Δημοκρίτειο Πανεπιστήμιο Θράκης

Σχολή Επιστημών Υγείας

Τμήμα Μοριακής Βιολογίας και Γενετικής

Διπλωματική Εργασία

<<Αναγνώριση της μεταβατικής κατάστασης του CLN025 μέσω του διαγράμματος T-Q>>

Χατζησάββας Αλέξανδρος (ΑΕΜ 1412)

Επιβλέπων Καθηγητής: Νικόλαος Μ. Γλυκός

Αλεξανδρούπολη Ιανουάριος 2018

Ευχαριστίες

Πρώτο από όλους θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Νικόλαο Μ. Γλυκό για την υπομονή του και την εμπιστοσύνη που μου έδειχνε καθ' όλη την διάρκεια εκπόνησης της διπλωματικής μου. Μου μετέδωσε τις γνώσεις και τις εμπειρίες του, αλλά κυρίως μου έμαθε να στέκομαι στα πόδια μου σαν επιστήμονας.

Έπειτα, θα ήθελα να ευχαριστήσω την οικογένεια μου για την στήριξη και το κουράγιο που μου προσέφεραν όλα αυτά τα χρόνια και κυρίως στις στιγμές που το είχα πιο πολύ ανάγκη.

Επίσης, θα ήθελα να ευχαριστήσω τους φίλους μου, που με συνόδευσαν σε αυτό το ταξίδι, αλλά και για την βοήθεια που μου έδωσε ο καθένας με τον δικό του τρόπο.

Τελευταίους αλλά εξίσου σημαντικούς θα ήθελα να ευχαριστήσω τους NMG groupies για όλη την βοήθεια που μου παρείχαν, αλλά και για τις όμορφες εμπειρίες που μου χάρισαν όλο το διάστημα που ήμουν και εγώ ενεργό μέλος.

«Σα βγεις στον πηγαιμό για την Ιθάκη,
να εύχεσαι νάναι μακρύς ο δρόμος,
γεμάτος περιπέτειες, γεμάτος γνώσεις.»

-Κ. Π. Καβάφης, Ιθάκη

Περιεχόμενα

Περίληψη

Η εύρεση των μεταβατικών καταστάσεων των μικρών πρωτεϊνών μας βοηθάει στην κατανόηση του τρόπου αναδίπλωσης των μεγάλων πρωτεϊνών. Στην παρούσα εργασία χρησιμοποιήθηκε η προσομοίωση της μίνι πρωτεΐνης CLN025, ενός τεχνητού μορίου 10 αμινοξέων με δομή β-φουρκέτας στον χώρο, στο δυναμικό πεδίο AMBER99SB-STAR-ILDN. Με βάση το διάγραμμα T-Q που δημιουργήθηκε από την προσομοίωση πραγματοποιήθηκε ανάλυση των δεδομένων, που έδειξε ότι στο πλήθος των μεταβατικών καταστάσεων του CLN025 κυριαρχούν δομές στις οποίες έχει σχηματιστεί ο υδροφοβικός πυρήνας (Hydrophobic collapse) και η στροφή.

Abstract

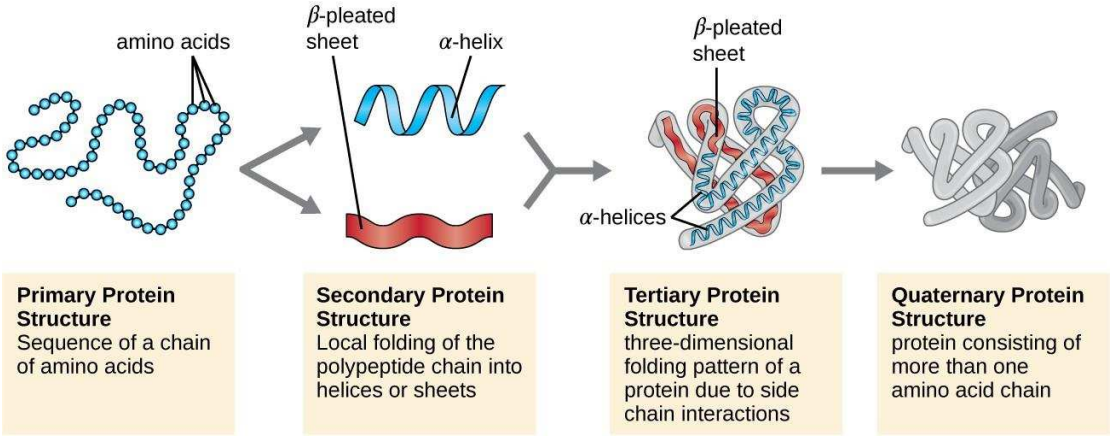
Analysing the transition states of mini-proteins helps us to understand the mechanisms undergoing the protein folding of complex proteins. In the present thesis, we used a molecular dynamic simulation of the mini-protein CLN025, a designed molecule consisting of 10 amino acids with a beta hairpin structure, in the AMBER99SB-STAR-ILDN force field. The analysis of the T-Q diagram showed that the transition state ensemble is consisting of structures which have undergone hydrophobic collapse and turn formation.

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

1.1 Πρωτεΐνες

Οι πρωτεΐνες είναι τα πιο άφθονα βιολογικά μακρομόρια, διότι υπάρχουν σε όλα τα κύτταρα και σε όλα τα μέρη των κυττάρων. Εμφανίζουν τεράστια ποικιλότητα βιολογικών λειτουργιών. Οι πρωτεΐνες είναι γραμμικά πολυμερή αμινοξέων, που συνδέονται μεταξύ τους με πεπτιδικούς δεσμούς. Συνήθως ορίζονται τέσσερα επίπεδα πρωτεϊνικής δομής: πρωτοταγής δομή, δευτεροταγής δομή, τριτοταγής δομή και τεταρτοταγής δομή. Η πρωτοταγής δομή είναι η αλληλουχία των αμινοξέων μιας πρωτεΐνης, που συνδέονται μεταξύ τους με ομοιοπολικούς δεσμούς (πεπτιδικούς και δισουλφιδικούς). Η δευτεροταγής δομή αναφέρεται σε ευδιάκριτα δομικά πρότυπα αμινοξέων (πχ. β-πτυχώσεις, α-έλικες κ.α.) που σταθεροποιούνται με υδρογονοδεσμούς μεταξύ των πεπτιδικών ομάδων N-H και C=O. Η τριτοταγής δομή περιγράφει όλες τις παραμέτρους της τρισδιάστατης πτύχωσης ενός πολυπεπτιδίου. Η τεταρτοταγής δομή ορίζεται στις πρωτεΐνες που αποτελούνται από δύο ή περισσότερες πολυπεπτιδικές αλυσίδες και περιγράφει την διάταξη τους στον χώρο [1][2].



Εικόνα 1: (Αναπαράγεται άνευ άδειας) Παρουσιάζεται η πρωτοταγής, δευτεροταγής, τριτοταγής και τεταρτοταγής δομή μιας πρωτεΐνης.

https://courses.lumenlearning.com/microbiology/chapter/proteins/)

Ενώ μερικές πρωτεΐνες αναδιπλώνονται αυθόρμητα στην φυσική τους κατάσταση, άλλες απαιτούν τη βοήθεια ενζύμων, ενώ άλλες χρειάζονται την βοήθεια ειδικών πρωτεϊνών που ονομάζονται σαπερόνες.

1.2 Πρωτεϊνική αναδίπλωση

Η λειτουργία της πρωτεΐνης εξαρτάται από την τρισδιάστατη δομή της, η οποία με την σειρά της προσδιορίζεται από την αλληλουχία των αμινοξέων της [3]. Οι πρωτεΐνες σε οποιαδήποτε από τις λειτουργικές διαμορφώσεις τους ονομάζονται φυσικές πρωτεΐνες (native proteins)[1]. Για την κατανόηση της λειτουργίας των βιολογικών μακρομορίων (πχ. πρωτεϊνών) απαιτείται η γνώση της δομής τους [4].

Αλλά υπάρχει ένα πρόβλημα, πως καταλήγει μια πρωτεΐνη από την ξεδιπλωμένη (unfolded) μορφή της στην αναδιπλωμένη, φυσική κατάσταση της (native state); Υπάρχουν δύο πιθανοί τρόποι, είτε αναδιπλώνεται σε όλες τις πιθανές δομές της, είτε αναδιπλώνεται μέσω ενδιάμεσων δομών στην φυσική της δομή. Σε αυτό το ερώτημα έδωσε την απάντηση ο C. Levinthal με το λεγόμενο παράδοξο του Levinthal. Υπολόγισε πως για μια μικρή πρωτεΐνη θα χρειαζόταν υπερβολικά πολύς χρόνος, σε σχέση με τον πραγματικό χρόνο αναδίπλωσης της, ώστε να αναδιπλωθεί στην φυσική της δομή δοκιμάζοντας όλες τις πιθανές δομές. Συνεπώς, οι πρωτεΐνες αναδιπλώνονται μέσω ενδιάμεσων δομών στην φυσική τους δομή [5].

Οι δευτεροταγείς, τριτοταγείς και τεταρτοταγείς δομές σταθεροποιούνται μέσω ασθενών μη ομοιοπολικών αλληλεπιδράσεων (πχ. αλληλεπιδράσεις van der Walls, υδροφοβικές αλληλεπιδράσεις). Οι μη ομοιοπολικές αλληλεπιδράσεις είναι ασθενείς σε σχέση με τους ομοιοπολικούς δεσμούς, αλλά ο συνδυασμός πολλών ασθενών αλληλεπιδράσεων αρκεί για να καθορίσει το πρότυπο αναδίπλωσης μιας πρωτεΐνης [6].

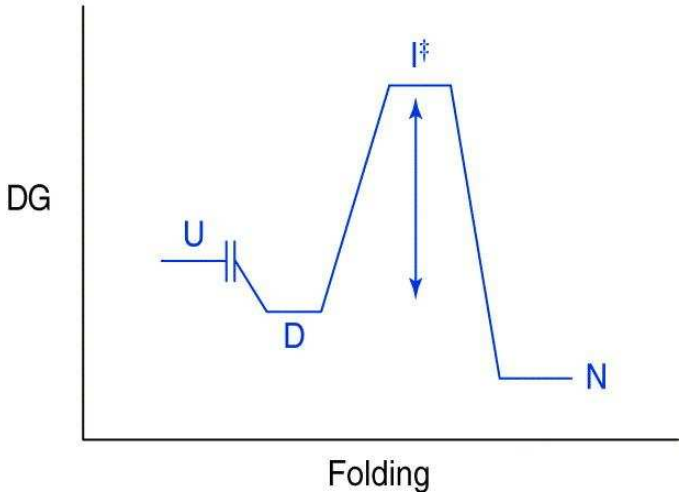
Υπάρχουν δύο μηχανισμοί που εξηγούν την κινητική αναδίπλωσης των πρωτεϊνών, κινητική των δύο σταδίων αναδίπλωσης (two-state folding) και κινητική των πολλών σταδίων αναδίπλωσης (multi-state folding). Η αναδίπλωση δύο σταδίων είναι μια γρήγορη και τελείως αντιστρεπτή διεργασία, κατά την οποία δεν παρατηρούνται ενδιάμεσα και σχηματισμός δισουλφιδικών δεσμών. Η αναδίπλωση πολλών σταδίων είναι μια σύνθετη διεργασία, κατά την οποία λαμβάνουν χώρα πολλές διαμορφώσεις της πρωτεΐνης και παρατηρείται τουλάχιστον ένα ή περισσότερα ενδιάμεσα [13][14].

1.3 Μεταβατική κατάσταση

Η πρωτεϊνική αναδίπλωση μπορεί να χαρακτηριστεί από τρείς όρους δανειζόμενους από τις χημικές αντιδράσεις: πορεία αντίδρασης, ενδιάμεσα και μεταβατική κατάσταση (transition state). Συνήθως, οι όροι αντίδραση αναδίπλωσης και μεταβατική κατάσταση συγχέονται. Ενώ η μεταβατική κατάσταση υποδηλώνει μια δομή, η αντίδραση αναδίπλωσης μπορεί να έχει ένα ευρύ πλήθος μεταβατικών καταστάσεων (transition state ensemble, TSE)[7].

Πρώτον, υπάρχει διαφορά στην μεταβατική κατάσταση μιας χημικής αντίδρασης και της πρωτεϊνικής αναδίπλωσης. Κατά την μεταβατική κατάσταση των χημικών αντιδράσεων σχηματίζονται και διασπώνται ελάχιστοι, μόνο, δεσμοί και επειδή είναι ισχυροί, χρειάζεται η κβαντική θεωρεία για να εξηγηθεί. Από την άλλη, οι μεταβατικές καταστάσεις, κατά την πρωτεϊνική αναδίπλωση, περιλαμβάνουν τον σχηματισμό και την διάσπαση πολλών ασθενών δεσμών, οι οποίοι μπορούν να εξηγηθούν με την χρήση κλασσικής μηχανικής στατιστικής και των εξισώσεων του Newton.

Δεύτερον, πρέπει να γίνει ξεκάθαρη η διαφορά της ενδιάμεσης δομής και της μεταβατικής κατάστασης. Μια ενδιάμεση δομή, κατά την αναδίπλωση, βρίσκεται σε ένα ελάχιστο σε μια πιθανή ενεργειακή επιφάνεια. Αντιθέτως, μια μεταβατική κατάσταση είναι μια δομή η οποία βρίσκεται σε ένα μέγιστο ενέργειας, κατά την αναδίπλωση. Σε μια ενεργειακή επιφάνεια, η μεταβατική κατάσταση βρίσκεται στο μέγιστο της ελεύθερης ενέργειας κατά την πορεία της αντίδρασης [8].



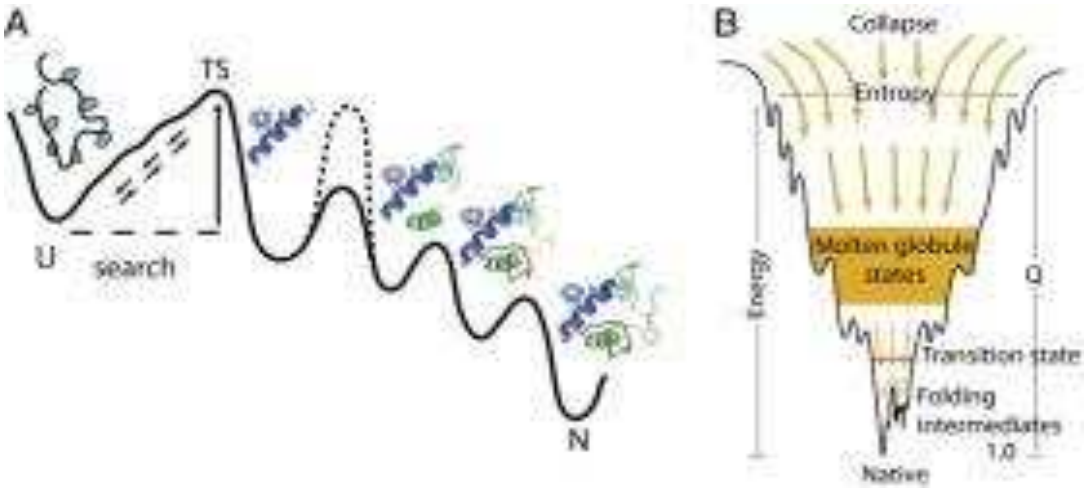
Εικόνα 2: (Αναπαράγεται άνευ άδειας)[7] Απεικόνιση διαγράμματος ελεύθερης ενέργειας ως προς την διαδικασία αναδίπλωσης. Το U είναι για τις ξεδιπλωμένες διαμορφώσεις της πρωτεΐνης (unfolded), το D είναι για τις μετουσιωμένες διαμορφώσεις της πρωτεΐνης (denatured), το $I^{\ddagger}$ είναι για το πλήθος μεταβατικών καταστάσεων (TSE) και το N είναι για τις φυσικές διαμορφώσεις της πρωτεΐνης (native). (http://www.sciencedirect.com/science/article/pii/S0968000498013450?_rdoc=1&_fmt=high&_origin=gateway&_docanchor=&md5=b8429449ccfc9c30159a5f9aeaa92ffb#FIGGR4)

1.4 Μίνι πρωτεΐνες και ενεργειακά τοπία

Τα πολυπεπτίδια με 50 ή λιγότερα αμινοξέα μπορούν να χαρακτηρισθούν και ως μίνι πρωτεΐνες (mini proteins), διότι έχουν ιδιότητες που σχετίζονται μόνο με τις πρωτεΐνες και διαθέτουν καλά καθορισμένες τριτοταγείς δομές. Λόγω του μικρού τους μεγέθους και της απλότητας των δομών τους, σε σχέση με την πολυπλοκότητα των μεγάλων πρωτεϊνών, χρίζουν ιδανικά μοντέλα για την χρήση τους σε μεγάλης

διάρκειας προσομοιώσεις μοριακής δυναμικής και για την μελέτη των μοριακών μηχανισμών της πρωτεϊνικής αναδίπλωσης [9][10].

