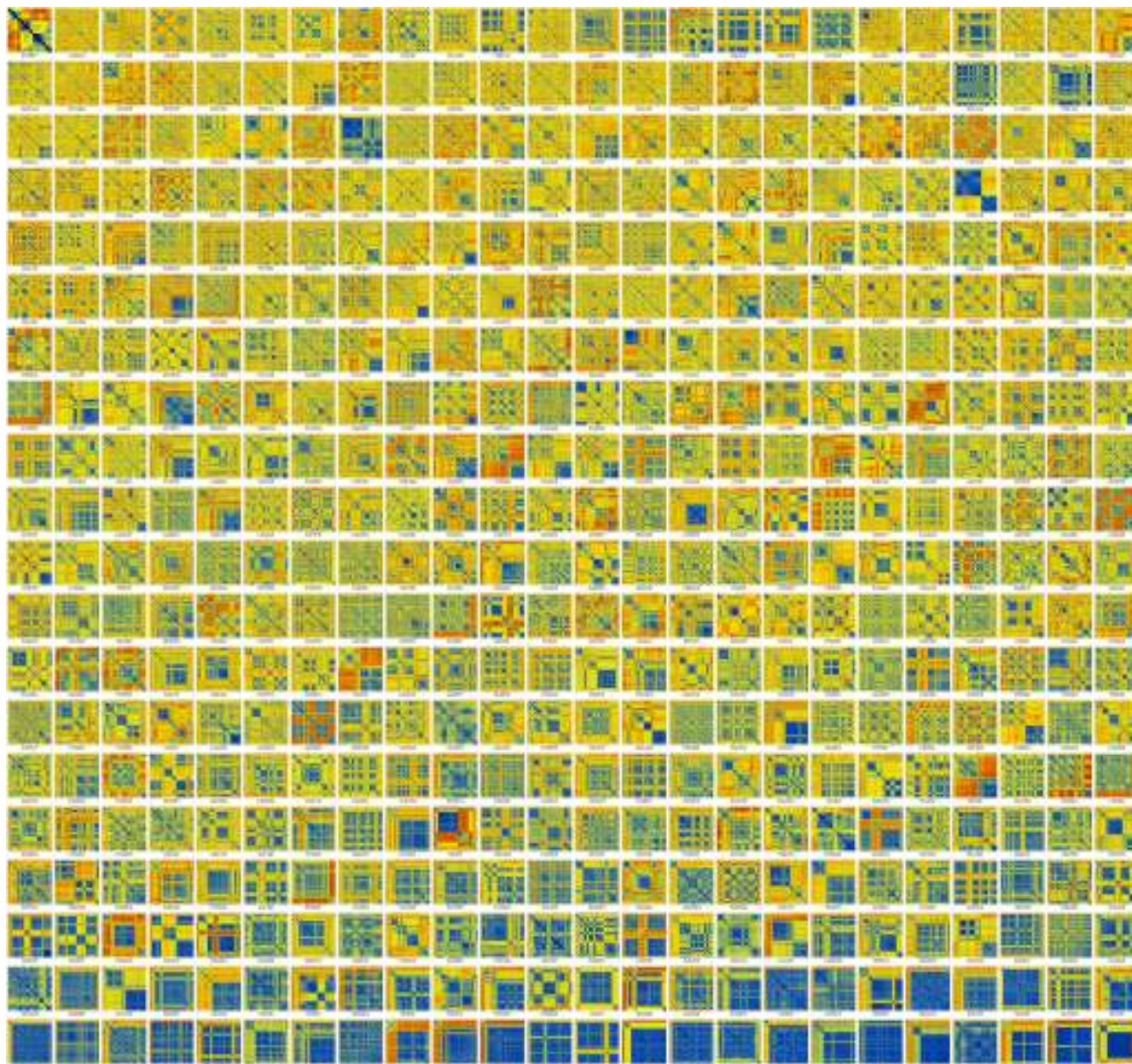
Τα 480 πενταπεπτίδια (*top480A*) που προέκυψαν από την απευθείας εφαρμογή της συνάρτησης TF2 των ατομικών αποστάσεων μελετήθηκαν με προσομοιώσεις διάρκειας 100ns οι οποίες διήρκεσαν 2 μήνες φυσικού χρόνου, απασχολώντας το 100% της συστοιχίας των υπολογιστών. Το πρωτόκολλο της προσομοίωσης είναι πανομοιότυπο με αυτό που βρίσκεται στο Παράρτημα (#13, NAMD script, all.namd) με μοναδική διαφορά τον τελικό αριθμό βημάτων (run -> 50.000.000 steps).

Για το σύνολο των πεπτιδίων αυτών έχουμε στη διάθεσή μας τους πίνακες RMSD (Εικόνα 4.4) και συνεπώς μπορούμε να προχωρήσουμε απευθείας στην εφαρμογή της συνάρτησης TF3, η οποία εμπεριέχει και τις ατομικές διακυμάνσεις της πλευρικής ομάδας της τρυπτοφάνης (Εικόνα 4.5). Για του λόγου το αληθές, στις Εικόνες 4.4 και 4.5 που ακολουθούν βλέπουμε τους πίνακες RMSD και για τα 480 πενταπεπτίδια, σε σειρά κατάταξης με βάση τη βαθμολογία τους (αλγόριθμος των "επεκτεινομένων παραθύρων") αλλά και με βάση τη βαθμολογία της συνάρτησης TF3 αντίστοιχα. Το επόμενο λοιπόν βήμα, με βάση την πορεία που ακολουθήσαμε στα τετραπεπτίδια, είναι να επιλέξουμε τα 32 πεπτίδια με την καλύτερη βαθμολογία με βάση τη συνάρτηση TF3 και να προχωρήσουμε σε προσομοιώσεις μεγαλύτερης διάρκειας, με την προϋπόθεση ότι έχουμε στη διάθεση μας ολόκληρη τη συστοιχία των υπολογιστών (32 πυρήνες). Όπως αναφέρθηκε στην προηγούμενη ενότητα, δημιουργήσαμε δύο λίστες, τις *top480A* και *top480B*, οι οποίες έχουν κοινό μόλις το 36% των πεπτιδίων.



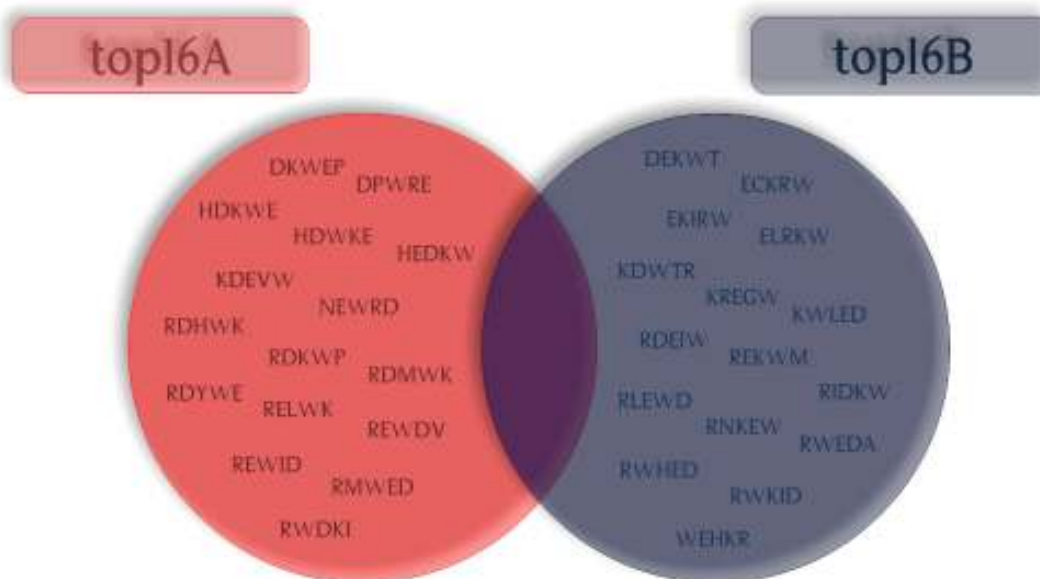
Εικόνα 4.4 Γραφική απεικόνιση των διδιάστατων πινάκων RMSD, χρησιμοποιώντας όλα τα βαριά άτομα για τον υπολογισμό. Η χρωματική κλίμακα είναι η ίδια και κυμαίνεται από σκούρο μπλε (0Å) μέχρι κόκκινο (μέγιστο rmsd 7.3Å). Η σειρά των γραφημάτων από αριστερά προς τα δεξιά (24) και από πάνω προς τα κάτω (20), όπως γίνεται η ανάγνωση κειμένου, ακολουθεί τη βαθμολογική κατάταξη με βάση τη βαθμολογία του πίνακα RMSD (αλγόριθμος των “επεκτεινομένων παραθύρων”).





Εικόνα 4.5 Γραφική απεικόνιση των διδιάστατων πινάκων RMSD, χρησιμοποιώντας όλα τα βαριά άτομα για τον υπολογισμό. Η χρωματική κλίμακα είναι η ίδια και κυμαίνεται από σκούρο μπλε (0Å) μέχρι κόκκινο (μέγιστο rmsd 7.3Å). Η σειρά των γραφημάτων από αριστερά προς τα δεξιά (24) και από πάνω προς τα κάτω (20), όπως γίνεται η ανάγνωση κειμένου, ακολουθεί τη βαθμολογική κατάταξη με βάση τη βαθμολογία της συνάρτησης TF3.

Για το λόγο αυτό προτιμήσαμε να επιλέξουμε τα 16 καλύτερα πεπτίδια της λίστας *top480A* (Εικόνα 4.5) από τις προσομοιώσεις διάρκειας 100ns (*top16A*) και τα υπόλοιπα 16 (*top16B*) επιλέχθηκαν με οπτική εξέταση των γραφικών παραστάσεων των ατομικών αποστάσεων από το σύνολο των 437 πενταπεπτιδίων, τα οποία είναι αυτά για τα οποία το κυρίαρχο cluster δομών διαρκεί περισσότερο από 40% του χρόνου της προσομοίωσης και δεν εμπεριέχονται στην λίστα *top480A* (Εικόνα 4.6). Από τα πενταπεπτίδια της λίστας *top16A* μόνο τα 7 πεπτίδια (DKWEP, HDWKE, HEDKW, NEWRD, RDKWP, RELWK, RWDKI) υπάρχουν στο σύνολο της περίπτωσης Δ, ενώ για 9 από αυτά το κυρίαρχο cluster διαρκεί λιγότερο. Πιο συγκεκριμένα τα πεπτίδια DPWRE, HDKWE, KDEVW, RDMWK, REWID ανήκουν στο σύνολο της περίπτωσης Γ, όπου το κυρίαρχο cluster διαρκεί για τουλάχιστον 30% του χρόνου της προσομοίωσης. Τα πεπτίδια RDYWE, REWDV, RMWED ανήκουν στο σύνολο της περίπτωσης Β, όπου το κυρίαρχο cluster διαρκεί για τουλάχιστον 20% του χρόνου της προσομοίωσης, ενώ το πεπτίδιο RDHWK ανήκει στο σύνολο της περίπτωσης Α, όπου το κυρίαρχο cluster διαρκεί για τουλάχιστον 10% του χρόνου της προσομοίωσης. Περαιτέρω μελέτη των 32 αυτών πεπτιδίων θα μας δείξει ποια από τις δύο πορείες, που οδήγησαν στις δύο λίστες της Εικόνας 4.6, είναι περισσότερο αποτελεσματική στην ανάδειξη αναδιπλούμενων πεπτιδίων.



Εικόνα 4.6 Διάγραμμα Venn των 32 πενταπεπτιδίων που επιλέχθηκαν για τον επόμενο κύκλο προσομοιώσεων.

*"Evolution is cleverer than you are."*

*Francis Crick*



*“Nothing exists except atoms and empty space; everything else is opinion.”*

*Democritus*



## 4.5 Μελέτη 32 πενταπεπτιδίων με τέσσερα force fields

Τα 32 πενταπεπτίδια (*top16A + top16B*) που προέκυψαν με τον τρόπο που αναλύθηκε στην προηγούμενη ενότητα, μελετήθηκαν με προσομοιώσεις διάρκειας 120ns. Την περίοδο αυτή, η σύγχρονη βιβλιογραφία στράφηκε προς τη συγκριτική μελέτη των διαφόρων force fields αναφορικά με την απόδοση τους στη μελέτη της αναδίπλωσης μικρών πρωτεϊνών και πεπτιδίων (Aliev et al., 2010, Best et al., 2008, Lange et al., 2010) (Ενότητα 3.7). Ακολουθώντας το ρεύμα αυτό, πραγματοποιήσαμε 4 ανεξάρτητες προσομοιώσεις για κάθε ένα από τα 32 πενταπεπτίδια χρησιμοποιώντας τα πιο διαδεδομένα force fields της εποχής, CHARMM-CMAP (MacKerell et al., 2004, Buck et al., 2006), OPLS-AA (Jorgensen et al., 1996, Kaminski et al., 2001), AMBER99SB (Hornak, et al., 2006, Wickstrom et al., 2009) και AMBER99SB-ILDN (Lindorff-Larsen et al., 2010). Το υπολογιστικό τμήμα των προσομοιώσεων αυτών ολοκληρώθηκε σε ~51 μέρες φυσικού χρόνου απασχολώντας το 50% της συστοιχίας των υπολογιστών. Το πρωτόκολλο της προσομοίωσης είναι πανομοιότυπο με αυτό που βρίσκεται στο Παράρτημα (#13, NAMD script, all.namd) με τις ακόλουθες διαφορές:

### CHARMM-CMAP force field:

- parameters -> par\_all22\_prot.inp
- wrapAll -> on
- run -> 60.000.000

### OPLS-AA force field:

- parameters -> par\_opls\_aa\_modified.inp
- vdwGeometricSigma -> yes
- wrapAll -> on
- run -> 60.000.000

AMBER99SB force field & AMBER99SB-ILDN force field:

```

❖ amber -> yes
❖ readexclusions -> yes
❖ parmfile -> pentapept., rmtop
❖ ambercoor -> pentapept.inpcrd
❖ 1-4scaling -> 0.833333
❖ PmeGridsizeX -> 24
❖ PmeGridsizeY -> 24
❖ PmeGridsizeZ -> 24
❖ wrapAll -> on
❖ cellBasisVector1 -> 24.00 0.00 0.00
❖ cellBasisVector2 -> -8.00 22.62 0.00
❖ cellBasisVector3 -> -8.00 -11.31 -19.59
❖ langevinHydrogen -> off
❖ run -> 60.000.000

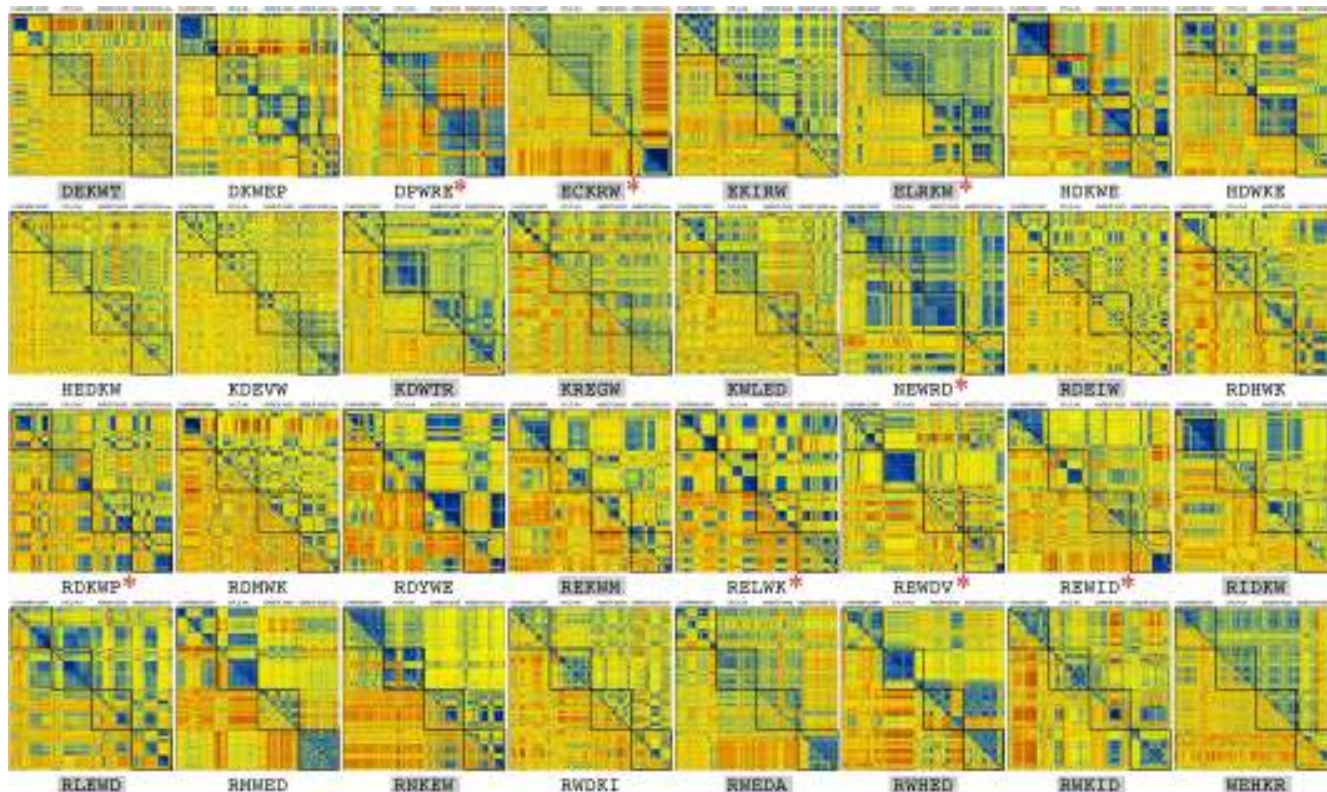
```

Η διεξαγωγή των προσομοιώσεων μέσω του προγράμματος NAMD για τα force fields CHARMM22, CHARMM-CMAP και OPLS-AA είναι παρόμοια με εξαίρεση φυσικά τα αρχεία ανάγνωσης (input files) της τοπολογίας και των παραμέτρων που εξαρτώνται από το εκάστοτε force field. Ειδικά για το force field OPLS χρειάζεται ο ορισμός μίας επιπλέον παραμέτρου στο πεδίο των basic dynamics (vdwGeometricSigma) που ορίζει τη χρήση γεωμετρικού αντί για αριθμητικό μέσο κατά τον συνδυασμό των Lennard-Jones sigma παραμέτρων για τους διάφορους ατομικούς τύπους.

Η χρήση των AMBER force field μέσω του προγράμματος NAMD είναι περισσότερο πολύπλοκη. Η προετοιμασία των συστημάτων γίνεται μέσω του προγράμματος LEAP (από το πακέτο προγραμμάτων AMBERTOOLS, Case et al., 2005) και τα αρχεία των παραμέτρων (pentapept.prmtop) και της τοπολογίας (pentapept.inpcrd) παράγονται εκ νέου για κάθε πεπτιδική αλληλουχία. Στο Παράρτημα (#20, systematic.AMBER-ildn.pl) παραθέτουμε το Perl script, κατάλληλα τροποποιημένο για τα AMBER force fields. Η επιλογή μεταξύ των διαφόρων εκδόσεων των AMBER force fields γίνεται στη γραμμή εντολής που γίνεται επίκληση του προγράμματος LEAP (Case et al., 2005). Οι υπόλοιπες παράμετροι που διαφοροποιούνται υποδεικνύονται παραπάνω.

Το κύριο μέλημα είναι κατά πόσο τα διάφορα force fields συγκλίνουν ή αποκλίνουν στις δομές που προβλέπουν για τις συγκεκριμένες πεπτιδικές αλληλουχίες. Ο πιο άμεσος και αποτελεσματικός τρόπος για να το εξετάσουμε αυτό, όπως έχουμε εξακριβώσει επανειλημμένως κατά την ανάλυση των αποτελεσμάτων των τετραπεπτιδίων, είναι να ενώσουμε τεχνητά τα

ανεξάρτητα τροχιακά και να υπολογίσουμε πίνακες RMSD χρησιμοποιώντας όλα τα βαριά άτομα. Στην Εικόνα 4.7 παρουσιάζουμε συγκριτικά (και σε αλφαβητική σειρά) τα αποτελέσματα και για τα 32 πενταπεπτίδια.



Εικόνα 4.7 Συνοπτική γραφική αναπαράσταση των πινάκων RMSD των 32 πενταπεπτιδίων (σε αλφαβητική σειρά). Τα τέσσερα ανεξάρτητα τροχιακά που προέκυψαν με τα τέσσερα force fields εσωκλείονται με τα μαύρα τετράγωνα επί της διαγωνίου και ακολουθούν τη σειρά CHARMM-CMAP, OPLS-AA, AMBER99SB, AMBER99SB-ildn. Σε κάθε πίνακα, βλέπουμε τον υπολογισμό τόσο για όλα τα βαριά άτομα (κάτω από τη διαγώνιο) όσο και μόνο για τα άτομα του πεπτιδικού σκελετού (πάνω από τη διαγώνιο). Η χρωματική κλίμακα κυμαίνεται από σκούρο μπλε (0Å) ως σκούρο κόκκινο (7.81Å). Με γκρι σκίαση δηλώνονται οι πεπτιδικές αλληλουχίες της λίστας *top16B* και με κόκκινο αστερίσκο (\*) τα πεπτίδια που επιλέχθηκαν για περαιτέρω προσομοιώσεις.

Το αξιοσημείωτο χαρακτηριστικό της Εικόνας 4.7 είναι η ποικιλομορφία των αποτελεσμάτων, καθώς βλέπουμε όλες τις περιπτώσεις, από πεπτίδια με πλήρη ασυμφωνία (REWID, RMWED, RWEDA) μεταξύ των force fields έως πεπτίδια με πλήρη συμφωνία (NEWRD, RELWK, RDKWP).

Για παράδειγμα, σε πεπτίδια όπως τα DEKWT, DKWEP, HDKWE, βλέπουμε τη δημιουργία δύο cluster δομών στο CHARMM-CMAP, τα οποία τα συναντάμε μεν αλλά πολύ σποραδικά στα άλλα force fields. Η πλειοψηφία των πεπτιδίων (EKIRW, HEDKW, KDEVW, KREGW, KWLED, RDEIW, REKWM, RLEWD, RWDKI, WEHKR) δείχνουν την αναμενόμενη συμπεριφορά για τόσο μικρού μήκους πεπτίδια, δείχνοντας μεγάλη αστάθεια και πολλαπλά και συχνά γεγονότα αναδίπλωσης/αποδιάταξης. Υπάρχουν επίσης πεπτίδια που δείχνουν δημιουργία δομής μόνο στα τροχιακά με τα δύο AMBER force fields, όπως τα DPWRE, ELRKW, RWKID.

Το πεπτίδιο ECKRW αποτελεί ιδιαίτερη περίπτωση καθώς δείχνει το σχηματισμό ενός πολύ συμπαγούς cluster δομών με το AMBER99SB-ILDN force field, ενώ δε σχηματίζεται κάποιος cluster σε κανένα από τα τροχιακά με τα άλλα force fields, ούτε καν με το AMBER99SB. Στην αλληλουχία του πεπτιδίου δε συναντάμε ωστόσο κανένα από τα κατάλοιπα I-L-D-N που να δικαιολογεί την απόκλιση αυτή, η οποία ίσως να πρέπει να αποδοθεί στον περιορισμένο χρόνο της προσομοίωσης.

Σε άλλα πεπτίδια (HDWKE) βλέπουμε συμφωνία μεταξύ των δομών που προβλέπονται από τα OPLS-AA-AMBER99SB-ILDN και CHARMM-CMAP-AMBER99SB αντίστοιχα. Σε μεμονωμένες περιπτώσεις προβλέπεται δημιουργία cluster δομών μόνο από το OPLS-AA (KDWTR, REWDV, RWHED) ή μόνο από το AMBER99SB-ILDN (RDHWK) ή μόνο από CHARMM-CMAP (RDMWK, RIDKW). Για το πεπτίδιο RNKEW βλέπουμε τις ίδιες δομές στα δύο τροχιακά με τα AMBER force fields, διαφορετική δομή στο τροχιακό με το OPLS και τη μη δημιουργία cluster στο τροχιακό με το CHARMM-CMAP. Παρόμοια εικόνα δείχνει και το πεπτίδιο RDYWE, με τη διαφορά ότι οι δομές του CHARMM-CMAP είναι πιο κοντά στις δομές που βλέπουμε με τα AMBER force fields.

Με βάση τις παραπάνω παρατηρήσεις επιλέξαμε 8 πεπτίδια (Εικόνα 4.8) που έδειξαν την περισσότερο ενδιαφέρουσα συμπεριφορά με βάση τα τέσσερα αυτά force fields (Εικόνα 4.7). Από αυτά, το 75% των πεπτιδίων ανήκει στη λίστα *top16A* και μόνο τα ECKRW και ELRKW ανήκουν στη λίστα *top16B*. Επίσης, 5 πεπτίδια (ECKRW, ELRKW, NEWRD, RDKWP, RELWK) ανήκουν στο κοινό τμήμα μεταξύ της λίστας *top480A* και της περίπτωσης Δ, όπου το κατώφλι για cluster analysis είναι 10.000 frames, ενώ τα πεπτίδια DPWRE, REWID ανήκουν στην περίπτωση Γ (κατώφλι 7.500 frames για cluster analysis) και το πεπτίδιο REWDV ανήκει στην περίπτωση Β (κατώφλι 5.000 frames για cluster analysis). Ωστόσο, το REWDV είναι το μοναδικό πεπτίδιο το οποίο σχηματίζει ένα μεγάλο και συμπαγές cluster δομών με το OPLS-AA force field,

το οποίο το συναντάμε και προς το τέλος της προσομοίωσης με το AMBER99SB-ILDN.

Βλέπουμε λοιπόν ότι η επέκταση του χρόνου της προσομοίωσης από 5ns στην περίπτωση των τετραπεπτιδίων σε 20ns στα πενταπεπτίδια βελτίωσε σαφώς τη διακριτική ικανότητα των βασισμένων σε ατομικές αποστάσεις συναρτήσεων εκτίμησης της αναδιπλωσιμότητας (TF2).



Εικόνα 4.8 Word-cloud των 8 πενταπεπτιδίων, όπου το μέγεθος της αλληλουχίας είναι ενδεικτικό της βαθμολογίας που έλαβε με βάση τον πίνακα RMSD για το τεχνητό τροχιακό που προέκυψε από την ένωση και των τεσσάρων ανεξάρτητων τροχιακών με τα τέσσερα force fields. Το τέχνασμα της ένωσης των τροχιακών και της βαθμολόγησής τους δίνει μία αίσθηση του μέτρου της συμφωνίας (ή ασυμφωνίας) στις προβλέψεις των τεσσάρων force fields για κάθε πεπτίδιο.

Για τα πεπτίδια της λίστας *top16A* άλλωστε έχουμε στη διάθεση μας και αποτελέσματα από ένα πέμπτο force field, το CHARMM22 με το οποίο έγιναν οι προσομοιώσεις διάρκειας 100ns (Ενότητα 4.4). Για τις προσομοιώσεις αυτές όμως δεν έχουμε (πλέον) τα ανεπεξέργαστα δεδομένα (DCD αρχεία) των τροχιακών, αλλά μόνο τους πίνακες RMSD (υπολογισμένοι για όλα τα βαριά άτομα) και ένα σύνολο αρχείων PDB με στιγμιότυπα του κυρίαρχου cluster δομών (Cartesian-PCA, heavy atoms). Προκειμένου να πραγματοποιήσουμε τη σύγκριση ακολουθήσαμε τα παρακάτω βήματα:





Εικόνα 4.9 Σύγκριση των τεσσάρων force fields των 16 πενταπεπτιδίων της λίστας top16A με το CHARM22, με το οποίο διεξήχθησαν οι προσομοιώσεις στο στάδιο των top480A.



- ★ Δημιουργία τεχνητού τροχιακού (DCD + PSF) από το σύνολο PDB δομών του cluster. Έτσι χρησιμοποιούμε όλη τη διαθέσιμη πληροφορία του cluster και όχι μόνο μία μέση δομή.
- ★ Υπολογισμός ενός πίνακα RMSD (intra-DCD) μεταξύ του τεχνητού τροχιακού που ανταποκρίνεται στο κυρίαρχο cluster και του συνενωμένου τεχνητού τροχιακού των τεσσάρων force fields. Ο πίνακας αυτός δεν είναι πλέον τετράγωνος αλλά έχει διάσταση 2.400x867. Η διάσταση 867 αντιστοιχεί στο τροχιακό του cluster και η ακριβής διάσταση εξαρτάται από το σύνολο δομών που αποθηκεύτηκαν για το cluster (το οποίο έχουμε ορίσει να εξαρτάται από την κατοχή του σε χρόνο προσομοίωσης καθώς οι δομές παράγονται με σταθερό βήμα μέσω του Perl script).
- ★ Ο ασύμμετρος αυτός πίνακας RMSD υφίσταται περαιτέρω επεξεργασία (ΠΑΡΑΡΤΗΜΑ, #21, find\_min.pl) δίνοντας ένα μονοδιάστατο αρχείο 2.400 γραμμών, όπου κάθε τιμή είναι η ελάχιστη τιμή που βρέθηκε στην αντίστοιχη γραμμή του πίνακα RMSD.

Στην Εικόνα 4.9 βλέπουμε την οπτική παρουσίαση των αποτελεσμάτων. Το CHARMM22 φαίνεται να δίνει συμβατά αποτελέσματα με όλα τα force fields κατά περίπτωση. Υπάρχουν περιπτώσεις που βλέπουμε τις ίδιες δομές μεταξύ μόνο CHARMM22 και CHARMM-CMAP (DKWE, RDMWK, RELWK\*, REWDV\*). Σε πεπτίδια (HDKWE, HDWKE, RDKWP\*) στα οποία βλέπουμε δημιουργία σταθερής δομής μόνο με το CHARMM-CMAP αλλά όχι με τα υπόλοιπα 3 force fields, το CHARMM22 συμφωνεί με τα τρία force fields και όχι με το CHARMM-CMAP. Πεπτίδια εξαιρετικά ασταθή και με τα 4 force fields (HEDKW, KDEVW, RWDKI) επιδεικνύουν εξίσου αστάθεια και με το CHARMM22. Ωστόσο υπάρχουν και οι αξιοσημείωτες περιπτώσεις (DPWRE\*, REWID\*, RMWED) όπου το CHARMM22 συμφωνεί με τα AMBER force fields χωρίς να συμφωνεί με το CHARMM-CMAP. Για ένα από τα καλύτερα πεπτίδια (NEWRD) κατ'ομοφωνία και των τεσσάρων force fields, το CHARMM22 δείχνει πλήρη ασυμφωνία.

Βλέπουμε λοιπόν πως υπάρχει μεγάλη ποικιλομορφία στις προβλέψεις τόσο μεταξύ των οικογενειών των force fields όσο και μεταξύ διαφορετικών εκδόσεων της ίδιας οικογένειας (CHARMM22 – CHARMM-CMAP και AMBER99SB – AMBER99SB-ILDN). Ως αιτία της διαφοροποίησης αυτής μπορεί να θεωρηθεί ο περιορισμένος χρόνος της προσομοίωσης, η αδυναμία των force fields για τα πεπτίδια εξαιρετικά μικρού μήκους της μελέτης μας ή ο συνδυασμός και των δύο. Σε κάθε περίπτωση οι τυχόν αδυναμίες των force fields μπορούν να αποκαλυφθούν μόνο κατόπιν σύγκρισης με πειραματικά δεδομένα.

*“I do not fear computers. I fear the lack of them.”*

*Isaac Asimov*

*“Physics is like sex: sure, it may give some practical results, but that’s not why we do it.”*

*Richard Feynman*

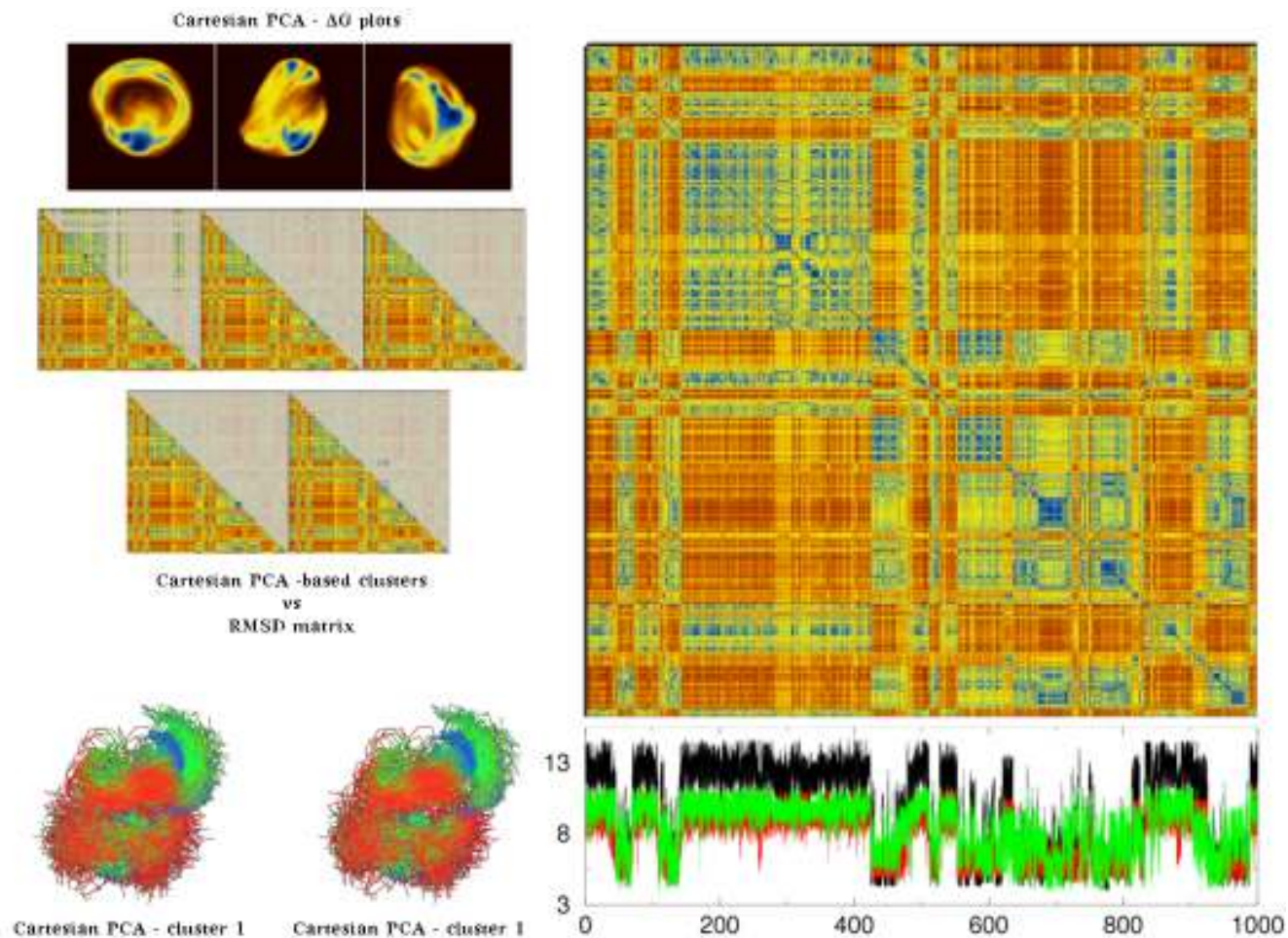


## 4.6 Μελέτη της αναδίπλωσης 8 πενταπεπτιδίων με το AMBER99SB-ILDN force field

Τα 8 πενταπεπτίδια που προέκυψαν από τη συγκριτική ανάλυση των προσομοιώσεων με τέσσερα force fields, μελετήθηκαν με προσομοιώσεις διάρκειας 1μs με το force field AMBER99SB-ILDN (Lindorff-Larsen et al., 2010). Το υπολογιστικό τμήμα των προσομοιώσεων αυτών ολοκληρώθηκε σε ~11 μέρες φυσικού χρόνου απασχολώντας το 50% της συστοιχίας των υπολογιστών. Το πρωτόκολλο της προσομοίωσης είναι πανομοιότυπο με αυτό που βρίσκεται στο Παράρτημα (#13, NAMD script, all.namd) προσαρμοσμένο για το AMBER99SB-ILDN force field όπως αναφέρεται στην Ενότητα 4.5.

Η μεγαλύτερη διάρκεια των προσομοιώσεων αυτών κατέδειξε την αστάθεια των περισσότερων πεπτιδίων, τουλάχιστον με βάση τις προβλέψεις του force field AMBER99SB-ILDN. Στις Εικόνες 4.10 – 4.18 που ακολουθούν συνοψίζονται τα κυριότερα αποτελέσματα (σε αλφαβητική σειρά) που στηρίζουν τη θέση αυτή.

Το πεπτιδίο DPWRE (Εικόνα 4.10) φαίνεται αρκετά ασταθές: το κυρίαρχο cluster διαρκεί μόλις 26% του χρόνου προσομοίωσης και οι δομές που ανήκουν σε αυτό έχουν μεγάλες ατομικές διακυμάνσεις (μέση τιμή RMSF για τα άτομα των πλευρικών ομάδων 2.5Å) καθώς το cluster εμφανίζει μεγάλη διασπορά. Οι κύριες αλληλεπιδράσεις που παρατηρούνται είναι μεταξύ των πλευρικών ομάδων της τρυπτοφάνης με του γλουταμικού οξέος και μεταξύ των πλευρικών ομάδων της αργινίνης με του ασπαρτικού οξέος. Ο πεπτιδικός σκελετός παραμένει σε σχετικά ανοιχτή διαμόρφωση (~8Å) όπως φαίνεται από την εξέλιξη των ατομικών αποστάσεων.