Η θεωρεία του τοπίου ενέργειας υποστηρίζει ότι η αναδίπλωση της πρωτεΐνης δεν συμβαίνει μέσω ενός συγκεκριμένου μονοπατιού, όπως προτείνει η "κλασσική άποψη" ("classical view"), αλλά να διέρχεται μέσω πολλών μονοπατιών, όπως προτείνει η "νέα άποψη" ("new view"), που περνάνε από πλήθος μεταβατικών καταστάσεων (TSE), σχηματίζοντας ένα ενεργειακό τοπίο σε σχήμα χωνιού [11][15][16]. Η γενική εικόνα του ενεργειακού τοπίου μας βοηθάει να κατανοήσουμε την κινητική των δύο σταδίων αναδίπλωσης και των πολλών σταδίων αναδίπλωσης [12]. Ο κάθετος άξονας του ενεργειακού τοπίου δείχνει την εσωτερική ελεύθερη ενέργεια της εκάστοτε πρωτεΐνης, ενώ οι πλευρικοί άξονες αντιπροσωπεύουν τις συντεταγμένες των διαμορφώσεων [15].



Εικόνα 3: (Αναπαράγεται άνευ άδειας) Α) Αναπαράσταση της κλασσικής άποψης της πρωτεϊνικής αναδίπλωσης και Β) η αναπαράσταση της νέας άποψης των πολλαπλών μονοπατιών μέσω του funneled landscape. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4234557/figure/fig01/)

Η θεωρεία των ενεργειακών τοπίων είναι χρήσιμη ποσοτικά για τις πρωτεΐνες που αναδιπλώνονται γρήγορα, επειδή τα γεγονότα γρήγορης αναδίπλωσης είναι λιγότερο ευαίσθητα σε μεμονωμένες ατομικές

λεπτομέρειες. Η κινητική των γρήγορων αναδιπλώσεων μπορεί να γίνει πιο κατανοητή με την συσχέτιση μεταξύ μιας πρωτεΐνης στο εργαστήριο και ενός πιο απλού υπολογιστικού μοντέλου, το οποίο έχει παρόμοια τοπογραφία χωνιού, σε στατιστικό επίπεδο, με την πρωτεΐνη [11].

1.5 Προσεγγίσεις για την αναδίπλωση πρωτεϊνών

Υπάρχουν αρκετές θεωρητικές προσεγγίσεις και τεχνικές για την μελέτη της δομής και της αναδίπλωσης των πρωτεϊνών, πειραματικές και υπολογιστικές. Ενώ οι πειραματικές τεχνικές θεωρούνται πιο αξιόπιστες, οι υπολογιστικές τεχνικές έρχονται να συμπληρώσουν σε πληροφορία τις πρώτες. Για την εύρεση δευτεροταγούς και τριτοταγούς δομής των πρωτεϊνών χρησιμοποιούνται κυρίως η κρυσταλλογραφία ακτίνων Χ και ο πυρηνικός μαγνητικός συντονισμός (NMR). Η κρυσταλλογραφία βασίζεται στην περίθλαση μιας δέσμης ακτίνων Χ από έναν καλά οργανωμένο κρύσταλλο πολλών μορίων πρωτεΐνης, και λόγω αυτής της περίθλασης παράγεται ένα φάσμα, το διάγραμμα περίθλασης (diffraction pattern), που μέσω της ανάλυσης του προσδιορίζεται η δομή της εκάστοτε πρωτεΐνης. Ο πυρηνικός μαγνητικός συντονισμός χρησιμοποιεί ως μέσο την ιδιοπεριστροφή των πυρήνων των ατόμων μιας πρωτεΐνης, κυρίως του ^{1}H, για την εξαγωγή πληροφορίας που αφορά τις αποστάσεις των ατόμων στην πρωτεΐνη, οι οποίες με την σειρά τους μπορούν να χρησιμοποιηθούν για εξαχθεί, υπολογιστικά, το τρισδιάστατο μοντέλο της πρωτεΐνης. Μια σημαντική διαφορά της κρυσταλλογραφίας ακτίνων Χ και του NMR είναι ότι στην τελευταία τεχνική δεν απαιτείται η χρήση πρωτεϊνικών κρυστάλλων [17]. Επίσης, άλλες πειραματικές τεχνικές για την μελέτη της δομής των πρωτεϊνών αποτελούν: ο κυκλικός διχρωισμός (CD), η ηλεκτρονική μικροσκοπία, η πρωτεϊνική μηχανική και η φασματομετρία μάζας. Οι υπολογιστικές τεχνικές έρχονται να συνδέσουν

τις θεωρητικές προσεγγίσεις με τις πειραματικές τεχνικές, μέσω της αναγνώρισης των μηχανισμών αναδίπλωσης των πρωτεϊνών και της καλύτερης κατανόησης της θερμοδυναμικής και της κινητικής των συστημάτων. Οι προσομοιώσεις μοριακής δυναμικής (molecular dynamics, MD) είναι η κύρια υπολογιστική τεχνική που χρησιμοποιείται για την εξέταση των πρωτεϊνών.

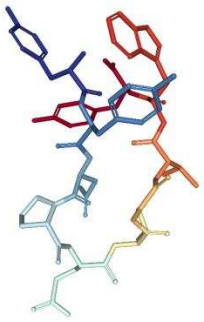
1.6 CLN025

Το CLN025 είναι ένα συνθετικό μόριο 10 αμινοξέων, YYDPETGTWY, το οποίο κατασκεύασαν οι Honda et al. Σχεδιάστηκε με βάση ένα άλλο συνθετικό μόριο 10 αμινοξέων, την chignolin, το οποίο κατασκεύασε ο ίδιος με μια άλλη ομάδα. Μετά από πολλές παραλλαγές των αμινοξέων Gly1και Gly10 είδαν ότι το CLN025 με τα Tyr1 και Tyr10 είναι από τις πιο σταθερές παραλλαγές και ότι έχει θερμοκρασία τήξης (melting temperature, T_m) τους 343K, 28 βαθμούς περισσότερο της chignolin [18].

Διατηρεί την ίδια διαμόρφωση σε υγρό διάλυμα και στην κρυσταλλική μορφή του. Επίσης, είναι δυνατή η αναστρέψιμη αναδίπλωση του. Προσομοιώσεις μοριακής δυναμικής έδειξαν ότι το μόριο αναδιπλώνεται με τέτοιο τρόπο, ώστε οι δομές να διανέμονται σε funnel-like ενεργειακό τοπίο. Όλα τα παραπάνω το καθιστούν ικανό να θεωρηθεί πρωτεΐνη παρά το μικρό μέγεθος του, συνεπώς μίνι πρωτεΐνη. Το προτείνουν ως "ιδανική πρωτεΐνη" παρότι δεν εντοπίζεται στην φύση.

Οι Honda et al. προσδιόρισαν την κρυσταλλική δομή του σε διακριτικότητα 1.11 Å και έδειξαν ότι υιοθετεί δομή β-φουρκέτας στην κεντρική περιοχή, επιτρέποντας έτσι τα άκρα να έλθουν σε επαφή μεταξύ τους. Εντόπισαν, επίσης, 6 ενδομοριακούς δεσμούς υδρογόνου και μια γέφυρα άλατος, που σταθεροποιούν αυτή τη δομή. Έπειτα, με ανάλυση NMR εντόπισαν 20 δομές του CLN025 των οποίων το root-mean-square-

deviation of backbone coordinates (bb-rmsd) με την κρυσταλλική δομή ήταν 1.75 Å και διατηρούσαν παρόμοια διαμόρφωση σε υδατικό διάλυμα στους 298K.

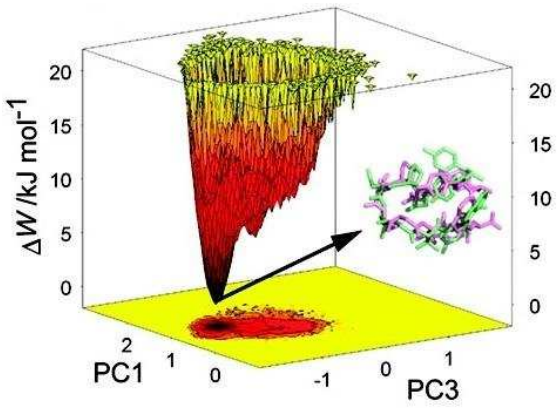


Εικόνα 4: (Αναπαράγεται άνευ άδειας) Η κρυσταλλική δομή του CLN025.

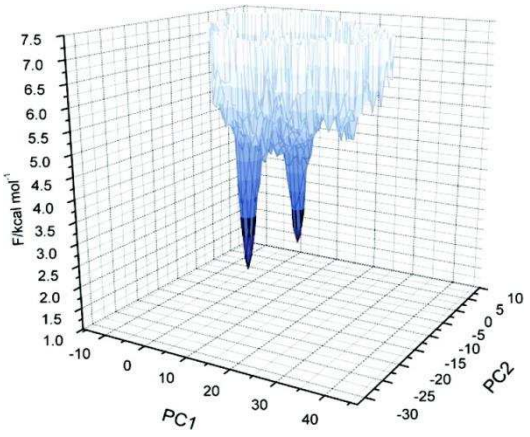
(https://www.rcsb.org/pdb/explore.do?structureId=5AWL)

Έρευνες των Hatfield et al. έδειξαν με VCD, ECD και προσομοιώσεις μοριακής δυναμικής ότι το CLN025 διατηρεί την δομή β-φουρκέτα σε υδατικό διάλυμα σε θερμοκρασίες και πάνω και κάτω από την T_m, σε περιβάλλον όπως TFE, MeOH, DMSO και σε αποδιατακτικά διαλύματα ουρίας και GdmCl [9][19][20]. Έτσι, λόγω του ότι το CLN025 δεν έχει καλά προσδιορισμένη τριτοταγή δομή, είναι ανθεκτικό σε αποδιατακτικούς παράγοντες, δεν έχει αυξημένη θερμοχωρητικότητα κοντά στην T_m και ότι πάνω από αυτή την θερμοκρασία διατηρεί διαμόρφωση β-φουρκέτας, τους οδήγησε να αμφισβητήσουν την θεώρηση του CLN025 ως μίνι πρωτεΐνη [20].

Ενώ οι Honda et al. έδειξαν ότι το funnel-like ενεργειακό τοπίο έχει μόνο ένα ενεργειακό ελάχιστο, μετέπειτα έρευνες έδειξαν ότι υπάρχει και δεύτερο ενεργειακό ελάχιστο [21].



Εικόνα 5: (Αναπαράγεται άνευ άδειας) Επιφάνεια ελεύθερης ενέργειας του CLN025 των Honda et. al. (http://pubs.acs.org/doi/abs/10.1021/ja8030533)



Εικόνα 6: (Αναπαράγεται άνευ άδειας) Επιφάνεια ελεύθερης ενέργειας του CLN025 των Rodriguez et .al. (http://pubs.acs.org/doi/abs/10.1021/jp106475c)

Μελέτες για την κατανόηση του μηχανισμού αναδίπλωσης του CLN025, έδειξαν ότι οι ηλεκτροστατικές αλληλεπιδράσεις μεταξύ των φορτισμένων άκρων του, παίζουν σημαντικό ρόλο στην σταθερότητα της β-φουρκέτας [22]. Επόμενες μελέτες έδειξαν ότι το CLN025 αναδιπλώνεται πολύ γρήγορα, διότι έχει μικρό φράγμα ελεύθερης ενέργειας και ότι ο μηχανισμός αναδίπλωσης του δεν μπορεί να περιγραφθεί από ένα απλό μοντέλο δύο σταδίων αναδίπλωσης, αλλά είναι μια ετερογενής διεργασία [23]. Εισαγωγή του CLN025 σε κάθε βρόγχο της Formin binding protein 28 (FBP28), έδειξε ότι ο ρυθμός αναδίπλωσης του CLN025 παρέμεινε ο ίδιος μέσα στη μεγαλύτερη πρωτεΐνη. Αυτό υποδηλώνει ότι οι υποπεριοχές πρωτεϊνών που αναδιπλώνονται γρήγορα μπορούν να χρησιμοποιηθούν για να αυξήσουν την ταχύτητα αναδίπλωσης πιο σύνθετων πρωτεϊνών, και ότι η δυναμική αναδίπλωσης αυτών των υποπεριοχών παραμένει σταθερή μέσα σε μεγαλύτερες πρωτεΐνες [24]. Η αμινοξική αλληλουχία του CLN025 εκτός από το ότι κάνει εφικτό το στενό πακετάρισμα του σκελετού (backbone) και των αρωματικών ομάδων του, διασφαλίζει και τον σχηματισμό των

ενδομοριακών υδρογονικών δεσμών κατά το “θάψιμο” ενός δότη και ενός δέκτη όταν ο σκελετός σχηματίζει την φυσική δομή [25].

Αρκετές θεωρητικές και πειραματικές μελέτες έχουν παρουσιάσει αντιφατικούς μηχανισμούς σχηματισμού β-φουρκέτας, και θα αναφερθούν μερικοί. Ένας από αυτούς προτείνει ότι στην β-φουρκέτα σχηματίζεται πρώτα η στροφή και μετά σχηματίζεται το β-φύλλο σαν “φερμουάρ” (“zips”) που κλείνει από την στροφή προς τα άκρα με τον σχηματισμό υδρογονοδεσμών μεταξύ των δύο αλυσίδων [26]. Ένας άλλος μηχανισμός προτείνει ότι πρώτα συμβαίνει hydrophobic collapse της πρωτεΐνης και μετά σχηματίζεται η στροφή. Όμως, υπάρχουν δύο αντιμαχόμενες θεωρίες για αυτόν τον μηχανισμό, η μία υποστηρίζει ότι ο σχηματισμός υδρογονοδεσμών συμβαίνει με το μοτίβο “φερμουάρ” από την στροφή προς τα άκρα [27], ενώ η άλλη θεωρεία υποστηρίζει ότι ο σχηματισμός των υδρογονοδεσμών ξεκινάει κοντά στο κέντρο των β-φύλλων και εξαπλώνεται και προς τις δύο μεριές [28]. Στο CLN025 πρώτα συμβαίνει hydrophobic collapse και έπειτα σχηματίζεται η στροφή [23] και οι υδρογονοδεσμοί σχηματίζονται με το μοντέλο “φερμουάρ” [29].

1.7 Στόχος της εργασίας

Στην παρούσα εργασία χρησιμοποιήθηκαν τα αποτελέσματα από την προσομοίωση μοριακής δυναμικής της μίνι πρωτεΐνης CLN025 από την εργασία των Serafeim et al. [30] σε δυναμικό πεδίο AMBER99SB-STAR-ILDN. Ο στόχος είναι η απομόνωση του πλήθους των μεταβατικών καταστάσεων του CLN025 μέσω του διαγράμματος θερμοκρασία-συντελεστή ομοιότητας με την φυσική δομή (T-Q) και η εξέταση τους με σκοπό την εξαγωγή συμπερασμάτων που ίσως

βοηθήσουν στην κατανόηση του μηχανισμού αναδίπλωσης της μίνι πρωτεΐνης.

ΚΕΦΑΛΑΙΟ 2

ΜΕΘΟΔΟΛΟΓΙΑ

2.1 Χαρακτηριστικά υπολογιστικού συστήματος

Το υπολογιστικό σύστημα στο οποίο πραγματοποιήθηκε η εργασία διέθετε λογισμικό Linux Ubuntu 16.04 Lts, τετραπύρηνο επεξεργαστή Intel Core i5-2400 στα 3.10 Ghz, 500 Gb αποθηκευτικό χώρο, φυσική μνήμη (RAM) 4 Gb και κάρτα γραφικών Nvidia Quadro K620 2 Gb.

2.2 Γλώσσες προγραμματισμού

Ολόκληρη η εργασία έγινε σε σύστημα UNIX και για αυτό χρησιμοποιήθηκαν γλώσσες προγραμματισμού που είναι εύκολα προσβάσιμες και αρκετά χρήσιμες για την εργασία, όπως Bash script και AWK, λόγω της αυτοματοποίησης και της εύκολης διαχείρισης των δεδομένων. Έπειτα, χρησιμοποιήθηκε η γλώσσα προγραμματισμού Perl [31] η οποία σχεδιάστηκε από τον Larry Wall και είναι ελεύθερο λογισμικό. Δεν χρειάζεται τη χρήση compiler, παρέχει δυνατότητες για χειρισμό κειμένου και έχει καθιερωθεί στον τομέα της Βιοπληροφορικής μέσω του εργαλείου Bioperl [32]. Επειδή χρειάστηκε να γίνει cluster analysis στα δεδομένα, η πιο χρήσιμη λύση ήταν η R, μια δωρεάν γλώσσα προγραμματισμού για στατιστικές αναλύσεις και γραφήματα. Είναι ιδανική για την διαχείριση μεγάλων δεδομένων και έχει έτοιμες λειτουργίες για στατιστικούς υπολογισμούς [33].

Ο πηγαίος κώδικας όλων των **Script** παρατίθεται στο **Παράρτημα**.

2.3 Ανάλυση κύριων συνιστωσών

Η ανάλυση κύριων συνιστωσών (principal component analysis, PCA) ή αλλιώς μέθοδος essential dynamics ή ανάλυση quasiharmonic, είναι μια από τις κύριες μεθόδους που χρησιμοποιείται για την μείωση των διαστάσεων πολύπλοκων συστημάτων, καθιστώντας, έτσι, εφικτή την

μελέτη τους. Υπάρχουν δύο κύριες εκδοχές που χρησιμοποιούνται στην μελέτη προσομοιώσεων μοριακής δυναμικής: Cartesian principal component analysis (cPCA) και dihedral angle principal component analysis (dPCA). Η cPCA χρησιμοποιεί τις καρτεσιανές συντεταγμένες των ατόμων που ορίζουν τις ατομικές μετατοπίσεις σε κάθε διαμόρφωση της πρωτεΐνης και θεωρεί ότι το κέντρο μάζας της πρωτεΐνης είναι σταθερό, καταλήγοντας έτσι σε όχι και τόσο σωστά αποτελέσματα. Ενώ η dPCA χρησιμοποιεί τις δίεδρες γωνίες, φ και ψ, της κύριας αλυσίδας [34][35]. Για το cluster analysis χρησιμοποιήθηκαν οι πρώτες πέντε συντεταγμένες dPCA των διαμορφώσεων που επιλέχθηκαν από την προσομοίωση μοριακής δυναμικής.

2.4 Επιλογή δεδομένων

Από τα τρία αρχεία που υπήρχαν για το CLN025 τα δύο από αυτά, cln025_adapt_STAR_ILDN_COMPLETE.dcd και cln025_pseudo.psf, περιέχουν όλες τις διαμορφώσεις της προσομοίωσης μοριακής δυναμικής και τις δομικές πληροφορίες, της μίνι πρωτεΐνης, αντίστοιχα. Το τρίτο αρχείο, Frame_Temp_Q_PC12345.dat, περιέχει ,σε στήλες, τον αριθμό (frame), την θερμοκρασία (T), τον συντελεστή ομοιότητας με την φυσική δομή (Q) και τις πρώτες πέντε συντεταγμένες dPCA (PC1, PC2, PC3, PC4, PC5) της κάθε διαμόρφωσης της προσομοίωσης μοριακής δυναμικής. Για λόγους ευκολίας στην συνέχεια αυτοί οι παράμετροι θα αναφέρονται με την ονομασία στις παρενθέσεις.

```
 1 368.5696 0.2243    0.5735481   0.4681100   0.4286123   0.9877259  -0.1362247
 2 367.5843 0.2497    0.8382158   0.1190629   0.1823447   0.7624885  -0.2128815
 3 379.9233 0.2486    0.9803393   0.4297339   0.1878107   0.9045210  -0.3489046
 4 398.9674 0.2241    0.7755528   0.5811364   0.1082077   0.8752595  -0.1003057
 5 421.0954 0.2568    0.9951302   0.6974563   0.2086003   0.5871567  -0.4877343
 6 391.8266 0.2512    0.7521077   0.8874243   0.3856905   0.6936994  -0.6313688
 7 378.3243 0.2413    0.7542454   0.0400746   0.3021234   0.9343871  -0.5528902
 8 378.6783 0.2307    0.7159548   0.6105074   0.4070489   1.0012257  -0.0818404
 9 357.6669 0.2256    0.7469440   0.1773570   0.2427839   0.9085780  -0.3563608
10 347.3311 0.2531    0.6990167   0.1349072   0.3924367   0.7163072  -0.4529548
```

Εικόνα 7: Μορφή του αρχείου Frame_Temp_Q_PC12345.dat.

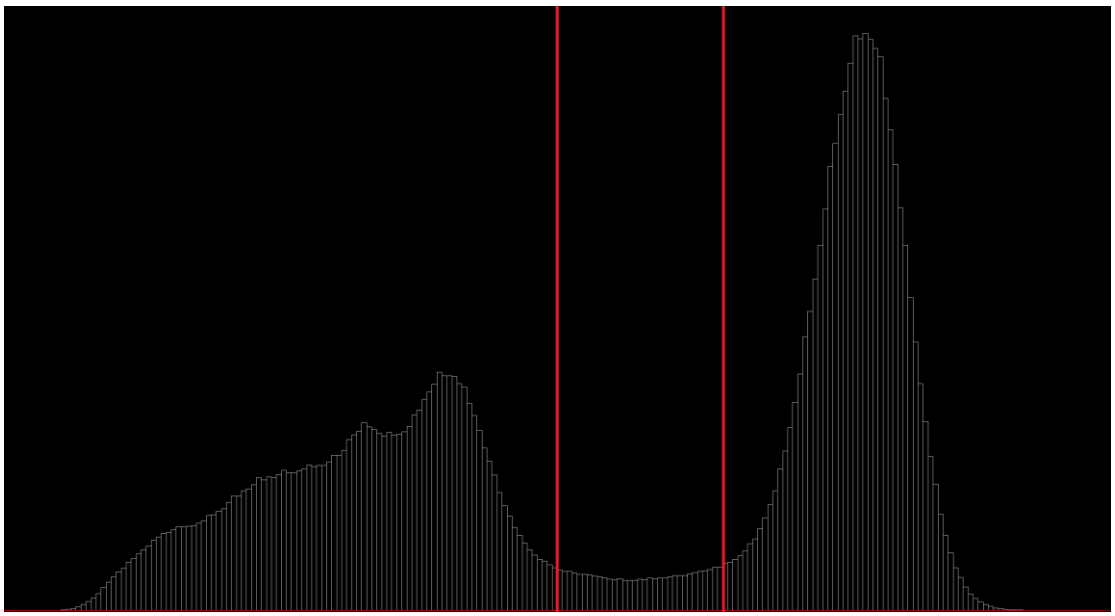
Απομονώνοντας, από το αρχείο Frame_Temp_Q_PC12345.dat, την στήλη των Q με την εντολή στο τερματικό:

~awk '{print $3}' Frame_Temp_Q_PC12345.dat > Q.dat

και έπειτα δημιουργώντας το ιστόγραμμα των τιμών αυτών με το πρόγραμμα plot [36] με την εντολή στο τερματικό:

~plot –h < Q.dat

εμφανίζοντας έτσι το ιστόγραμμα. Από εκεί επιλέχθηκαν οι τιμές για $Q \leq 0.63$ και $Q \geq 0.49$ που οριοθετούν το σύνολο των διαμορφώσεων μεταξύ των offset διαμορφώσεων και των φυσικών διαμορφώσεων (native state).



Εικόνα 8: Παρουσιάζεται το ιστόγραμμα των τιμών Q και με κόκκινες γραμμές τα όρια των τιμών που επιλέχθηκαν.

Έπειτα πραγματοποιήθηκε η απομόνωση όλων των διαμορφώσεων μεταξύ των τιμών Q, που αναφέρθηκαν προηγουμένως, και σε ολόκληρο το εύρος των τιμών T από το αρχείο Frame_Temp_Q_PC12345.dat με τα **Script 1** (Take_away.pl) και **Script 2** (perQ.sh) με τις εντολές στο τερματικό:

~./Take_away.pl Frame_T_Q_PC12345.dat File0.dat 0.49 0.63 ~perQ.sh

Ενώ το **Script 1** έχει δημιουργήσει ένα αρχείο, File0, το οποίο περιέχει όλες τις διαμορφώσεις σε όλο το εύρος των τιμών T και Q, το **Script 2** έχει δημιουργήσει 19 αρχεία, που περιέχουν τις διαμορφώσεις μεταξύ των τιμών Q με βήμα 0.07. Το **Script 2** χρησιμοποιεί δύο άλλα **Script**, τα **Script 1** (Take_away.pl) και **Script 3** (Take_away1.pl). Στη συνέχεια, το αρχείο File0 επεξεργάστηκε με το **Script 4** (pick.sh) με την εντολή στο τερματικό:

~pick.sh

ώστε να δημιουργηθούν 4 αρχεία, που περιέχουν διαμορφώσεις στο ίδιο εύρος Q, αλλά σε διαφορετικό εύρος T. Το **Script 4** χρησιμοποιεί τέσσερα άλλα **Script**, τα **Script 5** (Pick_T1.pl), **Script 6** (Pick_T1.pl), **Script 7** (Pick_T1.pl), **Script 8** (Pick_T1.pl). Επειδή, όμως αρκετά από αυτά τα αρχεία διαθέτουν μεγάλο αριθμό δεδομένων που η φυσική μνήμη του υπολογιστή δεν αρκεί για την ανάλυση, έγινε μείωση του αριθμού των δεδομένων με το **Script 9** (select.pl), που όταν εκτελείται πρέπει να επιλεχθεί το αρχείο που προορίζεται για επεξεργασία, το αρχείο που θα αποθηκευτούν τα νέα δεδομένα και το βήμα με το οποίο θα γίνεται η επιλογή των διαμορφώσεων, και με την εντολή *awk* έγινε η απομόνωση των PC αυτών των αρχείων, για παράδειγμα:

~./select.pl fileX fileY 3

~awk '{print S4,$5,$6,$7$8}' fileX > fileX_PC.dat

Τα αρχεία που τελειώνουν σε "PC.dat" περιέχουν μόνο τα PC1, PC2, PC3, PC4 και PC5 των διαμορφώσεων, ενώ τα αρχεία που ξεκινάνε με

“Frames” περιέχουν μόνο τα frames των διαμορφώσεων και θα χρησιμοποιηθούν σε επόμενες αναλύσεις.

Έτσι έχουν δημιουργηθεί δύο ομάδες αρχείων, μια που περιέχει αρχεία με το ίδιο εύρος Q αλλά με διαφορετικό εύρος T, και την άλλη που περιέχει αρχεία με το ίδιο εύρος T αλλά με διαφορετικό εύρος Q.

2.5 Cluster analysis

Με το cluster analysis, σε μια ομάδα δεδομένων, διαχωρίζονται τα δεδομένα σε μικρότερες ομάδες. Τα δεδομένα που εντοπίζονται στην ίδια ομάδα (cluster) μοιάζουν περισσότερο μεταξύ τους σε σχέση με τα δεδομένα άλλων ομάδων. Για το cluster analysis χρησιμοποιήθηκαν τα αρχεία που τελειώνουν σε “PC.dat”, ώστε να πραγματοποιηθεί ιεραρχικό (hierarchical) clustering με βάση τα PC.

Έπειτα για την εύρεση του ιδανικού αριθμού cluster για το κάθε αρχείο που τελειώνει σε “PC.dat” πραγματοποιήθηκε ανάλυση στην R με 3 διαφορετικές μεθόδους: WSS, fviz_nbclust(kmeans) και fviz_nbclust(pam) [37][38][39]. Για όλα τα αρχεία η ανάλυση έγινε με το **Script10** “cluster.sh” με την εντολή στο τερματικό:

~patch.sh

με το οποίο δημιουργήθηκαν αρχεία “.pdf” με τα διαγράμματα των cluster. Από αυτά τα διαγράμματα, βλέπε ενότητα **3.2**, σαν ιδανική λύση, επιλέχθηκε ο μεγαλύτερος αριθμός cluster για το hierarchical clustering. Η απόσταση, για το hierarchical clustering, μεταξύ των principal component όλων των frame υπολογίστικε με βάση την ευκλείδεια απόσταση (Euclidean distance) που για δύο διανύσματα ορίζεται από τη εξίσωση (1):

$$d(y, x) = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2 + \ldots + (y_n - x_n)^2} = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2} \quad (1)$$

που στην περίπτωσή μας είναι για δύο διανύσματα με πέντε συνιστώσες και άρα μετατρέπεται στην εξίσωση (2):

$$\text{για } x = (x_{PC1}, x_{PC2}, x_{PC3}, x_{PC4}, x_{PC5})$$

$$\text{και}$$

$$\text{για } y = (y_{PC1}, y_{PC2}, y_{PC3}, y_{PC4}, y_{PC5})$$

$$d(y, x) = \sqrt{\sum_{i=1}^{n} \{(y_{iPC1} - x_{iPC1})^2 + (y_{iPC2} - x_{iPC2})^2 + (y_{iPC3} - x_{iPC3})^2 + (y_{iPC4} - x_{iPC4})^2 + (y_{iPC5} - x_{iPC5})^2\}}$$

(2)

Για τον υπολογισμό της απόστασης μεταξύ των cluster χρησιμοποιήθηκε η μέθοδος “complete” κατά την οποία η απόσταση μεταξύ δύο cluster X και Y είναι η μέγιστη απόσταση μεταξύ δύο σημείων x και y, όπου $x \in X$ και $y \in Y$:

$$D(X, Y) = \max_{x \in X, y \in Y} d(x, y) \quad (3)$$

Το hierarchical clustering έγινε με το **Script 11** “plot_dcd.sh” για όλα τα αρχεία, το οποίο κατά την εκτέλεση του, όταν θα έχει ολοκληρώσει το hierarchical clustering όλων των αρχείων, θα ζητήσει το όνομα του αρχείου που θα επεξεργαστεί και τον αριθμό των cluster, για να δημιουργήσει τα νέα αρχεία που θα περιέχουν μόνο τις διαμορφώσεις των ίδιων cluster αντίστοιχα. Ένα παράδειγμα εκτέλεσης του:

~plot_dcd.sh

>fileX

>1 2 3 4 5 6 0 0

Έτσι δημιουργείται ο αντίστοιχος αριθμός αρχείων ".dcd" για κάθε cluster διαμορφώσεων, που θα χρησιμοποιηθούν στην συνέχεια. Το **Script 11** αξιοποιεί και άλλα **Script**, όπως το **Script 12** "converge.pl", το οποίο ενώνει τα δύο αρχεία με τα frames και τα cluster σε ένα αρχείο και το **Script 13** "Extract_clusters.pl", το οποίο δημιουργεί τόσα αρχεία όσα είναι και τα cluster και που περιέχει το καθένα μόνο τις διαμορφώσεις του ίδιου cluster.

2.6 Carma analysis

Η δημιουργία και ανάλυση των αρχείων ".dcd" για την εύρεση των δομών των cluster και των χαρακτηριστικών τους πραγματοποιήθηκε με τα προγράμματα carma και grcarma [40][41]. Πρώτα έγινε least square superposition ("Fitting") του backbone των δομών, μετά πραγματοποιήθηκε ανάλυση δευτεροταγούς δομής ("Stride") και τέλος εύρεση των representative και superpositioned δομών με βάση το backbone ("Covariance, average and representative structures").

2.7 Εύρεση υδρογονοδεσμών

Για την εύρεση των υδρογονοδεσμών που χαρακτηρίζουν τα πιθανά πλήθη των μεταβατικών καταστάσεων (TSE) εξετάστηκαν όλα τα κυρίαρχα cluster με το πρόγραμμα vmd [42]. Για να γίνει αυτό δημιουργήθηκαν ".pdb" αρχεία του κάθε κυρίαρχου cluster με την βοήθεια του carma, με το **Script 14** "pdb.sh", με την εντολή στο τερματικό:

~pdp.sh

ΚΕΦΑΛΑΙΟ 3

ΑΠΟΤΕΛΕΣΜΑΤΑ

3.1 Διάγραμμα T-Q

Η βασική ιδέα που κρύβεται πίσω από την παρούσα εργασία είναι να εξεταστεί το ενδεχόμενο εάν μπορεί να βρεθεί το πλήθος των μεταβατικών καταστάσεων από ένα διάγραμμα T-Q και εάν ναι σε ποιο εύρος T και Q. Έτσι, επιλέχθηκε ένα εύρος T και Q, πιθανό να περιέχει το πλήθος των μεταβατικών καταστάσεων, και από αυτό δημιουργήθηκαν δύο ομάδες αρχείων, μια που περιέχει αρχεία με το ίδιο εύρος Q αλλά με διαφορετικό εύρος T, και την άλλη που περιέχει αρχεία με το ίδιο εύρος T αλλά με διαφορετικό εύρος Q. Με αυτόν τον τρόπο μπορούν να πραγματοποιηθούν αναλύσεις στα χαρακτηριστικά του πλήθους των διαμορφώσεων, όπως ο χαρακτηρισμός της δευτεροταγούς δομής τους και η εύρεση των σχηματισμένων υδρογονοδεσμών, ώστε να ελεγχθεί η σημασία του T και του Q στον μηχανισμό αναδίπλωσης του CLN025.