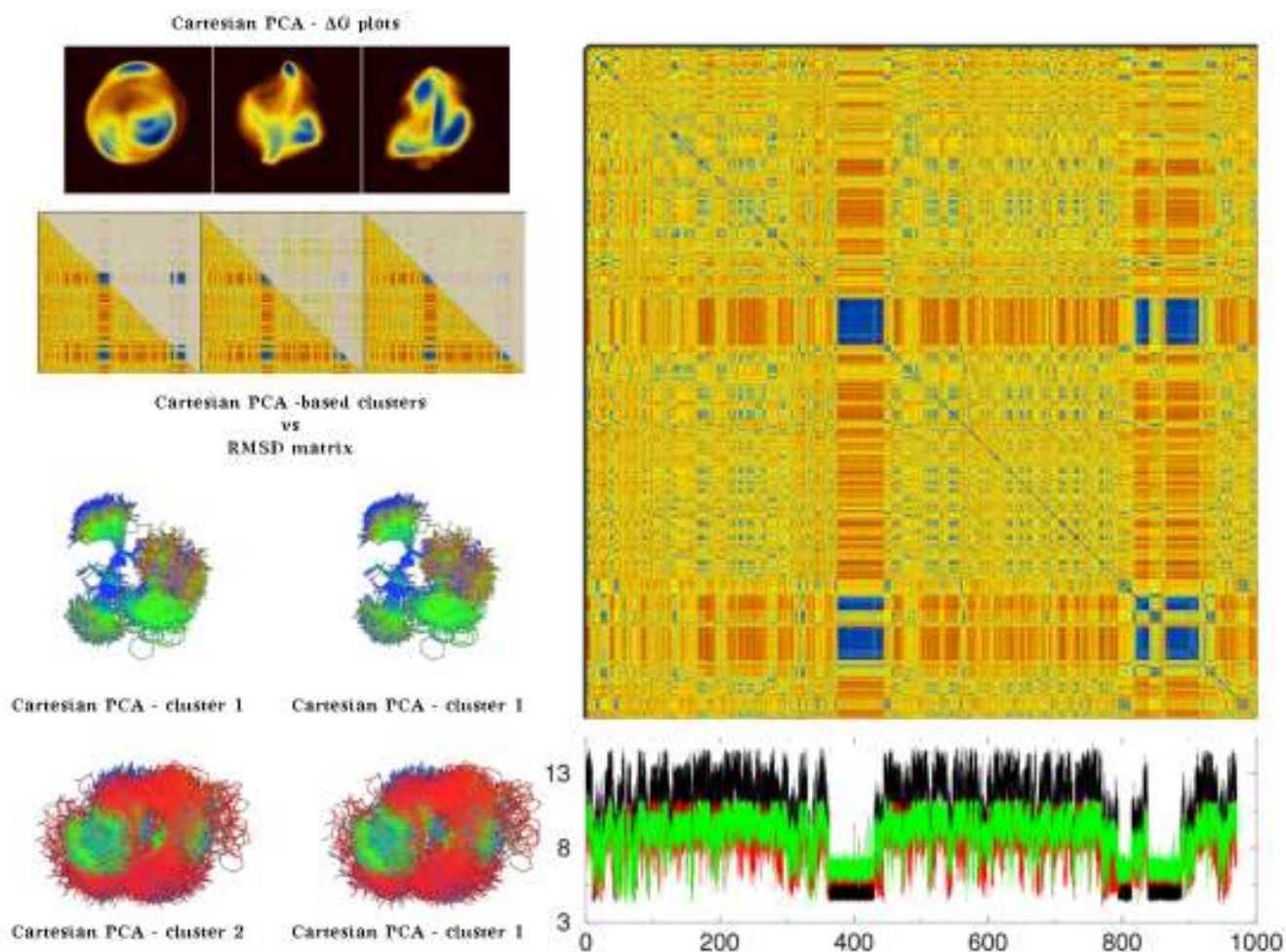


Εικόνα 4.10 Συγκεντρωτικά αποτελέσματα για το πεπτίδιο DPWRE. Από αριστερά προς τα δεξιά βλέπουμε: την αντιπροσωπευτική δομή του κυρίαρχου cluster με βάση την ανάλυση Cartesian-PCA (stereo αναπαράσταση), τα ενεργειακά τοπία ( $\Delta G$  energy plots) της προβολής του τροχιακού στους τρεις principal components, την προβολή όλων των cluster που προέκυψαν από Cartesian-PCA πάνω στον πίνακα RMSD (με RMSD cut-off 1.9 και variance-explained 0.84), τον πίνακα RMSD με βάση όλα τα βαριά άτομα (μέγιστη τιμή RMSD  $7.5\text{\AA}$ ) και τις τρεις αποστάσεις μεταξύ ατόμων Ca 1-5 (μαύρο χρώμα), 1-4 (κόκκινο χρώμα), 2-4 (πράσινο χρώμα). Ο χρωματισμός των δομών έγινε με βάση τις τιμές RMSF από μπλε σε κόκκινο (μέγιστη τιμή RMSF  $3.46\text{\AA}$ ). Για τα cluster με διάρκεια μικρότερη από 10% του χρόνου προσομοίωσης δεν παρουσιάζονται αντιπροσωπευτικές δομές.

Το πεπτίδιο ECKRW (Εικόνα 4.11) σχηματίζει ένα ιδιαίτερα συμπαγές cluster αλλά με κατοχή σε χρόνο προσομοίωσης μόλις 10% και ένα δεύτερο cluster με κατοχή 17% του οποίου τα frames είναι διάσπαρτα σε όλο το τροχιακό. Οι δομές του πρώτου cluster έχουν χαμηλές διακυμάνσεις



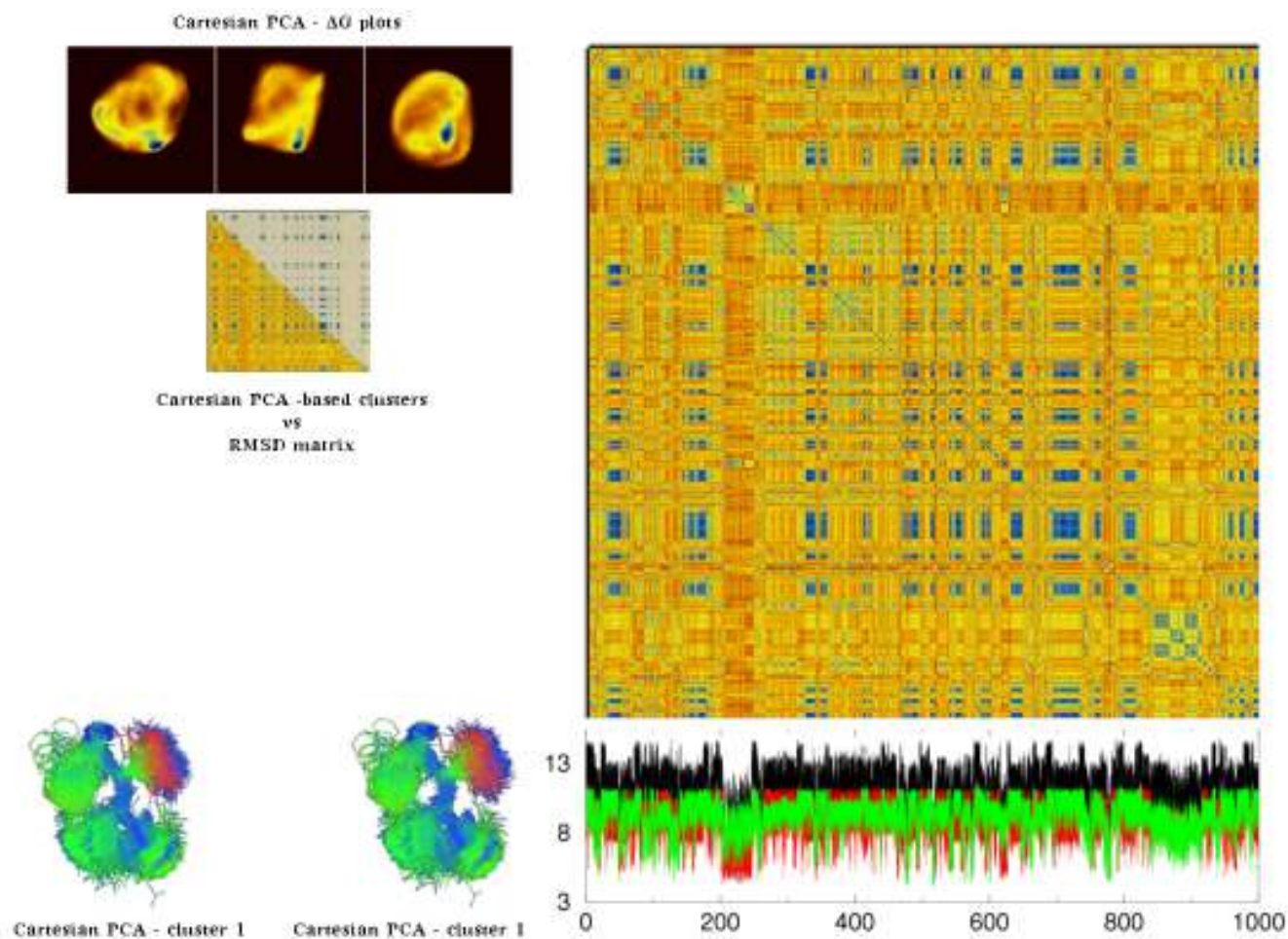
(μέση τιμή RMSF για όλα τα βαριά άτομα 1.1Å). Ο πεπτιδικός σκελετός παίρνει δομή θηλιάς η οποία σταθεροποιείται από αλληλεπιδράσεις μεταξύ των ελεύθερων φορτισμένων άκρων, παράλληλα με αλληλεπιδράσεις μεταξύ των πλευρικών ομάδων (Trp-Arg και Cys-Glu). Το δεύτερο cluster στην ουσία περιλαμβάνει ένα σύνολο δομών με τον πεπτιδικό σκελετό σε ανοιχτή διαμόρφωση (μέση τιμή RMSF για τα άτομα του πεπτιδικού σκελετού 0.9Å) και τις πλευρικές ομάδες μη σταθεροποιημένες (μέση τιμή RMSF για τα άτομα των πλευρικών ομάδων 3.6Å).



Εικόνα 4.11 Συγκεντρωτικά αποτελέσματα για το πεπτίδιο ECKRW. Από αριστερά προς τα δεξιά βλέπουμε: την αντιπροσωπευτική δομή των δύο κυρίαρχων cluster με βάση την ανάλυση Cartesian-PCA (stereo αναπαράσταση), τα ενεργειακά τοπία ( $\Delta G$  energy plots) της προβολής του τροχιακού στους τρεις principal components, την προβολή όλων των cluster που προέκυψαν από Cartesian-PCA πάνω στον πίνακα RMSD (με RMSD cut-off 2.4 και variance-explained 0.77), τον πίνακα RMSD με βάση όλα τα βαριά άτομα (μέγιστη τιμή RMSD 7.5Å) και τις τρεις αποστάσεις μεταξύ ατόμων Ca 1-5 (μαύρο χρώμα), 1-4

(κόκκινο χρώμα), 2-4 (πράσινο χρώμα). Ο χρωματισμός των δομών έγινε με βάση τις τιμές RMSF από μπλε σε κόκκινο (μέγιστη τιμή RMSF 5.76Å). Για τα cluster με διάρκεια μικρότερη από 10% του χρόνου προσομοίωσης δεν παρουσιάζονται αντιπροσωπευτικές δομές.

Το πεπτίδιο ELRKW (Εικόνα 4.12) σχηματίζει ένα cluster με κατοχή 21% του χρόνου της προσομοίωσης. Οι δομές που ανήκουν στο cluster σταθεροποιούνται μέσω πολλών αλληλεπιδράσεων μεταξύ των πλευρικών ομάδων των Trp-Arg-Glu (μέση τιμή RMSF για τα άτομα των πλευρικών ομάδων 1.7Å) που βρίσκονται όλες στην ίδια πλευρά του πεπτιδικού σκελετού, ο οποίος παραμένει σε εκτεταμένη διαμόρφωση. Οι δομές αυτές σχηματίζονται πολύ νωρίς στο τροχιακό αλλά έχουν πολύ μικρή διάρκεια και επανεμφανίζονται πολλές φορές μέχρι και το τέλος της διάρκειας του τροχιακού.

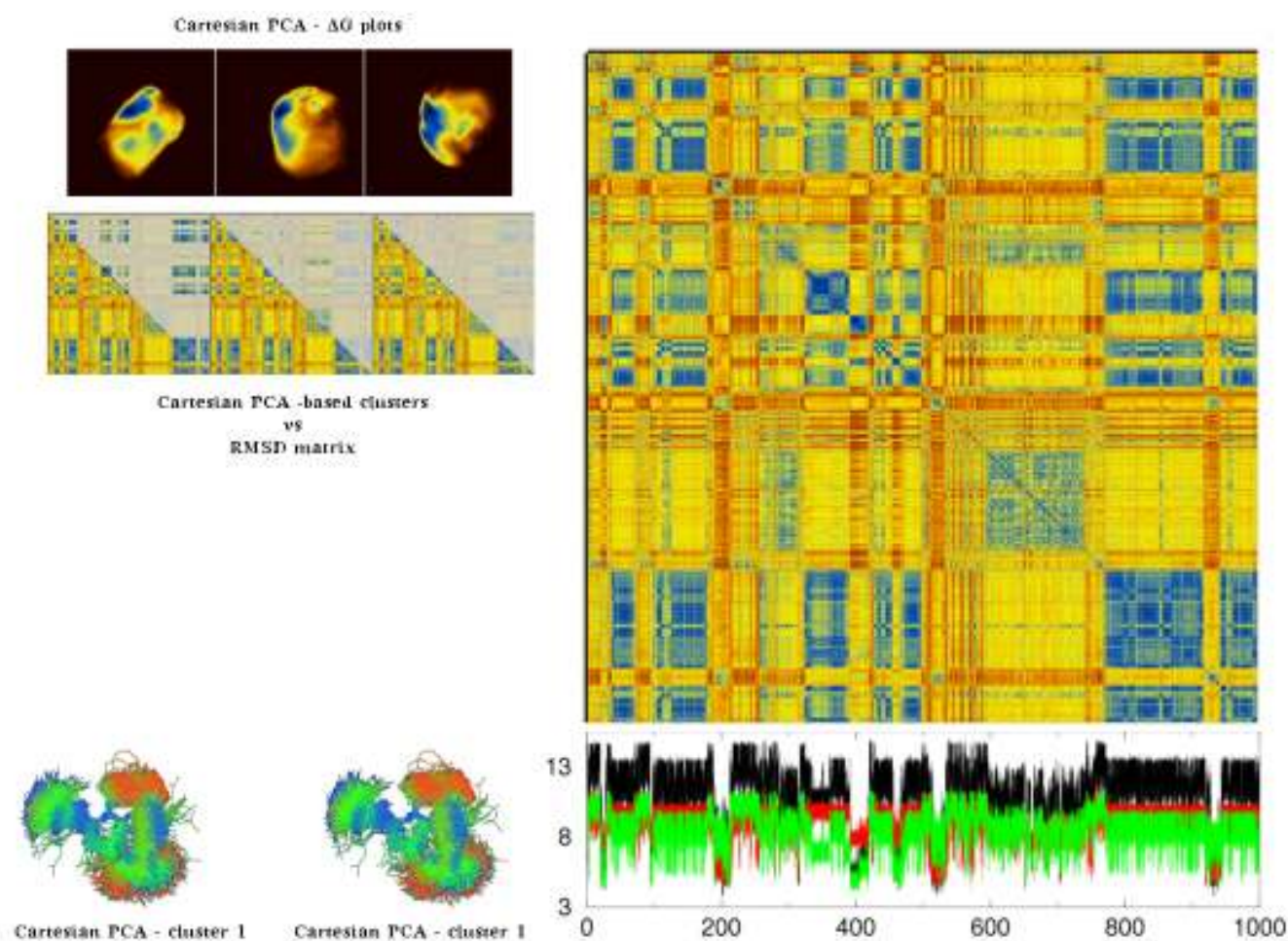


Εικόνα 4.12 Συγκεντρωτικά αποτελέσματα για το πεπτίδιο ELRKW. Από αριστερά προς τα δεξιά



βλέπουμε: την αντιπροσωπευτική δομή του κυρίαρχου cluster με βάση την ανάλυση Cartesian-PCA (stereo αναπαράσταση), τα ενεργειακά τοπία ( $\Delta G$  energy plots) της προβολής του τροχιακού στους τρεις principal components, την προβολή όλων των cluster που προέκυψαν από Cartesian-PCA πάνω στον πίνακα RMSD (με RMSD cut-off 2.7 και variance-explained 0.93), τον πίνακα RMSD με βάση όλα τα βαριά άτομα (μέγιστη τιμή RMSD 7.5Å) και τις τρεις αποστάσεις μεταξύ ατόμων Ca 1-5 (μαύρο χρώμα), 1-4 (κόκκινο χρώμα), 2-4 (πράσινο χρώμα). Ο χρωματισμός των δομών έγινε με βάση τις τιμές RMSF από μπλε σε κόκκινο (μέγιστη τιμή RMSF 3.65Å).

Το πεπτίδιο NEWRD (Εικόνα 4.13) σχηματίζει ένα κυρίαρχο cluster με κατοχή σε χρόνο προσομοίωσης 29%. Οι δομές του cluster σταθεροποιούνται μέσω αλληλεπίδρασης των αντίθετα φορτισμένων καταλοίπων Arg-Glu και το πακετάρισμα της Trp πάνω από αυτά (μέση τιμή RMSF για τα άτομα των πλευρικών ομάδων 1.9Å).



Εικόνα 4.13 Συγκεντρωτικά αποτελέσματα για το πεπτίδιο NEWRD. Από αριστερά προς τα δεξιά

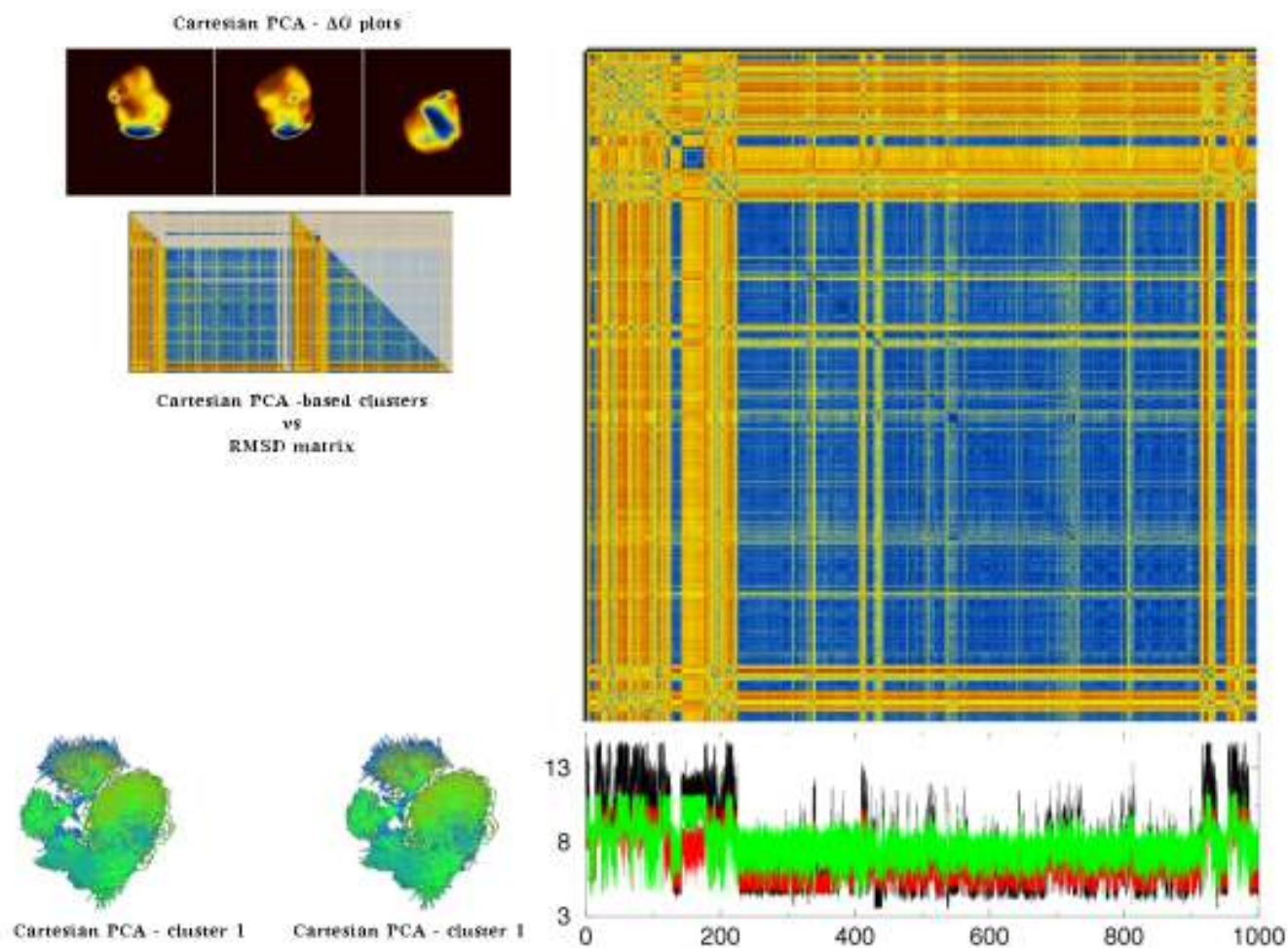


βλέπουμε: την αντιπροσωπευτική δομή του κυρίαρχου cluster με βάση την ανάλυση Cartesian-PCA (stereo αναπαράσταση), τα ενεργειακά τοπία ( $\Delta G$  energy plots) της προβολής του τροχιακού στους τρεις principal components, την προβολή όλων των cluster που προέκυψαν από Cartesian-PCA πάνω στον πίνακα RMSD (με RMSD cut-off 2.0 και variance-explained 0.83), τον πίνακα RMSD με βάση όλα τα βαριά άτομα (μέγιστη τιμή RMSD 7.5Å) και τις τρεις αποστάσεις μεταξύ ατόμων Ca 1-5 (μαύρο χρώμα), 1-4 (κόκκινο χρώμα), 2-4 (πράσινο χρώμα). Ο χρωματισμός των δομών έγινε με βάση τις τιμές RMSF από μπλε σε κόκκινο (μέγιστη τιμή RMSF 3.51Å). Για τα cluster με διάρκεια μικρότερη από 10% του χρόνου προσομοίωσης δεν παρουσιάζονται αντιπροσωπευτικές δομές.

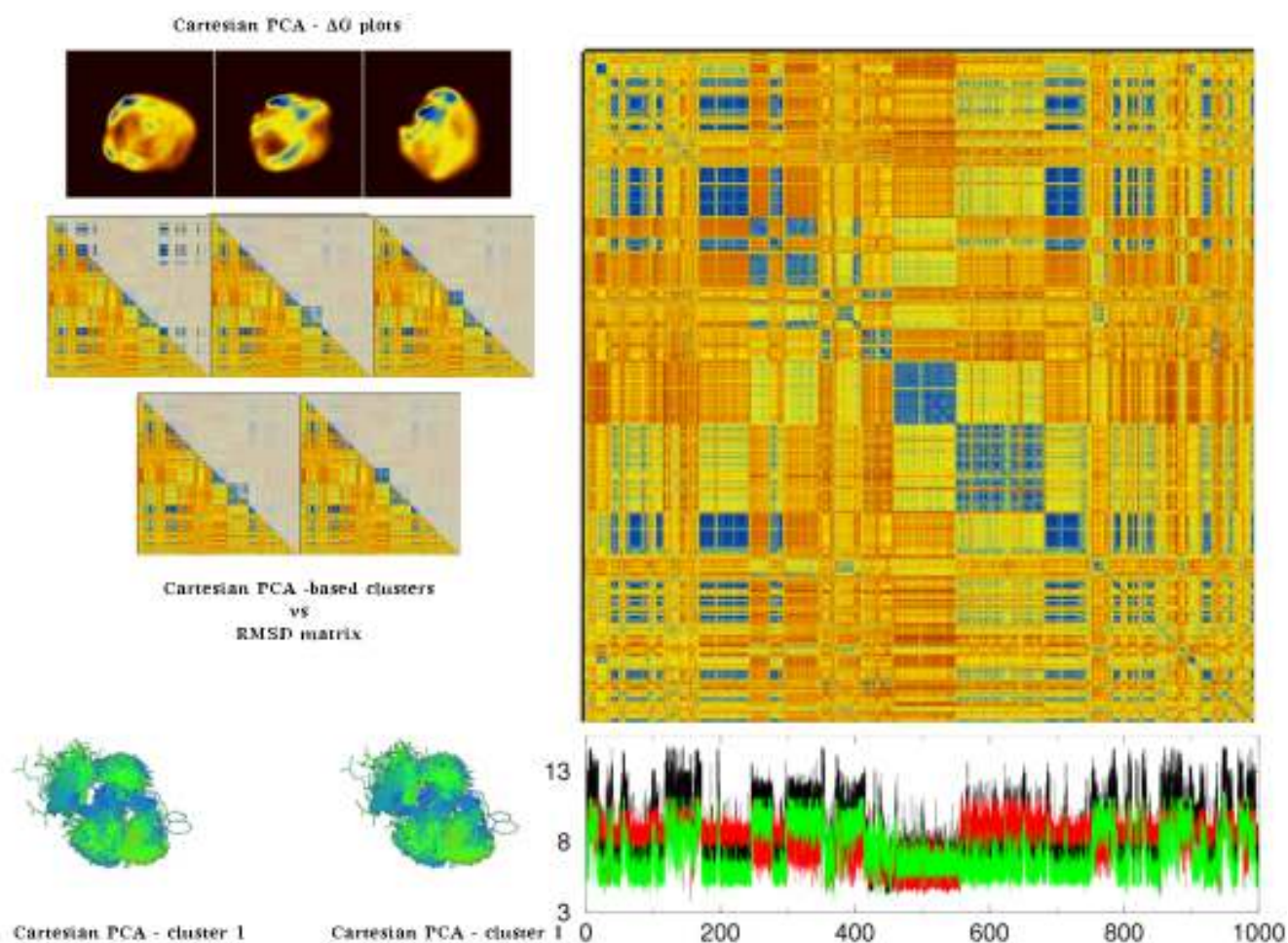
Παρά την πρώτη εντύπωση που δίνει ο πίνακας RMSD ότι το κυρίαρχο cluster έχει αξιοσημείωτη διάρκεια, πιο προσεκτική εξέταση μας επιτρέπει να δούμε τόσο τη διασπορά των frames όσο και το γεγονός ότι οι δομές δεν είναι τόσο εγγύς μεταξύ τους (cross-vectors εκτός της διαγωνίου με μπλε-κίτρινο χρώμα που συνδέουν τα cluster που φαίνονται ως μπλε τετράγωνα επί της διαγωνίου).

Το πεπτίδιο RDKWP (Εικόνα 4.14) σχηματίζει ένα cluster με κατοχή σε χρόνο προσομοίωσης 58% ολόκληρου του τροχιακού. Οι δομές του cluster εμφανίζονται πολύ νωρίς στην αρχή, και μετά από μία ενδιάμεση παροδική εμφάνιση, παραμένουν μέχρι και το τέλος της προσομοίωσης. Πρόκειται για δομές αρκετά κοντινές μεταξύ τους (μέση τιμή RMSF για τα άτομα των πλευρικών ομάδων 1.6Å και 0.6Å για τα άτομα του πεπτιδικού σκελετού) που προκύπτουν από την αλληλεπίδραση της πλευρικής ομάδας της αργινίνης με το ελεύθερο καρβοξυτελικό άκρο και το πακετάρισμα της τρυπτοφάνης με την προλίνη, στο οποίο συνεισφέρει και η αργινίνη. Η πλευρική ομάδα της λυσίνης βρίσκεται επίσης πάνω από την πλευρική ομάδα της τρυπτοφάνης από την άλλη πλευρά. Ο πεπτιδικός σκελετός παίρνει μία διαμόρφωση που μοιάζει με θηλιά αλλά έχει σχήμα τύπου “Λ” (με τη λυσίνη στην κορυφή) ίσως λόγω της παρουσίας της προλίνης. Το πεπτίδιο RELWK (Εικόνα 4.15) σχηματίζει κυρίαρχο cluster με κατοχή σε χρόνο προσομοίωσης 17% και μεγάλη διασπορά κατά μήκος ολόκληρου του τροχιακού, ενώ εμφανίζει και πολλά ακόμα διάσπαρτα cluster αλλά όλα με κατοχή μικρότερη από 4% του τροχιακού και τελείως διακριτές δομές (απουσία cross-vectors στον πίνακα RMSD). Το κυρίαρχο cluster αν και μικρό περιλαμβάνει ένα σύνολο δομών με μικρές ατομικές διακυμάνσεις (μέση τιμή RMSF για τα άτομα των πλευρικών ομάδων 1.5Å και 1.0Å για όλα τα βαριά άτομα). Η διαμόρφωση του πεπτιδικού σκελετού σταθεροποιείται από την αλληλεπίδραση της αργινίνης με το ελεύθερο καρβοξυτελικό άκρο ενώ η τρυπτοφάνη πακετάρεται πάνω από τις πλευρικές ομάδες της

λυσίνης και του γλουταμικού οξέος.



Εικόνα 4.14 Συγκεντρωτικά αποτελέσματα για το πεπτίδιο RDKWP. Από αριστερά προς τα δεξιά βλέπουμε: την αντιπροσωπευτική δομή του κυρίαρχου cluster με βάση την ανάλυση Cartesian-PCA (stereo αναπαράσταση), τα ενεργειακά τοπία ( $\Delta G$  energy plots) της προβολής του τροχιακού στους τρεις principal components, την προβολή όλων των cluster που προέκυψαν από Cartesian-PCA πάνω στον πίνακα RMSD (με RMSD cut-off 1.0 και variance-explained 0.95), τον πίνακα RMSD με βάση όλα τα βαριά άτομα (μέγιστη τιμή RMSD 7.5Å) και τις τρεις αποστάσεις μεταξύ ατόμων Ca 1-5 (μαύρο χρώμα), 1-4 (κόκκινο χρώμα), 2-4 (πράσινο χρώμα). Ο χρωματισμός των δομών έγινε με βάση τις τιμές RMSF από μπλε σε κόκκινο (μέγιστη τιμή RMSF 3.36Å). Για τα cluster με διάρκεια μικρότερη από 10% του χρόνου προσομοίωσης δεν παρουσιάζονται αντιπροσωπευτικές δομές.

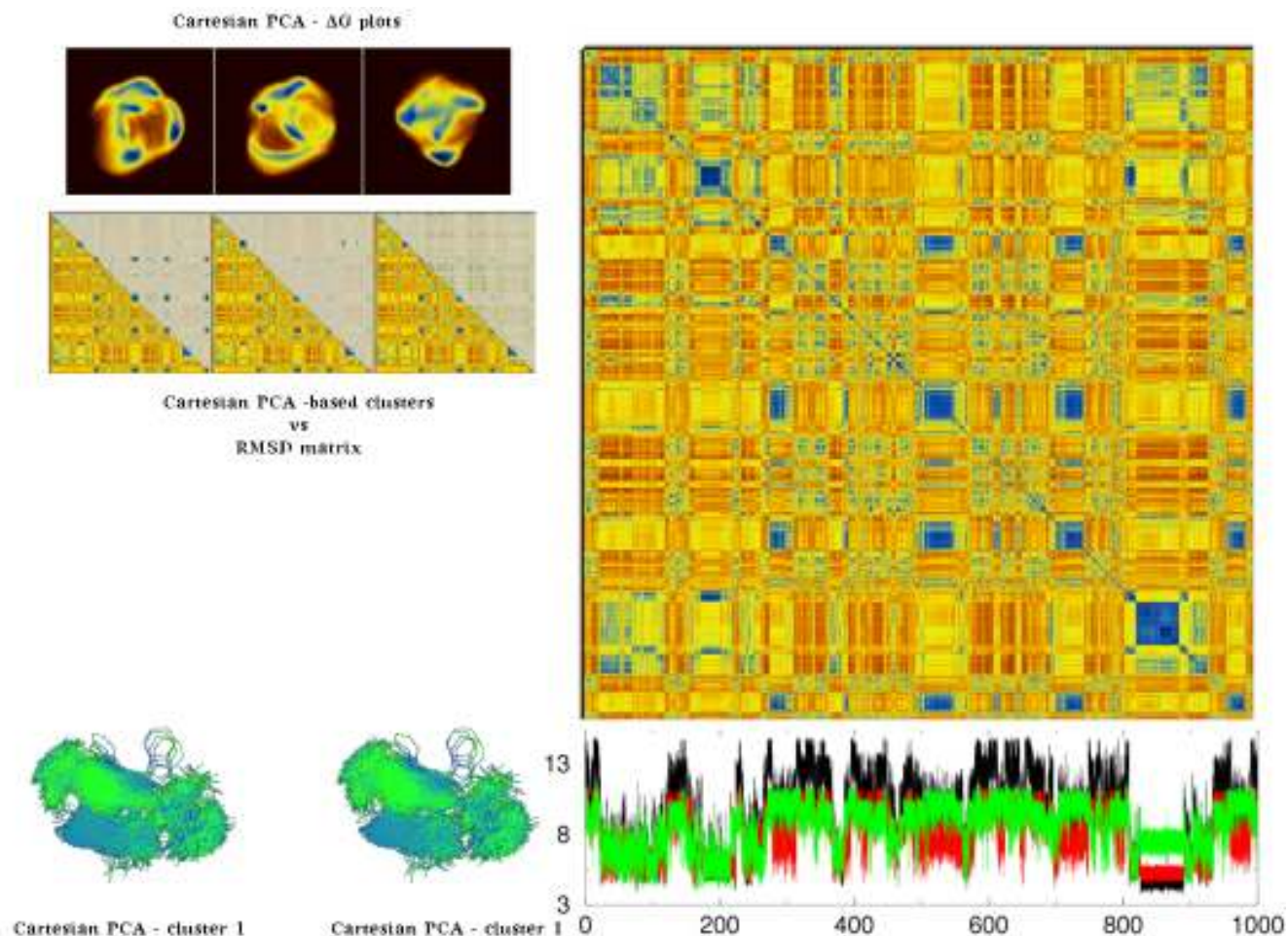


Εικόνα 4.15 Συγκεντρωτικά αποτελέσματα για το πεπτίδιο RELWK. Από αριστερά προς τα δεξιά βλέπουμε: την αντιπροσωπευτική δομή του κυρίαρχου cluster με βάση την ανάλυση Cartesian-PCA (stereo αναπαράσταση), τα ενεργειακά τοπία ( $\Delta G$  energy plots) της προβολής του τροχιακού στους τρεις principal components, την προβολή όλων των cluster που προέκυψαν από Cartesian-PCA πάνω στον πίνακα RMSD (με RMSD cut-off 3.2 και variance-explained 0.85), τον πίνακα RMSD με βάση όλα τα βαριά άτομα (μέγιστη τιμή RMSD 7.5Å) και τις τρεις αποστάσεις μεταξύ ατόμων Ca 1-5 (μαύρο χρώμα), 1-4 (κόκκινο χρώμα), 2-4 (πράσινο χρώμα). Ο χρωματισμός των δομών έγινε με βάση τις τιμές RMSF από μπλε σε κόκκινο (μέγιστη τιμή RMSF 3.03Å). Για τα cluster με διάρκεια μικρότερη από 10% του χρόνου προσομοίωσης δεν παρουσιάζονται αντιπροσωπευτικές δομές.

Το πεπτίδιο REWDV (Εικόνα 4.16) φαίνεται ασταθές με τα cluster να έχουν διάρκεια μικρότερη από 10% του χρόνου προσομοίωσης. Το πρώτο cluster έχει κατοχή 7% και επανασηματίζεται τουλάχιστον 5 φορές στη διάρκεια του τροχιακού. Οι δομές είναι ασταθείς με υψηλές ατομικές

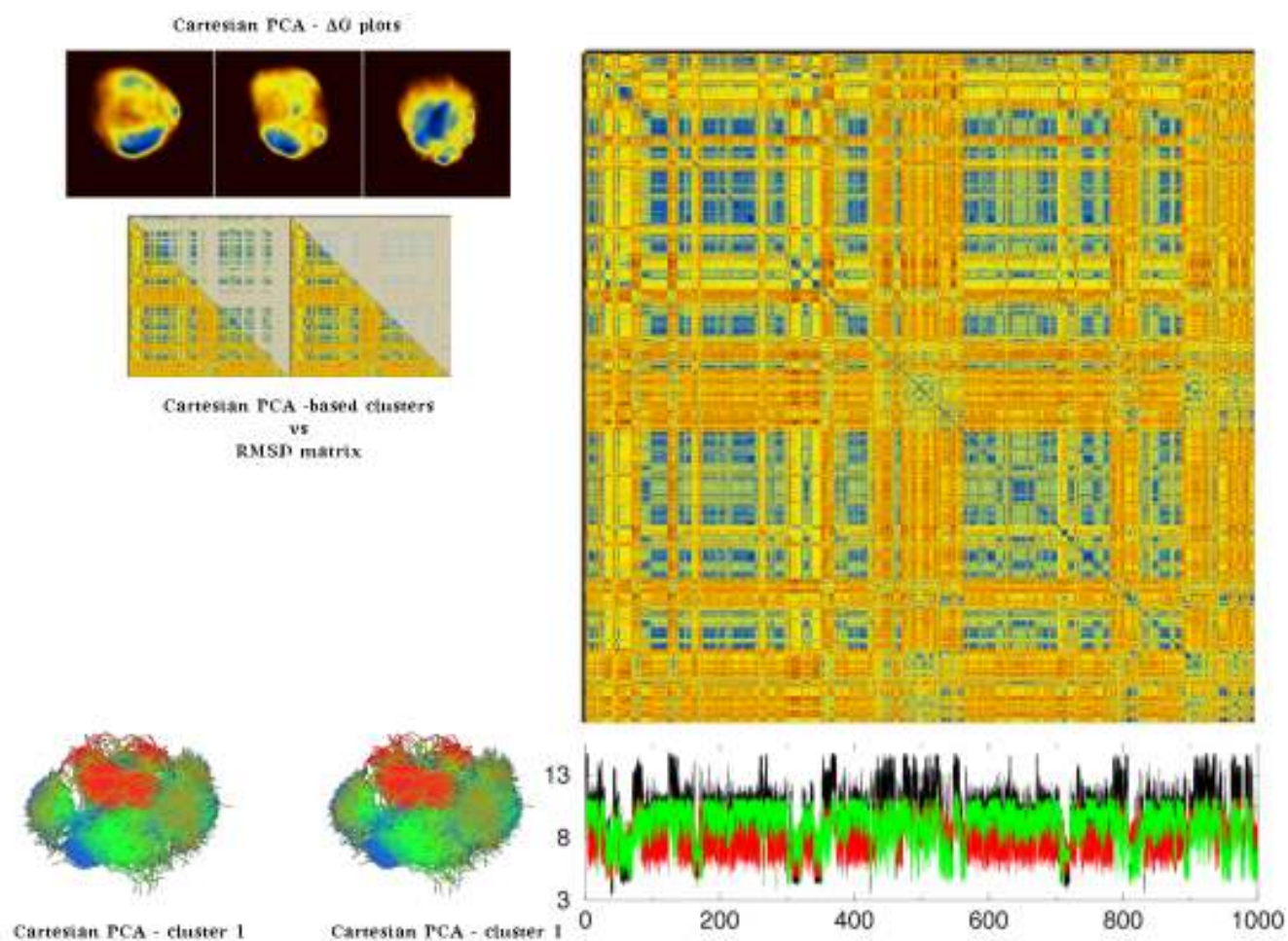


διακυμάνσεις (μέση τιμή RMSF για τα άτομα των πλευρικών ομάδων 2.2Å και 1.4Å για όλα τα βαριά άτομα) και η μόνη υποτυπώδης αλληλεπίδραση που αναπτύσσεται είναι μεταξύ της τρυπτοφάνης και της αργινίνης.



Εικόνα 4.16 Συγκεντρωτικά αποτελέσματα για το πεπτίδιο REWDV. Από αριστερά προς τα δεξιά βλέπουμε: την αντιπροσωπευτική δομή του κυρίαρχου cluster με βάση την ανάλυση Cartesian-PCA (stereo αναπαράσταση), τα ενεργειακά τοπία ( $\Delta G$  energy plots) της προβολής του τροχιακού στους τρεις principal components, την προβολή όλων των cluster που προέκυψαν από Cartesian-PCA πάνω στον πίνακα RMSD (με RMSD cut-off 1.0 και variance-explained 0.87), τον πίνακα RMSD με βάση όλα τα βαριά άτομα (μέγιστη τιμή RMSD 7.5Å) και τις τρεις αποστάσεις μεταξύ ατόμων Ca 1-5 (μαύρο χρώμα), 1-4 (κόκκινο χρώμα), 2-4 (πράσινο χρώμα). Ο χρωματισμός των δομών έγινε με βάση τις τιμές RMSF από μπλε σε κόκκινο (μέγιστη τιμή RMSF 4.77Å). Για τα cluster με διάρκεια μικρότερη από 5% του χρόνου προσομοίωσης δεν παρουσιάζονται αντιπροσωπευτικές δομές.

Το πεπτίδιο REWID (Εικόνα 4.17) εμφανίζει τη χαρακτηριστική εικόνα των πολλαπλών γεγονότων αναδίπλωσης/αποδιάταξης που επιδεικνύουν τα μικρά ασταθή πεπτίδια. Το κυρίαρχο cluster διαρκεί για περίπου 20% του χρόνου προσομοίωσης και τα frames που ανήκουν σε αυτό είναι διάσπαρτα σε όλο το μήκος του τροχιακού. Το σύνολο δομών που ανήκει στο cluster έχει μεγάλο εύρος ατομικών διακυμάνσεων (μέση τιμή RMSF για τα άτομα των πλευρικών ομάδων 2.0Å) κυρίως οι πλευρικές ομάδες των καταλοίπων Trp-Arg-Ile, μεταξύ των οποίων βλέπουμε και τη δημιουργία της όποιας σταθεροποιητικής αλληλεπίδρασης.