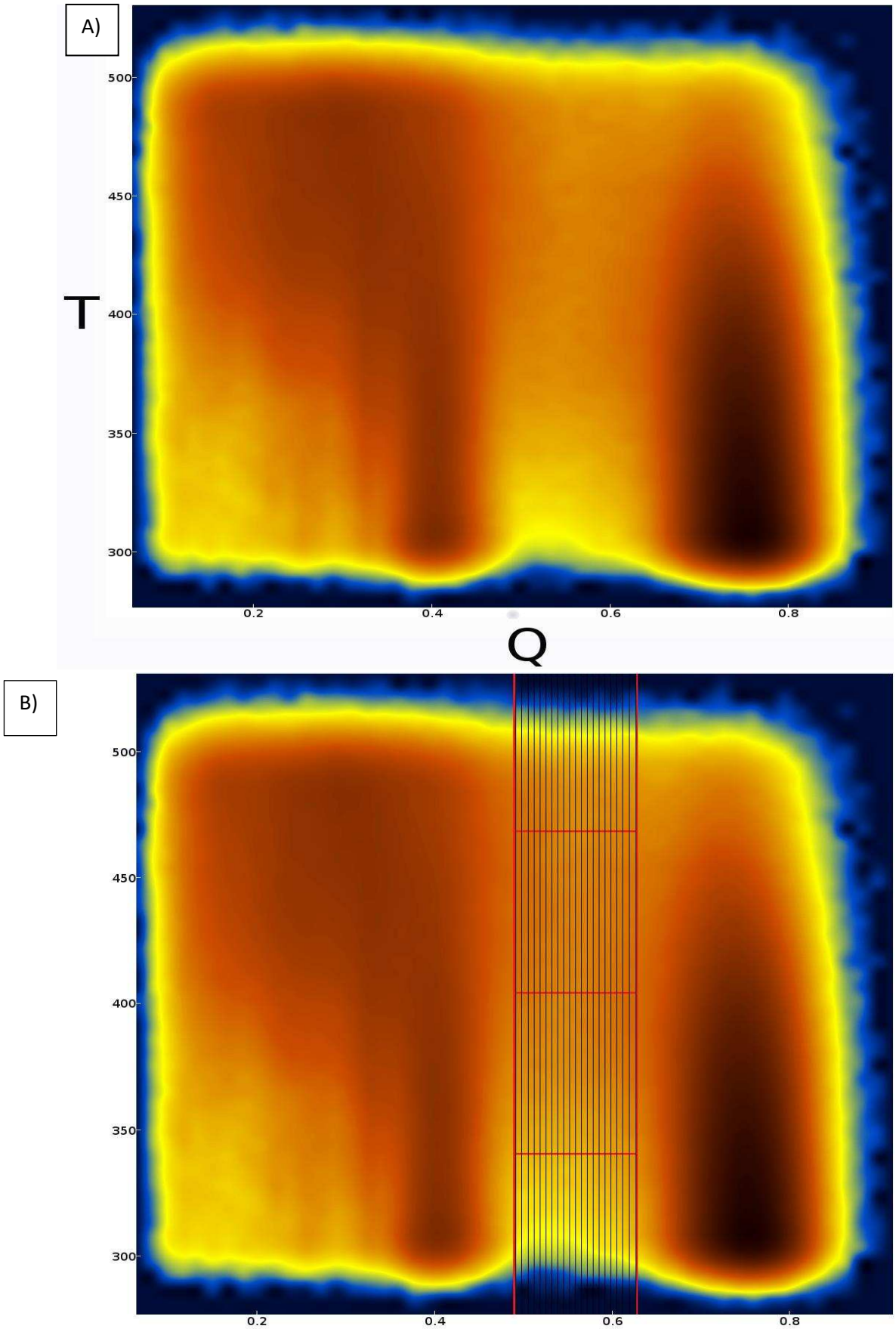
3.2 Ομάδες δεδομένων

Μετά την χρήση των **Script 1**, **Script 2**, **Script 4** και **Script 9** δημιουργήθηκαν 29 ομάδες δεδομένων των οποίων τα χαρακτηριστικά παρατίθενται στον **Πίνακα 1**.

ΠΙΝΑΚΑΣ 1: Παρουσιάζεται το εύρος T, το εύρος Q και το πλήθος των διαμορφώσεων κάθε ομάδας δεδομένων. (Συνεχίζεται και στην επόμενη σελίδα)

Ομάδα δεδομένων	T	Q	Πλήθος διαμορφώσεων
File1	$467.4721<T\leq530.9874$	$0.49\leq Q\leq0.63$	17011
File2	$467.4721<T\leq530.9874$	$0.49\leq Q\leq0.63$	17012
File3	$403.9569<T\leq467.4721$	$0.49\leq Q\leq0.63$	21220
File4	$403.9569<T\leq467.4721$	$0.49\leq Q\leq0.63$	21221

File5	403.9569<T≤467.4721	0.49≤Q≤0.63	21221
File6	340.4417<T≤403.9569	0.49≤Q≤0.63	19735
File7	340.4417<T≤403.9569	0.49≤Q≤0.63	19736
File8	340.4417<T≤403.9569	0.49≤Q≤0.63	19736
File9	276.9265≤T≤340.4417	0.49≤Q≤0.63	22931
File10	276.9265≤T≤530.9874	0.49≤Q<0.497	17178
File11	276.9265≤T≤530.9874	0.497≤Q<0.504	16495
File12	276.9265≤T≤530.9874	0.504≤Q<0.511	15641
File13	276.9265≤T≤530.9874	0.511≤Q<0.518	15343
File14	276.9265≤T≤530.9874	0.518≤Q<0.525	14761
File15	276.9265≤T≤530.9874	0.525≤Q<0.532	14106
File16	276.9265≤T≤530.9874	0.532≤Q<0.539	13452
File17	276.9265≤T≤530.9874	0.539≤Q<0.546	13345
File18	276.9265≤T≤530.9874	0.546≤Q<0.553	12975
File19	276.9265≤T≤530.9874	0.553≤Q<0.560	12951
File20	276.9265≤T≤530.9874	0.560≤Q<0.567	13250
File21	276.9265≤T≤530.9874	0.567≤Q<0.574	13567
File22	276.9265≤T≤530.9874	0.574≤Q<0.581	13793
File23	276.9265≤T≤530.9874	0.581≤Q<0.588	14354
File24	276.9265≤T≤530.9874	0.588≤Q<0.595	14718
File25	276.9265≤T≤530.9874	0.595≤Q<0.602	15198
File26	276.9265≤T≤530.9874	0.602≤Q<0.609	15849
File27	276.9265≤T≤530.9874	0.609≤Q<0.616	16688
File28	276.9265≤T≤530.9874	0.616≤Q<0.623	17545
File29	276.9265≤T≤530.9874	0.623≤Q≤0.630	18429

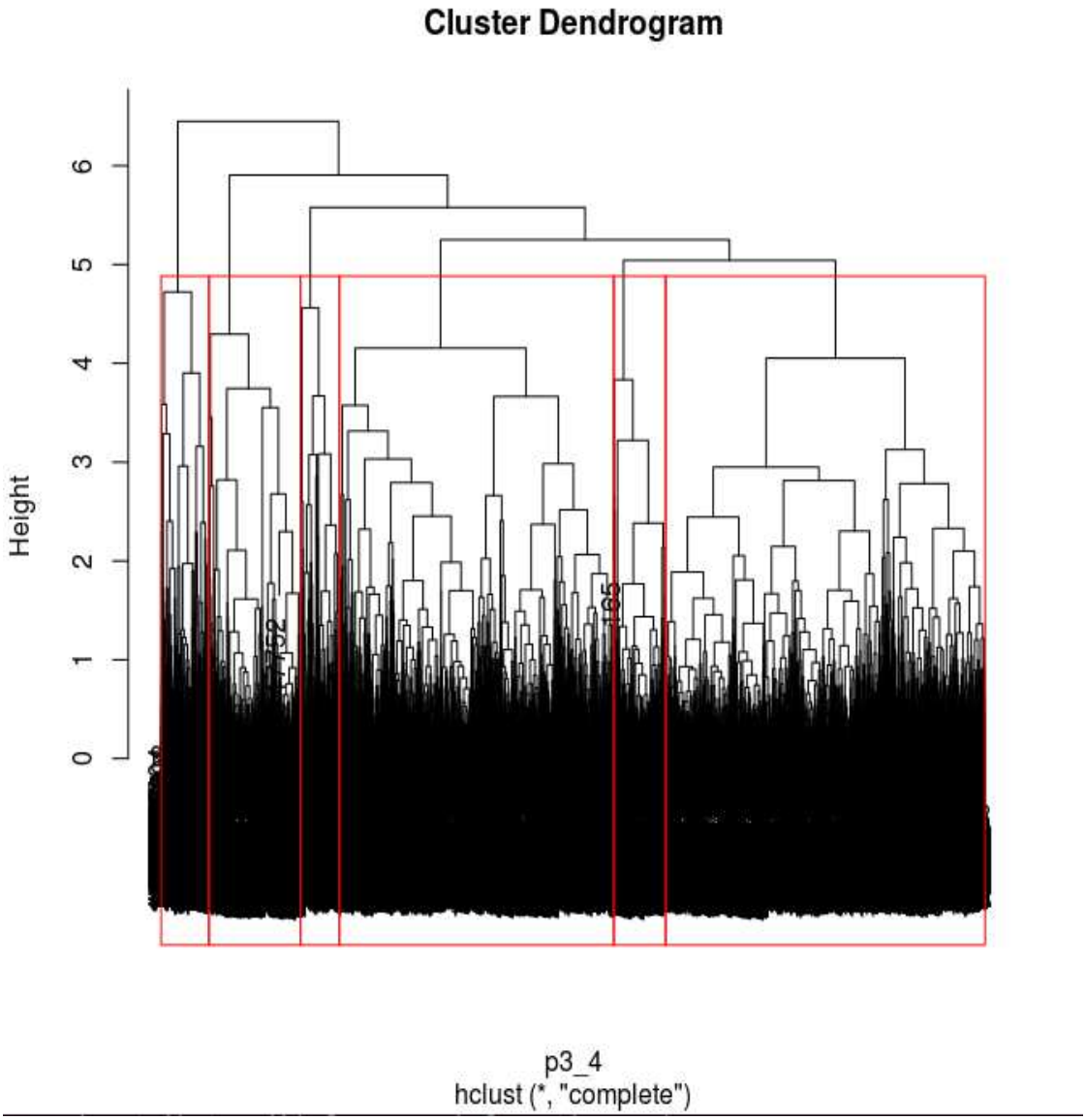


Εικόνα 9: Α) Παρουσιάζεται το αρχικό διάγραμμα T-Q. Β) Παρουσιάζονται τα όρια των File1-9 με κόκκινες γραμμές και τα όρια των File10-29 με μαύρες γραμμές.

3.3 Ιδανικός αριθμός Cluster

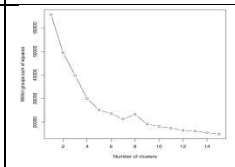
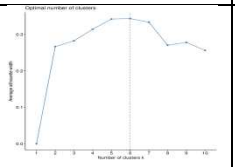
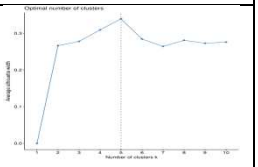
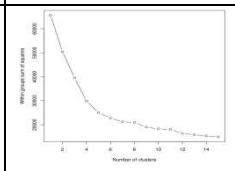
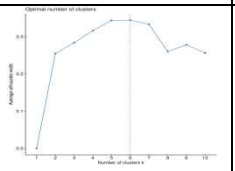
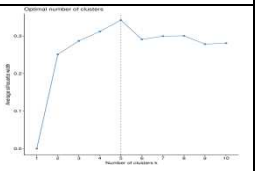
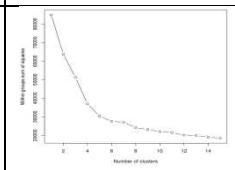
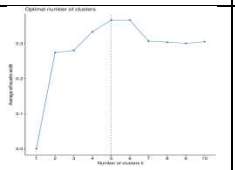
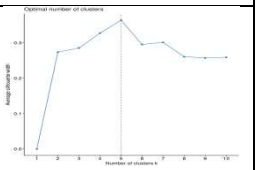
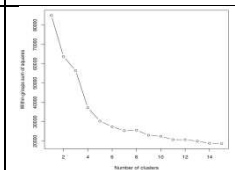
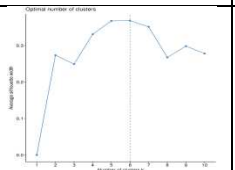
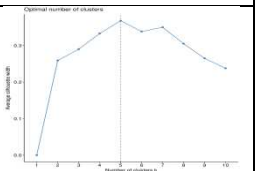
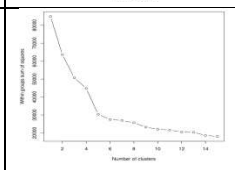
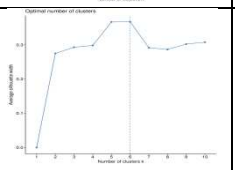
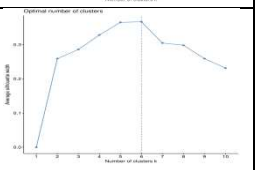
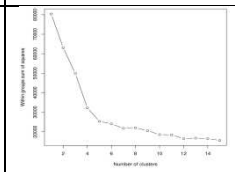
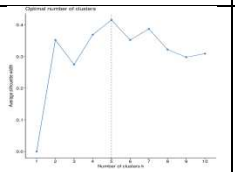
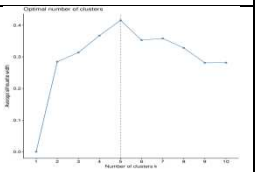
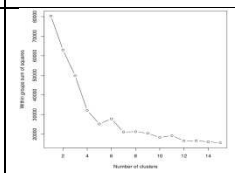
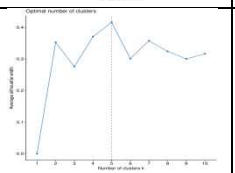
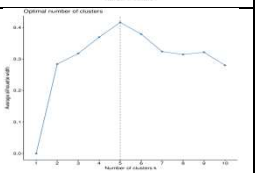
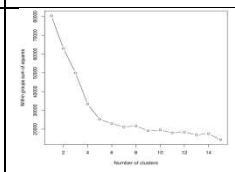
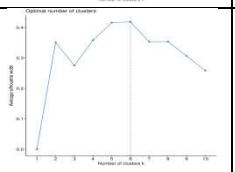
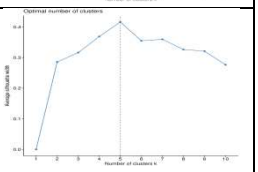
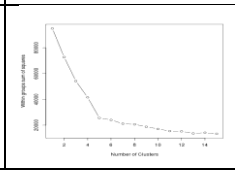
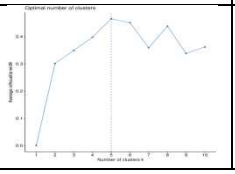
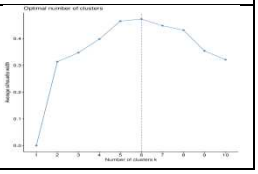
Όπως αναφέρθηκε προηγουμένως, για την επιλογή του ιδανικού αριθμού cluster για το hierarchical clustering, πρέπει να γίνει ανάλυση των δεδομένων με διάφορες μεθόδους. Τα αποτελέσματα από την χρήση του **Script 10** παρατίθενται στον **Πίνακα 2**. Επιλέχθηκε ο μεγαλύτερος αριθμός cluster για κάθε ομάδα δεδομένων, άρα ο ιδανικός αριθμός cluster για τα αρχεία File1, File2, File3, File4, File5, File6, File7, File8 και File9 είναι 6, 6, 5, 6, 6, 5, 5, 6, 6 αντίστοιχα, και για τα αρχεία File10, File11, File12, File13, File14, File15, File16, File17, File18, File19, File20, File21, File22, File23, File24, File25, File26, File27, File28 και File29 είναι 8, 8, 8, 8, 8, 6, 6, 6, 6, 5, 5, 5, 5, 5, 4, 4, 4, 4, 4, 4 αντίστοιχα. Φαίνεται ότι όταν μεταβάλλεται το εύρος T ενώ παραμένει σταθερό το εύρος Q, για τα αρχεία File1- File9, ο ιδανικός αριθμός cluster παραμένει σχετικά ίδιος, ενώ όταν παραμένει σταθερό το εύρος T αλλά αλλάζει το εύρος Q, για τα αρχεία File10- File29, τότε παρατηρείται μείωση του αριθμού cluster καθώς η τιμή Q μετακινείται προς το 1, δηλαδή καθώς οι διαμορφώσεις πλησιάζουν προς την φυσική δομή.

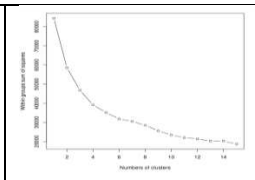
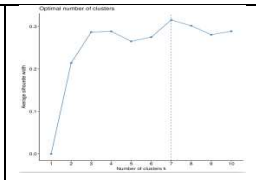
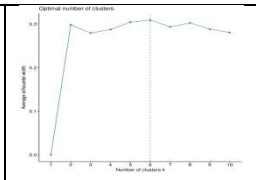
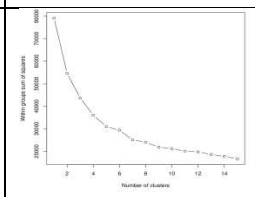
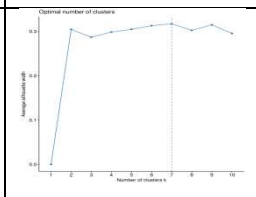
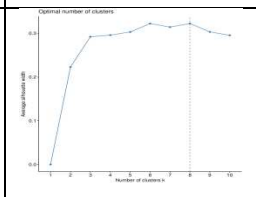
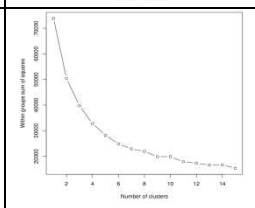
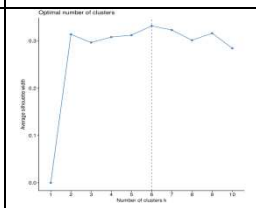
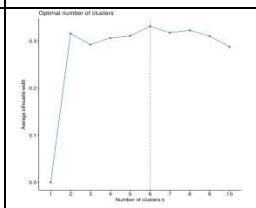
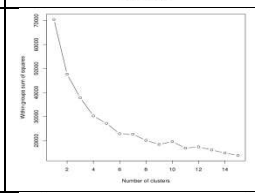
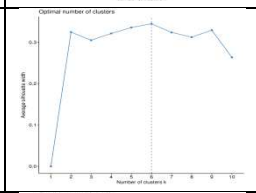
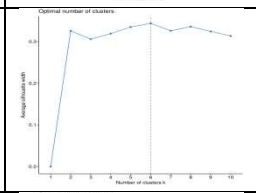
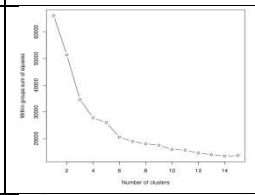
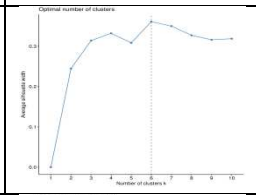
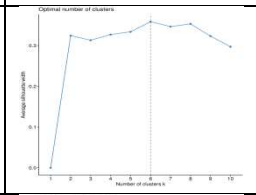
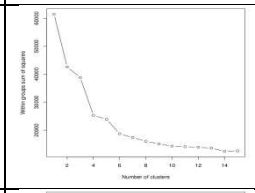
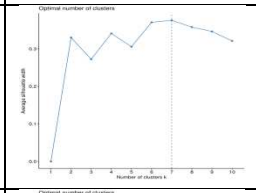
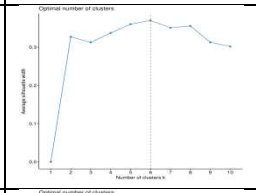
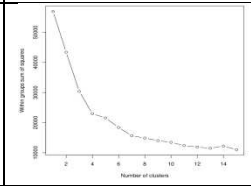
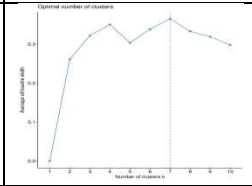
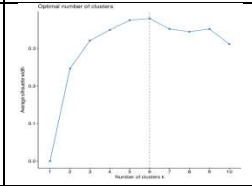
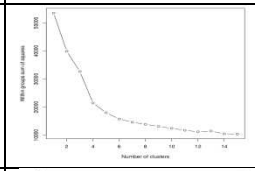
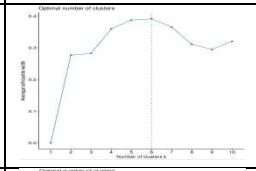
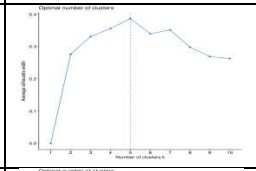
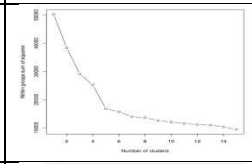
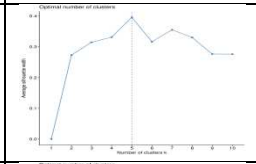
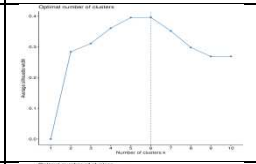
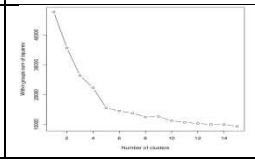
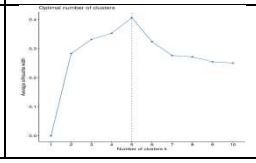
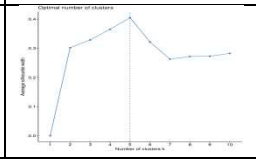
Έτσι, πραγματοποιήθηκε το hierarchical clustering με το **Script 11**, βλέπε ενότητα **2.5**, και δημιουργήθηκαν αρχεία ".dcd" για κάθε cluster για όλες τις ομάδες δεδομένων.

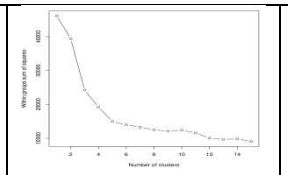
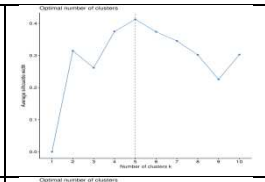
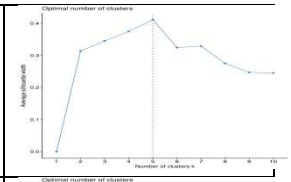
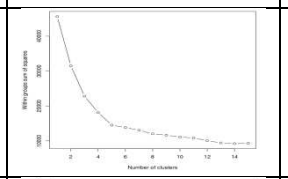
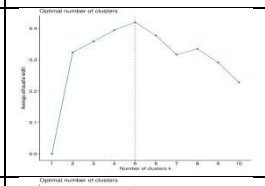
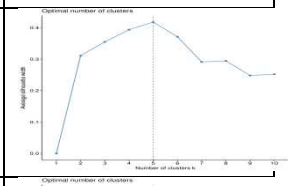
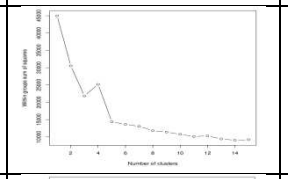
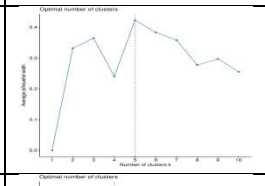
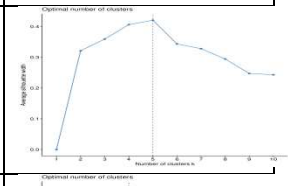
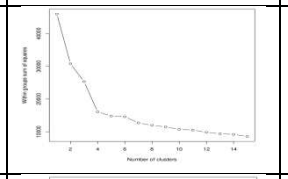
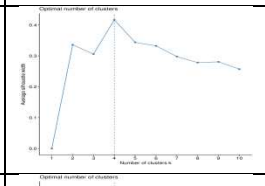
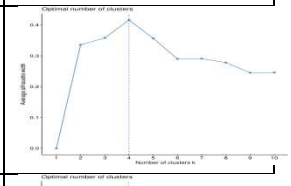
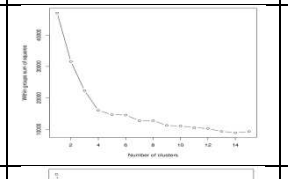
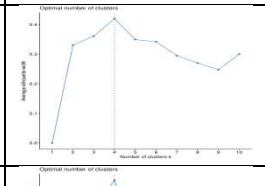
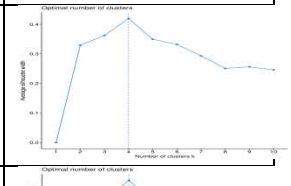
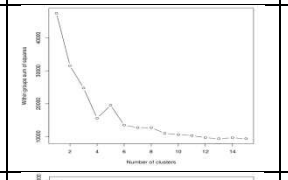
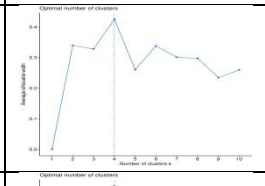
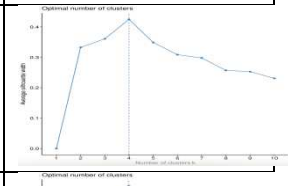
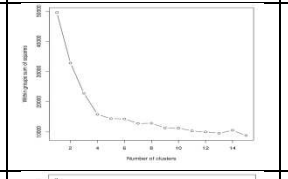
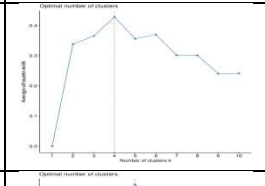
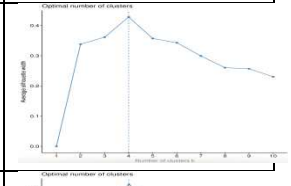
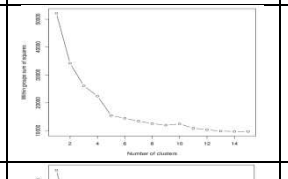
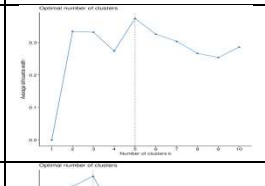
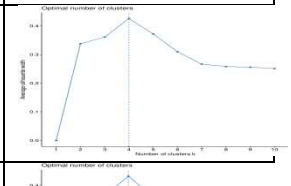
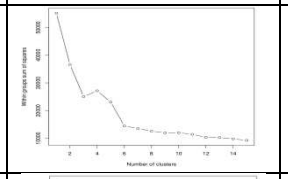
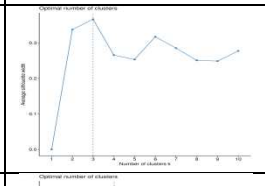
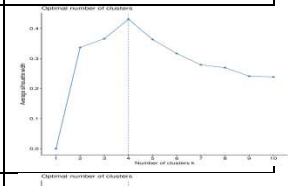
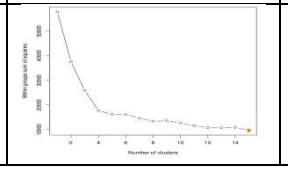
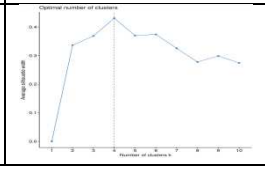
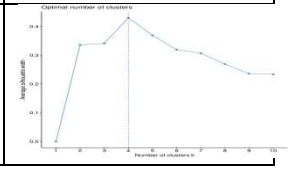


Εικόνα 10: Plotting του hierarchical clustering του αρχείου File9.

ΠΙΝΑΚΑΣ 2: Σχεδιαγράμματα WSS, fviz_nbclust(kmeans), fviz_nbclust(pam) κάθε ομάδας δεδομένων. (Συνεχίζεται και στις επόμενες σελίδες)

Ομάδα δεδομένων	WSS	Kmeans	Pam
File1			
File2			
File3			
File4			
File5			
File6			
File7			
File8			
File9			

File10			
File11			
File12			
File13			
File14			
File15			
File16			
File17			
File18			
File19			

File20			
File21			
File22			
File23			
File24			
File25			
File26			
File27			
File28			
File29			

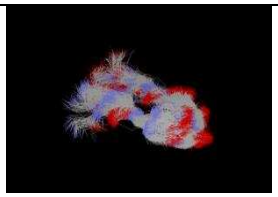
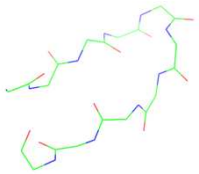
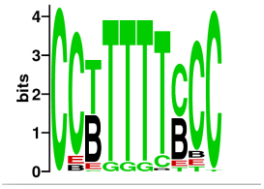
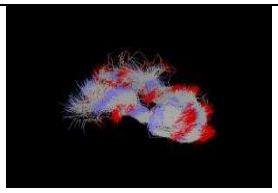
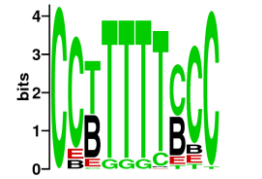
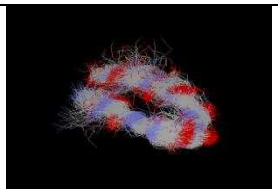
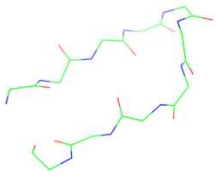
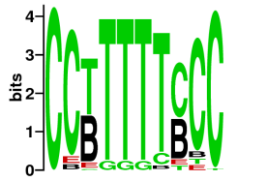
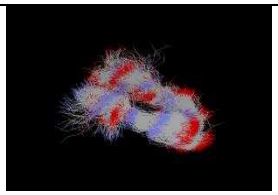
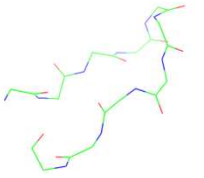
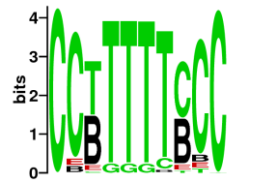
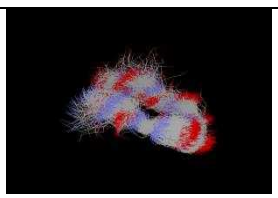
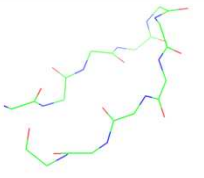
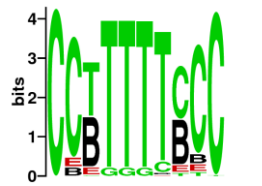
3.4 Χαρακτηριστικά δομών