Εικόνα 4.17 Συγκεντρωτικά αποτελέσματα για το πεπτίδιο REWID. Από αριστερά προς τα δεξιά βλέπουμε: την αντιπροσωπευτική δομή του κυρίαρχου cluster με βάση την ανάλυση Cartesian-PCA (stereo αναπαράσταση), τα ενεργειακά τοπία ( $\Delta G$  energy plots) της προβολής του τροχιακού στους τρεις principal components, την προβολή όλων των cluster που προέκυψαν από Cartesian-PCA πάνω στον πίνακα RMSD (με RMSD cut-off 3.6 και variance-explained 0.72), τον πίνακα RMSD με βάση όλα τα βαριά άτομα



(μέγιστη τιμή RMSD 7.5Å) και τις τρεις αποστάσεις μεταξύ ατόμων Ca 1-5 (μαύρο χρώμα), 1-4 (κόκκινο χρώμα), 2-4 (πράσινο χρώμα). Ο χρωματισμός των δομών έγινε με βάση τις τιμές RMSF από μπλε σε κόκκινο (μέγιστη τιμή RMSF 3.97Å). Για τα cluster με διάρκεια μικρότερη από 5% του χρόνου προσομοίωσης δεν παρουσιάζονται αντιπροσωπευτικές δομές.

Βλέπουμε πως η πλειοψηφία των πεπτιδίων έχει ασταθή συμπεριφορά με πολλαπλά γεγονότα αναδίπλωσης/αποδιάταξης, υψηλές ατομικές διακυμάνσεις και κατοχή σε χρόνο προσομοίωσης μικρότερη του 20%. Εάν συγκρίνουμε τους πίνακες RMSD που προέκυψαν από τα τροχιακά διάρκειας 1μs με το force field AMBER99SB-ILDN με τα τροχιακά διάρκειας 0.12μs με τα τέσσερα force fields της προηγούμενης ενότητας (Εικόνα 4.7) συμπεραίνουμε ότι η επιμήκυνση του χρόνου δεν ήταν ο καθοριστικός παράγοντας της διαφοροποίησης του αποτελέσματος αλλά η επιλογή του force field. Τα 2 πεπτίδια, RDKWP και NEWRD, με την σταθερότερη συμπεριφορά μελετήθηκαν περαιτέρω για την επίδραση της θερμοκρασίας στην αναδιπλωσιμότητά τους.

*“Everyone thinks of changing the world,  
but no one thinks of changing himself.”*

*Tolstoy*

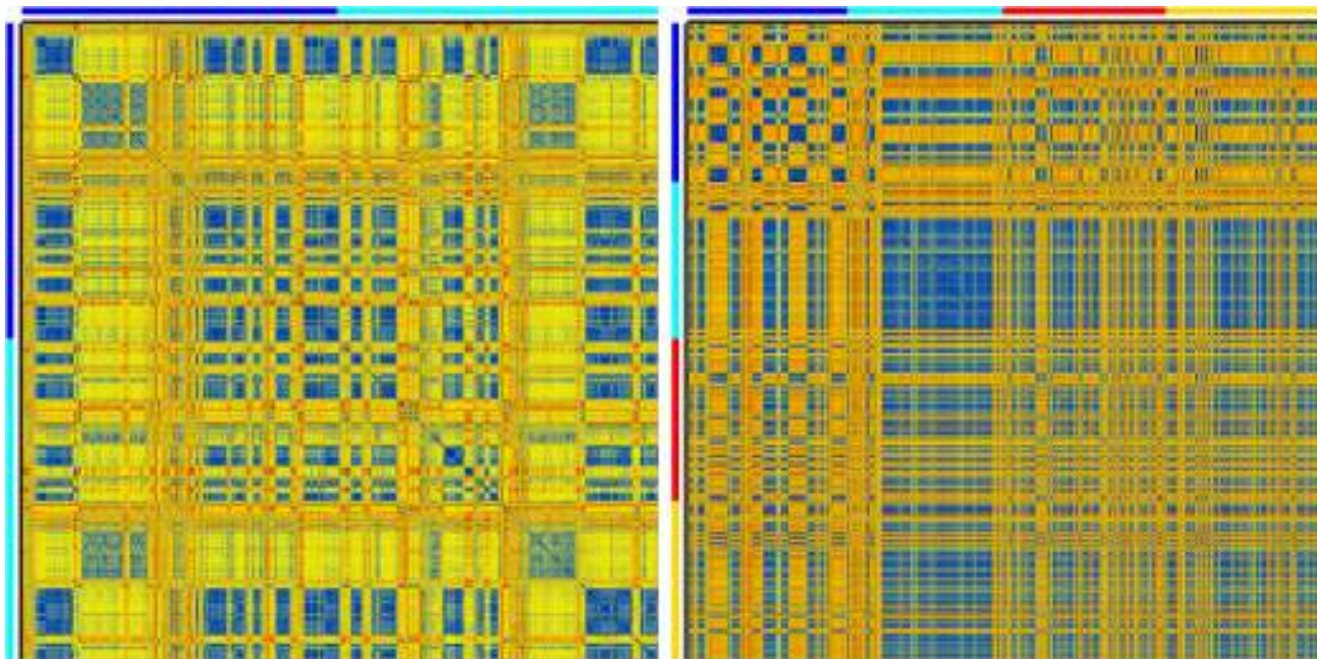
*“Many people would rather die than think;  
in fact, most do.”*

*Bertrand Russell*



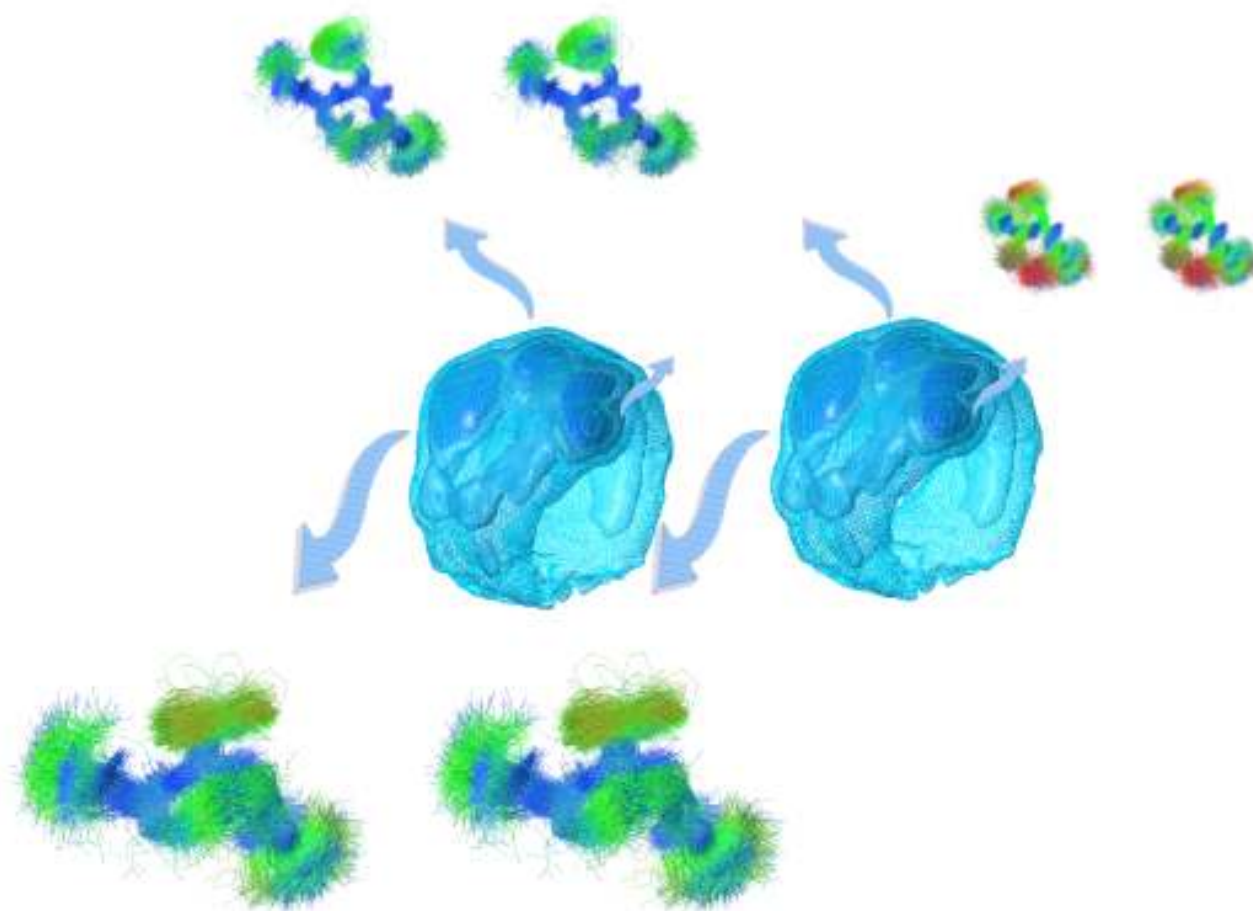
## 4.7 Μελέτη της επίδρασης της θερμοκρασίας στην αναδίπλωση των RDKWP και NEWRD

Τα πενταπεπτίδια RDKWP και NEWRD έδειξαν την πιο σταθερή συμπεριφορά στην πορεία που ακολουθήσαμε. Όπως και στην περίπτωση των τετραπεπτιδίων, εξετάσαμε την επίδραση της θερμοκρασίας στη δυναμική συμπεριφορά τους με προσομοιώσεις διάρκειας 2μs και με το force field AMBER99SB-ILDN, με απώτερο σκοπό την παραγωγή αποτελεσμάτων για σύγκριση με πειραματικά δεδομένα. Στην προηγούμενη ενότητα (Ενότητα 4.6) οι προσομοιώσεις (διάρκειας 1μs) διεξήχθησαν στους 320K (47°C). Για το πεπτίδιο NEWRD δοκιμάσαμε επιπλέον τη θερμοκρασία δωματίου, 298K (25°C). Για το πεπτίδιο RDKWP που έχει δείξει τη μεγαλύτερη σταθερότητα στις μέχρι τώρα προσομοιώσεις (Εικόνες 4.4, 4.5, 4.7, 4.9, 4.14) και με όλα τα force fields, δοκιμάσαμε τη θερμοκρασία δωματίου 298K (25°C) αλλά και τις θερμοκρασίες 340K (67°C) και 360K (87°C) για να ελέγξουμε τη σταθερότητα του. Το υπολογιστικό τμήμα των προσομοιώσεων ολοκληρώθηκε σε ~4 μέρες φυσικού χρόνου απασχολώντας 4 κόμβους της συστοιχίας των υπολογιστών. Το πρωτόκολλο της προσομοίωσης είναι πανομοιότυπο με αυτό του Παραρτήματος (#14, NAMD script, heat.namd και #15, NAMD script, equi.namd) με τις διαφορές στις παραμέτρους της θερμοκρασίας όπως περιγράφονται στο Κεφάλαιο 3, Ενότητα 3.6 και την προσαρμογή του πρωτοκόλλου για τη χρήση του AMBER99SB-ILDN force field, όπως περιγράφεται στο Κεφάλαιο 4, Ενότητα 4.5.



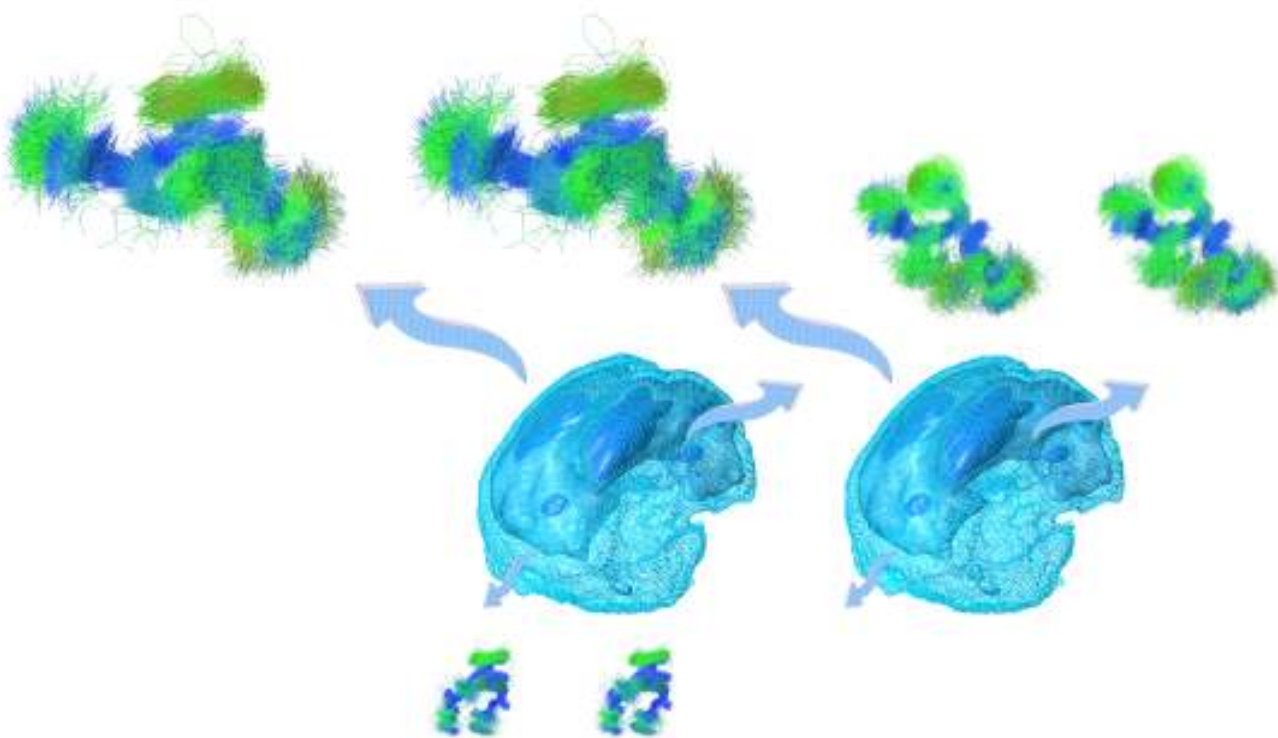
Εικόνα 4.18 Γραφική απεικόνιση του ενιαίου πίνακα RMSD μετά την ένωση των τροχιακών των 2 και 4 θερμοκρασιών του πεπτιδίου NEWRD (αριστερά) και RDKWP (δεξιά), χρησιμοποιώντας όλα τα βαριά άτομα. Η χρωματική κλίμακα κυμαίνεται από σκούρο μπλε (0Å) έως σκούρο κόκκινο (7.3Å). Οι οριζόντιες και κάθετες χρωματιστές μπάρες οριοθετούν τα δύο και τέσσερα ανεξάρτητα τροχιακά αντίστοιχα, όπου με μπλε απεικονίζεται το τροχιακό των 298K, με γαλάζιο των 320K, με κόκκινο των 340K και με πορτοκαλί των 360K. Το βήμα κατά τον υπολογισμό του πίνακα RMSD προσαρμόστηκε κατάλληλα ώστε κάθε τροχιακό να έχει το ίδιο ποσοστό αντιπροσώπευσης.

Για το πεπτίδιο NEWRD δεν παρουσιάζεται διαφοροποίηση στην κινητικότητα και τη σταθερότητα του με την άνοδο της θερμοκρασίας από τους 298K στους 320K (Εικόνα 4.18). Η παρατήρηση αυτή στηρίζεται τόσο από τους πίνακες RMSD όσο και από την ανάλυση PCA. Με βάση τον πίνακα RMSD βλέπουμε τις ίδιες ομάδες δομών και με παρόμοια αντιπροσώπευση στα τροχιακά με τις δύο θερμοκρασίες. Πράγματι, από την ανάλυση Cartesian-PCA προκύπτουν 3 cluster δομών για το τροχιακό των 298K (με RMSD cut-off 1.80 και variance-explained 0.90) με κατοχή σε χρόνο προσομοίωσης 32%, 5% και 5% του συνολικού χρόνου προσομοίωσης (Εικόνα 4.19). Για το τροχιακό των 320K (με RMSD cut-off 2.00 και variance-explained 0.83) προκύπτουν επίσης 3 cluster δομών με κατοχή σε χρόνο προσομοίωσης 28%, 5% και 0.5% του συνολικού χρόνου προσομοίωσης (Εικόνα 4.20).



Εικόνα 4.19 Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου NEWRD. Στο κέντρο φαίνεται η προβολή του τροχιακού των 298K στους τρεις principal components της ανάλυσης Cartesian-PCA χρησιμοποιώντας όλα τα βαριά άτομα. Υποδεικνύονται τρία επίπεδα ισοεπιφάνειας (μέση τιμή,  $1\sigma$ ,  $6\sigma$  του χάρτη κατανομής). Κάθε ισχυρή κορυφή του ενεργειακού τοπίου αντιστοιχεί σε ένα cluster δομών (οι οποίες φαίνονται με τα βέλη), το μέγεθος αναπαράστασης των οποίων είναι ανάλογο της κατοχής του cluster σε χρόνο προσομοίωσης. Η υπέρθεση των δομών έγινε χρησιμοποιώντας τα άτομα του πεπτιδικού σκελετού για να δοθεί έμφαση στην κινητικότητα των πλευρικών ομάδων, ενώ ο χρωματισμός των ατόμων από μπλε σε κόκκινο (μέγιστη τιμή  $4.85\text{\AA}$ ) έγινε με βάση τις ατομικές διακυμάνσεις (RMSFs).

Το κυρίαρχο cluster δομών είναι παρόμοιο μεταξύ των δύο τροχιακών, με μέσο RMSD μεταξύ των δύο ομάδων δομών  $2.2\text{\AA}$ , λαμβάνοντας υπόψιν όλα τα βαριά άτομα. Το δεύτερο σε αντιπροσώπευση cluster δομών είναι επίσης κοινό, με μέσο RMSD μεταξύ των δύο ομάδων δομών  $2.0\text{\AA}$ , λαμβάνοντας υπόψιν όλα τα βαριά άτομα.

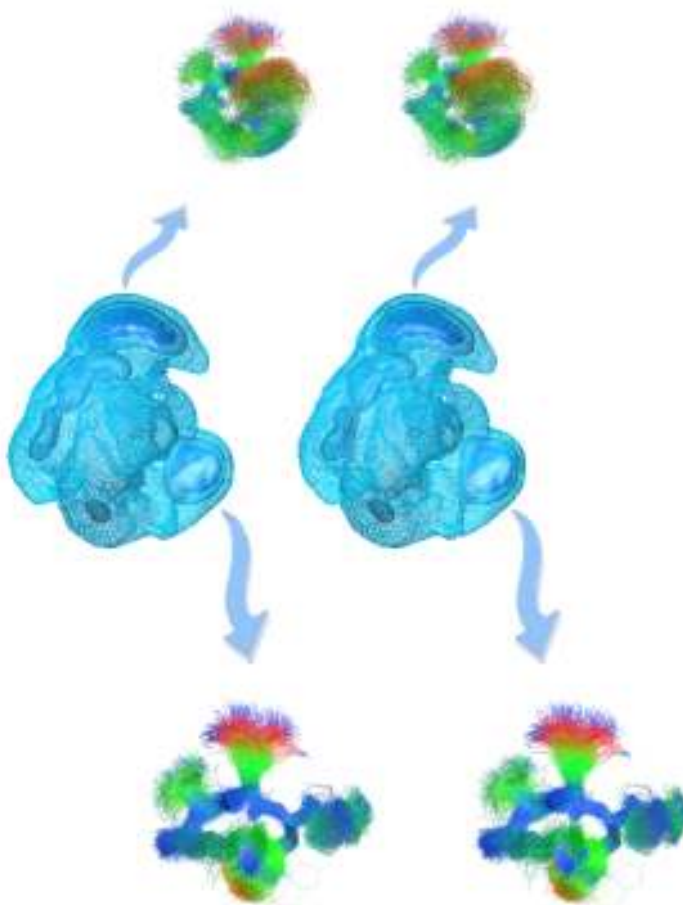


Εικόνα 4.20 Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου NEWRD για το τροχιακό των 320K, σε αντιστοιχία με την Εικόνα 4.19 (αντιπαράβολή με τα αποτελέσματα της Εικόνας 4.13).

Το τρίτο cluster δομών στο τροχιακό των 298K είναι πιο κοντά στις δομές του δεύτερου cluster (με μέσο RMSD μεταξύ των δύο ομάδων δομών 2.4Å, λαμβάνοντας υπόψιν όλα τα βαριά άτομα) σε σχέση με του πρώτου (με μέσο RMSD μεταξύ των δύο ομάδων δομών 3.6Å, λαμβάνοντας υπόψιν όλα τα βαριά άτομα), αλλά με πολύ υψηλότερες διακυμάνσεις. Το τρίτο cluster δομών στο τροχιακό των 320K διαφέρει από τα προηγούμενα (με μέσο RMSD μεταξύ των ομάδων δομών 3.2Å και 3.9Å, από το πρώτο και δεύτερο αντίστοιχα, λαμβάνοντας υπόψιν όλα τα βαριά άτομα) με υψηλές ατομικές διακυμάνσεις και σημαντικά μειωμένη αντιπροσώπευση.

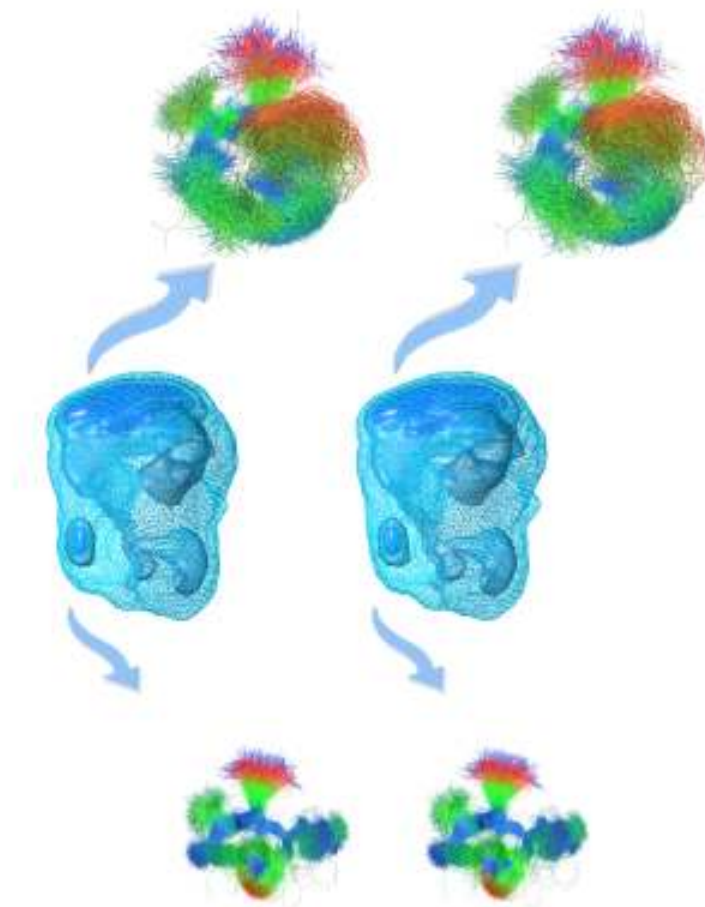
Συνολικά, το πεπτίδιο αυτό φαίνεται να μη σχηματίζει σταθερή δομή για περισσότερο από 20%-30% του χρόνου προσομοίωσης και αυτή η συμπεριφορά δεν μεταβάλλεται με την επιμήκυνση του τροχιακού ή του force field που χρησιμοποιείται (Εικόνες 4.4, 4.5, 4.7, 4.9, 4.13). Η παρατήρηση αυτή ισχύει καθολικά για όλα τα force fields (Εικόνα 4.8).





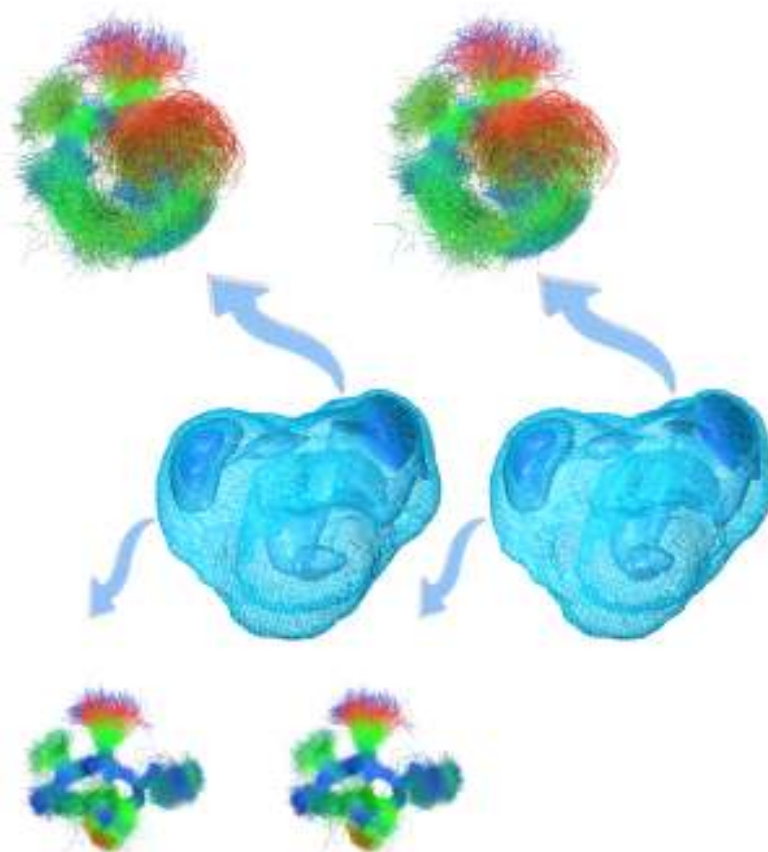
Εικόνα 4.21 Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RDKWP. Στο κέντρο φαίνεται η προβολή του τροχιακού των 298K στους τρεις principal components της ανάλυσης Cartesian PCA χρησιμοποιώντας όλα τα βαριά άτομα. Υποδεικνύονται τρία επίπεδα ισοεπιφάνειας (μέση τιμή,  $1\sigma$ ,  $6\sigma$  του χάρτη κατανομής). Κάθε ισχυρή κορυφή του ενεργειακού τοπίου αντιστοιχεί σε ένα cluster δομών (οι οποίες φαίνονται με τα βέλη), το μέγεθος αναπαράστασης των οποίων είναι ανάλογο της κατοχής του cluster σε χρόνο προσομοίωσης. Η υπέρθεση των δομών έγινε χρησιμοποιώντας τα άτομα του πεπτιδικού σκελετού για να δοθεί έμφαση στην κινητικότητα των πλευρικών ομάδων, ενώ ο χρωματισμός των ατόμων από μπλε σε κόκκινο (μέγιστη τιμή  $3.43\text{\AA}$ ) έγινε με βάση τις ατομικές διακυμάνσεις (RMSFs).

Οι κυριότερες αλληλεπιδράσεις που παρατηρούνται στις κυρίαρχες ομάδες δομών είναι μεταξύ Glu-Trp-Arg, ενώ τα ακραία κατάλοιπα παραμένουν ελεύθερα στο διάλυμα. Στις δομές που παρατηρούνται για μικρότερο χρονικό διάστημα, αναπτύσσονται άλλες αλληλεπιδράσεις μεταξύ Asn-Trp και Glu-Arg.



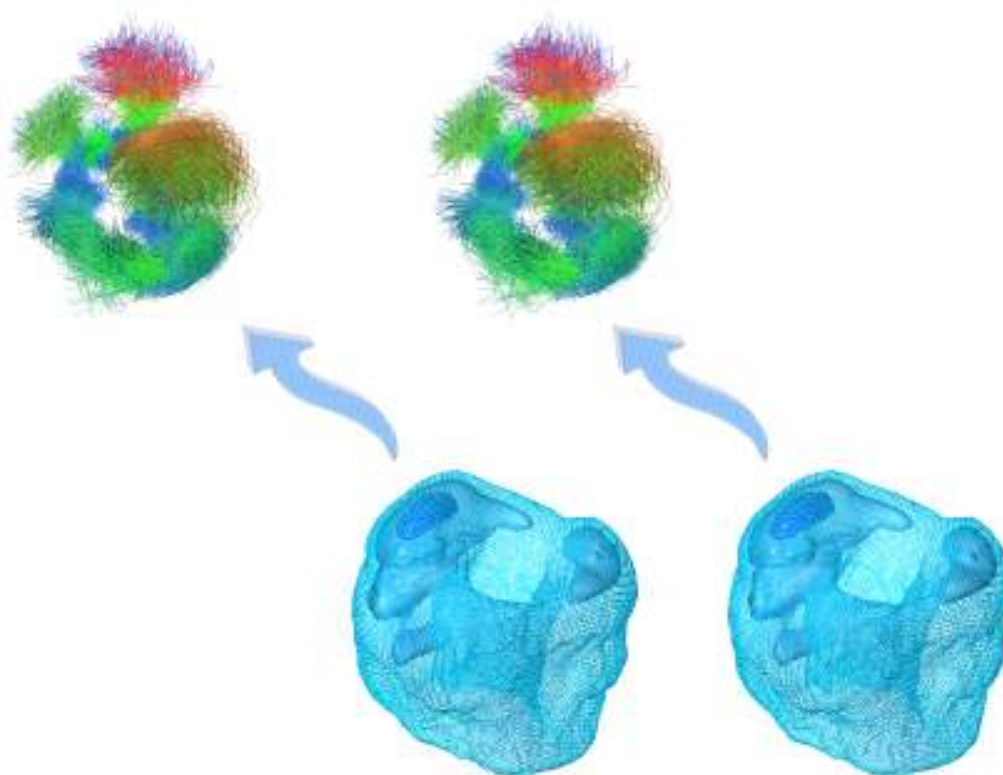
Εικόνα 4.22 Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RDKWP για το τροχιακό των 320K, σε αντιστοιχία με την Εικόνα 4.21 (αντιπαράβολή με τα αποτελέσματα της Εικόνας 4.14).

Για το πεπτίδιο RDKWP βλέπουμε με μία πρώτη εντύπωση με βάση τη γραφική απεικόνιση των πινάκων RMSD (Εικόνα 4.18), ότι υπάρχει μεγαλύτερη σταθερότητα του τροχιακού στους 320K σε σχέση με τα τροχιακά στις υπόλοιπες θερμοκρασίες. Μία πιο προσεκτική παρατήρηση μας επιτρέπει να δούμε ότι στο τροχιακό των 298K υπάρχουν 2 ομάδες δομών με συνεχείς μεταβάσεις από το ένα cluster στο άλλο, ενώ στα υπόλοιπα 3 φαίνεται μόνο ένα κυρίαρχο cluster του οποίου η σταθερότητα μειώνεται προοδευτικά με την άνοδο της θερμοκρασίας. Πράγματι, με βάση την ανάλυση Cartesian-PCA προκύπτουν 2 cluster δομών για το τροχιακό των 298K (με RMSD cut-off 1.90 και variance-explained 0.95) με κατοχή σε χρόνο προσομοίωσης 30% και 24% του συνολικού χρόνου προσομοίωσης.



Εικόνα 4.23 Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RDKWP για το τροχιακό των 340K, σε αντιστοιχία με την Εικόνα 4.21

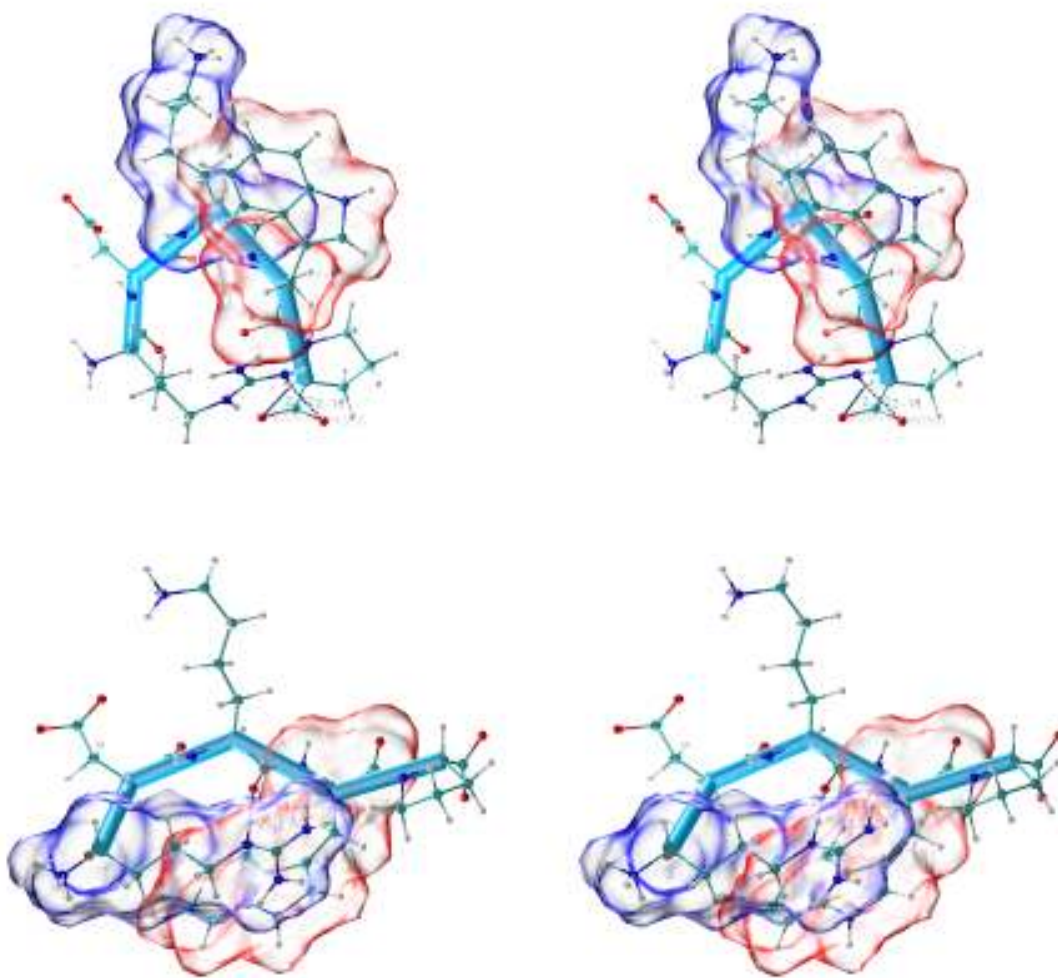
Για το τροχιακό των 320K προκύπτουν 2 cluster δομών (με RMSD cut-off 1.00 και variance-explained 0.95) με κατοχή σε χρόνο προσομοίωσης 58% και 2%. Για το τροχιακό των 340K προκύπτουν 2 cluster δομών (με RMSD cut-off 1.30 και variance-explained 0.94) με κατοχή σε χρόνο προσομοίωσης 41% και 12%. Τέλος, για το τροχιακό των 360K προκύπτει 1 cluster δομών (με RMSD cut-off 5.40 και variance-explained 0.87) με κατοχή σε χρόνο προσομοίωσης 38%. Οι δομές των cluster που παρατηρούνται στις διάφορες θερμοκρασίες, όπως προκύπτουν από την ανάλυση Cartesian-PCA, είναι πολύ κοντινές, όπως και αναμένεται από τον πίνακα RMSD (Εικόνα 4.18). Οι δομές του πρώτου σε αντιπροσώπευση cluster του τροχιακού των 298K αντιστοιχούν στις δομές του δεύτερου cluster των τροχιακών στους 320K και 340K ενώ δεν εμφανίζονται στις θερμοκρασίες των 360K. Οι δομές του δεύτερου cluster του τροχιακού των 298K αντιστοιχούν στις δομές των κυρίαρχων cluster των υψηλότερων θερμοκρασιών.



Εικόνα 4.24 Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RDKWP για το τροχιακό των 360K, σε αντιστοιχία με την Εικόνα 4.21

Για την ακρίβεια οι δομές του πρώτου cluster των 298K έχουν μέσο RMSD (για όλα τα βαριά άτομα)  $2.6\text{\AA}$  και  $1.2\text{\AA}$  από τις δομές του δεύτερου cluster των 320K και 340K, αντίστοιχα και οι δομές του δεύτερου cluster των 298K έχουν μέσο RMSD (για όλα τα βαριά άτομα)  $2.3\text{\AA}$  και  $1.8\text{\AA}$  από τις δομές του πρώτου cluster των 320K και 340K. Οι κυρίαρχες δομές στις θερμοκρασίες 320K, 340K και 360K έχουν μεταξύ τους μέσο RMSD (για όλα τα βαριά άτομα) μόλις  $1.7\text{\AA}$ - $1.9\text{\AA}$ . Συνολικά, το πεπτίδιο RDKWP φαίνεται να παίρνει σταθερή δομή για 50%-60% του χρόνου προσομοίωσης και η συμπεριφορά αυτή υποστηρίζεται από τα force field της οικογένειας CHARMM και AMBER (Εικόνες 4.4, 4.5, 4.7, 4.9, 4.14) και ειδικότερα από τις προσομοιώσεις μεγαλύτερης διάρκειας ( $1\mu\text{s}$ - $2\mu\text{s}$ ) που πραγματοποιήθηκαν με το force field AMBER99SB-ILDN.

Στην Εικόνα 4.25 βλέπουμε μία απεικόνιση των δύο αντιπροσωπευτικών δομών, δηλαδή τα στιγμιότυπα του τροχιακού που είναι πλησιέστερα στην υπολογιζόμενη μέση δομή των δύο κυρίαρχων cluster, όπως αυτά προέκυψαν με βάση την ανάλυση Cartesian-PCA και παρουσιάστηκαν παραπάνω.



Εικόνα 4.25 Γραφική απεικόνιση (σε stereo αναπαράσταση) των 2 διακριτών δομών του πεπτιδίου RDKWP. Όλα τα άτομα απεικονίζονται με σφαίρες και ράβδους και χρωματίζονται με βάση το όνομα του ατόμου. Με αναπαράσταση σωλήνα και γαλάζιο χρώμα υποδεικνύεται η διεύθυνση του πεπτιδικού σκελετού. Οι δεσμοί υδρογόνου σημειώνονται με στικτή μπλε γραμμή και οι πλευρικές ομάδες που πακετάρονται επισημαίνονται με διαφανή επιφάνεια και κόκκινο χρώμα για την τρυπτοφάνη και μπλε για τα φορτισμένα κατάλοιπα, λυσίνη και αργινίνη αντίστοιχα.



Η αντιπροσωπευτική δομή του πρώτου σε αντιπροσώπευση cluster (cluster 2 για το τροχιακό των 298K και cluster 1 για τα τροχιακά των 320K και 340K) σταθεροποιείται από το πακετάρισμα της λυσίνης με την τρυπτοφάνη και τον δεσμό υδρογόνου μεταξύ της NH ομάδας της αργινίνης και του ελεύθερου καρβοξυλικού άκρου του πεπτιδίου.

Η αντιπροσωπευτική δομή του δεύτερου σε αντιπροσώπευση cluster (cluster 1 για το τροχιακό των 298K και cluster 2 για τα τροχιακά των 320K και 340K) σταθεροποιείται από το πακετάρισμα της αργινίνης απέναντι από την τρυπτοφάνη, το δεσμό υδρογόνου μεταξύ του NE της αργινίνης και του O του πεπτιδικού σκελετού της λυσίνης και της γέφυρας άλατος μεταξύ του ασπαρτικού οξέος και της λυσίνης.