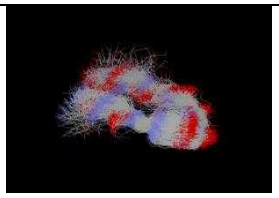
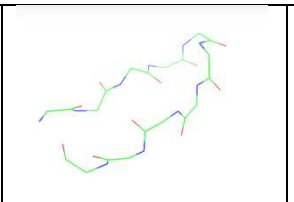
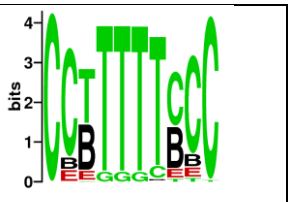
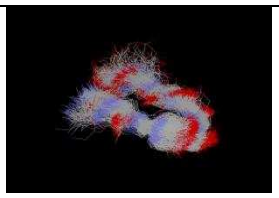
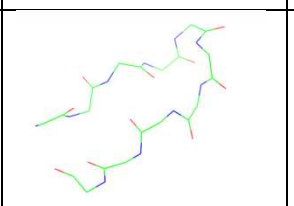
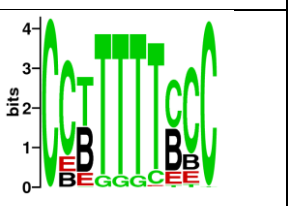
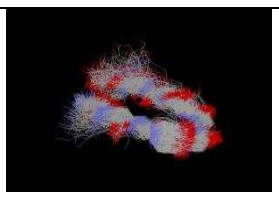
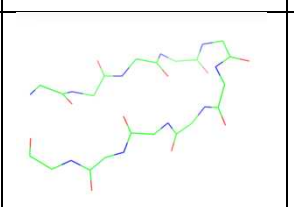
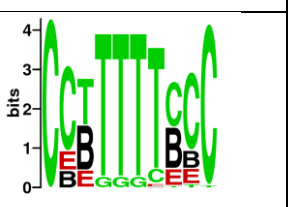
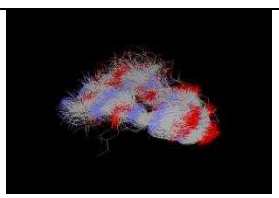
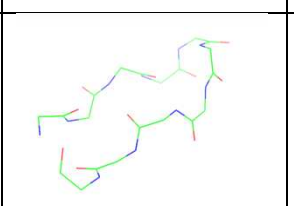
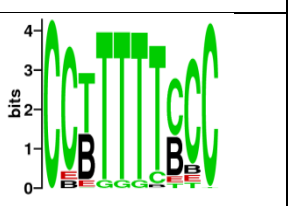
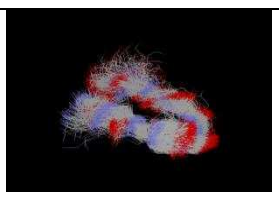
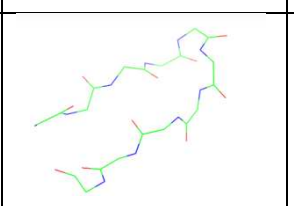
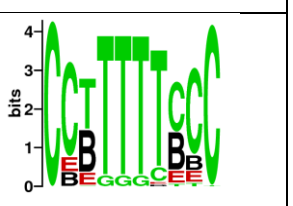
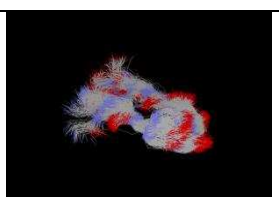
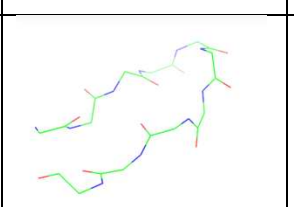
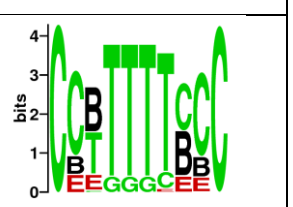
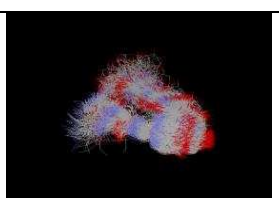
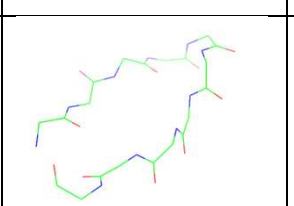
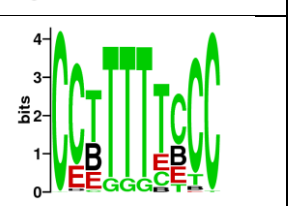
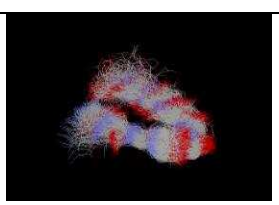
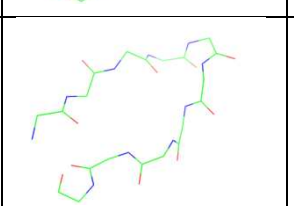
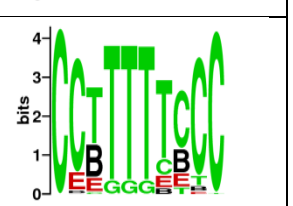
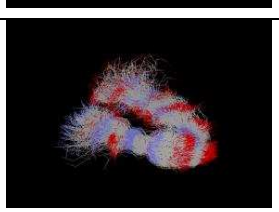
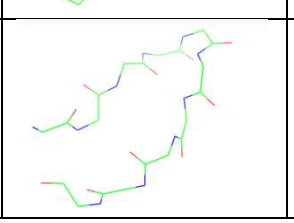
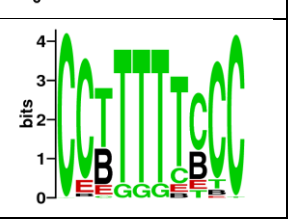
Τα αρχεία ".dcd" αναλύθηκαν με το πρόγραμμα grcarma και τα προγράμματα STRIDE (STRuctural IDEntification), ένα πρόγραμμα που προσδιορίζει την δευτεροταγή δομή, και WebLogo [43][44] και βρέθηκαν οι δομές των cluster και τα χαρακτηριστικά της δευτεροταγούς δομής τους. Το WebLogo παράγει μια αναπαράσταση της αλληλουχίας από στοιβαγμένα γράμματα σε έναν πίνακα. Κάθε γράμμα αντιστοιχεί σε ένα αμινοξύ της αλληλουχίας και το γράμμα με την μεγαλύτερη συχνότητα παρατηρείται στο επάνω μέρος της στοίβας. Τα αποτελέσματα παρατίθενται στον **Πίνακα 3**. Η αντιστοίχηση των γραμμάτων με τα στοιχεία δευτεροταγούς δομής είναι η εξής: C για το τυχαίο σπείραμα (coil), T για την στροφή (turn), G για την 3_{10} έλικα, B για κατάλοιπο σε απομονωμένη β-γέφυρα (β-bridge), E για το β-φύλλο (β-sheet), H για την α-έλικα.

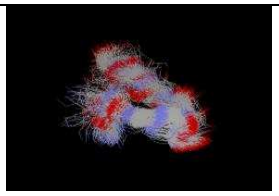
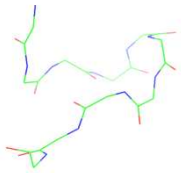
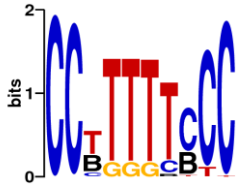
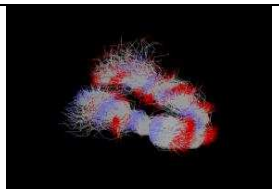
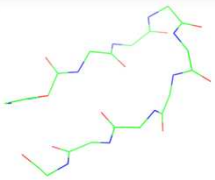
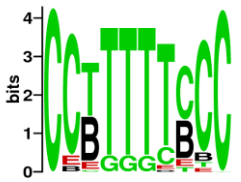
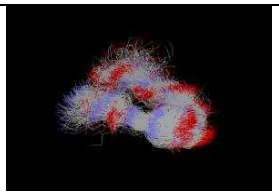
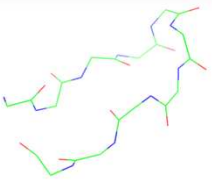
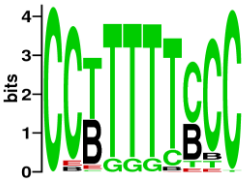
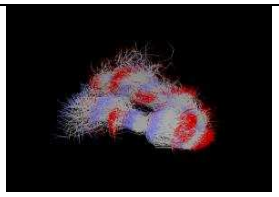
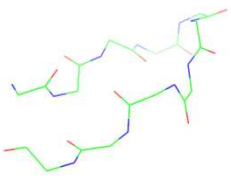
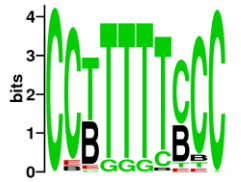
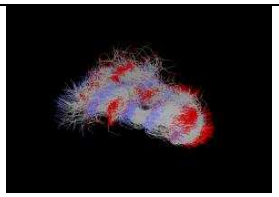
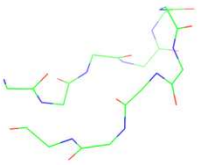
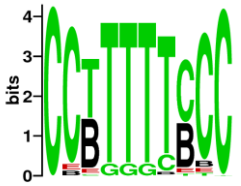
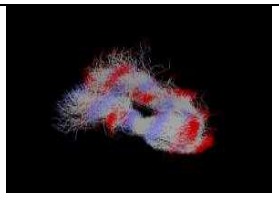
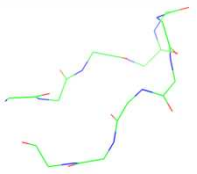
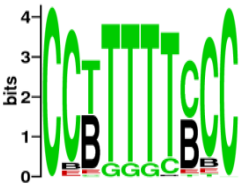
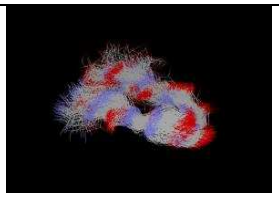
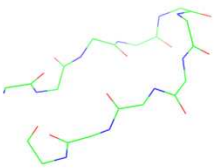
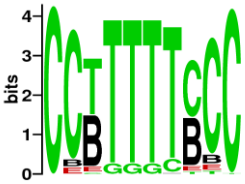
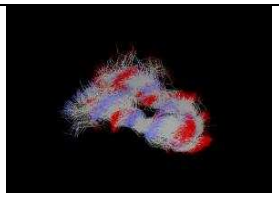
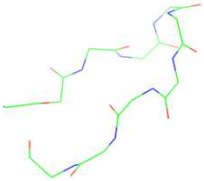
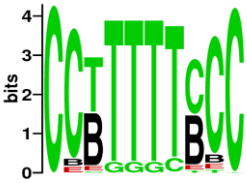
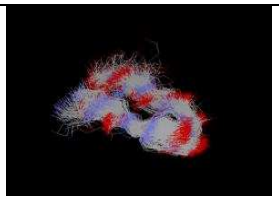
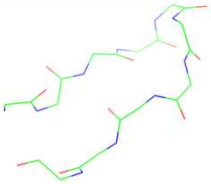
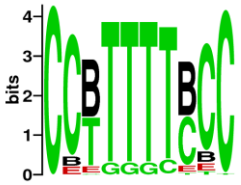
Τα αποτελέσματα αυτά δείχνουν ότι στο κυρίαρχο cluster του αρχείου File13 επικρατεί διαφορετική διαμόρφωση σε σχέση με τα άλλα cluster αλλά από το weblogo του φαίνεται ότι τα κατάλοιπα διατηρούν σχεδόν το ίδιο πρότυπο με τα κατάλοιπα των υπολοίπων cluster. Παρατηρείται ότι σε όλα τα κυρίαρχα cluster τα άκρα της μίνι πρωτεΐνης, δηλαδή τα αμινοξέα Tyr1 και Tyr10, διατηρούν διαμόρφωση τυχαίου σπειράματος, αυτό φαίνεται και στις superpositioned δομές όπου φαίνεται ότι τα άκρα δεν διατηρούν ένα σταθερό μοτίβο, αλλά καταλαμβάνουν αρκετό χώρο με τυχαίες διαμορφώσεις. Τα αμινοξέα Tyr2 και Trp9 έχουν και αυτά κυρίως χαρακτηριστικά τυχαίου σπειράματος, αλλά μπορεί να παρουσιάσουν και χαρακτηριστικά β-γέφυρας ή β-φύλλου, κυρίως όσο οι διαμορφώσεις πλησιάζουν την φυσική διαμόρφωση. Το αμινοξύ Asp3 διατηρεί κυρίως χαρακτηριστικά στροφής, όμως καθώς η θερμοκρασία

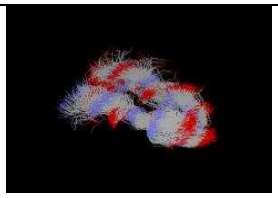
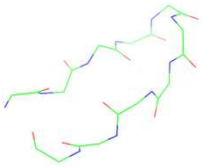
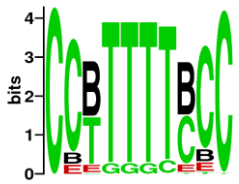
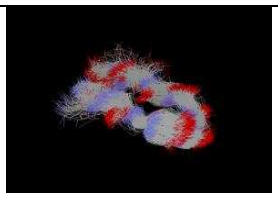
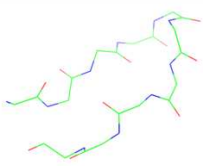
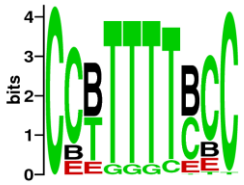
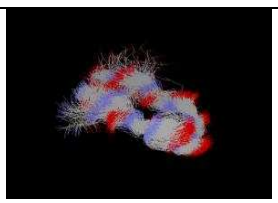
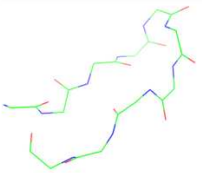
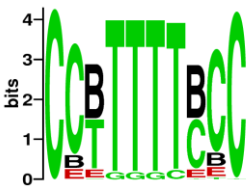
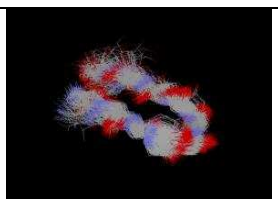
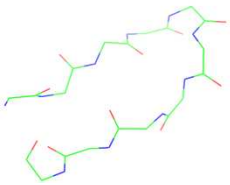
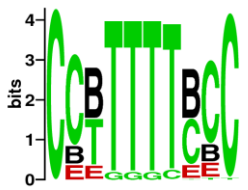
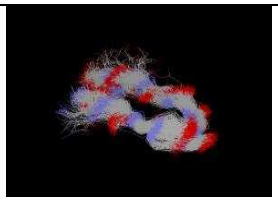
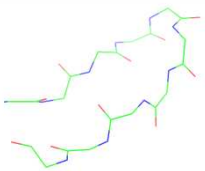
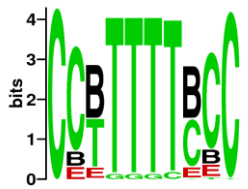
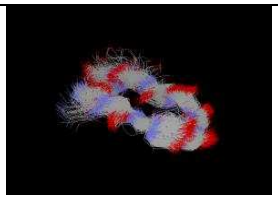
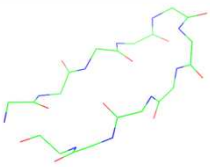
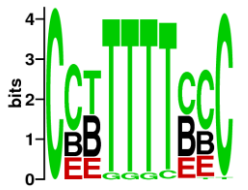
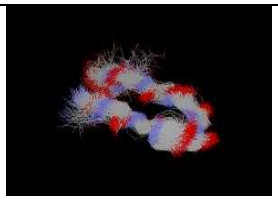
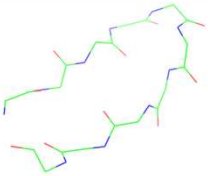
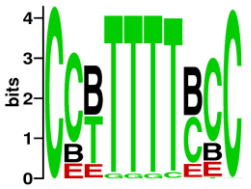
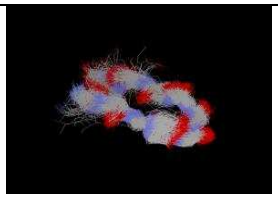
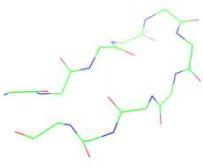
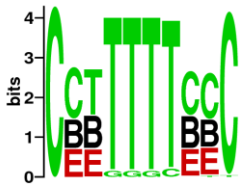
μειώνεται ή οι διαμορφώσεις πλησιάζουν την φυσική δομή, υιοθετεί χαρακτηριστικά είτε β-γέφυρας είτε β-φύλλου. Παρόμοια, και στο αμινοξύ Thr8 ενώ κυριαρχούν τα χαρακτηριστικά τυχαίου σπειράματος, καθώς η θερμοκρασία μειώνεται ή οι διαμορφώσεις πλησιάζουν την φυσική δομή, αρχίζει να υιοθετεί χαρακτηριστικά είτε β-γέφυρας είτε β-φύλλου. Στα αμινοξέα Pro4, Glu5, Thr6 κυριαρχούν τα χαρακτηριστικά στροφής σε όλες τις διακυμάνσεις των T και Q, ενώ το ίδιο συμβαίνει σε μικρότερη συχνότητα για το Gly7.

ΠΙΝΑΚΑΣ 3: Εικόνες των superpositioned δομών, των representative δομών και των Weblogo των κυρίαρχων cluster κάθε ομάδας δεδομένων. (Συνεχίζεται και στις επόμενες σελίδες)

Κυρίαρχο cluster	Superposition structures	Representative structure	WebLogo
0.49≤Q≤0.63 467.4721<T≤530.9874 Cluster 2 (8117 δομές)			
0.49≤Q≤0.63 467.4721<T≤530.9874 Cluster 1 (5959 δομές)			
0.49≤Q≤0.63 467.4721<T≤530.9874 Cluster 4 (5998 δομές)			
0.49≤Q≤0.63 403.9569<T≤467.4721 Cluster 3 (7385 δομές)			
0.49≤Q≤0.63 403.9569<T≤467.4721 Cluster 2 (9615 δομές)			

0.49≤Q≤0.63 403.9569<T≤467.4721 Cluster 2 (7713 δομές)			
0.49≤Q≤0.63 340.4417<T≤403.9569 Cluster 1 (6315 δομές)			
0.49≤Q≤0.63 340.4417<T≤403.9569 Cluster 1 (6070 δομές)			
0.49≤Q≤0.63 340.4417<T≤403.9569 Cluster 3 (8625 δομές)			
0.49≤Q≤0.63 340.4417<T≤403.9569 Cluster 1 (9712 δομές)			
0.49≤Q≤0.63 276.9265≤T≤340.4417 Cluster 2 (8890 δομές)			
0.49≤Q<0.497 276.9265≤T≤530.9874 Cluster 3 (4029 δομές)			
0.497≤Q<0.504 276.9265≤T≤530.9874 Cluster 1 (4924 δομές)			
0.504≤Q<0.511 276.9265≤T≤530.9874 Cluster 1 (4842 δομές)			

0.511≤Q<0.518 276.9265≤T≤530.9874 Cluster 3 (4384 δομές)			
0.518≤Q<0.525 276.9265≤T≤530.9874 Cluster 1 (4019 δομές)			
0.525≤Q<0.532 276.9265≤T≤530.9874 Cluster 2 (5660 δομές)			
0.532≤Q<0.539 276.9265≤T≤530.9874 Cluster 3 (3933 δομές)			
0.539≤Q<0.546 276.9265≤T≤530.9874 Cluster 2 (3809 δομές)			
0.546≤Q<0.553 276.9265≤T≤530.9874 Cluster 3 (3763 δομές)			
0.553≤Q<0.560 276.9265≤T≤530.9874 Cluster 2 (7630 δομές)			
0.560≤Q<0.567 276.9265≤T≤530.9874 Cluster 2 (8038 δομές)			
0.567≤Q<0.574 276.9265≤T≤530.9874 Cluster 2 (6792 δομές)			

0.574≤Q<0.581 276.9265≤T≤530.9874 Cluster 2 (6959 δομές)			
0.581≤Q<0.588 276.9265≤T≤530.9874 Cluster 1 (5470 δομές)			
0.588≤Q<0.595 276.9265≤T≤530.9874 Cluster 2 (7531 δομές)			
0.595≤Q<0.602 276.9265≤T≤530.9874 Cluster 1 (12440 δομές)			
0.602≤Q<0.609 276.9265≤T≤530.9874 Cluster 2 (8876 δομές)			
0.609≤Q<0.616 276.9265≤T≤530.9874 Cluster 1 (7268 δομές)			
0.616≤Q<0.623 276.9265≤T≤530.9874 Cluster 1 (9437 δομές)			
0.623≤Q≤0.63 276.9265≤T≤530.9874 Cluster 3 (9296 δομές)			

3.5 Υδρογονοδεσμοί

Μετά από την ανάλυση των κυρίαρχων cluster με το πρόγραμμα vmd, με cutoff 3.2 για την απόσταση μεταξύ δότη και δέκτη και γωνία N-H…O μεγαλύτερη από 130^{o} [22][45], βρέθηκαν ορισμένοι υδρογονοδεσμοί που εμφανίζονται σε όλα τα cluster ή σχεδόν σε όλα. Και στις ομάδες δεδομένων όπου αλλάζει το εύρος Τ αλλά και στις ομάδες όπου αλλάζει το εύρος του Q, παρατηρούνται υδρογονοδεσμοί που είτε μένουν σταθεροί, παρά την αλλαγή Τ ή Q, είτε αλλάζει η συχνότητα εμφάνισής τους.