*“What you see depends upon the perspective  
from which you look.”*

*George Rose*

*“There are two major products that come out of Berkeley: LSD and UNIX. We don’t believe this to be a coincidence .”*

*Jeremy S. Anderson*



## 4.8 Μελέτη της *cis/trans* ισομερείωσης του πεπτιδικού δεσμού της προλίνης στο RDKWP μέσω adaptive tempering

Η απόδειξη της επαρκούς διερεύνησης όλων των διαμορφώσεων για την αναπαράσταση του ενεργειακού τοπίου των πρωτεϊνών αλλά και των πεπτιδίων (sufficient sampling) είναι ένα μείζον ζήτημα για τις προσομοιώσεις μοριακής δυναμικής και μία από τις σημαντικότερες αδυναμίες τους (βλέπε Κεφάλαιο 1, Ενότητα 1.1). Από την πλευρά των αλγόριθμων προσομοίωσης, οι λεγόμενες tempering μέθοδοι όπου αλλάζει δυναμικά η θερμοκρασία του συστήματος είναι οι ευρύτερα αποδεκτές και χρησιμοποιούν είτε ένα αντίγραφο (single-copy, simulated tempering) είτε πολλά αντίγραφα (multiple-copy, replica-exchange). Η ιδέα πίσω από τη συνεχή αλλαγή της θερμοκρασίας του συστήματος είναι να ξεπεραστούν τα ενεργειακά φράγματα ώστε να εξεταστούν και άλλες διαμορφώσεις του ενεργειακού τοπίου (Voter, 1997). Μία μέθοδος που εμπίπτει στην κατηγορία αυτή (single-copy) είναι το adaptive tempering (Zhang et al., 2010), όπου στο σύστημα εφαρμόζεται ένα εύρος συνεχών τιμών θερμοκρασίας: όταν η δυναμική ενέργεια (που υπολογίζεται για το τρέχον βήμα της προσομοίωσης) έχει χαμηλότερη τιμή από την τρέχουσα μέση τιμή, η θερμοκρασία μειώνεται, και αντίστοιχα όταν η δυναμική ενέργεια έχει υψηλότερη τιμή, η θερμοκρασία αυξάνεται. Με τον τρόπο αυτό επιτρέπεται στο σύστημα να επισκεφτεί ενεργειακά ελάχιστα, για τα οποία χρειάζεται εξαιρετικά μεγάλος χρόνος προσομοίωσης με το κλασικό πρωτόκολλο των προσομοιώσεων μοριακής δυναμικής. Η αλλαγή της θερμοκρασίας επιτυγχάνεται είτε μέσω του θερμοστάτη (Langevin thermostat) είτε μέσω των ταχυτήτων (velocity rescaling).

Η μέθοδος αυτή χρησιμοποιήθηκε με επιτυχία στην αναδίπλωση μίνι-πρωτεϊνών (ultra-fast folders) που χρησιμοποιούνται ευρέως ως μοντέλα, τις trpzip2, trp-cage, villin headpiece, με απόκλιση 0.2Å, 0.4Å, 0.4Å (Ca-RMSD) από τη native δομή σε μόλις 0.5-1.0μs (και σε ένα εύρος θερμοκρασίας 290K-600K) (Zhang et al., 2010). Η μέθοδος αυτή μειώνει σημαντικά το υπολογιστικό κόστος σε σχέση με άλλες μεθόδους (replica-exchange) αλλά παρατηρούνται αποκλίσεις στις ενέργειες και στις θερμοκρασίες αναδίπλωσης (folding enthalpy, folding temperature) (Zhang et al., 2010).

Στην περίπτωση των πεπτιδίων της δικής μας μελέτης η μέθοδος αυτή μπορεί να χρησιμοποιηθεί για μελέτη φαινομένων όπως η *cis/trans* ισομερείωση του πεπτιδικού δεσμού σε αλληλουχίες που περιέχουν προλίνη, όπως του σταθερότερου πεπτιδίου RDKWP στο οποίο καταλήξαμε με την πορεία που ακολουθήσαμε μέχρι τώρα. Πρόκειται για ένα εξαιρετικά αργό φαινόμενο της τάξης των 20sec σε θερμοκρασία 22°C για μικρού μήκους πεπτίδια (Kreiger et al., 2003) και συνεπώς είναι δύσκολο να το προσεγγίσει κανείς με κλασσικές προσομοιώσεις μοριακής δυναμικής. Η *cis/trans* ισομερείωση είναι ένας μηχανισμός ετερογένειας που οδηγεί σε διακριτές native διαμορφώσεις που αλληλομετατρέπονται, με συνέπειες στη βιολογική λειτουργία (Evans et al., 1987, Chazin et al., 1989) ενώ έχει ενοχοποιηθεί για την επιβράδυνση της ταχύτητας της αναδίπλωσης (Brandts et al., 1975). Η προλίνη είναι το μοναδικό κατάλοιπο για το οποίο η *cis* διαμόρφωση του πεπτιδικού σκελετού είναι ενεργειακά προσιτή λόγω της αποσταθεροποίησης της *trans* διαμόρφωσης από την υποκατάσταση της αμινομάδας με το Cδ άτομο της πλευρικής ομάδας. Για το λόγο αυτό συναντάμε τη *cis* διαμόρφωση σε μόλις 0.05% των κρυσταλλογραφικά προσδιορισμένων δομών, ποσοστό που ανέρχεται σε 6.5% σε δεσμούς τύπου X-Pro που συναντώνται κυρίως σε δομικές περιοχές καμπής (bends) και στροφής (turns) (Stewart et al., 1990). Τα ποσοστά αυτά είναι μικρότερα από αυτά που αναμένονται βάσει ενθαλπίας, και θεωρούνται αμφισβητήσιμα λόγω της περιορισμένης διακριτικότητας των κρυσταλλογραφικά προσδιορισμένων δομών, η οποία δεν επιτρέπει την ανίχνευση όλων των *cis* δεσμών (Weiss et al., 1998).

Το ενεργειακό φράγμα μεταξύ *cis/trans* διαμόρφωσης είναι 20kcal/mol λόγω του μερικώς διπλού χαρακτήρα του πεπτιδικού δεσμού, αλλά ειδικά στην περίπτωση της προλίνης η διαφορά στην ενέργεια των δύο διαμορφώσεων είναι μόλις 0.5kcal/mol και το ενεργειακό φράγμα γίνεται 13kcal/mol (Wedemeyer et al., 2002). Μάλιστα υπάρχει ισχυρή προτίμηση για την παρουσία ενός αρωματικού αμινοξέος (με σειρά προτίμησης Tyr-Phe-Trp) πριν την προλίνη λόγω της ανάπτυξης

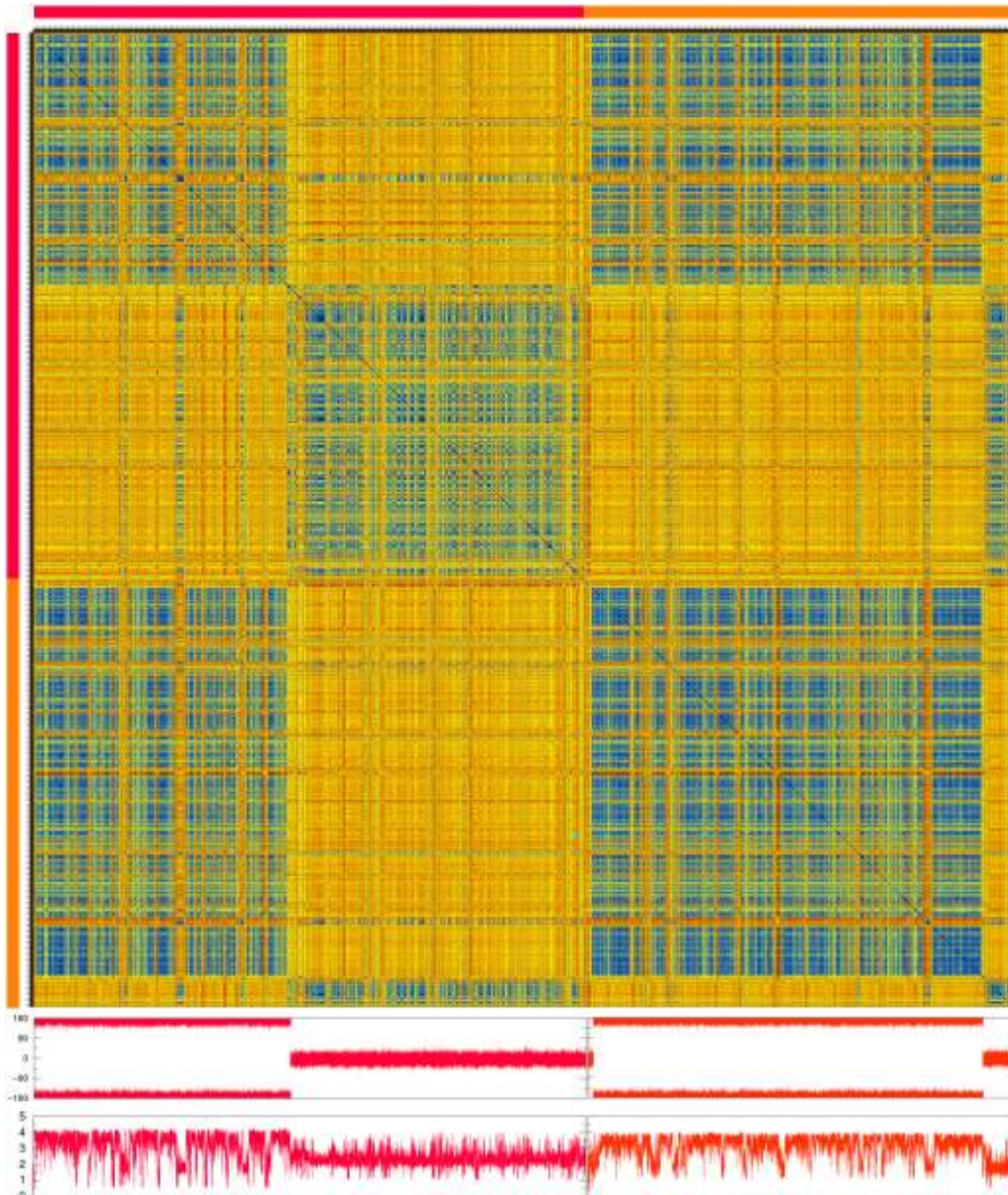
ισχυρών αλληλεπιδράσεων με την προλίνη, ακόμα και παρουσία ισχυρού αποδιατακτικού (8M ουρίας) (Wu et al., 1998). Πειραματικά δεδομένα από NMR στο πενταπεπτίδιο YGGPL δείχνουν την ύπαρξη και των δύο διαμορφώσεων αλλά και την αλλαγή της σχετικής αναλογίας τους με βάση το διαλύτη (πολικό/μη-πολικό) (Ramaprasad et al., 1993).

Το φαινόμενο αυτό μελετήθηκε μέσω accelerated dynamics σε τετραπεπτίδια με την αλληλουχία Ace-TSPI-Nme (σε φωσφορυλιωμένη και μη μορφή), δείχνοντας ότι η ισομερείωση γύρω από τον ω πεπτιδικό δεσμό της προλίνης είναι ασύμμετρη και εξαρτάται από τη διαμόρφωση της δίεδρης γωνίας  $\psi$  της προλίνης (Hamelberg et al., 2005). Στο ίδιο συμπέρασμα οδήγησαν και μελέτες κβαντικής μηχανικής σε ένα διπεπτίδιο προλίνης (N-acetylproline methylamide) (Fischer et al., 1994).

Εμείς χρησιμοποιήσαμε τη μέθοδο adaptive tempering σε δύο προσομοιώσεις διάρκειας 1μs με το force field AMBER99SB-ILDN στο πεπτίδιο RDKWP ξεκινώντας από εκτεταμένες διαμορφώσεις και με τον πεπτιδικό δεσμό της προλίνης στην *trans* (trans τροχιακό) και *cis* διαμόρφωση (*cis* τροχιακό). Το υπολογιστικό κομμάτι των προσομοιώσεων διήρκεσε ~26 μέρες σε 2 κόμβους της συστοιχίας των υπολογιστών. Για τη διεξαγωγή των προσομοιώσεων με τη μέθοδο αυτή ακολουθήσαμε τα πρωτόκολλα που βρίσκονται στο Παράρτημα (#14, NAMD script, heat.namd και #15, NAMD script, equi.namd) με την προσαρμογή για χρήση του AMBER99SB-ILDN force field όπως περιγράφεται στο Κεφάλαιο 4, Ενότητα 4.5. Η αναπροσαρμογή της θερμοκρασίας γίνεται μέσω του θερμοστάτη (Langevin thermostat). Για την επίτευξη αυτού, προστίθενται στο κλασσικό πρωτόκολλο οι παράμετροι που ακολουθούν:

- AdaptTempMD -> on
- AdaptTempRestartFile -> output/restart.tempering
- AdaptTempInfile -> restart.tempering
- AdaptTempRestartFreq -> 10000
- AdaptTempRescaling -> off
- AdaptTempOutFreq -> 400
- AdaptTempLangevin -> on

Ο κύριος σκοπός των προσομοιώσεων αυτών ήταν να δούμε εάν με τη μέθοδο adaptive tempering μπορούμε να παρατηρήσουμε μετάβαση από την *trans* στη *cis* διαμόρφωση ή και το αντίστροφο. Στην Εικόνα 4.26 βλέπουμε τους πίνακες RMSD που έχουν υπολογιστεί μετά την συνένωση του *trans* και του *cis* τροχιακού.



Εικόνα 4.26 Γραφική απεικόνιση του ενιαίου πίνακα RMSD (χρησιμοποιώντας όλα τα βαριά άτομα) μετά τη συνένωση των δύο τροχιακών trans και cis του πεπτιδίου RDKWP. Η χρωματική κλίμακα κυμαίνεται από σκούρο μπλε έως σκούρο κόκκινο (7.0Å). Οι οριζόντιες και κάθετες χρωματιστές μπάρες οριοθετούν τα τροχιακά όπου με πορφυρό χρώμα απεικονίζεται το trans τροχιακό και με πορτοκαλί το cis τροχιακό. Από κάτω απεικονίζονται η εξέλιξη στο χρόνο της γωνίας  $\omega$  της προλίνης και του rmsd (για τα άτομα του πεπτιδικού σκελετού) από την αρχική δομή της προσομοίωσης.



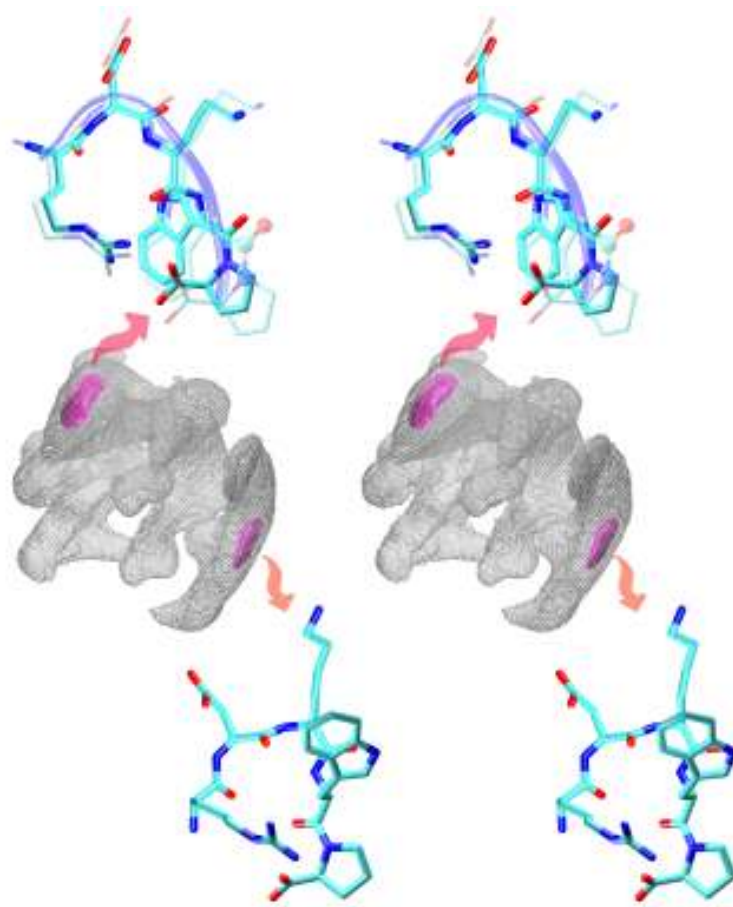
Και στις δύο προσομοιώσεις βλέπουμε τη δημιουργία δύο cluster δομών που αντιστοιχούν στις *trans* και *cis* διαμορφώσεις του πεπτιδικού σκελετού, όπως επιβεβαιώνεται και από την γωνία  $\omega$  της προλίνης. Στην περίπτωση του *trans* τροχιακού η μετάβαση στη *cis* διαμόρφωση γίνεται μετά από περίπου 460ns και παραμένει μέχρι το τέλος της προσομοίωσης. Στην περίπτωση του *cis* τροχιακού βλέπουμε μία πολύ γρήγορη μετάβαση στην *trans* διαμόρφωση μετά από μόλις 12ns προσομοίωσης αλλά και επαναφορά στην αρχική *cis* διαμόρφωση μετά από περίπου 900ns και ως το τέλος της προσομοίωσης. Η μετάβαση από τη μία διαμόρφωση στην άλλη επηρεάζει αναπόφευκτα και τη διαμόρφωση όλου του υπόλοιπου πεπτιδικού σκελετού, η δομή του οποίου παρουσιάζει μεγάλες διακυμάνσεις όπως φαίνεται και από την εξέλιξη στο χρόνο του RMSD σε σχέση με την αρχική δομή (Εικόνα 4.26). Οι έντονες διακυμάνσεις του πεπτιδικού σκελετού αποτρέπουν την περαιτέρω σταθεροποίηση των πλευρικών ομάδων (Εικόνα 4.26, πίνακας RMSD). Ωστόσο, οι διακυμάνσεις αυτές δικαιολογούνται από τις μεταβολές της θερμοκρασίας που εφαρμόζεται για να ξεπεραστεί το υψηλό ενεργειακό φράγμα που διαχωρίζει τη *cis* από την *trans* διαμόρφωση.

Στη συνέχεια πραγματοποιήσαμε ανάλυση Dihedral-PCA για το προσδιορισμό cluster δομών με βάση τις διέδρες γωνίες. Στην Εικόνα 4.27 βλέπουμε το ενεργειακό τοπίο που προκύπτει μετά την προβολή του *trans* τροχιακού στους τρεις κυρίαρχους principal components.

Τα δύο ισχυρότερα cluster (με RMSD cut-off 3.6 και variance-explained 0.85), με κατοχή 32% και 8% του συνολικού χρόνου προσομοίωσης αντίστοιχα, περιλαμβάνουν διακριτές διαμορφώσεις όπου ο πεπτιδικός δεσμός της προλίνης βρίσκεται στην *trans* (cluster 2) και στη *cis* διαμόρφωση (cluster 1).

Στην Εικόνα 4.28 βλέπουμε τα αντίστοιχα αποτελέσματα για το *cis* τροχιακό. Στην περίπτωση αυτή σχηματίζονται 3 κυρίαρχα cluster (με RMSD cut-off 4.2 και variance-explained 0.73) με κατοχή σε χρόνο προσομοίωσης 21% (*trans* διαμόρφωση), 9% (*trans* διαμόρφωση) και 2% (*cis* διαμόρφωση) του συνολικού χρόνου προσομοίωσης αντίστοιχα. Τα cluster 1 και 2 περιλαμβάνουν πολύ κοντινές δομές, ακόμα και για τις διαμορφώσεις των πλευρικών ομάδων (με RMSD 1.2Å για όλα τα βαριά άτομα).

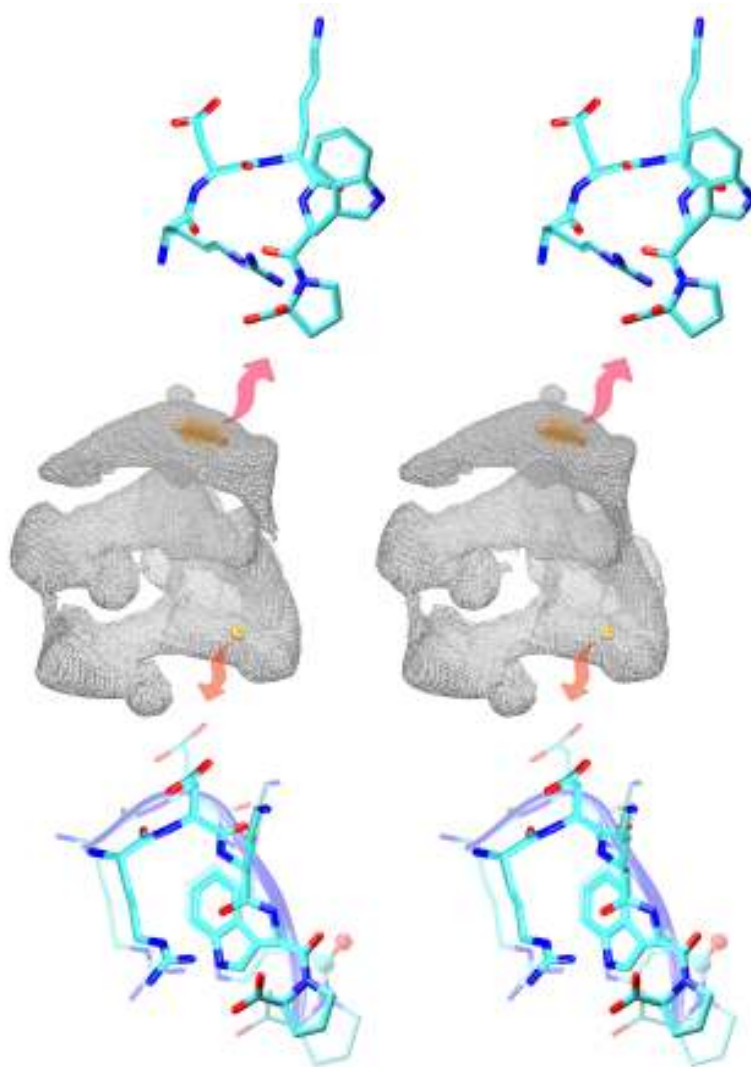
Ωστόσο διαφέρουν στις φ/ψ γωνίες των μεσαίων (2-3) καταλοίπων (Εικόνα 4.31), με αποτέλεσμα να διαφοροποιούνται στην ανάλυση Dihedral-PCA αλλά όχι στην ανάλυση Cartesian-PCA και στους πίνακες RMSD.



Εικόνα 4.27 Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RDKWP ξεκινώντας από *trans* διαμόρφωση. Στο κέντρο φαίνεται η προβολή του *trans* τροχιακού στους τρεις principal components της ανάλυσης Dihedral-PCA χρησιμοποιώντας όλα τα βαριά άτομα.

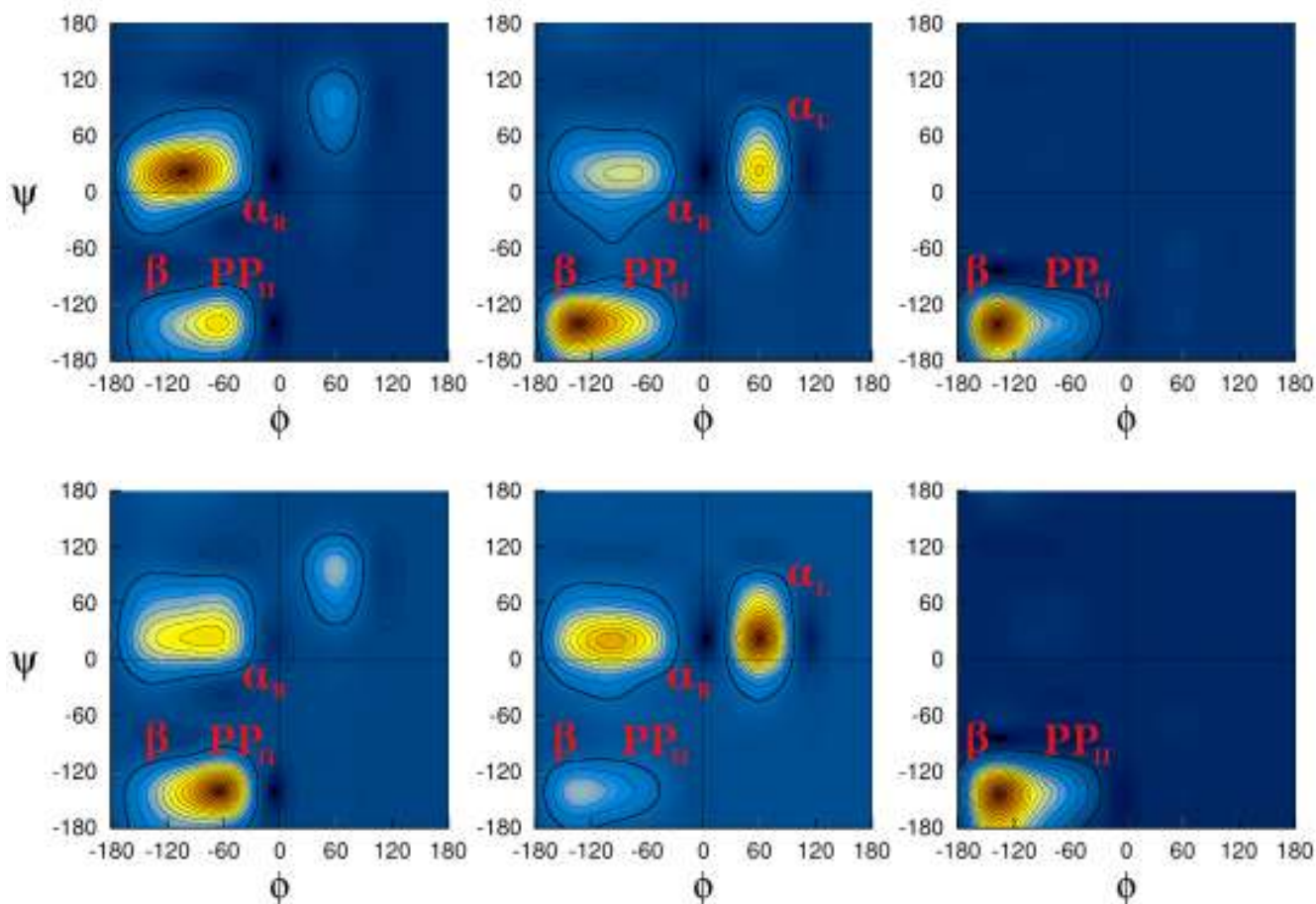
Υποδεικνύονται δύο επίπεδα ισοεπιφάνειας (μέση τιμή και  $10\sigma$  του χάρτη κατανομής) για την υπόδειξη των δύο ισχυρών κορυφών του ενεργειακού τοπίου που αντιστοιχούν σε διακριτά cluster δομών (οι οποίες φαίνονται με τα βέλη). Για κάθε cluster φαίνεται η αντιπροσωπευτική δομή (σε stereo αναπαράσταση) με χρωματισμό με βάση το όνομα του ατόμου. Η δομή με την *cis* διαμόρφωση επισημαίνεται με την υπέρθεση της ίδιας δομής σε διαφάνεια, με τα άτομα του *cis* πεπτιδικού δεσμού σε cprk αναπαράσταση και τον πεπτιδικό σκελετό σε αναπαράσταση με μπλε κορδέλα.

Οι δομές που παρατηρούνται στα *cis* και *trans* τροχιακά είναι πολύ κοντινές όπως διαπιστώνεται τόσο από τους πίνακες RMSD όσο και από τις αντιπροσωπευτικές δομές κάθε cluster: οι δομές των δύο τροχιακών με την *trans* διαμόρφωση έχουν RMSD  $0.6\text{\AA}$  και οι δομές με τη *cis* διαμόρφωση έχουν RMSD  $0.8\text{\AA}$  (για όλα τα βαριά άτομα).



Εικόνα 4.28 Τρισδιάστατο ενεργειακό τοπίο (σε stereo αναπαράσταση) της αναδίπλωσης του πεπτιδίου RDKWP ξεκινώντας από *cis* διαμόρφωση, σε αντιστοιχία με την Εικόνα 4.27.

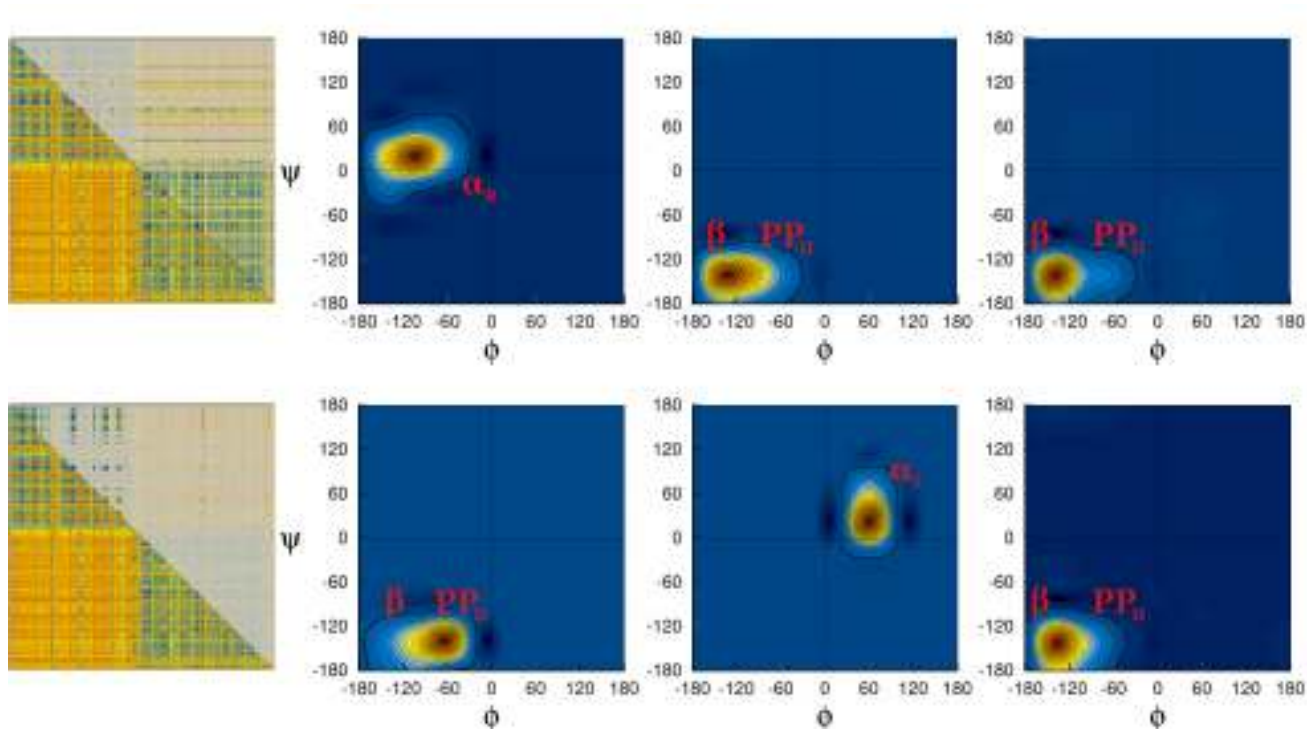
Συνολικά, στο *trans* τροχιακό, παρά την αρχική δομή, βλέπουμε πολύ μεγαλύτερη παραμονή στη *cis* διαμόρφωση. Στο *cis* τροχιακό βλέπουμε μεγαλύτερη παραμονή στην *trans* διαμόρφωση αλλά και επαναφορά στη *cis*. Η παρατήρηση αυτή μπορεί να οφείλεται στην ασυμμετρία του φαινομένου, δηλαδή της συχνότερης μετάβασης από *trans* σε *cis*, όπως παρατηρήθηκε σε πεπτίδια με το μοτίβο XSPX (Hamelberg et al., 2005). Αλλά για να εξακριβωθεί αυτό, χρειάζεται περισσότερος χρόνος προσομοίωσης ώστε να παρατηρηθούν περισσότερες μεταβάσεις. Στις μελέτες αυτές παρατηρήθηκε επίσης ισχυρή συσχέτιση του φαινομένου της ισομερείωσης με τη  $\psi$  γωνία της προλίνης.



Εικόνα 4.29 Διαγράμματα Ramachandran για τα εσωτερικά κατάλοιπα (2-4) του πεπτιδίου RDKWP για το trans (πάνω) και cis (κάτω) τροχιακό.

Προκειμένου να εξετάσουμε την ύπαρξη κάποιας σχέσης μεταξύ της ισομερείωσης του δεσμού και των δίδρων  $\phi/\psi$  γωνιών, υπολογίσαμε διαγράμματα Ramachandran για όλο το τροχιακό (Εικόνα 4.29) και για κάθε cluster (Εικόνες 4.30 και 4.31).

Υπάρχει ξεκάθαρη διάκριση μεταξύ του trans και cis τροχιακού: Το δεύτερο κατάλοιπο παίρνει κυρίως τιμές στην  $\alpha$  περιοχή στο trans τροχιακό ενώ στο cis τροχιακό παίρνει τιμές στην περιοχή  $PP_{II}$ . Το τρίτο κατάλοιπο παίρνει τιμές κυρίως στη  $\beta$  περιοχή στο trans τροχιακό ενώ στο cis τροχιακό παίρνει τιμές στην περιοχή  $\alpha_L$ . Το τέταρτο κατάλοιπο που προηγείται της προλίνης κινείται αποκλειστικά στην περιοχή  $\beta$  του διαγράμματος Ramachandran.



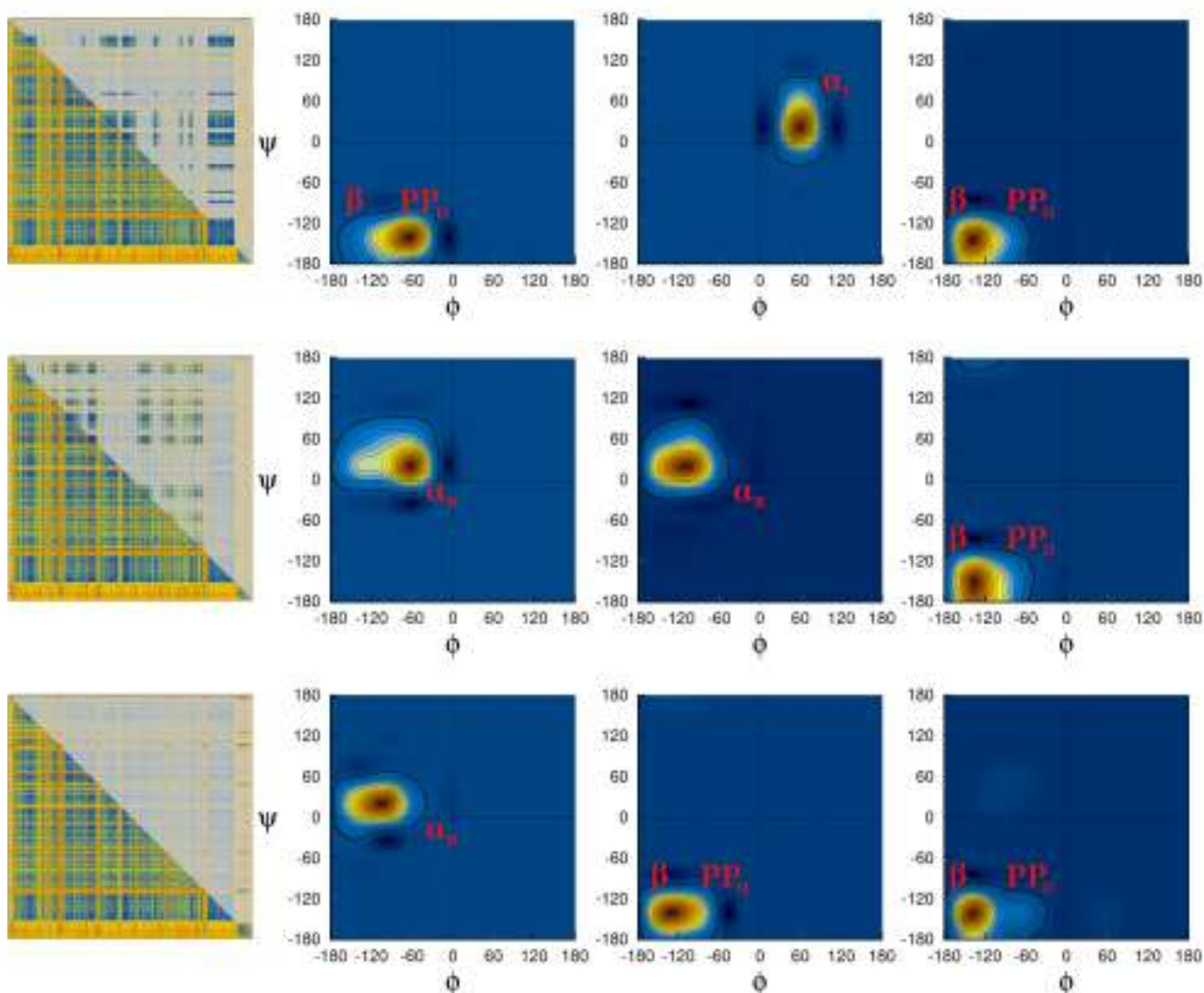
Εικόνα 4.30 Διαγράμματα Ramachandran για το cluster 1 (πάνω) και το cluster 2 (κάτω) που προέκυψαν από την ανάλυση Dihedral-PCA για το *trans* τροχιακό. Αριστερά φαίνεται η προβολή των frames του κάθε cluster πάνω στον πίνακα RMSD.

Οι προτιμήσεις των καταλοίπων για τις διάφορες περιοχές του Ramachandran γίνονται ξεκάθαρες όταν τις δούμε σε επίπεδο cluster:

- Όταν το πεπτίδιο βρίσκεται στην *trans* διαμόρφωση (cluster 2 του *trans* τροχιακού και cluster 1 του *cis* τροχιακού), το δεύτερο κατάλοιπο βρίσκεται αποκλειστικά στην περιοχή  $PP_{II}$ , το τρίτο κατάλοιπο βρίσκεται αποκλειστικά στην  $\alpha_L$  περιοχή και το τέταρτο κατάλοιπο αποκλειστικά στην περιοχή  $\beta$ .
- Όταν το πεπτίδιο βρίσκεται στην *cis* διαμόρφωση (cluster 1 του *trans* τροχιακού και cluster 3 του *cis* τροχιακού), το δεύτερο κατάλοιπο βρίσκεται αποκλειστικά στην περιοχή  $\alpha$ , το τρίτο κατάλοιπο βρίσκεται αποκλειστικά στην περιοχή  $\beta$  και το τέταρτο κατάλοιπο αποκλειστικά στην περιοχή  $\beta$ . Το δεύτερο cluster του *cis* τροχιακού, με τη μεγαλύτερη μέση απόκλιση RMSD ( $\sim 2\text{\AA}$ ) από όλα τα υπόλοιπα φαίνεται να αντιπροσωπεύει μία ενδιάμεση κατάσταση κατά τη μετάβαση από την *trans* στη *cis* διαμόρφωση.

Να σημειωθεί ότι η κυρίαρχη δομή στην *trans* διαμόρφωση που προβλέπεται για το RDKWP με





Εικόνα 4.31 Διαγράμματα Ramachandran για το cluster 1 (πάνω), το cluster 2 (μέση) και το cluster 3 (κάτω) που προέκυψαν από την ανάλυση Dihedral-PCA για το cis τροχιακό. Αριστερά φαίνεται η προβολή των frames του κάθε cluster πάνω στον πίνακα RMSD.

τη μέθοδο adaptive tempering και το συγκεκριμένο force field είναι παρόμοια με τη μία από τις διακριτές δομές (αυτή με την μεγαλύτερη αντιπροσώπευση) που έχουμε παρατηρήσει με τις κλασσικές προσομοιώσεις μοριακής δυναμικής (Εικόνα 4.25). Η μέση τιμή του RMSD είναι 3.1Å για όλα τα βαριά άτομα και 2.1Å για τα άτομα του πεπτιδικού σκελετού, με πλησιέστερη δομή αυτήν του δεύτερου cluster του cis τροχιακού.

*“Religion is flawed, but only because man is flawed.”*

*Dan Brown*







# Κεφάλαιο 5

# ΣΥΖΗΤΗΣΗ

*“Someday there will be a computer labeled “A Cell”,  
and it will accurately predict all details of the behavior of a normal cell,  
as well as that perturbed by exogenous regulatory influences,  
drugs, mutations, and so on.  
I think I still believe this premise  
but my time line for the prediction has expanded considerably.”*  
Alfred G. Gilman

**Σ**τις ενόπτες που προηγήθηκαν παρουσιάστηκε η έρευνα που πραγματοποιήθηκε προς αναζήτηση αναδιπλούμενων πεπτιδίων μέσω προσομοιώσεων μοριακής δυναμικής, ένα μικρό κομμάτι της ευρύτερης έρευνας για την κατανόηση του προβλήματος της αναδίπλωσης των πρωτεϊνών.

Ιδιαίτερα έντονη ήταν η ενασχόληση μας με την έννοια της “αναδιπλωσιμότητας” και την εύρεση ενός τρόπου ανίχνευσης και συστηματικής εκτίμησής της μέσω συναρτήσεων. Στα κεφάλαια που προηγήθηκαν αναλύσαμε εκτενώς τις διάφορες παραμέτρους που επιλέχθηκαν και μελετήθηκαν στα πλαίσια της παρούσας εργασίας και την ικανότητά τους να βαθμολογούν τις τετραπεπτιδικές και τις πενταπεπτιδικές αλληλουχίες ως προς την αναδιπλωσιμότητά τους. Η ανάλυση αυτή οδήγησε στη δημιουργία δύο συναρτήσεων για την εκτίμηση της αναδιπλωσιμότητας με συστηματικό τρόπο: η μία βασίζεται σε ατομικές αποστάσεις και η δεύτερη σε πίνακες RMSD μεταξύ διαδοχικών δομών του τροχιακού και ατομικές διακυμάνσεις. Οι συναρτήσεις αυτές δεν προέκυψαν αβίαστα και άλλαξαν πολλές φορές για να φτάσουν στην τελική μορφή που παρουσιάστηκε εδώ, ώστε να κατατάσσουν σωστά τα πεπτίδια ως προς την αναδίπλωσή τους με ένα αντικειμενικό και συστηματικό τρόπο.

Στα πλαίσια της διατριβής αυτής πραγματοποιήθηκαν πάνω από 15.000 ανεξάρτητες προσομοιώσεις μοριακής δυναμικής που αντιστοιχούν σε 0.27ms υπολογιστικού χρόνου, παρήχθησαν 500Gb δεδομένων και εξετάστηκαν σχεδόν 80.000 γραφήματα για τη μελέτη 1.440 και 7.200 αλληλουχιών μήκους τεσσάρων και πέντε καταλοίπων. Επιστρέφοντας στην αρχική μας ερώτηση, προέκυψε κάποιο αναδιπλούμενο πεπτίδιο από την πορεία που ακολουθήσαμε και βάσει των περιορισμών που θέσαμε στην αλληλουχία τους;

Για την περίπτωση των τετραπεπτιδίων, υποδείξαμε δύο πεπτίδια τα DTRW και RWPD τα



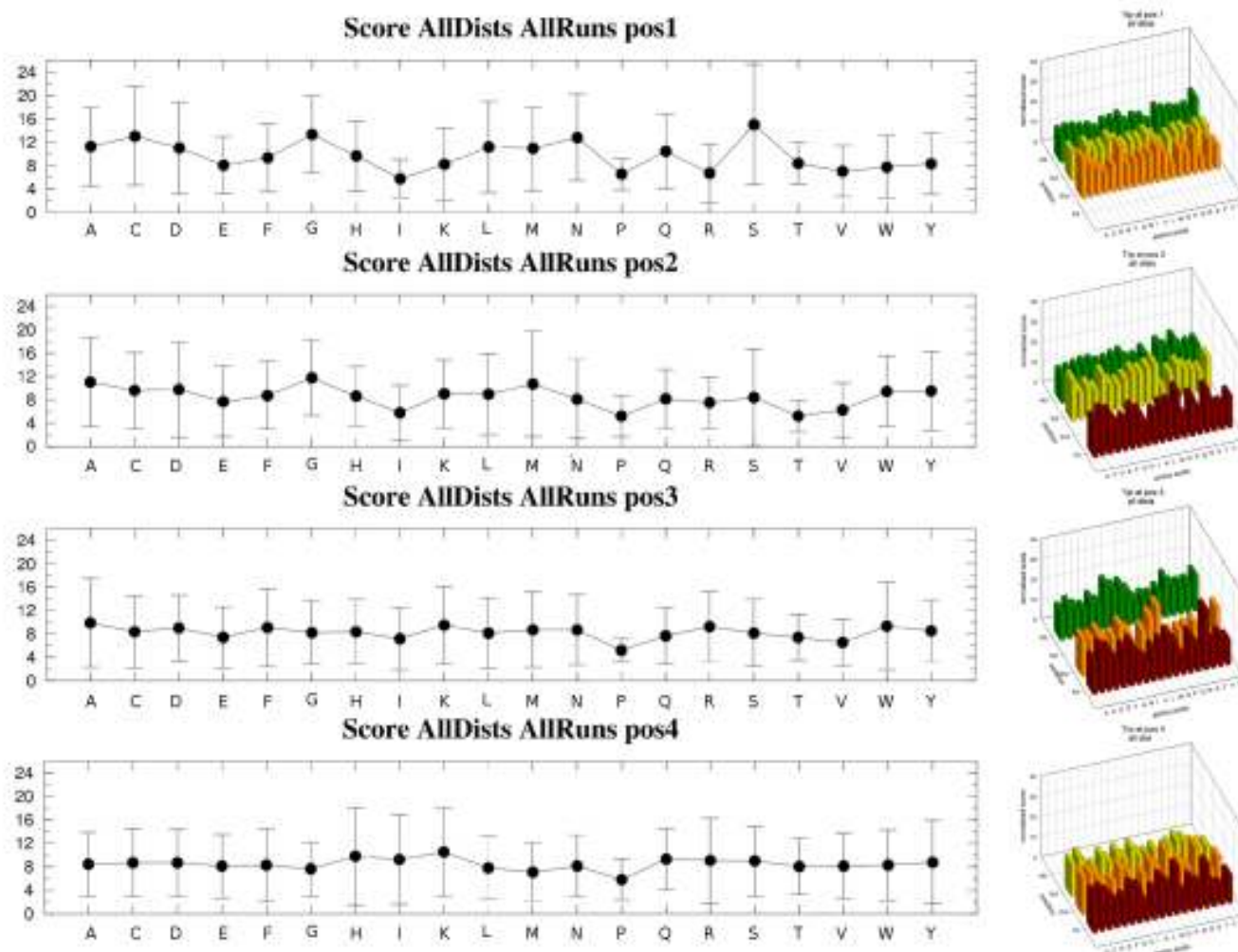
οποία παραμένουν σε αναδιπλωμένη κατάσταση κατά μέσο όρο για 30% και 50% του χρόνου προσομοίωσης αντίστοιχα, ανάλογα με το force field, το μήκος του τροχιακού και τη θερμοκρασία διεξαγωγής της προσομοίωσης. Και τα δύο όμως είναι ασταθή, χαρακτηρίζονται από πολλαπλά γεγονότα αναδίπλωσης/αποδιάταξης και δε μένουν για σημαντικό χρόνο στη κυρίαρχη δομή που προβλέπεται από τα force fields και υπέδειξε η ανάλυσή μας. Επιπλέον χαρακτηρίζονται από μία πληθώρα ασταθών διαμορφώσεων, ο χαρακτηρισμός των οποίων μεταξύ των διαφόρων force fields εμφανίζει μεγάλη ποικιλομορφία. Οι αδυναμίες των force fields να περιγράψουν με αξιοπιστία το σύνολο των μη-αναδιπλωμένων καταστάσεων (disordered state) φαίνεται ότι εξαρτάται σημαντικά από τον τρόπο αντιμετώπισης (cut-off, PME) των long-range αλληλεπιδράσεων (Piana et al., 2012).

Για τα μεγαλύτερου μήκους πενταπεπτίδια, η συντριπτική πλειοψηφία έδειξε εξίσου ασταθή συμπεριφορά με αυτή των τετραπεπτιδίων, με μόνο δύο πεπτίδια, τα NEWRD και RDKWP, να δείχνουν αναδίπλωση σε μία σταθερή δομή για περίπου 30% και 60% του χρόνου προσομοίωσης αντιστοίχως, ανάλογα με τη θερμοκρασία και το force field. Μάλιστα το RDKWP φαίνεται να διατηρείται σε αναδιπλωμένη κατάσταση για σχεδόν 40% του χρόνου προσομοίωσης ακόμα και σε θερμοκρασίες 340K-360K.

Η ενασχόλησή μας με ένα τέτοιο πλήθος αλληλουχιών μας οδήγησε αναπόφευκτα και στην αναζήτηση κάποιας στατιστικά σημαντικής σχέσης μεταξύ αναδιπλωσιμότητας και αλληλουχίας (sequence-structure relationships). Για το σκοπό αυτό πραγματοποιήσαμε μία ανάλυση όπου κάθε αμινοξύ παίρνει μία βαθμολογία βάσει των συναρτήσεων εκτίμησης της αναδιπλωσιμότητας και τη θέση του στην αλληλουχία. Τα προγράμματα που χρησιμοποιήθηκαν για το σκοπό αυτό παραθέτονται στο Παράρτημα (#22-#24). Πιο αναλυτικά, για το σύνολο των προσομοιώσεων των τετραπεπτιδίων έχουμε τη βαθμολογία τους βάσει της συνάρτησης TF2. Κάθε ένα από τα 20 αμινοξέα παίρνει μία συγκεντρωτική βαθμολογία που είναι το άθροισμα των βαθμολογιών των πεπτιδικών αλληλουχιών στις οποίες εμφανίζεται το εκάστοτε αμινοξύ, κανονικοποιημένη ως προς τη συχνότητα εμφάνισης του αμινοξέος. Ο υπολογισμός γίνεται για κάθε θέση της πεπτιδικής αλληλουχίας ξεχωριστά, αλλά και συναρτήσει της θέσης της τρυπτοφάνης, της οποίας επιβάλαμε την παρουσία. Έτσι βλέπουμε τις προτιμήσεις για την παρουσία των αμινοξέων σε κάθε θέση του τετραπεπτιδίου, όπως διαμορφώνονται με βάση τις προσομοιώσεις που πραγματοποιήσαμε και τη συνάρτηση TF2.

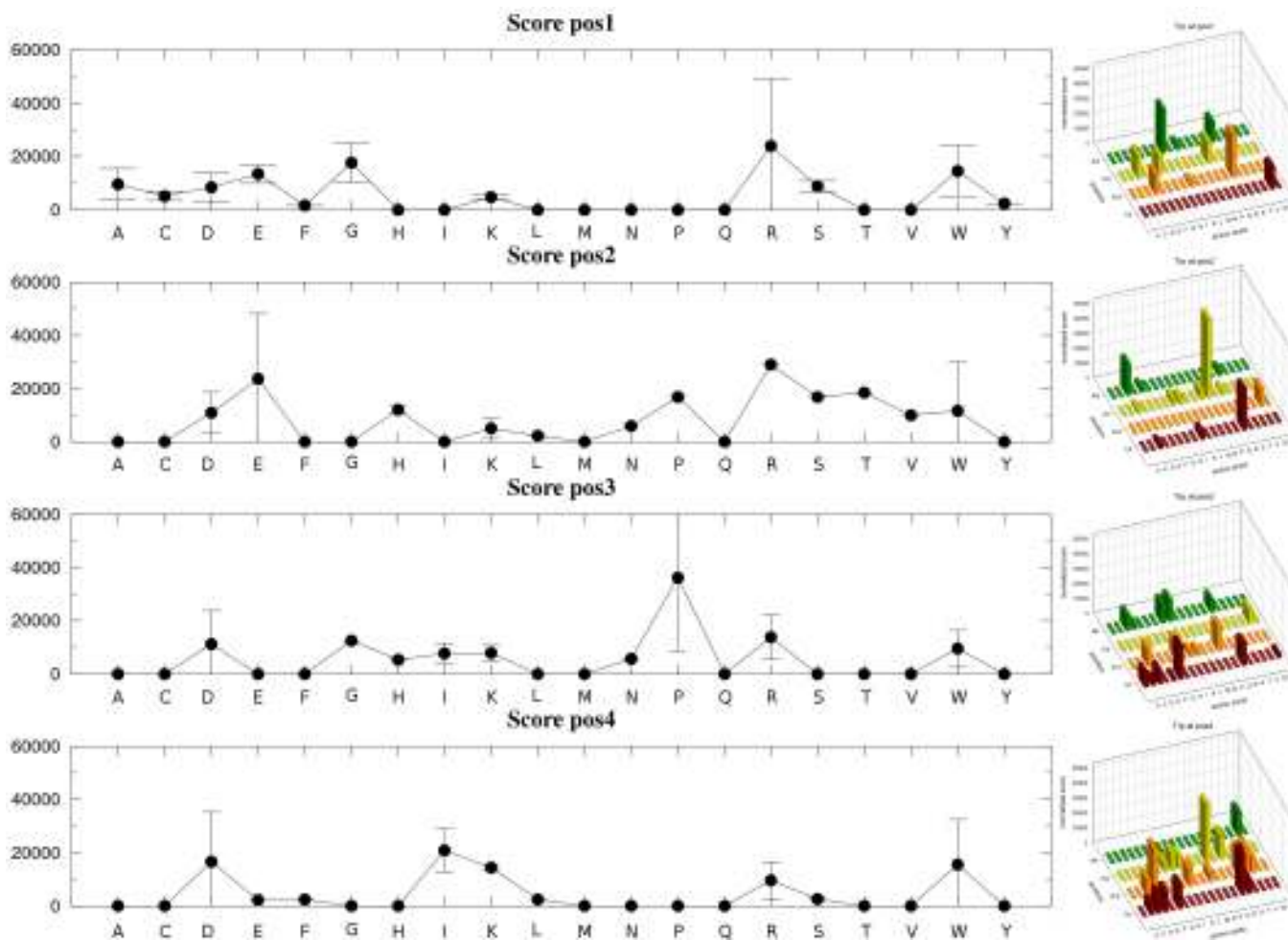
Από την Εικόνα 5.1 όπου βλέπουμε την ανάλυση για το σύνολο των 1440 τετραπεπτιδίων και

τις αρχικές προσομοιώσεις (Ενότητα 3.3) είναι φανερό ότι η πραγματοποίηση μίας τέτοιας συσχέτισης είναι ανούσια. Απουσία κάποιας στατιστικά σημαντικής διαφοροποίησης, χρησιμοποιήσαμε στατιστικά τεστ ([Grubb's test](#)) για την υπόδειξη ακραίων τιμών (outliers) που να υποδεικνύουν την παρουσία κάποιας προτίμησης για συγκεκριμένα αμινοξέα χωρίς ιδιαίτερο αποτέλεσμα πέρα από μία μικρή προτίμηση προς Ala, Cys, Lys, Ser και μειωμένη προτίμηση προς Pro (το οποίο οφείλεται στην αδυναμία σωστής βαθμολόγησης με τη συνάρτηση TF2 των τροχιακών διάρκειας 5ns που έχουμε ήδη υποδείξει).



Εικόνα 5.1 Αριστερά: Κατανομές των μέσων βαθμολογιών και rmsd (error-bars) για κάθε αμινοξύ και κάθε θέση στην αλληλουχία για το σύνολο των 5.760 προσομοιώσεων των τετραπεπτιδίων. Δεξιά: Τρισδιάστατη αναπαράσταση των βαθμολογιών όπως αυτές διαμορφώνονται συναρτήσει της θέσης της τρυπτοφάνης.

Στην Εικόνα 5.2 βλέπουμε τα ίδια αποτελέσματα όπως διαμορφώνονται για το σύνολο των 130 τετραπεπτιδίων (Ενότητα 3.4). Ωστόσο, το εύρος των rmsd (error-bars) σε σχέση με το μέση βαθμολογία είναι τέτοιο ώστε μία μειοψηφία πεπτιδίων με υψηλή βαθμολογία να μπορεί να είναι υπεύθυνη για την προτίμηση που βλέπουμε. Η εικόνα διαφοροποιείται σημαντικά αν συνυπολογίσουμε τη θέση της τρυπτοφάνης στην αλληλουχία: Όταν η τρυπτοφάνη βρίσκεται στην πρώτη θέση, δίπλα της συναντώνται κυρίως πολικά κατάλοιπα R, E, K, στη τρίτη θέση τα κατάλοιπα D, P, G και στην τέταρτη θέση τα κατάλοιπα I, L, R.



Εικόνα 5.2 Αριστερά: Κατανομές των μέσων βαθμολογιών και rmsd (error-bars) για κάθε αμινοξύ και κάθε θέση στην αλληλουχία για τα 130 τετραπεπτιδία. Δεξιά: Τρισδιάστατη αναπαράσταση των βαθμολογιών όπως αυτές διαμορφώνονται συναρτήσει της θέσης της τρυπτοφάνης.

Όταν η τρυπτοφάνη είναι στη δεύτερη θέση, βλέπουμε στη πρώτη και τρίτη θέση τα κατάλοιπα R, K, D και στην τέταρτη θέση τα κατάλοιπα P, N, I, K, D, με ισχυρή προτίμηση για P. Όταν η τρυπτοφάνη είναι στην τρίτη θέση, στις διπλανές θέσεις, δύο (P, H, D, K, L) και τέσσερα (K, I, D, R, E) βλέπουμε φορτισμένα κατάλοιπα ή κατάλοιπα με δακτύλιο. Όταν η τρυπτοφάνη βρίσκεται στην τέταρτη θέση, στη τρίτη θέση βλέπουμε τα κατάλοιπα P, R, I, K, H με τις υπόλοιπες θέσεις να έχουν μεγαλύτερη ποικιλομορφία. Δεδομένου ότι η παρουσία των φορτισμένων καταλοίπων έχει επιβληθεί από εμάς κατά τον αρχικό σχεδιασμό των αλληλουχιών, η ανάλυση αυτή δεν προσφέρει κάποια πληροφορία πέραν της ισχυρής προτίμησης για παρουσία προλίνης ή γλυκίνης, η οποία έρχεται σε συμφωνία με στατιστικά στοιχεία του αμινοξικού περιεχομένου των δομών θηλιάς σε πρωτεΐνες (με μέσο μέγεθος 6-10 κατάλοιπα) που είναι ιδιαίτερος πλούσια στα κατάλοιπα αυτά (Leszczynski et al., 1986).

Με βάση την ανάλυση αυτή για τις προτιμήσεις συγκεκριμένων αμινοξικών καταλοίπων που προκύπτουν από τη συνάρτηση εκτίμησης της αναδιπλωσιμότητας TF2, δημιουργήθηκε η λίστα πεπτιδίων *sequence-based* κατά την ανάλυση που παρουσιάστηκε στην Ενότητα 3.3. Με τον όρο προτίμηση εννοούμε ότι το αμινοξύ για τη συγκεκριμένη θέση έλαβε βαθμολογία πάνω από ένα συγκεκριμένο κατώφλι ( $2\sigma$  της κατανομής) της συγκεντρωτικής βαθμολογίας για κάθε αμινοξύ και κάθε θέση στην αλληλουχία (Εικόνα 5.1, αριστερά). Τα 16 πεπτίδια που ανήκουν στη λίστα αυτή (Εικόνα 3.7) αναμένεται να έχουν υψηλή αναδιπλωσιμότητα.

Η συσχέτιση αυτή έγινε περισσότερο για λόγους πληρότητας παρά για την ανάδειξη κάποιας σχέσης μεταξύ αμινοξικής αλληλουχίας και αναδιπλωσιμότητας, η οποία φαίνεται πως δεν υπάρχει. Για το λόγο αυτό η ανάλυση περιορίστηκε και μόνο στα τετραπεπτίδια, καθώς με τους περιορισμούς της αλληλουχίας που θέσαμε στα πενταπεπτίδια (1 τρυπτοφάνη, 3 φορτισμένα κατάλοιπα και όλα διαφορετικά) δεν θα είχε κάποιο νόημα.

Εάν ανατρέξουμε πίσω στις λίστες των τετραπεπτιδίων (Εικόνα 3.7, σελ.84), διαπιστώνουμε ότι από το σύνολο των 36 τετραπεπτιδίων (Ενότητα 3.5), το 28% προέκυψε με συστηματικό τρόπο από τις συναρτήσεις εκτίμησης της αναδιπλωσιμότητάς TF1 και TF2 (cluster-based), το 33% προέκυψε από την οπτική εξέταση των γραφικών παραστάσεων των ατομικών αποστάσεων (graph-based), και το 23% προέκυψε από την ανάλυση των προτιμήσεων των αμινοξέων (sequence-based). Ακόμα και τα τέσσερα δυνητικά αναδιπλούμενα τετραπεπτίδια (Ενότητα 3.5) προέρχονται από διαφορετικές λίστες.

Για τα πενταπεπτίδια και τη σύνδεση τους με την πληροφορία που υπάρχει ήδη στην PDB

(Ενότητα 4.1, PDB και nonPDB, Πίνακας 4.1) προσπαθήσαμε να κάνουμε κάποια σύνδεση μεταξύ της αναδιπλωσιμότητας μίας αλληλουχίας και της παρουσίας της σε πειραματικά προσδιορισμένη δομή. Στο στάδιο των 32 (Ενότητα 4.5) που η διάρκεια των προσομοιώσεων ήταν αρκετή ώστε να εξαγάγουμε συμπεράσματα αναφορικά με την σταθερότητα της αναδίπλωσής τους, βλέπουμε το 56% των πεπτιδίων (18 από τα 32) να ανήκουν στη λίστα NonPDB και παρομοίως το 50% στο στάδιο των 8 πενταπεπτιδίων (Ενότητα 4.6). Από τα δύο καλύτερα πενταπεπτίδια που προσδιορίσαμε, το NEWRD ανήκει στη λίστα PDB και το RDKWP ανήκει στη λίστα NonPDB.

Η ασάφεια των αποτελεσμάτων των αναλύσεων αυτών μας οδηγεί στο εξής συμπέρασμα: τα κριτήρια για την αναδιπλωσιμότητα δε βρίσκονται σε επίπεδο αλληλουχίας, αλλά είναι θέμα θερμοδυναμικής σταθερότητας (Irbäck et al., 1997) και κατά συνέπεια η αναδίπλωση των πεπτιδίων και των πρωτεϊνών μπορεί να προσεγγιστεί όχι μέσω εμπειρικών αλγόριθμων αλλά μέσω της αναλυτικής περιγραφής της ενέργειας του συστήματος, με χρησιμότερο εργαλείο τις προσομοιώσεις μοριακής δυναμικής (Best, 2012).

Από τη δική μας έρευνα προέκυψε ότι οι βασισμένες σε ατομικές αποστάσεις συναρτήσεις είχαν περιορισμένη διακριτική ικανότητα για μικρού μήκους τροχιακά (5ns) και μήκος αλληλουχίας 4 καταλοίπων. Ωστόσο η αδυναμία αυτή δεν παρατηρήθηκε όταν περάσαμε σε μεγαλύτερης διάρκειας τροχιακά (20ns) και μεγαλύτερου μήκους πεπτίδια (5 κατάλοιπα). Τα δύο καλύτερα πενταπεπτίδια NEWRD και RDKWP που προέκυψαν από την εφαρμογή της συνάρτησης TF2, αποδεικνύουν την ορθότητα του παραπάνω ισχυρισμού και δείχνουν την αποτελεσματικότητα της συνάρτησης TF2 ακόμα και παρουσία της προλίνης. Η διακριτική ικανότητα της συνάρτησης TF3 στην ανίχνευση της δημιουργίας δομής από τα πεπτίδια (*ab initio*) και η αποτελεσματικότητά της στην εκτίμηση της σταθερότητας αυτής διαφαίνεται από το γεγονός ότι καταφέραμε να υποδείξουμε πεπτίδια με σταθερή αναδίπλωση, ακόμα και με τους περιορισμούς της αλληλουχίας που θέσαμε.

Η πορεία που ακολουθήσαμε στο σύνολο των 1.440 τετραπεπτιδίων και 7.200 πενταπεπτιδίων και οι συναρτήσεις εκτίμησης της αναδιπλωσιμότητας που αναπτύξαμε θα μπορούσαν να εφαρμοστούν σε μεγαλύτερου μήκους πεπτίδια (εξαπεπτίδια, επταπεπτίδια) ώστε να εξεταστεί περαιτέρω η αποτελεσματικότητά τους. Δυστυχώς θα μπορούσε να εφαρμοστούν και σε οποιοδήποτε σύνολο πεπτιδίων δεδομένου μήκους που προκύπτουν βάσει ενός μοτίβου αλληλουχίας (και ίσως και δομής). Μία επιπλέον παράμετρος που πρέπει να τεθεί υπό μελέτη



είναι κατά πόσο η παρουσία πολλαπλών αντιγράφων των πεπτιδίων οδηγεί σε παρόμοια αποτελέσματα ή οδηγεί σε άμορφα συσσωματώματα και κατακρήμνιση από τη μεταξύ τους αλληλεπίδραση στο διάλυμα, αποτρέποντας την περαιτέρω πειραματική τους μελέτη. Παρότι η έρευνα αυτή χρειάζεται σημαντικά μεγαλύτερη υπολογιστική ισχύ είναι σημαντική τόσο αυτή όσο και η υπολογιστική εκτίμηση της ενέργειας διαλυτοποίησης (solvation energy) προτού προχωρήσει κανείς σε πειραματικές διαδικασίες.

Τα συμπεράσματα της παρούσας εργασίας για τα δυνητικά αναδιπλούμενα πεπτίδια που υποδείξαμε θα αποκτήσουν υπόσταση με την περαιτέρω πειραματική πλέον μελέτη τους. Μία ενδεικτική πορεία που θα μπορούσε να ακολουθηθεί είναι: (1) σύνθεση σε υψηλή καθαρότητα (>95%) και μεγάλη ποσότητα (>50mg) και μέτρηση διαλυτότητας, (2) αρχικός βιοφυσικός χαρακτηρισμός, μέσω φασμάτων κυκλικού διχρωϊσμού (near-UV (250-350nm) CD) και φασματοσκοπίας φθορισμού (απλής και single-molecule FRET) για την επιβεβαίωση της σταθεροποίησης της τρυπτοφάνης και (3) πειραματικός προσδιορισμός της δομής μέσω φασματοσκοπίας NMR με παράλληλες προσπάθειες για κρυσταλλογραφικές μελέτες. Η σύγκριση μεταξύ των θεωρητικών προγνώσεων μας με πειραματικά αποτελέσματα δε μπορεί παρά να είναι παραγωγική και να οδηγήσει σε βελτιώσεις και των δύο.

*“ As for the future,  
your task is not to foresee it,  
but to enable it. ”*

*Antoine de Saint - Exupery*







- (1) Adhikari B. & Banerjee A. (2011). Self-assembling peptides: from molecules to nanobiomaterials. *J. Ind. Inst. Sci.* **91**, 471-483.
- (2) Aliev A.E. & Courtier-Murias D. (2010). Experimental verification of force fields for molecular dynamics simulations using Gly-Pro-Gly-Gly. *J. Phys. Chem. B.*, XXX, A-R.
- (3) Aliev A.E., Courtier-Murias D., Bhandal S. & Zhou S. (2010). A combined NMR/MD/QM approach for structure and dynamics elucidations in the solution state: pilot studies using tetrapeptides. *Chem. Commun.* **46**, 695-697.
- (4) Allen L.R. & Paci E. (2009). Orientational averaging of dye molecules attached to proteins in Förster resonance transfer measurements: insights from a simulation study. *J. Chem. Phys.* **131**, 065101.
- (5) Altis A., Nguyen P.H., Hegger R. & Stock G. (2007). Dihedral angle principal component analysis of molecular dynamics simulations. *J. Chem. Phys.* **126**, 244111.
- (6) Altis A., Otten M., Nguyen P.H., Hegger R. & Stock G. (2008). Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis. *J. Chem. Phys.* **128**, 245102.
- (7) Amadei A., Daidone I., Di Nola A. & Aschi M. (2010). Theoretical-computational modelling of infrared spectra in peptides and proteins: a new frontier for combined theoretical-experimental investigations. *Curr. Opin. Struct. Biol.* **20**, 155-161.
- (8) Amadei A., Linssen A.B.M. & Berendsen H.J.C. (1993). Essential dynamics of proteins. *Proteins* **17**, 412-425.
- (9) Anfinsen C.B. (1973). Principles that govern the folding of protein chains. *Science* **181**, 223-230.
- (10) Antonosova Z., Mackova M., Kral V. & Macek T. (2009). Therapeutic applications of peptides and proteins: parenteral forever? *Trends Biotechnol.* **27**, 628-635.
- (11) Balle S.M. & Palermo D. (2007). Enhancing an Open Source Resource Manager with Multi-Core/Multithreaded Support. Job Scheduling Strategies for Parallel Processing.
- (12) Balsera M., Wriggers W., Dono Y. & Schulten K. (1996). Principal component analysis and long time protein dynamics. *J. Phys. Chem.* **100**, 2567-2572.
- (13) Barrick D. (2009). What have we learned from the studies of two-state folders, and what are the unanswered questions about two-state protein folding? *Phys. Biol.* **6**, 015001
- (14) Bashford D., Case D.A., Choi C. & Giipert G.P. (1997). A computational study of the role of solvation effects in reverse turn formation in tetrapeptides APGD and APGN. *J. Am. Chem. Soc.* **119**,



4964-4971.

(15) Beberg A.L., Ensign D.L., Jayachandran G., Khaliq S. & Pande V.S. (2009). Folding@Home: Lessons from eight years of volunteer distributed computing. IEEE International Symposium on Parallel & Distributed Processing, Rome, Italy

(16) Berezhkovskii A.M., Tofoleanu F. & Buchete N.-V. (2011). Are peptides good two-state folders? *J. Chem. Theory Comput.* **7**, 2370-2375.

(17) Bernstein F.C., Koetzle T.F., Williams G.J.B., Meyer E.F. Jr., Brice M.D., Rodgers J.R., Kennard O., Shimanouchi T. & Tasumi M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.

(18) Bernstein H.J. (1999). Rasmol 2.7.1. Molecular Graphics Visualization Tool.

(19) Best R.B. (2012). Atomistic molecular simulations of protein folding. *Curr. Opin. Struct. Biol.* **22**, 52-61.

(20) Best R.B., Buchete N.V. & Hummer G. (2008). Are current molecular dynamics force fields too helical? *Biophys. J.* **95**, L07-09.

(21) Best R.B. & Hummer G. (2009). Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *J. Phys. Chem. B* **113**, 9004-9015.

(22) Best R.B. & Mittal J. (2010). Balance between alpha and beta structures in *ab initio* protein folding. *J. Phys. Chem. B* **114**, 8790-8798.

(23) Bieri, O., Wirz, J., Hellrung, B., Schutkowski, M., Drewello, M. & Kiefhaber, T. (1999). The speed limit for protein folding measured by triplet-triplet energy transfer. *Proc Natl. Acad. Sci. USA* **96**, 9597-9601.

(24) Boned R., van Gunsteren W.F. & Daura, X. (2008). Estimating the temperature dependence of peptide folding entropies and free enthalpies from total energies in molecular dynamics simulations. *Chemistry* **14**, 5039-5046.

(25) Borhani D.W. & Shaw D.E. (2012). The future of molecular dynamics simulations in drug discovery. *J. Comput. Aided Mol. Des.* **26**, 15-26.

(26) Bowers K.J., Chow E., Xu H., Dror R., Eastwood M.P., Gregersen B.A., Klepeis J.L., Kolossvary I., Moraes M.A., Sacerdoti F.D., Salmon J.K., Shan Y. & Shaw D. (2006). Scalable algorithms for molecular dynamics simulations on commodity clusters. Proceedings of the ACM/IEEE Conference on

Supercomputing (SC06), Tampa, Florida.

(27) Bowie J.U., Lüthy R. & Eisenberg D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164-170.

(28) Bowman G.R. & Pande V. S. (2010). Protein folded states are kinetic hubs. *Proc Natl. Acad. Sci. USA* **107**, 10890-10895.

(29) Bowman G.R., Voelz V.A. & Pande V.S. (2011). Taming the complexity of protein folding. *Curr. Opin. Struct. Biol.* **21**, 4-11.

(30) Brandts J.F., Halvorson H.R. & Brennan M. (1975). Consideration of the Possibility that the slow step in protein denaturation reactions is due to cis-trans isomerism of proline residues. *Biochemistry* **14**, 4953-4963.

(31) Bryngelson J.D., Onuchic J.N., Socci N.D. & Wolynes P.G. (1995). Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **21**, 167-195.

(32) Buck M., Bouguet-Bonnet S., Pastor R.W. & Mackerell A.D. Jr. (2006). Importance of the CMAP correction to the CHARMM22 protein force field: dynamics of hen lysozyme. *Biophys. J.* **15**, L36-L38.

(33) Buscaglia M., Lapidus L.J., Eaton W.A. & Hofrichter J. (2006). Effects of denaturants on the dynamics of loop formation in polypeptides. *Biophys. J.* **91**, 276-288.

(34) Caflisch A. (2004). Protein folding: simple models for a complex process. *Structure* **12**, 1750-1752.

(35) Caflisch A. (2012). Complexity in protein folding: simulation meets experiment. *Curr. Phys. Chem.* **2**, 4-11.

(36) Cai Y.D., Li Y.X. & Chou K. C. (1999) Classification and prediction of b-turn types by neural network. *Adv. Eng. Software* **30**, 347-352.

(37) Case D.A., Cheatham T.E. III, Darden T., Gohlke H., Luo R., Merz K.M. Jr., Onufriev A., Simmerling C., Wang B. & Woods R.J. (2005). The AMBER Biomolecular Simulation Programs. *J. Comput. Chem.* **26**, 1668-1688.

(38) Caves L.S.D., Evanseck J.D. & Karplus M. (1998). Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. *Prot. Sci.* **7**, 649-666.

(39) Cellmer T., Buscaglia M., Henry E.R., Hofrichter J. & Eaton W.A. (2010). Making connections between ultrafast protein folding kinetics and molecular dynamics simulations. *Proc Natl. Acad. Sci. USA* **108**, 6103-6108.

- (40) Chan C.-K., Hu Y., Takahashi S., Rousseau D.L., Eaton W.A., Hofrichter J. (1997). Submillisecond protein folding kinetics studied by ultrarapid mixing. *Proc Natl. Acad. Sci. USA* **94**, 1779-1784.
- (41) Chazin W.J., Kördel J., Drakenberg T., Thulin E., Brodin P., Grunsström T & Forsén S. (1989). Proline isomerism leads to multiple folded conformations of calbindin D9k: direct evidence from two-dimensional <sup>1</sup>H NMR spectroscopy. *Proc Natl. Acad. Sci. USA* **86**, 2195-2198.
- (42) Chiarabelli C., Vrijbloed J.W., Lucrezia D., Thomas R.M., Stano P., Polticelli F., Ottone T., Papa E. & Luisi P.L. (2006). Investigation of de novo totally random biosequences. *Chem Biodivers.* **3**, 840-859.
- (43) Chiti F., Calamai M., Taddei N., Stefani M., Ramponi G. & Dobson C.M. (2002). Studies of the aggregation of mutant proteins *in vitro* provide insights into the genetics of amyloid diseases. *Proc Natl. Acad. Sci. USA* **99**, 16419-16426.
- (44) Chothia C. & Lesk A.M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **327**, 711-717.
- (45) Chou K. C. (1997) Prediction of b-Turns. *J. Pept. Res.* **49**, 120-144.
- (46) Chou K. C. (2000). Prediction of tight turns and their types in proteins. *Anal. Biochem.* **286**, 1-16.
- (47) Chou K. C. & Blinn, J. R. (1997) Classification and prediction of b-turn types. *J. Protein Chem.* **16**, 575-595.
- (48) Chou P.Y. & Fasman G.D. (1978). Empirical predictions of protein conformation. *Annu. Rev. Biochem.* **47**, 251-276.
- (49) Chou P. Y. & Fasman, G. D. (1979) Prediction of b-turns. *Biophys. J.* **26**, 367-384.
- (50) Chowdhury S., Lee M.C., Xiong G. & Duan Y. (2003). Ab initio folding simulation of the Trp-cage mini-protein approaches NMR resolution. *J. Mol. Biol.* **327**, 711-717.
- (51) Cid H. & Arellano, A. (1982) Secondary structure prediction of protamines. *Int. J. Biol. Macromol.* **4**, 3-8.
- (52) Cochran A.G., Skelton N.J. & Starovasnik M. (2001). Tryptophan zippers: stable, monomeric  $\beta$ -hairpins. *Proc Natl. Acad. Sci. USA* **98**, 5578-5583.
- (53) Cohen F. E., Abarbanel R. M., Kuntz I. D., and Fletterick, R. J. (1986) Turn prediction in proteins using a pattern-matching approach. *Biochemistry* **25**, 266-275.
- (54) Compiani M., Farisseli P, Martelli P.L. & Casadio R. (1998). An entropy criterion to detect minimally frustrated intermediates in native proteins. *Proc Natl. Acad. Sci. USA* **95**, 9290-9294.

- (55) Cooper S., Khatib F., Treville A., Barbero J., Lee J., Beenen M., Leaver-Fay A., Baker D., Popović Z. & Foldit players. (2010) Predicting protein structures with a multiplayer online game. *Nature* **466**, 756-760.
- (56) Cootes A.P., Curmi P.M.G. & Torda A.E. (2000). Automated protein design and sequence optimization: scoring functions and the search problem. *Curr. Protein Pept. Sci.* **1**, 255-271.
- (57) Cornell W.D., Cieplak P., Bayly C.I., Gould I.R., Merz K.M. Jr., Ferguson D.M., Spellmeyer D.C., Fox T., Caldwell J.W. & Kollman P.A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**, 5179-5197.
- (58) Cossio P., & Laio A. & Pietrucci F. (2011). Which similarity measure is better for analyzing protein structures in a molecular dynamics trajectory? *Phys. Chem. Chem. Phys.* **13**, 10421-10425.
- (59) Craik D.J. (2006). Seamless proteins tie up their loose ends. *Science* **211**, 1563-1567.
- (60) Creighton T.E. (1993). *Proteins: Structures and Molecular Properties*. Freeman, New York.
- (61) Daggett V. & Fersht A.R. (2003). Is there a unifying mechanism for protein folding? *Trends Biochem. Sci.* **28**, 1750-1752.
- (62) Daidone I., Neuweiler H., Doose S., Sauer M. & Smith J.C. (2010). Hydrogen-bond driven loop-closure kinetics in unfolded polypeptide chains. *PLoS Comput. Biol.* **6**, e1000645:1-9.
- (63) Dallüge R., Oschmann J., Birkenmeier O., Lücke C., Lilie H., Rudolph R. & Lange C. (2007). A tetrapeptide fragment-based design method results in highly stable artificial proteins. *Proteins* **68**, 839-849.
- (64) Darden T., York D. & Pedersen L. (1993). An  $N \log(N)$  method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089-10092.
- (65) Daura X., Gademann K., Javn B., Seebach D., van Gunsteren W.F. & Mark A.E. (1999). Peptide folding: when simulation meets experiment. *Angew. Chem. Int. Ed.* **38**, 236-240.
- (66) Daura X., Gademann K., Schafer H., Javn B., Seebach D. & van Gunsteren W.F. (2001). The beta-peptide hairpin in solution: conformational study of a beta-hexapeptide in methanol by NMR spectroscopy and MD simulation. *J. Am. Chem. Soc.* **123**, 2393-2404.
- (67) Daura X., Javn B., Seebach D., van Gunsteren W.F. & Mark. A.E. (1998). Reversible peptide folding in solution by molecular dynamics simulation. *J. Mol. Biol.* **280**, 925-932.
- (68) Daura X., van Gunsteren W.F. & Mark. A.E. (1999). Folding-Unfolding thermodynamics of a  $\beta$

- heptapeptide from equilibrium simulations. *Proteins* **34**, 269-280.
- (69) Debe D.A., Carlson M.J., Goddard W.A. (1999). The topomer-sampling model of protein folding. *Proc Natl. Acad. Sci. USA* **96**, 2596-2601.
- (70) DeMarco M.L., Alonso D.O.V. & Daggett V. (2004). Diffusing and colliding: the atomic level folding/unfolding pathway of a small helical protein. *J. Mol. Biol.* **341**, 1109-1124.
- (71) Demchuck E., Bashford D. & Case D.A. (1997). Dynamics of a type VI reverse turn in a linear peptide in aqueous solution. *Fold. Des.* **2**, 35-46.
- (72) Dill K.A. (1985). Theory for the folding and stability of globular proteins. *Biochemistry* **24**, 1501-1509.
- (73) Dill K.A. (1990). Dominant force in protein folding. *Biochemistry* **29**, 7133-7155.
- (74) Dill K.A., Fiebig K.M. & Chan H.S. (1993). Cooperativity in protein-folding kinetics. *Proc Natl. Acad. Sci. USA* **90**, 1942-1946.
- (75) Dill K.A. & Chan H.S. (1997). From Levinthal to pathways to funnels. *Nat. Struct. Biol.* **4**, 10-19.
- (76) Dill K.A., Ozkam S.B., Shell M.S & Weikl T.R. (2008). The protein folding problem. *Annu. Rev. Biophys.* **37**, 289-316.
- (77) Dill K.A., Ozkam S.B., Weikl T.R., Chodera J.D. & Voelz V.A. (2007). The protein folding problem: when will it be solved? *Curr. Opin. Struct. Biol.* **17**, 342-346.
- (78) Doig A.J. & Sternberg M.J.E. (1995). Side-chain conformational entropy in protein folding. *Protein Sci.* **4**, 2247-2251.
- (79) Dror R.O., Dirks R.M., Grossman J.P., Xu H. & Shaw D.E. (2012). Biomolecular simulation: a computational microscope for molecular biology. *Annu. Rev. Biophys.* **41**, 429-452.
- (80) Duan Y. & Kollman P.A. (1998). Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **282**, 740-744.
- (81) Duan Y., Wu C., Chowdhury S., Lee M.C., Xiong G., Zhang W., Yang R., Cieplak P., Luo R., Lee T., Caldwell J., Wang J. & Kollman P. (2003). A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **24**, 1999-2012.
- (82) Eaton W.A., Munoz V., Thompson P.A., Henry E.R. & Hofrichter J. (1998). Kinetics and dynamics of loops,  $\alpha$ -helices,  $\beta$ -hairpins and fast-folding proteins. *Acc. Chem. Res.* **31**, 745-753.