Οι υδρογονοδεσμοί που παρατηρούνται στις ομάδες μεταβολής Τ και η συμπεριφορά της συχνότητας εμφάνισής τους είναι οι εξής:

- Gly7 N – Asp3 O: η συχνότητα εμφάνισής του παραμένει σχετικά σταθερή σε όλο το εύρος Τ.
- Tyr1 N – Tyr10 O: όσο ελαττώνεται η θερμοκρασία, ελαττώνεται και η συχνότητα εμφάνισής του.
- Tyr1 N – Tyr10 OXT: όσο ελαττώνεται η θερμοκρασία, ελαττώνεται και η συχνότητα εμφάνισής του.
- Thr6 OG1 – Asp3 OD1: σε υψηλές θερμοκρασίες η συχνότητα εμφάνισής του είναι αρκετά μικρή, ενώ σε μικρότερες θερμοκρασίες είναι μεγαλύτερη και σταθερή.
- Thr6 OG1 – Asp3 OD2: παρατηρείται η ίδια συμπεριφορά με τον υδρογονοδεσμό Thr6 OG1 – Asp3 OD1.
- Asp3 N – Thr8 O: όσο ελαττώνεται η θερμοκρασία, αυξάνεται η συχνότητα εμφάνισής του.
- Thr6 N – Asp3 OD1: μικρότερη η συχνότητα εμφάνισής του σε υψηλές θερμοκρασίες σε σχέση με την συχνότητα του σε μικρότερες θερμοκρασίες.

- Glu5 N – Asp3 OD1: παρατηρείται η ίδια συμπεριφορά με τον υδρογονοδεσμό Thr6 N – Asp3 OD1.
- Thr8 N – Asp3 O: παρατηρείται η ίδια συμπεριφορά με τους υδρογονοδεσμούς Thr6 N – Asp3 OD1 και Glu5 N – Asp3 OD1.

Οι υδρογονοδεσμοί που παρατηρούνται στις ομάδες μεταβολής Q και η συμπεριφορά της συχνότητάς εμφάνισής τους είναι οι εξής:

- Thr6 OG1 – Asp3 OD2: παρατηρείται σε ολόκληρο το εύρος Q με σχεδόν σταθερή συχνότητα.
- Thr6 OG1 – Asp3 OD1: εμφανίζεται με μεγαλύτερη και σχετικά σταθερή συχνότητα για τις τιμές $0.511 \leq Q \leq 0.63$.
- Gly7 N – Asp3 O: εμφανίζεται σε ολόκληρο το εύρος Q με λίγο μεγαλύτερη συχνότητα στις ομάδες δεδομένων που έχουν μεγαλύτερο Q.
- Tyr1 N – Tyr10 O: αρκετά μικρή συχνότητα στις ομάδες με χαμηλό Q, μεγαλύτερη και σταθερή συχνότητα στο ενδιάμεσο εύρος Q, ενώ έχει αρκετά μεγαλύτερη συχνότητα στις ομάδες με το μεγαλύτερο Q, δεν υπάρχει στο File13 λόγω της διαφορετικής δομής του.
- Tyr1 N – Tyr10 OXT: σταθερή σχετικά συχνότητα σε όλο το εύρος Q, εκτός των ομάδων με πολύ μικρό Q και του File13 λόγω της δομής του.
- Asp3 N – Thr8 O: όσο μεγαλώνει η τιμή Q, αυξάνει και η συχνότητα εμφάνισής του με σχεδόν διπλάσια συχνότητα στις ομάδες με τις μεγαλύτερες τιμές Q, δεν εμφανίζεται στο File13 λόγω της δομής του.
- Thr6 N – Asp3 OD1: παρατηρείται για τιμές Q μεγαλύτερες του 0.532 με σταθερή σχετικά συχνότητα.

- Thr6 N – Asp3 OD2: παρατηρείται η ίδια συμπεριφορά με τον υδρογονοδεσμό Thr6 N – Asp3 OD1.
- Glu5 N – Asp3 OD1: εμφανίζεται με μεγαλύτερη συχνότητα κυρίως μετά την τιμή Q 0.588.
- Glu5 N – Asp3 OD2: παρατηρείται η ίδια συμπεριφορά με τον υδρογονοδεσμό Glu5 N – Asp3 OD1.
- Thr8 OG1 – Tyr1 O: εμφανίζεται με μεγαλύτερη συχνότητα στο εύρος $0.539 \leq Q \leq 0.567$.
- Tyr10 N – Tyr1 O: εμφανίζεται με μεγάλη συχνότητα μόνο στις ομάδες με $0.610 \leq Q$.

Οι υδρογονοδεσμοί Gly7 N – Asp3 O, Thr6 N – Asp3 OD1, Thr6 N – Asp3 OD2, Glu5 N – Asp3 OD1 και Glu5 N – Asp3 OD2 συμβάλουν στον σχηματισμό της στροφής, ενώ οι υδρογονοδεσμοί Tyr1 N – Tyr10 OXT, Tyr10 N – Tyr1 O, Asp3 N – Thr8 O και Tyr1 N – Tyr10 O συμβάλουν στον σχηματισμό του β-φύλλου [46]**.** Επίσης, οι υδρογονοδεσμοί Tyr10 N – Tyr1 O, Asp3 N – Thr8 O, Thr6 N – Asp3 OD1, Glu5 N – Asp3 OD1 και Gly7 N – Asp3 O εντοπίζονται στην κρυσταλλική δομή του CLN025, ενώ οι υδρογονοδεσμοί Asp3 N – Thr8 O, Thr8 N – Asp3 O, Gly7 N – Asp3 O και Thr6 N – Asp3 OD1 εντοπίζονται στην δομή του CLN025 σε υγρό διάλυμα. Ο υδρογονοδεσμός Tyr1 N – Tyr10 OXT θεωρείται γέφυρα άλατος στην κρυσταλλική δομή του [47].

ΚΕΦΑΛΑΙΟ 4

ΣΥΖΗΤΗΣΗ

Η παρούσα εργασία είχε ως σκοπό να διερευνήσει εάν είναι εφικτή η εύρεση των μεταβατικών καταστάσεων της μίνι πρωτεΐνης CLN025 μέσω του διαγράμματος T-Q, το οποίο κατασκευάστηκε από προσομοίωση μοριακής δυναμικής. Το CLN025 είναι ένα αρκετά σταθερό και γρήγορα αναδιπλούμενο μόριο, και η μελέτη των μεταβατικών καταστάσεων του θα βοηθήσει στην μελέτη, και ίσως αναγνώριση, του μηχανισμού αναδίπλωσης των β-φύλλων.

Έχοντας απομονώσει τις διαμορφώσεις στο εύρος των τιμών Q, από το διάγραμμα T-Q, που πιστεύουμε ότι βρίσκονται οι μεταβατικές καταστάσεις, διαχωρίσαμε αυτές τις διαμορφώσεις σε ομάδες που είτε παραμένει σταθερό το εύρος T και μεταβάλλεται το εύρος Q, είτε παραμένει σταθερό το εύρος Q και μεταβάλλεται το εύρος T. Αυτό έγινε για να για να συγκριθούν τα αποτελέσματα από τις δύο ομάδες και να φανερωθούν τυχόν διαφορές που θα υποδηλώνουν την σημασία της θερμοκρασίας ή του Q στον μηχανισμό αναδίπλωσης. Έπειτα πραγματοποιήθηκε hierarchical cluster analysis σε αυτές τις ομάδες με σκοπό να διαχωριστούν οι διαφορετικές διαμορφώσεις και να φανεί το κυρίαρχο cluster, δηλαδή οι διαμορφώσεις που επικρατούν σε κάθε ομάδα. Αναλύοντας την δομή και τα στοιχεία των δευτεροταγών διαμορφώσεων που υιοθετούν οι διαμορφώσεις των κυρίαρχων cluster, φάνηκε ότι οι διαμορφώσεις από όλα τα κυρίαρχα cluster έχουν υποστεί hydrophobic collapse και έχουν σχηματισμένη την στροφή, ενώ στις χαμηλότερες θερμοκρασίες ή καθώς το Q αυξάνεται αρχίζουν να αποκτούν την διαμόρφωση του β-φύλλου. Τέλος, αναλύσαμε τις διαμορφώσεις για να βρούμε τους υδρογονοδεσμούς που διέπουν τα κυρίαρχα cluster. Η ανάλυση των ομάδων που μεταβάλλεται το εύρος T δείχνει ότι όσο μειώνεται η θερμοκρασία αποσταθεροποιούνται οι υδρογονοδεσμοί των άκρων και σταθεροποιούνται οι υδρογονοδεσμοί

της στροφής και αυτοί που βρίσκονται κοντά σε αυτή. Η ανάλυση των ομάδων που μεταβάλλεται το εύρος Q δείχνει ότι οι διαμορφώσεις με μικρό Q έχουν σχηματίσει κυρίως υδρογονοδεσμούς στην στροφή, ενώ όσο αυξάνεται το Q αρχίζουν να σχηματίζουν τους υδρογονοδεσμούς του β-φύλλου. Συνεπώς, τα αποτελέσματα αυτά δείχνουν ότι η αύξηση της θερμοκρασίας βοηθά τον σχηματισμό των υδρογονοδεσμών που συμβάλουν στον σχηματισμό του β-φύλλου, ενώ κατά την πορεία των διαμορφώσεων από τις αποδιατεταγμένες διαμορφώσεις προς την φυσική διαμόρφωση η β-φουρκέτα σχηματίζεται, και επομένως οι υδρογονοδεσμοί, με το μοτίβο "φερμουάρ", δηλαδή από την στροφή προς τα άκρα.

Αυτά τα αποτελέσματα συμφωνούν, εν μέρη, με τα αποτελέσματα μιας πρόσφατης έρευνας των McKieman et al., που μελέτησαν τον μηχανισμό αναδίπλωσης του CLN025 και έδειξαν ότι πρώτα συμβαίνει hydrophobic collapse της δομής και μετά από λίγο σχηματίζεται η στροφή, με τελικό στάδιο τον σχηματισμό των υδρογονοδεσμών με το μοτίβο "φερμουάρ" [29]. Δεν συμφωνεί στο ότι πρώτα συμβαίνει hydrophobic collapse της δομής και έπειτα σχηματίζεται η στροφή, διότι δεν αρκούν τα αποτελέσματα μας για να γίνει αυτή η διάκριση καθώς σε αυτά οι διαμορφώσεις έχουν ήδη σχηματισμένη την στροφή και έχουν υποστεί hydrophobic collapse.

Επομένως, η ανάλυση του διαγράμματος T-Q μας έδειξε ποιοί υδρογονοδεσμοί σχηματίζονται πρώτα κατά την διαδικασία της αναδίπλωσης και τον μηχανισμό με τον οποίο σχηματίζονται, αλλά δεν ήταν αρκετή ώστε να διαχωρίσει την χρονική σειρά που συμβαίνει το hydrophobic collapse των δομών και ο σχηματισμός της στροφής. Αυτός ο διαχωρισμός, ίσως, να είναι εφικτός με περαιτέρω ανάλυση του διαγράμματος T-Q.

Βιβλιογραφια

1. Cox, M. M., & Nelson, D. L. (2005). Principles of Biochemistry: Lehninger(4th ed.). New York: W. H. Freeman and Company.

2. Berg, J. M., Tymoczko, J. L., & Stryer, L. (2012). Biochemistry(7th ed.). New York: W.H. Freeman and Company.

3. Anfinsen, C. B., Haber, E., Sela, M., & White, F. H. (1961). The Kinetics Of Formation Of Native Ribonuclease During Oxidation Of The Reduced Polypeptide Chain. Proceedings of the National Academy of Sciences ,47(9), 1309-1314. https://doi.org/10.1073/pnas.47.9.1309

4. Karplus, M., & Kuriyan, J. (2005). Molecular dynamics and protein function. Proceedings of the National Academy of Sciences of the United States of America, 102(19), 6679–6685. https://doi.org/10.1073/pnas.0408930102

5. Levinthal, C. (1998). Are there pathways for protein folding? Journal de Chimie Physique, 65, 44-45. https://doi.org/10.1051/jcp/1968650044

6. Tropp, B. E. (2014). Principles of molecular biology(1¬¬¬st ed.). Burlington, MA: Jones & Bartlett Learning.

7. Baldwin, R. L., & Rose, G. D. (1999). Is protein folding hierarchic? II. Folding intermediates and transition states. Trends in Biochemical Sciences, 24(2), 77-83. https://doi.org/10.1016/S0968-0004(98)01345-0

8. Fersht, A. R. (1995). Characterizing transition states in protein folding: an essential step in the puzzle. Current Opinion in Structural Biology, 5(1), 79-84. https://doi.org/10.1016/0959-440X(95)80012-P

9. Hatfield, M. P., Murphy, R. F., & Lovas, S. (2010). Molecular Dynamics Analysis of the Conformations of a β-Hairpin Miniprotein. The Journal of Physical Chemistry B, 114(8), 3028-3037. https://doi.org/10.1021/jp910465e

10. Polticelli, F., Raybaudi-Massilia, G., & Ascenzi, P. (2001). Structural determinants of mini-protein stability. Biochemistry and Molecular Biology Education, 29(1), 16-20.

https://doi.org/10.1016/S1470-8175(00)00066-7

11. Onuchic, J. N., Socci, N. D., Luthey-Schulten, Z., & Wolynes, P. G. (1996). Protein folding funnels: the nature of the transition state ensemble. Folding and Design, 1(6), 441-450. https://doi.org/10.1016/S1359-0278(96)00060-0

12. Dill, K. A., & Chan, H. S. (1997). From Levinthal to pathways to funnels. Nature Structural & Molecular Biology, 4(1), 10-19. https://doi.org/10.1038/nsb0197-10

13. Huang, J., & Cheng, J. (2008). Differentiation between two-state and multi-state folding proteins based on sequence. Proteins: Structure, Function, and Bioinformatics, 72(1), 44-49. https://doi.org/10.1002/prot.21893

14. Zwanzig, R. (1997). Two-state models of protein folding kinetics. Proceedings of the National Academy of Sciences of the United States of America, 94(1), 148–150. https://doi.org/10.1073/pnas.94.1.148

15. Ahluwalia, U., Katyal, N., & Deep, S. (2012). MODELS OF PROTEIN FOLDING. Journal Of Proteins & Proteomics, 3(2), 85-93. Retrieved from http://jpp.org.in/index.php/jpp/article/view/17

16. 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: Structure, Function, and Genetics, 21(3), 167-195. https://doi.org/10.1002/prot.340210302

17. Branden, C., & Tooze, J. (2009). Introduction to protein structure(2nd ed.). New York: Garland Publishing.

18. Honda, S., Akiba, T., Kato, Y. S., Sawada, Y., Sekijima, M., Ishimura, M., et al. (2008). Crystal Structure of a Ten-Amino Acid Protein. Journal of the American Chemical Society, 130(46), 15327-15331. https://doi.org/10.1021/ja8030533

19. Hatfield, M. P. D., Murphy, R. F., & Lovas, S. (2010). VCD Spectroscopic Properties of the β-hairpin Forming Miniprotein CLN025 in Various Solvents. Biopolymers, 93(5), 442–450. https://doi.org/10.1002/bip.21356

20. Hatfield, M. P., Murphy, R. F., & Lovas, S. (2011). The CLN025 Decapeptide Retains a β-Hairpin Conformation in Urea and Guanidinium Chloride. The Journal of Physical Chemistry B, 115(17), 4971-4981. https://doi.org/10.1021/jp111076j

21. Rodriguez, A., Mokoema, P., Corcho, F., Bisetty, K., & Perez, J. J. (2011). Computational Study of the Free Energy Landscape of the Miniprotein CLN025 in Explicit and Implicit Solvent. The Journal of Physical Chemistry B, 115(6), 1440-1449. https://doi.org/10.1021/jp106475c

22. Zhao, G., & Cheng, C. (2011). Molecular dynamics simulation exploration of unfolding and refolding of a ten-amino acid miniprotein.