- (83) Editorial (2005): So much more to know. *Science* **309**, 78-102.
- (84) Edwards C.M.B., Cohen M.A. & Bloom S.R. (1999). Peptides as drugs. *Q. J. Med (Editorial)* **92**, 1-4.
- (85) Eidenschink L., Kier B.L., Huggins K.N.L. & Andersen N.H. (2009). Very short peptides with stable folds: building on the interrelationship of Trp/Trp, Trp/cation, and Trp/backbone-amide interaction geometries. *Proteins* **75**, 308-322.
- (86) Ensign D.L. & Pande V.S. (2009). The Fip35 WW domain folds with structural and mechanistic heterogeneity in molecular dynamics simulations. *Biophys. J.* **96**, L53-L55.
- (87) Ensign D.L., Kasson P.M. & Pande V.S. (2007). Heterogeneity even at the speed limit of folding: large-scale molecular dynamics study of a fast-folding variant of the villin headpiece. *J. Mol. Biol.* **374**, 806-816.
- (88) Evans P.A., Dobson C.M., Kautz R.A., Hartfield G. & Fox R.O. (1987). Proline isomerism in staphylococcal nuclease characterized by NMR and site-directed mutagenesis. *Nature* **329**, 266-268.
- (89) Faver J.C., Benson M.L., He X., Roberts B.P., Wang B., Marshall M.S., Sherrill C.D. & Merz K.M. Jr. (2011). The energy computation paradox and *ab initio* protein folding. *PLoS one* **6**, e18868.
- (90) Feenstra K.A., Peter C., Scheek R.M., van Gunsteren W.F. & Mark A.E. (2002). A comparison of methods for calculating NMR cross-relaxation rates (NOESY and ROESY intensities) in small peptides. *J. Biomol. NMR* **23**, 181-194.
- (91) Feige M.J. & Paci E. (2008). Rate of loop formation in peptides: a simulation study. *J. Mol. Biol.* **382**, 556-565.
- (92) Feng Y. & Luo L. (2008). Use of tetrapeptide signals for protein secondary-structure prediction. *Amino Acids* **35**, 607-641.
- (93) Ferrara P., Apostolakis J. & Caflisch A. (2000) Thermodynamics and kinetics of folding of two model peptides investigated by molecular dynamics simulations. *J. Phys. Chem. B.* **104**, 5000-5010.
- (94) Ferrara P. & Caflisch A. (2000) Folding simulations of a three-stranded antiparallel beta-sheet peptide. *Proc Natl. Acad. Sci. USA* **20**, 10780-10785.
- (95) Fersht A.R. (1997). Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* **7**, 3-9.
- (96) Fersht, A.R. (2002). On the simulation of protein folding by short time scale molecular dynamics and distributed computing. *Proc Natl. Acad. Sci. USA* **99**, 14122-14125.
- (97) Fierz B. & Kiefhaber T. (2007). End-to-end vs interior loop formation kinetics in unfolded

- polypeptide chains. *J. Am. Chem. Soc.* **129**, 672-679.
- (98) Fierz B., Satzger H., Root C., Gilch P., Zinth W. & Kiefhaber T. (2007). Loop formation in unfolded polypeptide chains on the picoseconds to microseconds time scale. *Proc Natl. Acad. Sci. USA* **104**, 2163-2168.
- (99) Fisher S., Dunbrack R.L. & Karplus M. (1994). Cis-Trans imide isomerization of the proline dipeptide. *J. Am. Chem. Soc.* **116**, 11931-11937.
- (100) Flöck D., Rossetti G., Daidone I., Amadei A. & Di Nola A. (2006). Aggregation of small peptides studied by molecular dynamics simulations. *Proteins* **65**, 914-921.
- (101) Freddolino P.L., Harrison C.B., Liu Y & Schulten K. (2010). Challenges in protein folding simulations: timescale, representation, and analysis. *Nat. Phys.* **6**, 751-758.
- (102) Freddolino P.L., Liu F., Grubele M. & Schulten K. (2008). Ten-microsecond MD simulation of a fast-folding WW domain. *Biophys. J.* **94**, L75-L77.
- (103) Freddolino P.L., Park S., Roux B & Schulten K. (2009). Force field bias in protein folding simulations. *Biophys. J.* **96**, 3772-3780.
- (104) Fuchs P.F.J., Bonvin A.M.J.J., Bochicchio B., Pepe A., Alix A.J.P. & Tamburro A.M. (2006). Kinetics and thermodynamics of type VIII  $\beta$ -turn formation: a CD, NMR, and microsecond explicit molecular dynamics study of the GDNP tetrapeptide. *Biophys. J.* **90**, 2745-2759.
- (105) Gao F., Wang Y., Giv Y., Li Y., Sha Y., Lai L. & Wu H. (2002)  $\beta$ -turn formation by a six-residue linear peptide in solution. *J. Pept. Res.* **60**, 75-80.
- (106) Garcia A.E. (1992). Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.* **68**, 2696-2699.
- (107) Garcia A.E. Sambonmatsu K.Y. (2002).  $\alpha$ -helical stabilization by side-chain shielding of backbone hydrogen bonds. *Proc Natl. Acad. Sci. USA* **99**, 2381-2391.
- (108) Gellman S.H. (1998). Foldamers: a manifesto. *Acc Chem. Res.* **31**, 173-180.
- (109) Gianni S., Guydosh N.R., Khan F., Caldas T.D., Mayor U., White G.W., DeMarco M.L., Daggett V. & Fersht A.R. (2003). Unifying features in protein-folding mechanisms. *Proc Natl. Acad. Sci. USA* **100**, 13286-13291.
- (110) Glättli A., Daura X., Bindshädler P., Jaun B., Mahajan Y.R., Mathad R.I., Rueping M., Seebach D. & van Gunsteren W.F. (2005). On the influence of charged side chains on the folding-unfolding equilibrium of  $\beta$ -peptides: A molecular dynamics simulation study. *Chem. Eur. J.* **11**, 7276-7293.

- (111) Glykos N.M. (2006). Software news and updates carma: a molecular dynamics analysis program. *J. Comput. Chem.* **27**, 1765-1768. <http://utopia.duth.gr/~glykos/Carma.html>
- (112) Gnanakaram S. & Garcia A.E. (2003). Validation of an all-atom protein force field: from dipeptides to larger peptides. *J. Phys. Chem. B* **107**, 12555-12557.
- (113) Gnanakaram S. & Garcia A.E. (2005). Helix-coil transition of alanine peptides in water: force field dependence on the folded and unfolded structures. *Proteins* **59**, 773-782.
- (114) Gnanakaram S., Nymeyer H., Portman J. Sanbonmatsu K.Y. & Garcia A.E. (2003). Peptide folding simulations. *Curr. Opin. Struct. Biol.* **13**, 168-174.
- (115) Gordon H.L. & Somorjai R.L. (1992). Fuzzy cluster analysis of molecular dynamics trajectories. *Proteins* **14**, 249-264.
- (116) Greenfield N.J. (2006). Using circular dichroism spectra to estimate protein secondary structure. *Nat. Protoc.* **1**, 2876-2890.
- (117) Guruprasad K., Pavan M.N., Rajkumar S. & Swaminathan S. (2000). Isolated and multiple  $\beta$ -turns with proline in the third position. *Curr. Sci.* **79**, 992-994.
- (118) Guruprasad K. & Rajkumar S. (2000).  $\beta$ - and  $\gamma$ -turns in proteins revisited: A new set of amino acid turn-type dependent positional preferences and potentials. *J. Biosci.* **25**, 143-156.
- (119) Guvench O. & Mackerell A.D. Jr. (2008). Comparison of protein force fields for molecular dynamics simulations. *Methods Mol. Biol.* **443**, 63-88.
- (120) Haile J.M. (1997). Molecular dynamics simulation. John Wiley & Sons.
- (121) Hagren T.A. & Damm W. (2001). Polarizable force fields. *Curr. Opin. Struct. Biol.* **11**, 236-242.
- (122) Hamelberg D., Mongan J. & McCammon J.A. (2004). Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J. Chem. Phys.* **120**, 11919-11929.
- (123) Hamelberg D., Shen T. & McCammon J.A. (2005). Phosphorylation effects on cis/trans isomerization and the backbone conformation of serine-proline motifs: accelerated molecular dynamics analysis. *J. Am. Chem. Soc.* **127**, 1969-1974.
- (124) Harano Y. & Kinoshita M. (2005). Translational-entropy gain of solvent upon protein folding. *Biophys. J.* **89**, 2701-2710.
- (125) Harder E., Anisimov V.M., Vorobyov I.V., Lopes P.E.M., Noskov S.Y., Mackerell A.D. Jr. & Roux B. (2006). Atomic level anisotropy in the electrostatic modeling of lone pairs for a polarizable force field

- based on the classical drude oscillator. *J. Chem. Theor. Comput.* **2**, 1587-1597.
- (126) Hemmer B., Kondo T., Gram B., Pinilla C., Cortese I., Pascal J., Tzou A, McFarland H.F., Houghten R., & Martin R. (2000). Minimal peptide length requirements for CD4+ T cell clones - implications for molecular mimicry and T cell survival. *Int. Immunol.* **12**, 375-383.
- (127) Hess B. (2000). Similarities between principal components of protein dynamics and random diffusion. *Phys. Rev. E* **62**, 8438-8448.
- (128) Hess B. (2002). Convergence of sampling in protein simulations. *Phys. Rev. E* **65**, 031910.
- (129) Hess B., Kutzner C., van der Spoel D., Lindahl E. (2009). Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Ther. Comput.* **4**, 435-447.
- (130) Ho B.K. & Dill K.A. (2006). Folding very short peptides using molecular dynamics. *PLoS Comput. Biol.* **2**, e27: 0228-0237.
- (131) Honda S., Yamasaki K., Sawada Y. & Morii H. (2004). 10 residue folded peptide designed by segment statistics. *Structure* **12**, 1507-1518.
- (132) Hornak V., Abel R., Okur A., Strockbine B., Roitberg, A., Simmerling, C. (2006). Comparison of multiple AMBER force fields and development of improved protein backbone parameters. *Proteins* **65**, 712-725.
- (133) Humphrey W., Dalke A. & Schulten K. (1996). VMD: Visual Molecular Dynamics. *J. Mol. Graphics* **14**, 33-38.
- (134) Hünenberger P.H., Mark A.E. & vanGunsteren W.F. (1995). Fluctuation and cross-correlation analysis of protein motions observed in nanosecond molecular dynamics simulations. *J. Mol. Biol.* **252**, 492-503.
- (135) Hutchinson E. G. & Thornton J. M. (1994) A revised set of potentials for b-turn formation in proteins. *Protein Sci.* **3**, 2207-2216.
- (136) Ichiye T., & Karplus M. (1991). Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins* **11**, 205-217.
- (137) Irbäck A., Peterson C. & Potthast F. (1997). Identification of amino acid sequences with good folding properties in an off-lattice model. *Phys. Rev. E* **55**, 860-867.
- (138) Israilewitz B., Gao M. & Schulten K., (2001). Steered molecular-dynamics and mechanical functions of proteins. *Curr. Opin. Struct. Biol.* **11**, 224-230.

- (139)Itzhaki L.S., Otzen D.E. & Fersht A.R. (1995). The structure of the transition state for folding of chymotrypsin inhibitor-2 analyzed by protein engineering methods-evidence for a nucleation - condensation mechanism for protein-folding. *J. Mol. Biol.* **254**, 260-288.
- (140)Ivarsson Y., Travaglini-Allocatelli C., Brunori M. & Gianni S. (2008). Mechanisms of protein folding. *Eur. Biophys. J.* **37**, 721-728.
- (141)Izaguirre J.A., Reich S. & Skeel R.D. (1999). Longer time steps for molecular dynamics. *J. Chem. Phys.* **110**, 9853-9864.
- (142)Jain A.K., Murty M.N. & Flynn P.J. (1999). Data clustering: a review. *ACM Computing Surveys* **31**, 264-323.
- (143)Jette M. & Grondona, M. (2003). SLURM: Simple Linux Utility for Resource Management. Proceedings of ClusterWorld Conference and Expo, San Jose, California.
- (144)Jones C.M., Henry E.R., Hu Y., Hochstrasser R.M. (1995). Fast events in protein folding initiated by nanosecond laser photolysis. *Proc Natl. Acad. Sci. USA* **90**, 11860-11864.
- (145)Jones D.T., Taylor W.R. & Thornton J.M. (1992). A new approach to protein fold recognition. *Nature* **358**, 86-89.
- (146)Jorgensen W.L., Chandrasekhar J., Madura J.D., Impey R. W. & Klein M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926-935.
- (147)Jorgensen W.L. & Duffy E.M. (2000). Prediction of drug solubility from Monte Carlo simulations. *Bioorg. Med. Chem. Lett.* **10**, 1155-1158.
- (148)Jorgensen W.L. & Duffy E.M. (2002). Prediction of drug solubility from structure. *Adv. Drug Deliv. Rev.* **54**, 355-366.
- (149)Jorgensen W. L., Maxwell D. S. & Tirado-Rives J. (1996). Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **118**, 11225-11236.
- (150)Kabsch W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Cryst.* **A32**, 922-923.
- (151)Kabsch W. (1994). LSQKAB, version 42 Collaborative Computational Project, Number 4. The CCP4 suite: Programs for Protein Crystallography. *Acta Cryst.* **D50**, 760-763.
- (152)Kabsch W. & Sander C. (1984). On the use of sequence homologies to predict protein structure:



identical pentapeptides can have completely different conformations. *Proc Natl. Acad. Sci. USA* **81**, 1075-1078.

(153)Kale L., Skeel R., Bhandarkar M., Brunner R., Gursoy A., Krawetz N., Phillips J., Shinozaki A., Varadarajan K. & Schulten K. (1999). NAMD2: Greater scalability for parallel molecular dynamics. *J. Comp. Phys.* **151**, 283-312.

(154)Kaminski G., Friesner R. A., Tirado-Rives J., Jorgensen W. L. (2001). Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* **105**, 6474-6487.

(155)Karplus M. (2011). Behind the folding funnel diagram. *Nat. Chem. Biol.* **7**, 401-404.

(156)Karplus M. & McCammon J.A. (2002). Molecular dynamics simulations of biomolecules. *Nature Struct. Biol.* **9**, 646-652.

(157)Karplus M. & Weaver D.L. (1994). Protein folding dynamics: the diffusion-collision model and experimental data. *Protein Sci.* **3**, 650-668.

(158)Kaur H. & Raghava G.P.S. (2002). BetaTpred: Prediction of beta-turns in a protein using statistical algorithms. *Bioinformatics* **18**, 498-499.

(159)Kaur H. & Raghava G.P.S. (2003). Prediction of beta-turns in proteins from multiple alignment using neural network. *Protein Sci.* **12**, 627-634.

(160)Kaur H. & Raghava G.P.S. (2004). A neural network method for prediction of beta-turn types in proteins using evolutionary information. *Bioinformatics* **20**, 2751-2758.

(161) Kaur H. & Sasidhar Y.U. (2012). For the sequence YKGG, the turn and extended conformational forms are separated by small barriers and the turn propensity persists even at high temperatures: implications for protein folding. *J. Phys. Chem. B* **116**, 3850-3860.

(162)Keller B., Daura X. & vanGunsteren W.F. (2010). Comparing geometric and kinetic cluster algorithms for molecular simulation data. *J. Chem. Phys.* **132**, 0741110.

(163) Keller T.H., Pichota A. & Yin Z. (2006). A practical view of 'druggability'. *Curr. Opin. Chem. Biol.* **10**, 357-361.

(164)Khatib F., Dimaio F., Foldit Contenders Group, Foldit Void Crushers Group, Cooper S., Kazmierczyk M., Gilski M., Krzywda S., Záborská H., Pichová I., Thompson J., Popović Z., Jaskolski M. & Baker D. (2011). Crystal structure of a monomeric retroviral protease solved by protein folding game players.

*Nature* **18**, 1175-1177.

(165) Khavinson V.Kh. (2005). Effect of tetrapeptide on insulin biosynthesis in rats with alloxan-induced diabetes. *Bull. Exp. Biol. Med.* **140**, 452-454.

(166) Khavinson V.Kh., Malinin V.V., Grigoriev E.I. & Ryzhak G.A. (2009). Tetrapeptide regulating blood glucose level in diabetes mellitus. U.S. Patent 7,491,703 B2, Feb. 17, 2009.

(167) Kier B.L. & Andersen N.H. (2008). Probing the lower size-limit for protein-like fold stability: ten-residue microproteins with specific, rigid structures in water. *J. Am. Chem. Soc.* **130**, 14675-14683.

(168) Kirschner A. & Frishman D. (2008). Prediction of  $\beta$ -turns and  $\beta$ -turn types by a novel bidirectional Elman-type recurrent neural network with multiple output layers (MOLEBRNN). *Gene* **422**, 22-29.

(169) Kirshenbaum K., Zuckermann R.N. & Dill K.A. (1999). Designing polymers that mimic biomolecules. *Curr. Opin. Struct. Biol.* **9**, 530-535.

(170) Klepeis J.L., Lindorff-Larsen K., Dror R.O. & Shaw D.E. (2009). Long-timescale molecular dynamics simulations of protein structure and function. *Curr. Opin. Struct. Biol.* **19**, 120-127.

(171) Kleywegt G.J., Zou J.Y., Kjeldgaard M. & Jones T.A. (2001). Around O. In: "International Tables for Crystallography, Vol. F. Crystallography of Biological Macromolecules" (Rossmann, M.G. & Arnold, E., Editors). Chapter 17.1, pp. 353-356, 366-367. Dordrecht: Kluwer Academic Publishers, The Netherlands.

(172) Kliger Y. (2010). Computational approaches to therapeutic peptide discovery. *Biopolymers* **94**, 701-710.

(173) Kortemme T., Morozov A.V. & Baker D. (2003). An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J. Mol. Biol.* **326**, 1239-1259.

(174) Koslover E.F. & Wales D.J. (2007). Geometry optimization for peptides and proteins: comparison of Cartesian and internal coordinates. *J. Chem. Phys.* **127**, 234105.

(175) Kountouris P. & Hirst J.D. (2010). Predicting  $\beta$ -turns and their types using predicted backbone dihedral angles and secondary structures. *BMC Bioinformatics* **11**, 407-417.

(176) Krieger F., Fierz B., Bieri O., Drewello M. & Kiefhaber T. (2003). Dynamics of unfolded polypeptide chains as model for the earliest steps in protein folding. *J. Mol. Biol.* **332**, 265-274.

(177) Krieger F., Möglich A. & Kiefhaber T. (2004). Effect of proline and glycine residues on dynamics and barriers of loop formation in polypeptide chains. *J. Am. Chem. Soc.* **127**, 3346-3352.

- (178) Kubelka, J., Hofrichter, J. & Eaton, W.A. (2004). The protein folding 'speed limit'. *Curr. Opin. Struct. Biol.* **14**, 76-88.
- (179) Kubelka, J., Henry, E. R., Cellmer, T., Hofrichter, J. & Eaton, W. A. (2008). Chemical, physical and theoretical kinetics of an ultrafast folding protein. *Proc Natl. Acad. Sci. USA* **105**, 18655-18662.
- (180) Kuhlman B. & Baker D. (2004). Exploring folding free energy landscapes using computational protein design. *Curr. Opin. Struct. Biol.* **14**, 89-95.
- (181) Lange D.F., Grubmüller H. & de Groot B.L. (2005). Molecular dynamics simulations of protein G challenge NMR-derived correlated backbone motions. *Angew Chem. Int. Ed.* **44**, 3394-3399.
- (182) Lange D.F., van der Spoel D. & de Groot B.L. (2010). Scrutinizing molecular mechanics force fields on the submicrosecond timescale with NMR data. *Biophys. J.* **99**, 647-655.
- (183) Lapidus, L.J., Eaton, W.A. & Hofrichter, J. (2000). Measuring the rate of intramolecular contact formation in polypeptides. *Proc Natl. Acad. Sci. USA* **97**, 7998-8002.
- (184) Lapidus, L.J., Eaton, W.A. & Hofrichter, J. (2001). Dynamics of intramolecular contact formation in polypeptides: distance dependence of quenching rates in a room-temperature glass. *Phys. Rev. Lett.* **87**, 258101.
- (185) Larson, S.M., Snow, C. & Pande, V.S. (2003). Folding@Home and Genome@Home: Using distributed computing to tackle previously intractable problems in computational biology. *Modern Methods in Computational Biology*, R. Grant, ed, Horizon Press.
- (186) Lattman E.E. & Rose G.D. (1993). Protein folding - what's the question? *Proc Natl. Acad. Sci. USA* **90**, 439-441.
- (187) Lavelle, D.T. & Pearson W.R. (2009). Globally, unrelated protein sequences appear random. *Bioinformatics* **26**, 310-318.
- (188) Layton J.B. (2009) Caos NSA and Perceus: All-in-one Cluster Software Stack, *Linux Magazine*.
- (189) Leszczynski J. & Rose G.D. (1986). Loops in globular proteins: a novel category of secondary structure. *Science* **234**, 849-855.
- (190) Levinthal C. (1968). Are there pathways for protein folding? *J. Chem. Phys.* **85**, 44-45.
- (191) Levinthal C. (1969). How to fold graciously. *Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois*, 22-24.
- (192) Lewis P.N., Momany F.A. & Scheraga H.A. (1971). Folding of polypeptide chains in proteins: A

- proposed mechanism of folding. *Proc Natl. Acad. Sci. USA* **68**, 2293-2297.
- (193) Lewis P.N., Momany F.A. & Scheraga H.A. (1973). Chain reversals in proteins. *Biochem. Biophys. Acta* **303**, 211-229.
- (194) Li D.-W., Khanlarzadeh M., Wang J., Huo S. & Brüschweiler R. (2007). Evaluation of configurational entropy methods from peptide folding-unfolding simulation. *J. Phys. Chem. B* **111**, 13807-13813.
- (195) Lindorff-Larsen K., Maragakis P., Piana S., Eastwood M.P., Dror R.O. & Shaw D.E. (2012). Systematic validation of protein force fields against experimental data. *PLoS One* **7**, e32131.
- (196) Lindorff-Larsen K., Piana S., Dror R.O. & Shaw E. (2011). How fast-folding proteins fold. *Science* **334**, 517-520.
- (197) Lindorff-Larsen K., Piana S., Palmo K., Maragakis P., Klepeis J.L., Dror R.O., & Shaw D.E. (2010). Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**, 1950-1958.
- (198) MacArthur M.W. & Thornton J.M. (1991). Influence of proline residues on protein conformation. *J. Mol. Biol.* **218**, 397-412.
- (199) Mackerell A.D. Jr., Feig M. & Brooks C.L. III (2004). Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* **25**, 1400-1415.
- (200) Mackerell A.D. Jr. (2004). Empirical force fields for biological macromolecules: overview and issues. *J. Comput. Chem.* **25**, 1584-1604.
- (201) Mackerell A.D. Jr., Brooks B., Brooks C.L. III, Nilsson L., Roux B., Won Y. & Karplus M. (1998). CHARMM: the energy function and its parameterization with an overview of the program. *The Encyclopedia of Computational Chemistry* **1**, 271-277.
- (202) Mackerell A.D. Jr., Bashford D., Bellott M., Dunbrack R.L. Jr., Evanseck J.D., Field M.J., Fischer S., Gao J., Guo H., Ha S., Joseph-McCarthy D., L. Kuchmir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, III, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, M. Karplus (1998). All-atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem.* **102**, 3586-3616.
- (203) Mackerell A. D., Feig M., Brooks C. L. III. (2004). Extending the Treatment of Backbone Energetics in Protein Force Fields: Limitations of Gas-Phase Quantum Mechanics in Reproducing Protein

- Conformational Distributions in Molecular Dynamics Simulations. *J. Comput. Chem.* **25**, 1400–1415.
- (204) Mahalakshmi R., Sengupta A., Raghobhama S., Shamala N. & Balaram P. (2006). Tryptophan rich peptides: influence of indole rings on backbone conformation. *Biopolymers* **88**, 36–54.
- (205) Maisuradze G.G. & Leitner D.M. (2007). Free energy landscape of a biomolecule in dihedral principal component space: sampling convergence and correspondence between structures and minima. *Proteins* **67**, 569–578.
- (206) Maity H., Maity M., Krishna M.M., Mayne L. & Englander S.W. (2005). Protein folding: the stepwise assembly of foldon units. *Proc Natl. Acad. Sci. USA* **102**, 4741–4746.
- (207) Makhataдзе G.I. & Privalov P.L. (1996). On the entropy of protein folding. *Protein Sci.* **5**, 507–510.
- (208) Matthes D. & de Groot B.L. (2009). Secondary structure propensities in peptide folding simulations: A systematic comparison of molecular mechanics interaction schemes. *Biophys. J.* **97**, 599–608.
- (209) Matysiak S. & Clementi C. (2007). Mapping folding energy landscapes with theory and experiment. *Arch. Biochem. Biophys.* **469**, 29–33.
- (210) Mayer K.L., Earley M.R., Gupta S., Pichumani K., Regan L. & Stone M.J. (2003). Covariation of backbone motion throughout a small protein domain. *Nat. Struct. Biol.* **10**, 962–965.
- (211) McCammon J.A. (1996). A speed limit for protein folding. *Proc Natl. Acad. Sci. USA* **93**, 11426–11427.
- (212) McGregor M. J., Flores T. P., and Sternberg M. J. E. (1989) Prediction of b-turns in proteins using neural networks. *Protein Eng.* **2**, 521–526.
- (213) Merchant K.A., Best R.B., Louis J.M., Gopich I.V. & Eaton W.A. (2007). Characterizing the unfolded states of proteins using single-molecule FRET spectroscopy and molecular dynamics. *Proc Natl. Acad. Sci. USA* **104**, 1528–1533.
- (214) Merritt E.A. & Bacon D.J. (1997). Raster3D photorealistic molecular graphics. *Methods Enzymol.* **277**, 505–524.
- (215) Meus, J., Brylinski, M., Piwowar, M., Piwowar P. Wisniowski Z., Stefaniak, J., Konieczny L., Surowka G. & Roterman I. (2006). A tabular approach to the sequence-to-structure relation in proteins (tetrapeptide representation) for *de novo* protein design. *Med. Sci. Monit.* **12**, BR208–214.
- (216) Mezei M. (1998). Chameleon sequences in the PDB. *Protein Eng.* **11**, 411–414.

- (217) Mimervini G., Evangelista G., Polticelli F., Piwowar M., Kochanczyk M., Flis L., Malawski M., Szepieniec T., Wisniowski Z., Matczynska E., Prymoula K. & Roterman I. (2008). Never born peptides as a test case for ab initio protein structure prediction. *Bioinformatics* **3**, 177-179.
- (218) Minor D.L. & Kim P.S. (1996). Context-dependent secondary structure formation of a designed protein sequence. *Nature* **380**, 730-734.
- (219) Mirsky A.E. & Pauling L. (1936). On the structure of native, denatured, and coagulated proteins. *Proc. Natl. Acad. Sci. USA* **22**, 439-447.
- (220) Mittal J. & Best R.B. (2010). Tackling force field bias in protein folding simulations: folding of Villin HP35 and Pin WW domains in explicit water. *Biophys. J.* **99**, L26-L28.
- (221) Möglich, A., Joder, K. & Kiefhaber, T. (2006). End-to-end distance distributions and intrachain diffusion constants in unfolded polypeptide chains indicate intramolecular hydrogen bond formation. *Proc Natl. Acad. Sci. USA* **103**, 12394-12399.
- (222) Mohanty D., Elber R., Thirumalai D., Beglov D. & Roux B. (1997). Kinetics of peptide folding: computer simulations of SYPFDV and peptide variants in water. *J.Mol.Biol.* **272**, 423-442.
- (223) Monticelli L., Sorin E.J., Tielman D.P., Pande V.S. & Colombo G. (2008). Molecular simulation of multistate peptide dynamics: a comparison between microsecond timescale sampling and multiple shorter trajectories. *J. Comp. Chem.* **29**, 1740-1752.
- (224) Moult J., Pedersen J.T., Judson R. & Fidelis K. (1995). A large-scale experiment to assess protein structure prediction methods. *Proteins* **23**, ii-v.
- (225) Moult J., Fidelis K., Kryshchuk A., Rost B., Tramontano A. (2009). Critical assessment of methods of protein structure prediction - round VIII. *Proteins* **77**, Suppl 9, 1-4.
- (226) Mu Y., Nguyen P.H. & Stock G. (2005). Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins* **58**, 45-52.
- (227) Munoz V., Thompson P.A., Hofrichter J.A. & Eaton W.A. (1997). Folding dynamics and mechanism of beta-hairpin formation. *Nature* **390**, 196-199.
- (228) Myers J.K. & Das T.G. (2001). Preorganized secondary structure as an important determinant of fast protein folding. *Nat. Struct. Biol.* **8**, 552-558.
- (229) Naganathan A.N. & Munoz V. (2005). Scaling of folding times with protein size. *J. Am. Chem. Soc.* **127**, 480-481.



- (230) Okur A., Strockbine B., Hornak V. & Simmerling C. (2002). Using PC clusters to evaluate the transferability of molecular mechanics force fields for proteins. *J. Comput. Chem.* **24**, 21-31.
- (231) Onuchic J.N., Schulten L.-Z. & Wolynes P.G. (1997). Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* **48**, 545-600.
- (232) Onuchic J.N., Socci N.D., Schulten L.-Z. & Wolynes P.G. (1996). Protein folding funnels: the nature of the transition state ensemble. *Fold Des.* **1**, 441-450.
- (233) Onuchic J.N. & Wolynes P.G. (2004). Theory of protein folding. *Curr. Opin. Struct. Biol.* **14**, 70-75.
- (234) Oostenbrink C., Villa A., Mark A.E. & van Gunsteren W.F. (2004). A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force field parameter sets 53A5 and 53A6. *J. Comput. Chem.* **25**, 1656-1676.
- (235) Ozkan S.B., Wu G.A., Chodera J.D. & Dill K.A. (2007). Protein folding by zipping and assembly. *Proc. Natl. Acad. Sci. USA* **104**, 11987-11992.
- (236) Paci E., Cavalli A., Vendruscolo M. & Caflisch A. (2003). Analysis of the distributed computing approach applied to the folding of a small  $\beta$  peptide. *Proc Natl. Acad. Sci. USA* **100**, 8217-8222.
- (237) Patel S. & Brooks C.L. (2006). Fluctuating charge force fields: recent developments and applications from small molecules to macromolecular biological systems. *Mol. Simul.* **32**, 231-249.
- (238) Peter C., Daura X. & van Gunsteren W.F. (2001). Calculation of NMR-relaxation parameters for flexible molecules from molecular dynamics simulations. *J. Biomol. NMR* **20**, 297-310.
- (239) Petersen B., Lundegaard C. & Petersen T.N. (2010). NetTurnP - neural network prediction of beta-turns by use of evolutionary information and predicted protein sequence features. *PLOS one* **5**, e15079.
- (240) Piana S., Lindorff-Larsen K., Dirks R.M., Salmon J.K., Dror R.O. & Shaw D.E. (2012). Evaluating the effects of cutoffs and treatment of long-range electrostatics in protein folding simulations. *PLOS one* **7**, e39918.
- (241) Piana S., Lindorff-Larsen K. & Shaw D.E. (2011). How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.* **100**, L47-L49.
- (242) Pitera J.W. & Swope W. (2003). Understanding folding and design: replica-exchange simulations of "Trp-cage" miniproteins. *Proc. Natl. Acad. Sci. USA* **100**, 7587-7592.
- (243) Phillips J.C., Braun R., Wang W., Gumbart J., Tajkhorshid E., Villa E., Chipot C., Skeel R.D. Kale L. &

- Schulten K. (2005). Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781-1802.
- (244) Plaxco K.W. & Dobson C.M. (1996). Time-resolved biophysical methods in the study of protein folding. *Curr. Opin. Struct. Biol.* **6**, 630-636.
- (245) Plaxco K.W. & Gross M. (1997). The importance of being unfolded. *Nature* **386**, 657-659.
- (246) Plaxco K.W. & Gross M. (2001). Unfolded, yes, but random? Never! *Nat. Struct. Biol.* **8**, 659-660.
- (247) Portman, J.J. (2003) Non-Gaussian dynamics from a simulation of a short peptide: loop closure rates and effective diffusion coefficients. *J. Chem. Phys.* **118**, 2381-2391.
- (248) Price D.J. & Brooks C.L. III (2002). Modern force fields behave comparably in molecular dynamics simulations. *J. Comput. Chem.* **23**, 1045-1057.
- (249) R Development Core Team (2004). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- (250) Rackovsky S. (1993). On the nature of the protein folding code. *Proc. Natl. Acad. Sci. USA* **90**, 644-648.
- (251) Ramaprasad S. & Compadre C. (1993). Solution conformation of a pentapeptide by NMR and molecular modelling studies. *Spec. Lett.* **26**, 639-660.
- (252) Rao F. & Caflisch A. (2003). Replica-exchange molecular dynamics simulations of reversible folding. *J. Chem. Phys.* **119**, 4035-4042.
- (253) Rao F., Settanni G. & Caflisch A. (2007). Estimation of folding probabilities and phi values from molecular dynamics simulations of reversible peptide folding. *Methods Mol. Biol.* **350**, 225-249.
- (254) Richardson, J. S. (1981) The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* **34**, 167-339.
- (255) Rhee Y.M., Sorim E.J., Jayachandran G., Lindahl E. & Pande V.S. (2004). Simulations of the role of water in the protein folding mechanism. *Proc. Natl. Acad. Sci. USA* **101**, 6456-6461.
- (256) Roder H. (1995). Watching protein folding unfold. *Nat. Struct. Biol.* **2**, 817-820.
- (257) Roder H. & Shastry M.C.R (1999). Methods for exploring early events in protein folding. *Curr. Opin. Struct. Biol.* **9**, 620-626.
- (258) Rohl C.A., Strauss C.E.M., Misura K.M.S. & Baker D. (2004). Protein structure prediction using Rosetta. *Methods Enzymol.* **383**, 66-93.

- (259) Rose, G. D., Gierasch, L. M., and Smith, J. A. (1985) Turns in peptides and proteins. *Adv. Protein Chem.* **37**, 1-109.
- (260) Rousseau R., Schreiner E., Kohlmeyer A. & Marx D. (2004). Temperature-dependent conformational transitions and hydrogen-bond dynamics of the elastin-like octapeptide GVG(VPGVG): a molecular-dynamics study. *Biophys. J.* **86**, 1393-1407.
- (261) Rueda M., Ferrer-Costa C., Meyer T., Perez A., Camps J., Hospital A., Gelpi J.L. & Orozco M. (2007). A consensus view of protein dynamics. *Proc. Natl. Acad. Sci. USA* **104**, 796-801.
- (262) Ryckaert J.P., Ciccoti G. & Berendsen H.J.C. (1977). Numerical integration of cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327-341.
- (263) Sanchez I.E. & Kiefhaber T. (2003). Evidence for sequential barriers and obligatory intermediates in apparent two-state protein folding. *J. Mol. Biol.* **325**, 367-376.
- (264) Šali A., Shakhmovich E. & Karplus M. (1994). How does a protein fold? *Nature* **369**, 248-251.
- (265) Schäfer H., Daura X., Mark A.E. & van Gunsteren W.F. (2001). Entropy calculations on a reversible folding peptide: changes in solute free energy cannot explain folding behavior. *Proteins* **43**, 45-56.
- (266) Schuetz P., Wuttke R., Schuler B. & Cafilisch A. (2010). Free energy surfaces from single-distance information. *J. Phys. Chem. B* **114**, 15227-15235.
- (267) Schuler B., Lipman E.A. & Eaton W.A. (2002). Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature* **419**, 743-747.
- (268) Schwartz R. & King J. (2006). Frequencies of hydrophobic and hydrophilic runs and alterations in proteins of known structure. *Protein Sci.* **15**, 102-112.
- (269) Seringhaus M. & Gerstein M. (2007). Chemistry Nobel rich in structure. *Science* **315**, 40-41.
- (270) Shannon, C. (1948) A mathematical theory of communication. *Bell System Tech. J.* **27**, 379-423.
- (271) Shaw D.E., Maragakis P., Lindorff-Larsen K., Piana S., Dror R.O., Eastwood M.P., Bank J.A., Jumper J.M., Salmon J.K., Shan Y. & Wrighers W. (2010). Atomic-level characterization of the structural dynamics of proteins. *Science* **330**, 341-346.
- (272) Shaw D.E., Dror R., Salmon J.K., Grossman J.P., Mackenzie K.M., Bank J.A., Young C., Deneroff M.M., Batson B., Bowers K.J., Chow E., Eastwood M.P., Ierardi D.J., Klepeis J.L., Kuskin J.S., Larson R.H., Lindorff-Larsen K., Maragakis P., Moraes M.A., Piana S., Shan Y. & Towles B. (2009). Millisecond-scale molecular dynamics simulations on Anton. Proceedings of the ACM/IEEE conference on supercomputing,

Portland, Oregon, 1-11.

(273) Shell M.S., Ozkan S.B., Voelz V., Wu G.A. & Dill K.A. (2009). Blind test of physics-based prediction of protein structures. *Biophys. J.* **96**, 917-924.

(274) Shell M.S., Ritterson R. & Dill K.A. (2008). A test on peptide stability of AMBER force fields with implicit solvation. *J. Phys. Chem. B* **112**, 6878-6886.

(275) Shenkin P.S. & McDonald D.G. (1994). Cluster analysis of molecular conformations. *J. Comput. Chem.* **15**, 899-916.

(276) Shepherd A.J., Gorse D. & Thornton J.M. (1999). Prediction of the location and type of  $\beta$ -turns in proteins using neural networks. *Protein Sci.* **8**, 1045-1055.

(277) Shirts M. & Pande V.S. (2000). Computing: screen savers of the world unite! *Science* **290**, 1903-1904.

(278) Shortle D. & Ackerman M. (2001). Persistence of native-like topology in a denatured protein in 8M urea. *Science* **293**, 487-489.

(279) Simmerling C.L. & Elber R. (1995). Computer determination of peptide conformations in water: different roads to structure. *Proc. Natl. Acad. Sci. USA* **92**, 3190-3193.

(280) Simmerling C., Strockbine B. & Roitberg A.E. (2002). All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.* **124**, 11258-11259.

(281) Smith L.J., Daura X. & van Gunsteren W.F. (2002). Assessing equilibrium and convergence in biomolecular simulations. *Proteins* **48**, 487-496.

(282) Snow C.D., Nguyen N., Pande V.S. & Gruebele M. (2002). Absolute comparison of simulated and experimental protein-folding dynamics. *Nature* **42**, 102-106.

(283) Sosnick T.R., Berry R.S., Colubri A. & Fernandez A. (2002). Distinguishing foldable proteins from non-folders: when and how do they differ? *Proteins* **49**, 15-23.

(284) Sosnick T.R., Dothager R.S. & Krantz B.A. (2004). Differences in the folding transition state of ubiquitin indicated by  $\phi$  and  $\psi$  analyses. *Proc Natl. Acad. Sci. USA* **101**, 17377-17382.

(285) Srinivasan, R. RIBOSOME. <http://www.roselab.jhu.edu/~raj/Manuals/ribosome.html>

(286) Steinbach P.-J. (2004). Exploring peptide energy landscapes: a test of force fields and implicit solvent models. *Proteins* **57**, 665-677.

(287) Stewart D.E., Sarkar A. & Wampler J.E. (1990). Occurrence and role of cis peptide bonds in protein

- structures. *J. Mol. Biol.* **214**, 253-260.
- (288) Stone J.E., Phillips J.C., Freddolino P.L., Hardy D.J., Trabuco L.G. & Schulten K. (2007). Accelerating molecular modelling applications with graphics processors. *J. Comput. Chem.* **28**, 2618-2640.
- (289) Sugita Y. & Okamoto Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314**, 141-151.
- (290) Taly J.-F., Marin A. & Gibrat J.-F. (2008). Can molecular dynamics simulations help in discriminate correct from erroneous protein 3D models? *BMC Bioinformatics* **9**, 6.
- (291) Thomas A., Deshayes S., Decaffmeyer M., Van Eyck M.H., Charlotiaux B.B. & Brasseur R. (2006). Prediction of peptide structure: How far are we? *Proteins* **65**, 889-897.
- (292) Thomas A., Deshayes S., Decaffmeyer M., Van Eyck M.H., Charlotiaux B.B. & Brasseur R. (2009). PepLook: an innovative in silico tool for determination of structure, polymorphism and stability of peptides. *Adv. Exp. Med. Biol.* **611**, 459-460.
- (293) Torrie G.M., Valleau J.P. (1977). Nonphysical sampling distributions in Monte-Carlo free-energy estimation: umbrella sampling. *J. Comput. Phys.* **23**, 187-199.
- (294) Tréhin R. & Merkle H.P. (2004). Chances and pitfalls of cell penetrating peptides for cellular drug delivery. *Eur. J. Pharm. Biopharm.* **58**, 209-223.
- (295) Tsoulos I.G. & Stavrakoudis A. (2011). eucb: a C++ program for trajectory analysis. *Comput. Phys. Commun.* **182**, 834-841. <http://stavrakoudis.econ.uoi.gr/eucb>.
- (296) Ueki N., Someya K., Matsuo Y., Wakamatsu K. & Mukai H. (2007). Cryptides: functional cryptic peptides hidden in protein structures. *Biopolymers* **88**, 190-198.
- (297) van Gunsteren W.F., Billeter S.R., Eising A.A., Hünenberger P.H., Krüger P. Mark A.E., Scott W.R.P. & Toroni I.G. (1996). The GROMOS96 manual and user guide. Biomolecular simulation.
- (298) van Gunsteren W.F., Bürgi R., Peter C. & Daura X. (2001). The key to solving the protein-folding problem lies in an accurate description of the denatured state. *Angew. Chem. Int. Ed.* **40**, 352-355.
- (299) Venkatachalam C. M. (1968) Stereo chemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers* **6**, 1425-1436.
- (300) Voter A.F. (1997). Hyperdynamics: accelerated molecular dynamics of infrequent events. *Phys. Rev. Lett.* **78**, 3908-3911.
- (301) Wedemeyer W.J., Welker E. & Scheraga H.A. (2002). Proline cis-trans isomerization and protein

- folding. *Biochemistry* **41**, 14637-14644.
- (302) Wei C.-C., Ho M.-H., Wang W.-H. & Sun Y.-C. (2005). Molecular dynamics simulation of folding of a short helical peptide with many charged residues. *J. Phys. Chem B* **109**, 19980-19986.
- (303) Weiss M.S., Jabs A. & Hilgenfeld R. (1998). Peptide bonds revisited. *Nat. Struct. Biol.* **5**, 676.
- (304) Wetlaufer D.B. (1973). Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl. Acad. Sci. USA* **70**, 697-701.
- (305) Wetlaufer D.B. (1990). Nucleation in protein folding - confusion of structure and process. *Trends Biochem. Sci.* **15**, 414-415.
- (306) Wickstrom L., Okur A., Simmerling C. (2009). Evaluating the Performance of the ff99SB Force Field Based on NMR Scalar Coupling Data. *Biophys. J.* **97**, 853-856.
- (307) Williams T. & Kelly C. (1986-1993, 1998, 2004, 2007-2011). Gnuplot: an interactive plotting program. URL: <http://gnuplot.info>
- (308) Williams S., Causgrove T.P., Gilmanshin R., Fang K.S., Callender R.H., Woodruff W.H., Dyer R.B. (1996) Fast events in protein folding: helix melting and formation in a small peptide. *Biochemistry* **35**, 691-697.
- (309) Wilmot C.M. & Thornton J. M. (1988) Analysis and prediction of the different types of b-turn in proteins. *J. Mol. Biol.* **203**, 221-232.
- (310) Wilmot C.M. & Thornton J.M. (1990). Beta-turns and their distortions: a proposed new nomenclature. *Prot. Eng.* **3**, 479-493.
- (311) Worth G.A., Nardi F. & Wade R.C. (1998). Use of multiple molecular dynamics trajectories to study biomolecules in solution: the YTGp peptide. *J. Phys. Chem. B* **102**, 6260-6272.
- (312) Wu W.-J. & Raleigh D.P. (1998). Local control of peptide conformation: stabilization of cis proline peptide bonds by aromatic proline interactions. *Biopolymers* **45**, 381-394.
- (313) Wu X. & Wang S. (2000). Folding studies of a linear pentamer peptide adopting a reverse turn conformation in aqueous solution through molecular dynamics simulation. *J. Phys. Chem.* **104**, 8023-8034.
- (314) Yang W.Y. & Gruebele M. (2003). Folding at the speed limit. *Nature* **423**, 193-197.
- (315) Yeh I.-C. & Hummer G. (2002). Peptide loop-closure kinetics from a microsecond molecular dynamics simulation in explicit solvent. *J. Am. Chem. Soc.* **124**, 6563-6568.
- (316) Yeh I.-C. & Wallqvist A. (2009). Structure and dynamics of end-to-end loop formation of the



- penta-peptide Cys-Ala-Gly-Gln-Trp in implicit solvents. *J. Phys. Chem B* **113**, 12382-12390.
- (317) Yoda T., Sugita Y. & Okamoto Y. (2004). Comparison of force fields for proteins by generalized-ensemble simulations. *Chem. Phys. Lett.* **386**, 460-467.
- (318) Yoda T., Sugita Y. & Okamoto Y. (2004). Secondary-structure preferences of force fields for proteins evaluated by generalized-ensemble simulations. *Chem. Phys.* **307**, 269-283.
- (319) Yoo A., Jette M. & Grondona M. (2003) SLURM: Simple Linux Utility for Resource Management, Job Scheduling Strategies for Parallel Processing, volume 2862 of Lecture Notes in Computer Science, pages 44-60, Springer-Verlag.
- (320) Zanuy D., Flores-Ortega A., Casanovas J., Curcó D., Nussinov R. & Alemán C. (2008). The energy landscape of a selective tumor-homing pentapeptide. *J. Phys. Chem. B* **112**, 8692-8700.
- (321) Zagrovic B., Jayachandran G., Millet I.S., Doniach S. & Pande V.S. (2005). How large is an  $\alpha$ -helix? Studies of the radii of gyration of helical peptides by small-angle X-ray scattering and molecular dynamics. *J. Mol. Biol.* **353**, 232-241.
- (322) Zagrovic B., Snow C.D., Shirts M.R. & Pande V.S. (2002). Simulation of folding of a small  $\alpha$ -helical protein in atomistic detail using worldwide-distributed computing. *J. Mol. Biol.* **323**, 927-937.
- (323) Zbilut J.P., Chua G.H., Krishnan A., Bossa C. Colafranceschi M. & Giuliani A. (2006). Entropic criteria for protein folding derived from recurrences: six residues patch as the basic protein word. *FEBS Letters* **580**, 4861-4864.
- (324) Zhang, C. T., and Chou, K. C. (1997) Prediction of b-turns in proteins by 1-4 and 2-3 correlation model. *Biopolymers* **41**, 673-702.
- (325) Zhang C. & Ma J. (2010). Enhanced sampling and applications in protein folding in explicit solvent. *J. Chem. Phys.* **132**, 244101.
- (326) Zhou R. (2003). Trp-cage: folding free energy landscape in explicit water. *Proc. Natl. Acad. Sci. USA* **100**, 13280-13285.
- (327) Zielkiewicz, J. (2005). Structural properties of water: comparison of the SPC, SPCE, TIP4P and TIP5P models of water. *J. Chem. Phys.* **123**, 104501.
- (328) Zimmermann, S.S. & Scheraga, H.A. (1977). Local interactions in bends of proteins. *Proc. Natl. Acad. Sci. USA* **74**, 4126-4129.
- (329) Zorko, M. & Langel, U. (2005). Cell-penetrating peptides: mechanisms and kinetics of cargo

delivery. *Adv. Drug Deliv. Rev.* **57**, 529-545.

(330) <https://computing.lln.gov/linux/slurm/slurm.htm>

(331) <http://en.wikipedia.org/wiki?title=Talk:Correlation>

(332) <http://www.biosiris.com/products-and-services/pepbook.htm>

(333) <http://www.imtech.res.in/raghava/pepstr>

(334) <http://robetta.bakerlab.org/>

(335) <http://comp.chem.nottingham.ac.uk/debt/>

(336) <http://www.biochem.ucl.ac.uk/bsm/btpred/index.html#references>

(337) <http://imtech.res.in/raghava/betatpred/>

(338) <http://www.imtech.res.in/raghava/betatpred2/index.html>

(339) <http://www.imtech.res.in/raghava/betaturms/>

(340) <http://webclu.bio.wzw.tum.de/predator-web/>

(341) <http://www.cbs.dtu.dk/services/NetTurnP/>

(342) ACE/gr development team (1998-05-10), *Xmgr user guide: introduction*, retrieved 2009-06-20

(343) Paul J Turner and ACE/gr development team (1998-05-13), *Xmgr: List of changes*, retrieved 2009-06-20

(344) Stambulchik, Evgeny (1997), *Xmgr*, retrieved 2009-06-20

(345) Stambulchik, Evgeny (1998-2000), *Grace*, retrieved 2009-06-20

(346) <http://www.gimp.org/>

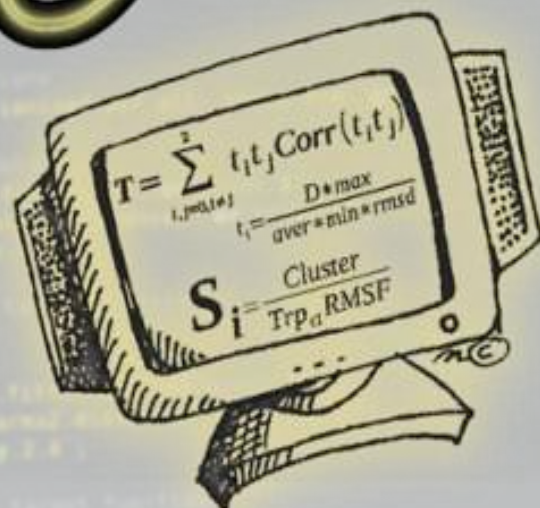
(347) <http://www.imagemagick.org/script/index.php>

(348) [http://www.smartdraw.com/specials/ppc/smartdraw.htm?id=104608&qclid=CM\\_R9MDP868CFQpj3wodsj9SVg](http://www.smartdraw.com/specials/ppc/smartdraw.htm?id=104608&qclid=CM_R9MDP868CFQpj3wodsj9SVg)





# Κεφάλαιο 7 Παράρτημα



## 1) MakeAll\_tetrapepts.c

```
#include <stdio.h>

main()
{
    int  a1, a2, a3, a4;
    int  aa[20] = { 1, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 16, 17, 18, 19, 20, 22, 23, 25 };

    for ( a1 = 0 ; a1 < 20 ; a1++ )
    for ( a2 = 0 ; a2 < 20 ; a2++ )
    for ( a3 = 0 ; a3 < 20 ; a3++ )
    for ( a4 = 0 ; a4 < 20 ; a4++ )

    printf( "%c%c%c%c\n", 64+aa[a1], 64+aa[a2], 64+aa[a3], 64+aa[a4] );
}
```

## 2) 1\_Trp.c

```
#include <stdio.h>

main()
{
    char  a[6];
    int   i;
    int   found;

    a[5] = 0;

    while ( scanf( "%c%c%c%c\n", &a[1], &a[2], &a[3], &a[4] ) == 4 )
    {
        found = 0;
        for ( i=1 ; i <= 4 ; i++ )

            if ( a[i] == 'W' )
                found++;

            if ( found == 1 )
                printf( "%s\n", &a[1] );
    }
}
```

## 3) 1pos\_1neg.c

```
#include <stdio.h>

main()
{
    char  a[6];
    int   i;
    int   pos;
    int   neg;

    a[5] = 0;
```



```

while ( scanf( "%c%c%c%c\n", &a[1], &a[2], &a[3], &a[4] ) == 4 )
{
    pos = 0;
    neg = 0;

    for ( i=1 ; i <= 4 ; i++ )
    {
        if ( a[i] == 'K' || a[i] == 'R' )
            pos++;

        if ( a[i] == 'D' || a[i] == 'E' )
            neg++;
    }

    if ( pos == 1 && neg == 1 )
        printf( "%s\n", &a[1] );
}
}

```

#### 4) all\_AA\_diff.c

```

#include <stdio.h>
main()
{
    char a[6];
    int i, k, m;
    int found;

    a[5] = 0;

    while ( scanf( "%c%c%c%c\n", &a[1], &a[2], &a[3], &a[4] ) == 4 )
    {
        found = 0;

        for ( i=1 ; i <= 3 ; i++ )
        for ( k=i+1 ; k <= 4 ; k++ )
        if ( a[i] == a[k] )
            found = 1;

        if ( found == 0 )
            printf( "%s\n", &a[1] );
    }
}

```

#### 5) 1\_Trp\_2\_3\_4.c

```

#include <stdio.h>
main()
{
    char a[7];
    int i;
    int found;

    a[6] = 0;

    while ( scanf( "%c%c%c%c%c\n", &a[1], &a[2], &a[3], &a[4], &a[5] ) == 5 )

```

```

{
    found = 0;

    for ( i=1; i <= 5; i++ )
        if ( a[i] == 'W' )
            found++;

    if ( found == 1 )

        if ( a[2] == 'W' || a[3] == 'W' || a[4] == 'W' )
            printf("%s\n", &a[1] );
}
}

```

## 6) 3\_charged.c

```

#include <stdio.h>

main()
{
    char a[7];
    int i;
    int found;

    a[6] = 0;

    while ( scanf("%c%c%c%c%c\n", &a[1], &a[2], &a[3], &a[4], &a[5]) == 5 )
    {
        found = 0;

        for ( i=1; i <= 5; i++ )

            if ( a[i] == 'K' || a[i] == 'R' || a[i] == 'D' || a[i] == 'E' )
                found++;

            if ( found == 3 )
                printf("%s\n", &a[1] );
    }
}

```

## 7) NoPro.c

```

#include <stdio.h>

main()
{
    char a[7];
    int i;
    int found;

    a[6] = 0;

    while ( scanf("%c%c%c%c%c\n", &a[1], &a[2], &a[3], &a[4], &a[5]) == 5 )

```

```

{
  found = 0;

  for ( i=1 ; i <= 5 ; i++ )

    if ( a[i] == 'P' )
      found++;

  if ( found == 0 )
    printf("%s\n", &a[1] );
}
}

```

## 8) Read\_fasta\_noXs.pl

```

#!/usr/bin/perl -w

$/ = undef;
$in = <STDIN>;

while ( $in =~ /(>.*?\n)([^\>]+)/ )
{
  $in = $';
  $seq = $2;

  if ( $seq =~ /.+X+.+?\n/ )
  {
    $seq =~ s/\n//g;
    $seq =~ s/X+//g;
  }

  elsif ( $seq =~ /.+X\n/ )
  {
    $seq =~ s/X+//g;
    $seq =~ s/\n//g;
  }

  else
  {
    $seq =~ s/\n//g;
  }
print "$seq\n";
}

exit();

```

## 9) Find\_pentapept.pl

```

#!/usr/bin/perl -w

$k = 0;
%seen = ();

while ( $line = <STDIN> )
{
  for ( $i=0 ; $i < (length $line) -5 ; $i++ )
  {
    $all[$k] = substr( $line, $i, 5 );
    $k++;
  }
}

```

```

foreach $pept (@all) {
    push(@unique, $pept) unless $seen{$pept}++;
}

@all = @unique;
$k = @all;
}

print "@unique\n";

```

## 10) systematic.pl

```

#!/usr/bin/perl -w
#
# This script exit()s. Should be run from crontab.
# To make sure that it will fill the queue, adjust periodicity
#
#
`renice +19 -p $$ >& /dev/null` or print $!;

$USER      = "georgoulia"; # User name
$target_PD = 8;           # How many pending jobs we aim to have
$max_pending = 16;       # If that many jobs are waiting, sleep
$pept2go    = "2GO";     # Filename with list of peptides
#$jobs_per_pept = 1;     # How many jobs per pept ?
#$pep_len    = 5;        # Residues per peptide

use constant WIDTH => 500;
use constant RMS_CUTOFF => 2.0;

# set the perl environment variable PATH
$ENV{'PATH'} = '/bin:/usr/bin:/usr/local/bin';

#
# Do we have a cluster ?
#
$cpus = `sinfo -h -o " %C " | awk -F '/' '{print \$4}'`;
if ( $cpus < 12 )
{
    exit;
}

###
# First thing first: are there any completed jobs to process ?
#
###

# Get list of directories
chdir ("done") || die $!;
opendir (DIR, '.') or die "Can't open current dir: $!\n";
@dirs = grep ( !/^\.\.?$/, readdir (DIR));
closedir (DIR);

if ( @dirs > 0 )
{
    # Something to process ...
    foreach $directory ( @dirs )
    {
        # Directories to process
        {
            chdir ("$directory") || die $!; # Go there

```

```

# Make fit.index
`/bin/grep -P '[2-4].*CA.*C..' ionized.psf > fit.index` or print $!;

# Remove waters-ions (using fit.index)
`/usr/local/bin/carma -v -fit -ind -atmid ALLID -segid A ionized.psf all\_atoms.dcd` or print $!;

`mv carma.fitted.dcd $directory.dcd` or print $!;

# Prepare some data ...
`/usr/local/bin/carma -v -dist 1 5 psfgen.psf $directory.dcd` or print $!;
`awk '{print \$2}' carma.distances > tmp ; mv tmp carma1.distances` or print $!;

# and calculate target function
$Target1 = TargetFunction( "carma1.distances" );

`/usr/local/bin/carma -v -dist 1 4 psfgen.psf $directory.dcd` or print $!;
`awk '{print \$2}' carma.distances > tmp ; mv tmp carma2.distances` or print $!;

# and calculate target function
$Target2 = TargetFunction( "carma2.distances" );

`/usr/local/bin/carma -v -dist 2 5 psfgen.psf $directory.dcd` or print $!;
`awk '{print \$2}' carma.distances > tmp ; mv tmp carma3.distances` or print $!;

# and calculate target function
$Target3 = TargetFunction( "carma3.distances" );

$Distcorr = CorrelationFunction( "carma1.distances", "carma2.distances", "carma3.distances",
                                $Target1, $Target2, $Target3 );

`/usr/local/bin/carma -v -rg -atmid HEAVY psfgen.psf $directory.dcd` or print $!;
`awk '{print \$2}' carma.Rgyration.dat > tmp ; mv tmp carma.Rg` or print $!;

# and calculate target function
$Target4 = TargetFunction( "carma.Rg" );

# Cartesian PCA
`/usr/local/bin/carma -v -w -col -cov -eigen -shannon -atmid HEAVY -proj 3 3 320 psfgen.psf $directory.dcd
> PCA_LOG` or print $!;

# configurational Entropy of PCA distribution
$Target5 = `cat PCA_LOG | grep 'entropy' | awk '{print \$7}'` or print $!;
chomp($Target5);

# number of Clusters
$Target6 = `cat PCA_LOG | grep 'Number.*clusters' | awk '{print \$4}'` or print $!;
chomp($Target6);

`awk '{ if (\$2 == 1) print \$1, \$3, \$4, \$5}' carma.clusters.dat > C_01.dat` or print $!;
`/usr/local/bin/carma -v -sort C_01.dat $directory.dcd` or print $!;
`/usr/local/bin/carma -v -fit -ind -atmid ALLID carma.reordered.dcd psfgen.psf` or print $!;

$Target7 = `wc C_01.dat | awk '{print \$1}'`;
chomp($Target7);

if ( $Target7 > 2500 )
{
  `/usr/local/bin/carma -v -w -col -cov -dot -atmid HEAVY carma.fitted.dcd psfgen.psf` or print $!;
  `/usr/local/bin/carma -v -pdb -step 100 -atmid HEAVY carma.fitted.dcd psfgen.psf` or print $!;

  # Rmsfs
  `cat carma.average.pdb | awk '{print \$11}' | sed '\$d' > $directory.heavy.rmsf` or print $!;
  `cat carma.average.pdb | grep TRP | sed -n '3,12p;12p' | awk '{print \$11}'
  > $directory.Trp-sidechain.rmsf` or print $!;
  `cat carma.average.pdb | grep -P " ([NCO]) | (CA) " | awk '{print \$11}'
  > $directory.backbone.rmsf` or print $!;
  `cat carma.average.pdb | grep -v -P " ([NCO]) | (CA) | TRP " | awk '{print \$11}' | sed '\$d'
  > $directory.rest-sidechains.rmsf` or print $!;
}

```



```

                                # Calculate average rmsfs
$Target8 = Aver( "$directory.heavy.rmsf");
$Target9 = Aver( "$directory.Trp-sidechain.rmsf");
$Target10 = Aver( "$directory.backbone.rmsf");
$Target11 = Aver( "$directory.rest-sidechains.rmsf");

                                # Prepare superpositionPDBs
`/bin/mkdir tmp` or print $!;
`/bin/mv carma.fitted.dcd.*.pdb tmp/` or print $!;

chdir ("tmp") || die $!;
opendir (DIR, '.') or die "Can't open current dir: $!\n";
@pdb = grep (!/^\\.\\.?$/, readdir (DIR));
closedir (DIR);

foreach $pdb (@pdb)
{
    `awk '{ printf "%-7s %3s %-3s %3s %-3s %s %11s %7s %7s %5s\n",
        \ $1, \ $2, \ $3, \ $4, \ $5, \ $6, \ $7, \ $8, \ $9, \ $10 }' "$pdb" > "$pdb.tmp" ` or print $!;
    `paste $pdb.tmp ../$directory.heavy.rmsf > $pdb.PDB ` or print $!;
}
`cat *.PDB > cluster1.$Target7.superposition.pdb` or print $!;
`/usr/bin/bzip2 cluster1.$Target7.superposition.pdb` or print $!;
`/bin/cp -f cluster1.$Target7.superposition.pdb.bz2 ../../../../graphs/
    $directory.cluster1.$Target7.superposition.pdb.bz2` or print $!;

chdir ("../") || die $!;
}

                                # crossDCD
`/bin/rm carma.fitted.dcd carma.fit-rms.dat` or print $!;
`/usr/local/bin/crossDCD psfgen.psf $directory.dcd $directory.dcd 250 "-atmid HEAVY"
    > crossDCD.log` or print $!;

                                # Calculate crossDCD-score
$Target12 = Expand_Windows( "crossDCD.matrix" );

                                # save some data ...
`/usr/bin/bzip2 carma1.distances` or print $!;
`/bin/cp -f carma1.distances.bz2 ../../../../graphs/$directory.1.5.bz2`;
`/usr/bin/bzip2 carma2.distances` or print $!;
`/bin/cp -f carma2.distances.bz2 ../../../../graphs/$directory.1.4.bz2`;
`/usr/bin/bzip2 carma3.distances` or print $!;
`/bin/cp -f carma3.distances.bz2 ../../../../graphs/$directory.2.5.bz2`;
`/usr/bin/bzip2 carma.Rg` or print $!;
`/bin/cp -f carma.Rg.bz2 ../../../../graphs/$directory.Rg.bz2`;
`/usr/bin/bzip2 carma.PCA.fluctuations.dat` or print $!;
`/bin/cp -f carma.PCA.fluctuations.dat.bz2 ../../../../graphs/$directory.eigen.bz2`;
`/usr/bin/bzip2 crossDCD.matrix` or print $!;
`/bin/cp -f crossDCD.matrix.bz2 ../../../../graphs/$directory.crossDCD.matrix.bz2`;

                                # Tidy-up
chdir ("../") || die $!;
`/bin/rm -rf $directory` or print $!;

                                # Write-out ...
if ( $Target7 > 2500 )
{
    open( OUT, ">>../results" ) || die $!;
    print OUT "$directory $Target1 $Target2 $Target3 $Distcorr
        $Target4 $Target5 $Target6 $Target7 $Target8 $Target9 $Target10 $Target11 $Target12\n";
    close( OUT );
}

```



```

else
{
open( OUT, ">>../results" ) || die $!;
print OUT "$directory $Target1 $Target2 $Target3 $Distcorr
$Target4 $Target5 $Target6 $Target7 $Target12\n";
close( OUT );
}
}
}

chdir ( ".." ) || die $!;

###
#
# Now: do we have to submit new jobs to the queue ?
#
###

# How many pending jobs are in the queue ?

$pending_all = `squeue -h -t PD | wc -l`;
$pending_user = `squeue -h -u $USER -t PD | wc -l`;

# If cluster too busy, go to sleep ...
if ( $pending_all > $max_pending )
{
exit;
}

# If less than target, submit $target_PD
if ( $pending_user < $target_PD )
{
$peptide = get_peptide(); # select randomly a peptide
chomp $peptide; # prepare files
prepare_MD_files( $peptide ); # submit job with NAMDjob

# chdir (" $peptide" ) || die $!;
# `usr/bin/NAMDjob 4 all.namd LOG` or print $!;
# chdir ( ".." ) || die $!;

open( SCRIPT, ">$peptide.sh" ) or print $!; # script for slurm
print SCRIPT "#!/bin/tcsh -f\n";
print SCRIPT "cd $peptide\n";
print SCRIPT "/usr/local/namdtest/namd2 +p4 all.namd\n";
print SCRIPT "cd ..\n";
print SCRIPT "/bin/mv -f $peptide done/\n";
print SCRIPT "/bin/rm -rf $peptide.sh\n";
close( SCRIPT );

`sbatch --no-requeue --mail-type=ALL -q -n4 -N 1 --exclusive -o $peptide/LOG $peptide.sh` or print $!;
}

###
#
# End of main()
#
###
exit;

```

```

###
#
# Open file containing peptides, select randomly one,
# put the rest back, return selected
#
###
sub get_peptide
{
  my $i;

  open( LIST, "$pept2go" ) or die $!;
  @all = <LIST>;
  close( LIST );

  if ( @all == 0 )      # We are done, e-mail user ...
  {
    print "No more peptides left. Done.\n";
    exit;
  }

  open( LIST, ">$pept2go" ) or die $!;
  $index = int(rand @all);
  for ( $i=0 ; $i < @all ; $i++ )
  {
    if ( $i != $index )
    {
      print LIST $all[$i];
    }
  }

  close( LIST );
  return( $all[$index] );
}

```

```

#####
#
# PrepSystem Function
#
#####

```

```

sub prepare_MD_files {
  my $seq;
  $seq = $_[0];
  @seq = split (',', $seq);

  my %AA_names = (
    'A' => 'Ala',
    'C' => 'Cys',
    'D' => 'Asp',
    'E' => 'Glu',
    'F' => 'Phe',
    'G' => 'Gly',
    'H' => 'His',
    'I' => 'Ile',
    'K' => 'Lys',
    'L' => 'Leu',
    'M' => 'Met',
    'N' => 'Asn',
    'P' => 'Pro',
    'Q' => 'Gln',
    'R' => 'Arg',
    'S' => 'Ser',
    'T' => 'Thr',
    'V' => 'Val',
    'W' => 'Trp',
    'Y' => 'Tyr',
  );
}

```

```

mkdir ("$seq", 0777) || die $!;
chdir ("$seq") || die $!;

```

```

open FILE, ">ribosome.script" or die $!;
print FILE "title $seq\n";
print FILE "default extended\n";
foreach $seq (@seq)
{
print FILE "res $AA_names{$seq}\n";
}
close(FILE);
system("../bin/ribosome ribosome.script starting.pdb ../bin/ribosome.dat") == 0 || die "system error $?";

open FILE, ">moleman.sh" or die $!;
print FILE "#!/bin/tcsh -f\n";
print FILE "../bin/lx_moleman2 >& moleman.log << eof\n";
print FILE "../bin/moleman2.lib\n";
print FILE "REad starting.pdb\n";
print FILE "Xyz Align_inertia_axes\n";
print FILE "WRite aligned.pdb\n";
print FILE "quit\n";
print FILE "eof\n";
print FILE "exit\n";
close(FILE);
system("chmod 755 moleman.sh") == 0 || die "system error $?";
system("../moleman.sh") == 0 || die "system error $?";

system("sed 's/HIS/HSP/g' < aligned.pdb > new.pdb ; mv new.pdb aligned.pdb") == 0 || die "system error $?";
system("../bin/psfgen.sh") == 0 || die "system error $?";

system("/usr/local/bin/vmd -dispdev text < ../bin/vmd.tcl > VMD_log") == 0 || die "system error $?";

`/bin/cp -f ../bin/all.namd ./' or print $!;
`/bin/cp -f ../bin/par_all27_prot_na.inp ./' or print $!;
`/bin/rm -rf starting.pdb VMD_log aligned.pdb combine.* hydrated.* moleman.* psfgen.pdb psfgen.log
ribosome.script` or print $!;

chdir ("..") || die $!;
return;
}

#####
#
# Average function
#
#####

sub Aver {
my @data;
my $len;
my $aver;

open (INFILE, "$_[0]" ) or die $!;
@data = <INFILE>;
close(INFILE);

$len = @data;
$aver = 0.0;

for ( $i=0 ; $i < $len ; $i++ )
{
$aver += $data[ $i ];
}

$aver /= $len;
return( $aver );
}

```

```

#####
#
# The target function
#
#####

sub TargetFunction {

my @data;
my $len;
my $min;
my $max;
my $i;
my $k;
my $D;
my $aver;
my $rmsd;
my @RMSDs;
my $target;

open (INFILE, "$_[0]" ) or die $!;
@data = <INFILE>;
close(INFILE);

$len = @data;

$min = $data[0];
$max = $data[0];
for ( $i = 0 ; $i < $len ; $i++ )
{
    if ( $data[$i] > $max )
    {
        $max = $data[$i];
    }

    if ( $data[$i] < $min )
    {
        $min = $data[$i];
    }
}

$D = $max - $min;

for ( $k=0 ; $k < $len - WIDTH ; $k++ )
{

$aver = 0.0;
for ( $i=$k ; $i < $k + WIDTH ; $i++ )
{
    $aver += $data[ $i ] ;
}

$aver /= WIDTH;

$rmsd = 0.0;
for ( $i=$k ; $i < $k + WIDTH ; $i++ )
{
    $rmsd += ( $data[ $i ] - $aver ) * ( $data[ $i ] - $aver );
}

$rmsd /= ( WIDTH - 1 );
$rmsd = sqrt( $rmsd );

$RMSDs[ $k ] = $rmsd;
}

$max = $RMSDs[0];
$min = $RMSDs[0];

```

```

for ( $i=0 ; $i < $len - WIDTH ; $i++ )
{
  if ( $RMSDs[ $i ] > $max )
  {
    $max = $RMSDs[ $i ];
  }

  if ( $RMSDs[ $i ] < $min )
  {
    $min = $RMSDs[ $i ];
  }
}

$aver = 0.0;
for ( $i=0 ; $i < $len ; $i++ )
{
  $aver += $data[ $i ];
}

$aver /= $len;

$rmsd = 0.0;
for ( $i=0 ; $i < $len ; $i++ )
{
  $rmsd += ( $data[ $i ] - $aver ) * ( $data[ $i ] - $aver );
}

$rmsd /= ( $len - 1 );
$rmsd = sqrt( $rmsd );

$target = ( $D * $max ) / ( $aver * $min * $rmsd );

return( $target );
}

#####
#
# Linear corr coeff between distances
#
#####

sub CorrelationFunction {

my @x;
my @y;
my @z;

my $sum_sq_x;
my $sum_sq_y;
my $sum_sq_z;
my $sum_coproduct_xy;
my $sum_coproduct_xz;
my $sum_coproduct_yz;
my $mean_x;
my $mean_y;
my $mean_z;
my $N;
my $i;
my $sweep;
my $Dx;
my $Dy;
my $Dz;
my $pop_sd_x;
my $pop_sd_y;
my $pop_sd_z;
my $cov_x_y;
my $cov_x_z;
my $cov_y_z;
my $correlation_xy;
my $correlation_xz;
my $correlation_yz;

```



```

open (INFILE, "$_[0]" ) or die $!;
@x = <INFILE>;
close(INFILE);

open (INFILE, "$_[1]" ) or die $!;
@y = <INFILE>;
close(INFILE);

open (INFILE, "$_[2]" ) or die $!;
@z = <INFILE>;
close(INFILE);

$sum_sq_x = 0;
$sum_sq_y = 0;
$sum_sq_z = 0;
$sum_coproduct_xy = 0;
$sum_coproduct_xz = 0;
$sum_coproduct_yz = 0;
$mean_x = $x[0];
$mean_y = $y[0];
$mean_z = $z[0];

$N = @x;

for ( $i=2 ; $i <= $N ; $i++)
{
    $sweep = ($i - 1.0) / $i ;
    $Dx = $x[$i-1] - $mean_x;
    $Dy = $y[$i-1] - $mean_y;
    $Dz = $z[$i-1] - $mean_z;

    $sum_sq_x += $Dx * $Dx * $sweep;
    $sum_sq_y += $Dy * $Dy * $sweep;
    $sum_sq_z += $Dz * $Dz * $sweep;

    $sum_coproduct_xy += $Dx * $Dy * $sweep;
    $sum_coproduct_xz += $Dx * $Dz * $sweep;
    $sum_coproduct_yz += $Dy * $Dz * $sweep;

    $mean_x += $Dx / $i;
    $mean_y += $Dy / $i;
    $mean_z += $Dz / $i;
}

$pop_sd_x = sqrt( $sum_sq_x/$N );
$pop_sd_y = sqrt( $sum_sq_y/$N );
$pop_sd_z = sqrt( $sum_sq_z/$N );

$cov_xy = $sum_coproduct_xy/$N;
$cov_xz = $sum_coproduct_xz/$N;
$cov_yz = $sum_coproduct_yz/$N;

$correlation_xy = $cov_xy/($pop_sd_x * $pop_sd_y);
$correlation_xz = $cov_xz/($pop_sd_x * $pop_sd_z);
$correlation_yz = $cov_yz/($pop_sd_y * $pop_sd_z);

return( ($Target1 * $Target2 * $correlation_xy) + ($Target1 * $Target3 * $correlation_xz) + ($Target2 *
$Target3 * $correlation_yz) );
}

###
#
# Expanding_Windows
#
#Second version. With a 501x501 matrix this is ~3000 times faster than the first version.
#
###

sub Expand_Windows {

$data = [];
$sums = [];

```



```

####
#
# Read rmsd matrix, convert to binary, stored in $$data (upper half only)
#
####

open (INFILE, "$_[0]" ) or die $!;
$i = 0;
while ( $line = <INFILE> )
{
    @numbers = split(' ', $line);
    $N= @numbers;

    for ( $j = 0 ; $j < $N ; $j++ )
    {
        if ( $numbers[ $j ] > RMS_CUTOFF )
        {
            $$data[ $i+1 ][ $j+1 ] = 0;
        }
        elsif ( $i > $j )
        {
            $$data[ $i+1 ][ $j+1 ] = 0;
        }
        else
        {
            $$data[ $i+1 ][ $j+1 ] = 1;
        }
    }
    $i++;
}
close(INFILE);

####
#
# Initialize summation matrix
#
####

for ( $i = 1 ; $i <= $N ; $i++ )
{
    for ( $j = 1 ; $j <= $N ; $j++ )
    {
        if ( $i == $j )
        {
            $$sums[ $i ][ $j ] = 1;
        }
        else
        {
            $$sums[ $i ][ $j ] = 0;
        }
    }
}

####
#
# Fill summation matrix
#
####

for ( $k = 1 ; $k < $N ; $k++ )
{
    $i = 1;
    for ( $j = $k+1 ; $j <= $N ; $j++ )
    {
        $$sums[ $i ][ $j ] = $$sums[ $i ][ $j-1 ] + $$sums[ $i+1 ][ $j ] + 2*$$data[ $i ][ $j ] - $$sums[ $i+1 ][
$j-1 ];
        $i++;
    }
}

```

```

####
#
# Initialize histogram values
#
####

for ( $i = 0 ; $i <= 100 ; $i++ )
{
    $hist[ $i ] = 0;
}

####
#
# Second pass through the summation matrix to calculate percentages
#
####

for ( $k = 1 ; $k < $N ; $k++ )
{
    $i = 1;
    for ( $j = $k+1 ; $j <= $N ; $j++ )
    {
        $percent = int ( 100 * ( $$sums[ $i ][ $j ] / (($k+1)*($k+1))) + 0.50 );
        $hist[ $percent ]++;
        $i++;
    }
}

$nof_data = 0;
for ( $i=0 ; $i <= 100 ; $i++ )
{
    $nof_data += $hist[ $i ];
}

####
#
# The rest ...
#
####

$N = @hist;

$max = $hist[0];
$mode = 0;

for ( $k = 0 ; $k < $N ; $k++ )
{
    if ( $hist[ $k ] > $max )
    {
        $max = $hist[ $k ];
        $mode = $k;
    }
}

$middle = int ( $nof_data / 2 );
$median=0;
$value = 0;

for ( $k = 0 ; $k < $N ; $k++ )
{
    $value += $hist[ $k ] ;

    if ( $value < $middle )
    {
        $median = $k + 1;
    }
}

return ( $median * $mode );

```

}

## 11) PSFGEN script

```
#!/bin/tcsh -f
/usr/local/namd/psfgen >& psfgen.log << END
topology ../bin/top_all127_prot_na.inp

segment A {
  pdb aligned.pdb
}
alias atom ILE CD1 CD
coordpdb aligned.pdb A

guesscoord
writepsf psfgen.psf
writepdb psfgen.pdb

END
exit
```

## 12) VMD script

```
#!/usr/local/bin/vmd

#
# Make water box
#
package require vexpr
package require toctsolvate
toctsolvate psfgen.psf psfgen.pdb -o hydrated -minmax {-14 -14 -14} {14 14 14} -b 1.80

#
# Add ions to neutralise charge
#
package require autoionize
autoionize -psf hydrated.psf -pdb hydrated.pdb -is 0.30

#
# Prepare restraints files
#
mol load psf ionized.psf pdb ionized.pdb
set all [atomselect top all]
set sel [atomselect top "protein and name CA"]
$all set beta 0
$sel set beta 0.5
$all writepdb restrain_ca.pdb
set all [atomselect top all]
set to_fix [atomselect top "protein and backbone"]
$all set beta 0
$to_fix set beta 1
$all writepdb fix_backbone.pdb
```

### 13) NAMD script: all.namd

```

#
# Input files
#
structure          ionized.psf
coordinates        ionized.pdb
parameters        par_all127_prot_na.inp
paraTypeCharmm    on
#
# Output files & writing frequency for DCD
# and restart files
#
outputname        heat_out
binaryoutput      off
dcdFile           all_atoms.dcd
dcdFreq           400
DCDunitcell      on

#
# Frequencies for logs and the xst file
#
outputEnergies    40
outputTiming      400
xstFreq           400

#
# Timestep & friends
#
timestep          2.0
stepsPerCycle     20
nonBondedFreq    2
fullElectFrequency 4

#
# Simulation space partitioning
#
switching         on
switchDist       7
cutoff            8
pairlistdist     9

#
# Basic dynamics
#
temperature       0
COMmotion        no
dielectric        1.0
exclude          scaled1-4
1-4scaling       1.0
rigidbonds       all

#
# Particle Mesh Ewald parameters.
#
Pme               on
PmeGridsizeX     27
PmeGridsizeY     27
PmeGridsizeZ     25

#
# Periodic boundary things
#

```

```

wrapWater          on
wrapNearest        on

cellBasisVector1   26.00   0.00   0.00
cellBasisVector2   0.00   26.00   0.00
cellBasisVector3   13.00   13.00   13.00
cellOrigin          0.00   0.00   0.00

#
# Fixed atoms for initial heating-up steps
#
fixedAtoms          on
fixedAtomsForces    on
fixedAtomsFile      fix_backbone.pdb
fixedAtomsCol       B

#
# Restrained atoms for initial heating-up steps
#
constraints         on
consRef             restrain_ca.pdb
consKFile           restrain_ca.pdb
consKCol            B

#
# Langevin dynamics parameters
#
langevin            on
langevinDamping     1
langevinTemp        320
langevinHydrogen    on

langevinPiston      on
langevinPistonTarget 1.01325
langevinPistonPeriod 200      #500
langevinPistonDecay 100       #200
langevinPistonTemp  320

useGroupPressure    yes

#####
# The actual minimisation and heating-up
# protocol follows. The number of steps
# shown below are too small for a real run
#####

#
# run one step to get into scripting mode
#
minimize            0

#
# turn off pressure control until later
#
langevinPiston      off

#
# minimize nonbackbone atoms
#
minimize            500
output              min_fix

#
# min all atoms
#
fixedAtoms          off
minimize            500

```

```

output                min_all

#
# heat with CAs restrained
#
set temp 20;
while { $temp < 321 } {
  langevinTemp         $temp
  run                  1000
  output               heat_ca
  set temp [expr $temp + 100]
}

#
# equilibrate volume with CAs restrained
#
run                   1000
output               equil_ca

#
# equilibrate volume without restraints
#
constraintScaling     0
langevinPiston        on
run                   10000000

```

#### 14) NAMD script: heat.namd

```

#
# Input files
#
structure              ionized.psf
coordinates            ionized.pdb
parameters             par_all127_prot_na.inp
paraTypeCharmm        on

#
# Output files & writing frequency for DCD
# and restart files
#
outputname             heat_out
binaryoutput          off
restartname            restart
restartfreq            10000
binaryrestart         yes
dcdFile               all_atoms.dcd
dcdFreq               200
DCDunitcell          on

#
# Frequencies for logs and the xst file
#
outputEnergies         40
outputTiming           400
xstFreq               400

#
# Timestep & friends
#
timestep               2.0
stepsPerCycle          20
nonBondedFreq         2
fullElectFrequency     4

```



```

#
# Simulation space partitioning
#
switching          on
switchDist        7
cutoff            8
pairlistdist      9

#
# Basic dynamics
#
temperature        0
COMmotion         no
dielectric         1.0
exclude           scaled1-4
1-4scaling        1.0
rigidbonds        all

#
# Particle Mesh Ewald parameters.
#
Pme               on
PmeGridsizeX     27
PmeGridsizeY     27
PmeGridsizeZ     25

#
# Periodic boundary things
#
wrapWater         on
wrapNearest       on
wrapAll           on

cellBasisVector1  26.00   0.00   0.00
cellBasisVector2   0.00  26.00   0.00
cellBasisVector3  13.00  13.00  13.00
cellOrigin         0.00   0.00   0.00

#
# Fixed atoms for initial heating-up steps
#
fixedAtoms        on
fixedAtomsForces  on
fixedAtomsFile    fix_backbone.pdb
fixedAtomsCol     B

#
# Restrained atoms for initial heating-up steps
#
constraints       on
consRef           restrain_ca.pdb
consKFile         restrain_ca.pdb
consKCol         B

#
# Langevin dynamics parameters
#
langevin          on
langevinDamping   10
langevinTemp      320
langevinHydrogen  off

langevinPiston    on
langevinPistonTarget 1.01325
langevinPistonPeriod 500
langevinPistonDecay 200
langevinPistonTemp 320

useGroupPressure  yes

#####
# The actual minimisation and heating-up
# protocol follows. The number of steps
# shown below are too small for a real run
#####

```

```

#
# run one step to get into scripting mode
#
minimize          0

#
# turn off pressure control until later
#
langevinPiston    off

#
# minimize nonbackbone atoms
#
minimize          500
output            min_fix

#
# min all atoms
#
fixedAtoms        off
minimize          500
output            min_all

#
# heat with CAs restrained
#
set temp 20;
while { $temp < 321 } {
  langevinTemp     $temp
  run              1000
  output           heat_ca
  set temp [expr $temp + 100]
}

#
# equilibrate volume with CAs restrained
#
run               1000
output            equil_ca

#
# equilibrate volume without restraints
#
constraintScaling 0
langevinPiston    on
run               20000

```

### 15) NAMD script: equi.namd

```

#
# Input files
#
structure          ionized.psf
coordinates         heat_out.coor
velocities          heat_out.vel
extendedSystem     heat_out.xsc
parameters         par_all27_prot_na.inp
paraTypeCharmm     on

#
# Output files & writing frequency for DCD
# and restart files
#
outputname         equi_out
binaryoutput       off

```

```

restartname          restart
restartfreq          10000
binaryrestart        yes
dcdFile              equi_out.dcd
dcdFreq              200
DCDunitcell         yes

#
# Frequencies for logs and the xst file
#
outputEnergies       40
outputTiming         400
xstFreq              400

#
# Timestep & friends
#
timestep             2.0
stepsPerCycle        20
nonBondedFreq        2
fullElectFrequency   4

#
# Simulation space partitioning
#
switching             on
switchDist           7
cutoff                8
pairlistdist         9

#
# Basic dynamics
#
COMmotion            no
dielectric            1.0
exclude              scaled1-4
1-4scaling           1.0
rigidbonds           all

#
# Particle Mesh Ewald parameters.
#
Pme                   on
PmeGridsizeX         27      # <===== CHANGE ME
PmeGridsizeY         27      # <===== CHANGE ME
PmeGridsizeZ         25      # <===== CHANGE ME

#
# Periodic boundary things
#
wrapWater             on
wrapNearest          on
wrapAll              on

#
# Langevin dynamics parameters
#
langevin              on
langevinDamping       1
langevinTemp          320     # <===== Check me
langevinHydrogen      off

langevinPiston        on
langevinPistonTarget  1.01325
langevinPistonPeriod  500
langevinPistonDecay   200
langevinPistonTemp    320     # <===== Check me

useGroupPressure      yes

firsttimestep         25000    # <===== CHANGE ME
run                   150000000 ;# <===== CHANGE ME

```

## 16) pickBestRun.pl

```
#!/usr/bin/perl -w

while (1) {
    $run1 = <STDIN> or exit;
    @first = split(' ', $run1);
    $run2 = <STDIN>;
    @second = split(' ', $run2);
    $run3 = <STDIN>;
    @third = split(' ', $run3);
    $run4 = <STDIN>;
    @fourth = split(' ', $run4);

    @scores = ($first[1], $second[1], $third[1], $fourth[1]);
    @best = sort { $b <=> $a } @scores;

    print "$first[0] $best[0]\n";
}

```

## 17) pickAverRun.pl

```
#!/usr/bin/perl -w

while (1) {
    $run1 = <STDIN> or exit;
    @first = split(' ', $run1);
    $run2 = <STDIN>;
    @second = split(' ', $run2);
    $run3 = <STDIN>;
    @third = split(' ', $run3);
    $run4 = <STDIN>;
    @fourth = split(' ', $run4);

    $aver = ($first[1] + $second[1] + $third[1] + $fourth[1]) / 4;

    print "$first[0] $aver\n";
}

```

## 18) dist\_matrix.pl

```
#!/usr/bin/perl -w

@scores = <STDIN>;
$len = @scores;

for ($i = 0 ; $i < $len ; $i++)
{
    for ($k = 0 ; $k < $len ; $k++)
    {
        $dist = $scores[$k] - $scores[$i];
        $dist = abs($dist);
        printf "%5.2f ", $dist;
    }
    print "\n";
}

```

```

    }
exit;

```

## 19) high\_blue.RMSDmatrix.pl

```

#!/usr/bin/perl -w

chdir ("res") || die $!;
opendir (DIR, '.') or die "Can't open current dir: $!\n";
@peptides = grep (!/^\\.\\.?$/, readdir (DIR));
closedir (DIR);

foreach $pept ( @peptides )
{
    $high_blue = 0;

    open( INFILE, "$pept") or die $!;
    while ($data = <INFILE>)
    {
        @value = split( ' ', $data);
        $N = @value;

        for ( $i = 0 ; $i < $N ; $i++ )
        {
            if ($value[ $i ] >= 0.90)
            {
                $high_blue++;
            }
        }
    }
    print "$pept $high_blue\n";
}

```

## 20) systematic.AMBER.pl

```

#!/usr/bin/perl -w
#
# This script exit()'s. Should be run from crontab.
# To make sure that it will fill the queue, adjust periodicity
#
#
`renice +19 -p $$ >& /dev/null` or print $!;

$USER          = "georgoulia"; # User name
$target_PD     = 8;           # How many pending jobs we aim to have
$max_pending   = 16;          # If that many jobs are waiting, sleep
$pept2go       = "2GO";      # Filename with list of peptides
#$jobs_per_pept = 1;         # How many jobs per pept ?
#$pep_len      = 5;          # Residues per peptide

use constant WIDTH => 500;
use constant RMS_CUTOFF => 2.0;

# set the perl environment variable PATH
$ENV{'PATH'} = '/bin:/usr/bin:/usr/local/bin';
$ENV{'LD_LIBRARY_PATH'} = '/usr/local/lib';

#
# Do we have a cluster ?
#

```



```

$cpus = `sinfo -h -o "%C" | awk -F '/' '{print \$4}`;
if ( $cpus < 12 )
{
  exit;
}
###
#
# First thing first: are there any completed jobs to process ?
#
###

                                # Get list of directories
chdir ("done") || die $!;
opendir (DIR, '.') or die "Can't open current dir: $!\n";
@dirs = grep (!/^\.\.?$/, readdir (DIR));
closedir (DIR);

if ( @dirs > 0 )                                # Something to process ...
{
  foreach $directory ( @dirs )                  # Directories to process
  {
    chdir ("$directory") || die $!;            # Go there

                                # Make fit.index
    `/bin/grep -P '[2-4] .*CA.*C..' ionized.psf > fit.index` or print $!;

                                # Remove waters-ions (using fit.index)
    `/usr/local/bin/carma -v -w -fit -ind -atmid ALLID -segid A ionized.psf all\_atoms.dcd` or print $!;

    `mv carma.fitted.dcd $directory.dcd` or print $!;
    `mv carma.selected_atoms.psf psfgen.psf` or print $!;
                                # Prepare some data ...
    `/usr/local/bin/carma -v -dist 1 5 psfgen.psf $directory.dcd` or print $!;
    `awk '{print \$2}' carma.distances > tmp ; mv tmp carma1.distances` or print $!;

                                # and calculate target function
    $Target1 = TargetFunction( "carma1.distances" );

    `/usr/local/bin/carma -v -dist 1 4 psfgen.psf $directory.dcd` or print $!;
    `awk '{print \$2}' carma.distances > tmp ; mv tmp carma2.distances` or print $!;

                                # and calculate target function
    $Target2 = TargetFunction( "carma2.distances" );

    `/usr/local/bin/carma -v -dist 2 5 psfgen.psf $directory.dcd` or print $!;
    `awk '{print \$2}' carma.distances > tmp ; mv tmp carma3.distances` or print $!;

                                # and calculate target function
    $Target3 = TargetFunction( "carma3.distances" );

    $Distcorr = CorrelationFunction( "carma1.distances", "carma2.distances", "carma3.distances", $Target1,
                                     $Target2, $Target3 );

    `/usr/local/bin/carma -v -rg -atmid HEAVY psfgen.psf $directory.dcd` or print $!;
    `awk '{print \$2}' carma.Rgyration.dat > tmp ; mv tmp carma.Rg` or print $!;

                                # and calculate target function
    $Target4 = TargetFunction( "carma.Rg" );

                                # Cartesian PCA
    `/usr/local/bin/carma -v -w -col -cov -eigen -shannon -atmid HEAVY -proj 3 3 320 psfgen.psf $directory.dcd >
                                     PCA_LOG` or print $!;

                                # configurational Entropy of PCA distribution
    $Target5 = `cat PCA_LOG | grep 'entropy' | awk '{print \$7}` or print $!;
    chomp($Target5);

                                # number of Clusters
    $Target6 = `cat PCA_LOG | grep 'Number.*clusters' | awk '{print \$4}` or print $!;
  }
}

```



```

chomp($Target6);

`awk '{ if ($2 == 1) print $1, $3, $4, $5}' carma.clusters.dat > C_01.dat` or print $!;
`/usr/local/bin/carma -v -sort C_01.dat $directory.dcd` or print $!;
`/usr/local/bin/carma -v -fit -ind -atmid ALLID carma.reordered.dcd psfgen.psf` or print $!;
$Target7 = `wc C_01.dat | awk '{print $1}'`;
chomp($Target7);

if ( $Target7 > 2500 )
{
  `/usr/local/bin/carma -v -w -col -cov -dot -atmid HEAVY carma.fitted.dcd psfgen.psf` or print $!;
  `/usr/local/bin/carma -v -pdb -step 100 -atmid HEAVY carma.fitted.dcd psfgen.psf` or print $!;

  # Rmsfs
  `cat carma.average.pdb | awk '{print $11}' | sed '\$d' > $directory.heavy.rmsf` or print $!;
  `cat carma.average.pdb | grep TRP | sed -n '3,12p;12p' | awk '{print $11}' > $directory.Trp-
  sidechain.rmsf` or print $!;
  `cat carma.average.pdb | grep -P " ([NCO]) | (CA) " | awk '{print $11}' > $directory.backbone.rmsf`
  or print $!;
  `cat carma.average.pdb | grep -v -P " ([NCO]) | (CA) | TRP " | awk '{print $11}' | sed '\$d' >
  $directory.rest-sidechains.rmsf` or print $!;

  # Calculate average rmsfs
  $Target8 = Aver( "$directory.heavy.rmsf");
  $Target9 = Aver( "$directory.Trp-sidechain.rmsf");
  $Target10 = Aver( "$directory.backbone.rmsf");
  $Target11 = Aver( "$directory.rest-sidechains.rmsf");

  # Prepare superpositionPDBs
  `/bin/mkdir tmp` or print $!;
  `/bin/mv carma.fitted.dcd.*.pdb tmp/` or print $!;
  chdir ("tmp") || die $!;
  opendir (DIR, '.') or die "Can't open current dir: $!\n";
  @pdirs = grep (!/\.\.?$/, readdir (DIR));
  closedir (DIR);
  foreach $pdb (@pdirs)
  {
    `awk '{ printf "%-7s %3s %-3s %3s %-3s %s %11s %7s %7s %5s\n", $1, $2, $3, $4, $5, $6, $7, \
    $8, $9, $10 }' "$pdb" > "$pdb.tmp" ` or print $!;
    `paste $pdb.tmp ../$directory.heavy.rmsf > $pdb.PDB ` or print $!;
  }
  `cat *.PDB > cluster1.$Target7.superposition.pdb` or print $!;
  `/usr/bin/bzip2 cluster1.$Target7.superposition.pdb` or print $!;
  `/bin/cp -f cluster1.$Target7.superposition.pdb.bz2 ../../../../graphs/
  $directory.cluster1.$Target7.superposition.pdb.bz2` or print $!;

  chdir ("../") || die $!;
}

# crossDCD
`/bin/rm carma.fitted.dcd carma.fit-rms.dat` or print $!;
`/usr/local/bin/crossDCD psfgen.psf $directory.dcd $directory.dcd 250 "-atmid HEAVY" > crossDCD.log`
or print $!;

# Calculate crossDCD-score
$Target12 = Expand_Windows( "crossDCD.matrix" );

# save some data ...
`/usr/bin/bzip2 carma1.distances` or print $!;
`/bin/cp -f carma1.distances.bz2 ../../graphs/$directory.1.5.bz2`;
`/usr/bin/bzip2 carma2.distances` or print $!;
`/bin/cp -f carma2.distances.bz2 ../../graphs/$directory.1.4.bz2`;
`/usr/bin/bzip2 carma3.distances` or print $!;
`/bin/cp -f carma3.distances.bz2 ../../graphs/$directory.2.5.bz2`;
`/usr/bin/bzip2 carma.Rg` or print $!;
`/bin/cp -f carma.Rg.bz2 ../../graphs/$directory.Rg.bz2`;
`/usr/bin/bzip2 carma.PCA.fluctuations.dat` or print $!;
`/bin/cp -f carma.PCA.fluctuations.dat.bz2 ../../graphs/$directory.eigen.bz2`;
`/usr/bin/bzip2 crossDCD.matrix` or print $!;
`/bin/cp -f crossDCD.matrix.bz2 ../../graphs/$directory.crossDCD.matrix.bz2`;
`/bin/cp -f $directory.dcd ../../graphs/`;
`/bin/cp -f pentapept.prmtop ../../graphs/$directory.prmtop`;
`/bin/cp -f psfgen.psf ../../graphs/$directory.psf`;

```

```

# Tidy-up
chdir ("../") || die $!;
`/bin/rm -rf $directory` or print $!;

# Write-out ...
if ( $Target7 > 2500 )
{
  open( OUT, ">>../results" ) || die $!;
  print OUT "$directory $Target1 $Target2 $Target3 $Distcorr $Target4 $Target5 $Target6 $Target7 $Target8
$Target9 $Target10 $Target11 $Target12\n";
  close( OUT );
}
else
{
  open( OUT, ">>../results" ) || die $!;
  print OUT "$directory $Target1 $Target2 $Target3 $Distcorr $Target4 $Target5 $Target6 $Target7
$Target12\n";
  close( OUT );
}
}
chdir ("..") || die $!;

###
#
# Now: do we have to submit new jobs to the queue ?
#
###

# How many pending jobs are in the queue ?
$pending_all = `squeue -h -t PD | wc -l`;
$pending_user = `squeue -h -u $USER -t PD | wc -l`;

# If cluster too busy, go to sleep ...
if ( $pending_all > $max_pending )
{
  exit;
}

# If less than target, submit $target_PD
if ( $pending_user < $target_PD )
{
  $peptide = get_peptide(); # select randomly a peptide
  chomp $peptide; # prepare files
  prepare_MD_files( $peptide ); # submit job with NAMDjob

# chdir ("$peptide") || die $!;
# `/usr/bin/NAMDjob 4 all.namd LOG` or print $!;
# chdir ("../") || die $!;

  open( SCRIPT, ">$peptide.sh" ) or print $!; # script for slurm
  print SCRIPT "#!/bin/tcsh -f\n";
  print SCRIPT "cd $peptide\n";
  print SCRIPT "/usr/local/namd_multicore/namd2 +p4 all.namd\n";
  print SCRIPT "cd ..\n";
  print SCRIPT "/bin/mv -f $peptide done/\n";
  print SCRIPT "/bin/rm -rf $peptide.sh\n";
  close( SCRIPT );

  `sbatch --no-requeue --mail-type=ALL -q -n4 -N 1 -p noncuda --exclusive -o $peptide/LOG $peptide.sh` or
  print $!;
}

##

```

```

#
# End of main()
#
##
exit;
###
#
# Open file containing peptides, select randomly one,
# put the rest back, return selected
#
###

sub get_peptide
{
  my $i;

  open( LIST, "$pept2go" ) or die $!;
  @all = <LIST>;
  close( LIST );

  if ( @all == 0 )      # We are done, e-mail user ...
  {
    print "No more peptides left. Done.\n";
    exit;
  }

  open( LIST, ">$pept2go" ) or die $!;
  $index = int(rand @all);
  for ( $i=0 ; $i < @all ; $i++ )
  {
    if ( $i != $index )
    {
      print LIST $all[$i];
    }
  }

  close( LIST );
  return( $all[$index] );
}

#####
#
# PrepSystem Function
#
#####
sub prepare_MD_files {
my $seq;
$seq = $_[0];
@seq = split (',', $seq);

my %AA_names = (
  'A' => 'Ala',
  'C' => 'Cys',
  'D' => 'Asp',
  'E' => 'Glu',
  'F' => 'Phe',
  'G' => 'Gly',
  'H' => 'His',
  'I' => 'Ile',
  'K' => 'Lys',
  'L' => 'Leu',
  'M' => 'Met',
  'N' => 'Asn',
  'P' => 'Pro',
  'Q' => 'Gln',
  'R' => 'Arg',
  'S' => 'Ser',
  'T' => 'Thr',
  'V' => 'Val',
  'W' => 'Trp',
  'Y' => 'Tyr',
);
}

```

```

mkdir ("${seq}", 0777) || die $!;
chdir ("${seq}") || die $!;

open FILE, ">ribosome.script" or die $!;
print FILE "title ${seq}\n";
print FILE "default extended\n";
foreach $seq (@seq)
{
print FILE "res $AA_names[${seq}]\n";
}
close(FILE);
system("../bin/ribosome ribosome.script starting.pdb ../bin/ribosome.dat") == 0 || die "system error $?";

open FILE, ">moleman.sh" or die $!;
print FILE "#!/bin/tcsh -f\n";
print FILE "../bin/lx_moleman2 >& moleman.log << eof\n";
print FILE "../bin/moleman2.lib\n";
print FILE "REad starting.pdb\n";
print FILE "XYZ ALign_inertia_axes\n";
print FILE "WRite aligned.pdb\n";
print FILE "quit\n";
print FILE "eof\n";
print FILE "exit\n";
close(FILE);
system("chmod 755 moleman.sh") == 0 || die "system error $?";
system("./moleman.sh") == 0 || die "system error $?";

system ("sed 's/HIS/HIP/g' < aligned.pdb > new.pdb ; mv new.pdb aligned.pdb") == 0 || die "system error $?";

open FILE, ">leap.script" or die $!;
print FILE "${seq} = loadPDB aligned.pdb\n";
print FILE "check ${seq}\n";
print FILE "addions ${seq} Na+ 1\n";
print FILE "solvateoct ${seq} TIP3PBOX 2.8\n";
print FILE "savePDB ${seq} ${seq}.pdb\n";
print FILE "saveamberparm ${seq} pentapept.prmtop pentapept.inpcrd\n";
print FILE "quit\n";
close(FILE);

open FILE, ">prmtop_to_psf" or die $!;
print FILE "mol new pentapept.prmtop waitfor all\n";
print FILE "animate dup 0\n";
print FILE "set sel [atomelect top all]\n";
print FILE "\$sel writepsf ionized.psf\n";
print FILE "quit\n";
close(FILE);

`/usr/local/amber10/bin/tleap -s -f /usr/local/amber10/dat/leap/cmd/leaprc.ff99SBildn -f leap.script` or print
$!;

system("/usr/local/bin/vmd -dispdev text < prmtop_to_psf > VMD_log") == 0 || die "system error $?";

`/bin/cp -f ../bin/all.namd ./` or print $!;

`/bin/rm -rf starting.pdb aligned.pdb moleman.* ribosome.script` or print $!;
chdir ("..") || die $!;
return;
}

#####
#
# Average function
#
#####

sub Aver {
my @data;
my $len;

```

```

my $aver;

open (INFILE, "$_[0]" ) or die $!;
@data = <INFILE>;
close(INFILE);

$len = @data;
$aver = 0.0;

for ( $i=0 ; $i < $len ; $i++ )
{
    $aver += $data[ $i ];
}

$aver /= $len;

return( $aver );
}

#####
#
# The target function
#
#####

sub TargetFunction {

my @data;
my $len;
my $min;
my $max;
my $i;
my $k;
my $D;
my $aver;
my $rmsd;
my @RMSDs;
my $target;

open (INFILE, "$_[0]" ) or die $!;
@data = <INFILE>;
close(INFILE);

$len = @data;

$min = $data[0];
$max = $data[0];
for ( $i = 0 ; $i < $len ; $i++ )
{
    if ( $data[$i] > $max )
    {
        $max = $data[$i];
    }

    if ( $data[$i] < $min )
    {
        $min = $data[$i];
    }
}

$D = $max - $min;

for ( $k=0 ; $k < $len - WIDTH ; $k++ )
{
    $aver = 0.0;
    for ( $i=$k ; $i < $k + WIDTH ; $i++ )
    {
        $aver += $data[ $i ];
    }
}
}

```



```

}
$aver /= WIDTH;
$rmsd = 0.0;
for ( $i=$k ; $i < $k + WIDTH ; $i++ )
{
    $rmsd += ( $data[ $i ] - $aver ) * ( $data[ $i ] - $aver );
}

$rmsd /= ( WIDTH - 1 );
$rmsd = sqrt( $rmsd );

$RMSDs[ $k ] = $rmsd;
}

$max = $RMSDs[0];
$min = $RMSDs[0];

for ( $i=0 ; $i < $len - WIDTH ; $i++ )
{
    if ( $RMSDs[ $i ] > $max )
    {
        $max = $RMSDs[ $i ];
    }

    if ( $RMSDs[ $i ] < $min )
    {
        $min = $RMSDs[ $i ];
    }
}

$aver = 0.0;
for ( $i=0 ; $i < $len ; $i++ )
{
    $aver += $data[ $i ];
}

$aver /= $len;

$rmsd = 0.0;
for ( $i=0 ; $i < $len ; $i++ )
{
    $rmsd += ( $data[ $i ] - $aver ) * ( $data[ $i ] - $aver );
}

$rmsd /= ( $len - 1 );
$rmsd = sqrt( $rmsd );

$target = ( $D * $max ) / ( $aver * $min * $rmsd );

return( $target );
}

#####
#
# Linear corr coeff between distances
#
#####

sub CorrelationFunction {

my @x;
my @y;
my @z;

my $sum_sq_x;
my $sum_sq_y;

```



```

my $sum_sq_x;
my $sum_coproduct_xy;
my $sum_coproduct_xz;
my $sum_coproduct_yz;
my $mean_x;
my $mean_y;
my $mean_z;
my $N;
my $i;
my $sweep;
my $Dx;
my $Dy;
my $Dz;
my $pop_sd_x;
my $pop_sd_y;
my $pop_sd_z;
my $cov_x_y;
my $cov_x_z;
my $cov_y_z;
my $correlation_xy;
my $correlation_xz;
my $correlation_yz;

open (INFILE, "$_[0]" ) or die $!;
@x = <INFILE>;
close(INFILE);

open (INFILE, "$_[1]" ) or die $!;
@y = <INFILE>;
close(INFILE);

open (INFILE, "$_[2]" ) or die $!;
@z = <INFILE>;
close(INFILE);

$sum_sq_x = 0;
$sum_sq_y = 0;
$sum_sq_z = 0;
$sum_coproduct_xy = 0;
$sum_coproduct_xz = 0;
$sum_coproduct_yz = 0;
$mean_x = $x[0];
$mean_y = $y[0];
$mean_z = $z[0];

$N = @x;

for ( $i=2 ; $i <= $N ; $i++)
{
    $sweep = ($i - 1.0) / $i ;
    $Dx = $x[$i-1] - $mean_x;
    $Dy = $y[$i-1] - $mean_y;
    $Dz = $z[$i-1] - $mean_z;

    $sum_sq_x += $Dx * $Dx * $sweep;
    $sum_sq_y += $Dy * $Dy * $sweep;
    $sum_sq_z += $Dz * $Dz * $sweep;

    $sum_coproduct_xy += $Dx * $Dy * $sweep;
    $sum_coproduct_xz += $Dx * $Dz * $sweep;
    $sum_coproduct_yz += $Dy * $Dz * $sweep;

    $mean_x += $Dx / $i;
    $mean_y += $Dy / $i;
    $mean_z += $Dz / $i;
}

$pop_sd_x = sqrt( $sum_sq_x/$N );
$pop_sd_y = sqrt( $sum_sq_y/$N );
$pop_sd_z = sqrt( $sum_sq_z/$N );

```

```

$cov_x_y = $sum_coproduct_xy/$N;
$cov_x_z = $sum_coproduct_xz/$N;
$cov_y_z = $sum_coproduct_yz/$N;

$correlation_xy = $cov_x_y/($pop_sd_x * $pop_sd_y);
$correlation_xz = $cov_x_z/($pop_sd_x * $pop_sd_z);
$correlation_yz = $cov_y_z/($pop_sd_y * $pop_sd_z);

return( ($Target1 * $Target2 * $correlation_xy) + ($Target1 * $Target3 * $correlation_xz) + ($Target2 *
$Target3 * $correlation_yz) );
}

####
#
# Expanding_Windows
#
#Second version. With a 501x501 matrix this is ~3000 times faster than the first version.
#
####

sub Expand_Windows {

$data = [];
$sums = [];

####
#
# Read rmsd matrix, convert to binary, stored in $$data (upper half only)
#
####
open (INFILE, "$_[0]" ) or die $!;
$i = 0;
while ( $line = <INFILE> )
{
    @numbers = split(' ', $line);
    $N= @numbers;

    for ( $j = 0 ; $j < $N ; $j++ )
    {
        if ( $numbers[ $j ] > RMS_CUTOFF )
        {
            $$data[ $i+1 ][ $j+1 ] = 0;
        }
        elsif ( $i > $j )
        {
            $$data[ $i+1 ][ $j+1 ] = 0;
        }
        else
        {
            $$data[ $i+1 ][ $j+1 ] = 1;
        }
    }
    $i++;
}
close(INFILE);

####
#
# Initialize summation matrix
#
####

for ( $i = 1 ; $i <= $N ; $i++ )
{
    for ( $j = 1 ; $j <= $N ; $j++ )
    {
        if ( $i == $j )
        {
            $$sums[ $i ][ $j ] = 1;
        }
    }
}

```

```

    }
    else
    {
        $$sums[ $i ][ $j ] = 0;
    }
}
}
###
#
# Fill summation matrix
#
###
for ( $k = 1 ; $k < $N ; $k++ )
{
    $i = 1;
    for ( $j = $k+1 ; $j <= $N ; $j++ )
    {
        $$sums[ $i ][ $j ] = $$sums[ $i ][ $j-1 ] + $$sums[ $i+1 ][ $j ] + 2*$$data[ $i ][ $j ] - $$sums[ $i+1 ][
$j-1 ];
        $i++;
    }
}

###
#
# Initialize histogram values
#
###
for ( $i = 0 ; $i <= 100 ; $i++ )
{
    $hist[ $i ] = 0;
}

###
#
# Second pass through the summation matrix to calculate percentages
#
###
for ( $k = 1 ; $k < $N ; $k++ )
{
    $i = 1;
    for ( $j = $k+1 ; $j <= $N ; $j++ )
    {
        $percent = int ( 100 * ( $$sums[ $i ][ $j ] / (( $k+1 )*( $k+1 )) ) + 0.50 );
        $hist[ $percent ]++;
        $i++;
    }
}

$nof_data = 0;
for ( $i=0 ; $i <= 100 ; $i++ )
{
    $nof_data += $hist[ $i ];
}

###
#
# The rest ...
#
###

$N = @hist;
$max = $hist[0];

```

```

$mode = 0;
for ( $k = 0 ; $k < $N ; $k++ )
{
  if ( $hist[ $k ] > $max )
  {
    $max = $hist[ $k ];
    $mode = $k;
  }
}

$middle = int ( $nof_data / 2 );
$median=0;
$value = 0;

for ( $k = 0 ; $k < $N ; $k++ )
{
  $value += $hist[ $k ] ;

  if ( $value < $middle )
  {
    $median = $k + 1;
  }
}

return ( $median * $mode );
}

```

## 21) find\_min.pl

```

#!/usr/bin/perl -w

while ( $data = <STDIN> )
{
  @raw = split(' ', $data);
  @sort = sort { $a <=> $b } @raw;
  print "$sort[0]\n";
}

```

## 22) Score\_aver-rms.pl

```

#!/usr/bin/perl -w

my %AA_names = (
  'A' => 0,
  'C' => 1,
  'D' => 2,
  'E' => 3,
  'F' => 4,
  'G' => 5,
  'H' => 6,
  'I' => 7,
  'K' => 8,
  'L' => 9,
  'M' => 10,
  'N' => 11,
  'P' => 12,
  'Q' => 13,
  'R' => 14,

```

```

        'S' => '15',
        'T' => '16',
        'V' => '17',
        'W' => '18',
        'Y' => '19',
    );

my $scores = [];
$index = 0;
for ( $j=0 ; $j < 6000 ; $j++)
{
    for ( $i=0 ; $i < 20 ; $i++)
    {
        $$scores[$i][$j] = -1.0;
    }
}

while ( $data = <STDIN> )
{
    if ( $data =~ /([A-Z])([A-Z])([A-Z])([A-Z])\s*([0-9]+\.[0-9]+)\s*/ )
    {
        $$scores[$AA_names{$1}][$index] = $5;
        $index++;
    }
}

for ( $i=0 ; $i < 20 ; $i++)
{
    $saver = 0.0;
    $nof_points = 0.0;
    for ( $j=0 ; $j < $index ; $j++)
    {
        if ( $$scores[$i][$j] >= 0.0 )
        {
            $saver += $$scores[$i][$j];
            $nof_points++;
        }
    }

    $saver /= $nof_points;

    $rmsd = 0.0;
    for ( $j=0 ; $j < $index ; $j++)
    {
        if ( $$scores[$i][$j] >= 0.0 )
        {
            $rmsd += ( $$scores[$i][$j] - $saver ) * ( $$scores[$i][$j] - $saver );
        }
    }

    $rmsd /= ( $nof_points - 1.0 );
    $rmsd = sqrt( $rmsd );

    print "$saver $rmsd\n";
}

#print "\n\n";
exit();

```

### 23) Score\_aver-rms.perpos.pl

```
#!/usr/bin/perl -w
```

```

my %AA_names = (
    'A' => 0,
    'C' => 1,
    'D' => 2,
    'E' => 3,
    'F' => 4,
    'G' => 5,
    'H' => 6,
    'I' => 7,
    'K' => 8,
    'L' => 9,
    'M' => 10,
    'N' => 11,
    'P' => 12,
    'Q' => 13,
    'R' => 14,
    'S' => 15,
    'T' => 16,
    'V' => 17,
    'W' => 18,
    'Y' => 19,
);

my $scores = [];
$index = 0;

for ( $j=0 ; $j < 6000 ; $j++ )
{
    for ( $i=0 ; $i < 20 ; $i++ )
    {
        for ( $k=0 ; $k < 4 ; $k++ )
        {
            $$scores[$i][$j][$k] = -1.0;
        }
    }
}

while ( $data = <STDIN> )
{
    if ( $data =~ /([A-Z])([A-Z])([A-Z])([A-Z])\s*([0-9]+\.[0-9]+)\s*/ )
    {
        $$scores[$AA_names{$1}][$index][0] = $5;
        $$scores[$AA_names{$2}][$index][1] = $5;
        $$scores[$AA_names{$3}][$index][2] = $5;
        $$scores[$AA_names{$4}][$index][3] = $5;
        $index++;
    }
}

for ( $k=0 ; $k < 4 ; $k++ )
{
    for ( $i=0 ; $i < 20 ; $i++ )
    {
        $saver = 0.0;
        $nof_points = 0.0;
        for ( $j=0 ; $j < $index ; $j++ )
        {
            if ( $$scores[$i][$j][$k] >= 0.0 )
            {
                $saver += $$scores[$i][$j][$k];
                $nof_points++;
            }
        }

        $saver /= $nof_points;

        $rmsd = 0.0;
        for ( $j=0 ; $j < $index ; $j++ )
        {
            if ( $$scores[$i][$j][$k] >= 0.0 )
            {
                $rmsd += ( $$scores[$i][$j][$k] - $saver ) * ( $$scores[$i][$j][$k] - $saver );
            }
        }
    }
}

```



```

    $rmsd /= ($nof_points - 1.0 );
    $rmsd = sqrt( $rmsd );
    print "$saver $rmsd\n";
}
print "\n\n";
}
exit();

```

## 24) Wfixed.pl

```

#!/usr/bin/perl -w

my %AA_names = (
    'A' => 0,
    'C' => 1,
    'D' => 2,
    'E' => 3,
    'F' => 4,
    'G' => 5,
    'H' => 6,
    'I' => 7,
    'K' => 8,
    'L' => 9,
    'M' => 10,
    'N' => 11,
    'P' => 12,
    'Q' => 13,
    'R' => 14,
    'S' => 15,
    'T' => 16,
    'V' => 17,
    'Y' => 18,
);

my $scores = [];
my $index = 0;

for ( $j=0 ; $j < 6000 ; $j++)
{
    for ( $i=0 ; $i < 19 ; $i++)
    {
        for ( $k=0 ; $k < 3 ; $k++)
        {
            $scores[$i][$j][$k] = -1.0;
        }
    }
}

while ( $data = <STDIN> )
{
    if ( $data =~ /([A-Z])([A-Z])([A-Z])\s*([0-9]+\.[0-9]+)\s*/ )
    {
        $scores[$AA_names{$1}][$index][0] = $4;
        $scores[$AA_names{$2}][$index][1] = $4;
        $scores[$AA_names{$3}][$index][2] = $4;
        $index++;
    }
}

for ( $k=0 ; $k < 3 ; $k++)
{
    for ( $i=0 ; $i < 19 ; $i++)
    {
        $saver = 0.0;
        $nof_points = 0.0;
        for ( $j=0 ; $j < $index ; $j++)
        {
            if ( $scores[$i][$j][$k] >= 0.0 )
            {
                $saver += $scores[$i][$j][$k];
                $nof_points++;
            }
        }
    }
}

```

```
    }  
  }  
  
  $aver /= $nof_points;  
  
  $rmsd = 0.0;  
  for ( $j=0 ; $j < $index ; $j++)  
  {  
    if ( $$scores[$i][$j][$k] >= 0.0 )  
    {  
      $rmsd += ( $$scores[$i][$j][$k] - $aver ) * ( $$scores[$i][$j][$k] - $aver );  
    }  
  }  
  
  $rmsd /= ( $nof_points - 1.0 );  
  $rmsd = sqrt( $rmsd );  
  print "$aver $rmsd\n";  
}  
print "\n\n";  
}  
  
exit();
```



Imagination is more important  
than knowledge.

For while knowledge defines  
all we currently know and understand,  
imagination points to all we might  
yet discover and create.

Albert Einstein 1879-1955