Amino Acids, 43(2), 557-565. https://doi.org/10.1007/s00726-011-1150-5

23. Davis, C. M., Xiao, S., Raleigh, D. P., & Dyer, R. B. (2012). Raising the Speed Limit for β-Hairpin Formation. Journal of the American Chemical Society, 134(35), 14476-14482. https://doi.org/10.1021/ja3046734

24. Davis, C. M., & Dyer, R. B. (2013). Dynamics of an Ultrafast Folding Subdomain in the Context of a Larger Protein Fold. Journal of the American Chemical Society, 135(51), 19260-19267. https://doi.org/10.1021/ja409608r

25. Yasuda, S., Hayashi, T., & Kinoshita, M. (2014). Physical origins of the high structural stability of CLN025 with only ten residues. The Journal of Chemical Physics, 141(10), 105103. https://doi.org/10.1063/1.4894753

26. Muñoz, V., Thompson, P. A., Hofrichter, J., & Eaton, W. A. (1997). Folding dynamics and mechanism of β-hairpin formation. Nature, 390(6656), 196-199. https://doi.org/10.1038/36626

27. Pande, V. S., & Rokhsar, D. S. (1999). Molecular dynamics simulations of unfolding and refolding of a -hairpin fragment of protein G. Proceedings of the National Academy of Sciences, 96(16), 9062-9067. https://doi.org/10.1073/pnas.96.16.9062

28. Dinner, A. R., Lazaridis, T., & Karplus, M. (1999). Understanding -hairpin formation. Proceedings of the National Academy of Sciences, 96(16), 9068-9073. https://doi.org/10.1073/pnas.96.16.9068

29. Mckiernan, K. A., Husic, B. E., & Pande, V. S. (2017). Modeling the Mechanism of CLN025 Beta-Hairpin Formation. The Journal of Chemical Physics, 147(10), 104107. https://doi.org/10.1063/1.4993207

30. Serafeim, A., Salamanos, G., Patapati, K. K., & Glykos, N. M. (2016). Sensitivity of Folding Molecular Dynamics Simulations to Even Minor Force Field Changes. Journal of Chemical Information and Modeling, 56(10), 2035-2041. https://doi.org/10.1021/acs.jcim.6b00493

31. Perl programming documentation. (n.d.). Retrieved January 14, 2018, from http://perldoc.perl.org/

32. Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., … Birney, E. (2002). The Bioperl Toolkit: Perl Modules for the Life Sciences. Genome Research, 12(10), 1611–1618. https://doi.org/10.1101/gr.361602

33. What is R? (n.d.). Retrieved January 14, 2018, from https://www.r-project.org/about.html

34. Mu, Y., Nguyen, P. H., & Stock, G. (2004). Energy landscape of a small peptide revealed by dihedral angle principal component analysis. Proteins: Structure, Function, and Bioinformatics, 58(1), 45-52. https://doi.org/10.1002/prot.20310

35. Altis, A., Nguyen, P. H., Hegger, R., & Stock, G. (2007). Dihedral angle principal component analysis of molecular dynamics simulations. The Journal of Chemical Physics, 126(24), 244111. https://doi.org/10.1063/1.2746330

36. Glykos, N. M. (n.d.). Plot. Retrieved January 14, 2018, from https://utopia.duth.gr/glykos/plot/#plot

37. Kabacoff , R. (n.d.). Cluster Analysis. Retrieved January 14, 2018, from https://www.statmethods.net/advstats/cluster.html

38. Determining The Optimal Number Of Clusters: 3 Must Know Methods. (2017, September 07). Retrieved January 14, 2018, from http://www.sthda.com/english/articles/29-cluster-validation-essentials/96-determining-the-optimal-number-of-clusters-3-must-know-methods/

39. Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: AnRPackage for Determining the Relevant Number of Clusters in a Data Set. Journal of Statistical Software, 61(6). https://doi.org/10.18637/jss.v061.i06

40. Glykos, N. M. (2006). Software news and updates carma: A molecular dynamics analysis program. Journal of Computational Chemistry, 27(14), 1765-1768. https://doi.org/10.1002/jcc.20482

41. Koukos, P. I., & Glykos, N. M. (2013). Grcarma: A fully automated task-oriented interface for the analysis of molecular dynamics trajectories. Journal of Computational Chemistry, 34(26), 2310-2312. https://doi.org/10.1002/jcc.23381

42. What is VMD? (n.d.). Retrieved January 14, 2018, from http://www.ks.uiuc.edu/Research/vmd/allversions/what_is_vmd.html

43. Heinig, M., & Frishman, D. (2004). STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Research, 32(Web Server issue), W500–W502. https://doi.org/10.1093/nar/gkh429

44. Crooks, G. E., Hon, G., Chandonia, J.-M., & Brenner, S. E. (2004). WebLogo: A Sequence Logo Generator. Genome Research, 14(6), 1188–1190. https://doi.org/10.1101/gr.849004

45. Kwan, E. (2009, September 11). Noncovalent interactions: An Introduction to Hydrogen Bonding [PDF]. Retrieved from http://evans.rc.fas.harvard.edu/pdf/smnr_2009_Kwan_Eugene.pdf

46. Mckiernan, K. A., Husic, B. E., & Pande, V. S. (2017). Modeling the Mechanism of CLN025 Beta-Hairpin Formation [Supplementary Material]. The Journal of Chemical Physics, 147(10), 104107. https://doi.org/10.1063/1.4993207

47. Honda, S., Akiba, T., Kato, Y. S., Sawada, Y., Sekijima, M., Ishimura, M., et al. (2008). Crystal Structure of a Ten-Amino Acid Protein [Supplementary Material]. Journal of the American Chemical Society, 130(46), 15327-15331. https://doi.org/10.1021/ja8030533

ΠΑΡΑΡΤΗΜΑ

Script 1: Take_away.pl

```
#!/usr/bin/perl -w

( @ARGV==4 ) or die "Usage Take_away.pl file.temps file.dat number number\n";

open (INPUT ,$ARGV[0]) or die "Unable to open $ARGV[0]\n";
open (OUTPUT , ">$ARGV[1]") or die "Unable to open $ARGV[1]\n";

while ($line = <INPUT>)
{
  if ( $line =~ /\d+\s(\d\d\d\W\d+)\s(\d\W\d+)/ )   {
    if ( $1 <= 530.9874 && $1 >= 276.9265)
    {
      if ($2 <= $ARGV[3] && $2 >= $ARGV[2] )
      {
        print OUTPUT "$line";
      }
    }
  }
}

close (OUTPUT);
close (INPUT);
```

Script 2: perQ.sh

```
#!/bin/bash

./Take_away1.pl Frame_Temp_Q_PC12345.dat File10.dat 0.4900 0.4970
./Take_away1.pl Frame_Temp_Q_PC12345.dat File11.dat 0.4970 0.5040
./Take_away1.pl Frame_Temp_Q_PC12345.dat File12.dat 0.5040 0.5110
./Take_away1.pl Frame_Temp_Q_PC12345.dat File13.dat 0.5110 0.5180
./Take_away1.pl Frame_Temp_Q_PC12345.dat File14.dat 0.5180 0.5250
./Take_away1.pl Frame_Temp_Q_PC12345.dat File15.dat 0.5250 0.5320
./Take_away1.pl Frame_Temp_Q_PC12345.dat File16.dat 0.5320 0.5390
./Take_away1.pl Frame_Temp_Q_PC12345.dat File17.dat 0.5390 0.5460
./Take_away1.pl Frame_Temp_Q_PC12345.dat File18.dat 0.5460 0.5530
./Take_away1.pl Frame_Temp_Q_PC12345.dat File19.dat 0.5530 0.5600
./Take_away1.pl Frame_Temp_Q_PC12345.dat File20.dat 0.5600 0.5670
./Take_away1.pl Frame_Temp_Q_PC12345.dat File21.dat 0.5670 0.5740
./Take_away1.pl Frame_Temp_Q_PC12345.dat File22.dat 0.5740 0.5810
./Take_away1.pl Frame_Temp_Q_PC12345.dat File23.dat 0.5810 0.5880
./Take_away1.pl Frame_Temp_Q_PC12345.dat File24.dat 0.5880 0.5950
./Take_away1.pl Frame_Temp_Q_PC12345.dat File25.dat 0.5950 0.6020
```

```
./Take_away1.pl Frame_Temp_Q_PC12345.dat File26.dat 0.6020 0.6090
./Take_away1.pl Frame_Temp_Q_PC12345.dat File27.dat 0.6090 0.6160
./Take_away1.pl Frame_Temp_Q_PC12345.dat File28.dat 0.6160 0.6230
./Take_away.pl Frame_Temp_Q_PC12345.dat File29.dat 0.6230 0.6300
```

Script 3: Take_away1.pl

```
#!/usr/bin/perl -w

( @ARGV==4 ) or die "Usage Take_away.pl file.temps file.dat number number\n";

open (INPUT ,$ARGV[0]) or die "Unable to open $ARGV[0]\n";
open (OUTPUT , ">$ARGV[1]") or die "Unable to open $ARGV[1]\n";

while ($line = <INPUT>)
{
  if ( $line =~ /\d+\s(\d\d\d\W\d+)\s(\d\W\d+)/ )   {
    if ( $1 <= 530.9874 && $1 >= 276.9265)
    {
      if ($2 < $ARGV[3] && $2 >= $ARGV[2] )
      {
        print OUTPUT "$line";
      }
    }
  }
}

close (OUTPUT);
close (INPUT);
```

Script 4: pick.sh

```
#!/bin/bash

echo Pick a \file you want to work with.

read pick

~/Ptuxiaki/Pick_T1.pl $pick.dat T1$pick.dat
~/Ptuxiaki/Pick_T2.pl $pick.dat T2$pick.dat
~/Ptuxiaki/Pick_T3.pl $pick.dat T3$pick.dat
~/Ptuxiaki/Pick_T4.pl $pick.dat T4$pick.dat

echo Done\!
```

Script 5: Pick_T1.pl

```
#!/usr/bin/perl -w

( @ARGV==2 ) or die "Usage Take_away.pl file.temps file.dat\n";

open (INPUT ,$ARGV[0]) or die "Unable to open $ARGV[0]\n";
open (OUTPUT , ">$ARGV[1]") or die "Unable to open $ARGV[1]\n";

while ($line = <INPUT>)
{
  if ( $line =~ /\d+\s(\d\d\d\W\d+)\s(\d\W\d+)/ )
  {
    if ( $1 <= 530.9874 && $1 > 467.4721)
    {
      print OUTPUT "$line";
    }
  }
}

close (OUTPUT);
close (INPUT);
```

Script 6: Pick_T2.pl

```
#!/usr/bin/perl -w

( @ARGV==2 ) or die "Usage Take_away.pl file.temps file.dat\n";

open (INPUT ,$ARGV[0]) or die "Unable to open $ARGV[0]\n";
open (OUTPUT , ">$ARGV[1]") or die "Unable to open $ARGV[1]\n";

while ($line = <INPUT>)
{
  if ( $line =~ /\d+\s(\d\d\d\W\d+)\s(\d\W\d+)/ )
  {
    if ($1 <= 467.4721 && $1 > 403.9569)
    {
      print OUTPUT "$line";
    }
  }
}
```

```
close (OUTPUT);
close (INPUT);
```

Script 7: Pick_T3.pl

```
#!/usr/bin/perl -w

( @ARGV==2 ) or die "Usage Take_away.pl file.temps file.dat\n";

open (INPUT ,$ARGV[0]) or die "Unable to open $ARGV[0]\n";
open (OUTPUT , ">$ARGV[1]") or die "Unable to open $ARGV[1]\n";

while ($line = <INPUT>)
{
  if ( $line =~ /\d+\s(\d\d\d\W\d+)\s(\d\W\d+)/ )
  {
    if ($1 <= 403.9569 && $1 > 340.4417)
    {
      print OUTPUT "$line";
    }
  }
}

close (OUTPUT);
close (INPUT);
```

Script 8: Pick_T4.pl

```
#!/usr/bin/perl -w

( @ARGV==2 ) or die "Usage Take_away.pl file.temps file.dat\n";

open (INPUT ,$ARGV[0]) or die "Unable to open $ARGV[0]\n";
open (OUTPUT , ">$ARGV[1]") or die "Unable to open $ARGV[1]\n";

while ($line = <INPUT>)
{
  if ( $line =~ /\d+\s(\d\d\d\W\d+)\s(\d\W\d+)/ )
  {
    if ($1 <= 340.4417 && $1 >= 276.9265)
    {
```

```
        print OUTPUT "$line";
      }
   }
}

close (OUTPUT);
close (INPUT);
```

Script 9: select.pl

```
#!/usr/bin/perl -w

( @ARGV==3 ) or die "Usage Take_away.py file.temps file.dat\n";

open (INPUT ,$ARGV[0]) or die "Unable to open $ARGV[0]\n";
open (OUTPUT , ">$ARGV[1]") or die "Unable to open $ARGV[1]\n";

$i=0;
while ($line = <INPUT>)
{
 $i++;
 {
  if ( $i%$ARGV[2]== 1 )
  {
   print OUTPUT "$line";
  }
 }
}

close (OUTPUT);
close (INPUT);
```

Script 10: cluster.sh

```
#!/bin/bash
#!/usr/bin/Rscript

R --no-save <<RSCRIPT

pca<-read.table("File1_PC.dat")
library(cluster)
library(factoextra)
pdf("File1.pdf")
```

```
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()

pca<-read.table("File2_PC.dat ")
library(cluster)
library(factoextra)
pdf("File2.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()

pca<-read.table("File3_PC.dat ")
library(cluster)
library(factoextra)
pdf("File3.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()

pca<-read.table("File4_PC.dat")
library(cluster)
library(factoextra)
pdf("File4.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()
```

```
pca<-read.table("File5_PC.dat")
library(cluster)
library(factoextra)
pdf("File5.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()

pca<-read.table("File6_PC.dat")
library(cluster)
library(factoextra)
pdf("File6.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()

pca<-read.table("File7_PC.dat")
library(cluster)
library(factoextra)
pdf("File7.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()

pca<-read.table("File8_PC.dat")
library(cluster)
library(factoextra)
pdf("File8.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
```

```
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()

pca<-read.table("File9_PC.dat")
library(cluster)
library(factoextra)
pdf("File9.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()

pca<-read.table("File10_PC.dat")
library(cluster)
library(factoextra)
pdf("File10.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()

pca<-read.table("File11_PC.dat")
library(cluster)
library(factoextra)
pdf("File11.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()

pca<-read.table("File12_PC.dat")
library(cluster)
library(factoextra)
pdf("File12.pdf")
```

```
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()

pca<-read.table("File13_PC.dat")
library(cluster)
library(factoextra)
pdf("File13.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()

pca<-read.table("File14_PC.dat")
library(cluster)
library(factoextra)
pdf("File14.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()

pca<-read.table("File15_PC.dat")
library(cluster)
library(factoextra)
pdf("File15.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()
```

```
pca<-read.table("File16_PC.dat")
library(cluster)
library(factoextra)
pdf("File16.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()

pca<-read.table("File17_PC.dat")
library(cluster)
library(factoextra)
pdf("File17.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()

pca<-read.table("File18_PC.dat")
library(cluster)
library(factoextra)
pdf("File18.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()

pca<-read.table("File19_PC.dat")
library(cluster)
library(factoextra)
pdf("File19.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
```

```
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()

pca<-read.table("File20_PC.dat")
library(cluster)
library(factoextra)
pdf("File20.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()

pca<-read.table("File21_PC.dat")
library(cluster)
library(factoextra)
pdf("File21.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()

pca<-read.table("File22_PC.dat")
library(cluster)
library(factoextra)
pdf("File22.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()

pca<-read.table("File23_PC.dat")
library(cluster)
library(factoextra)
pdf("File23.pdf")
```

```
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()

pca<-read.table("File24_PC.dat")
library(cluster)
library(factoextra)
pdf("File24.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()

pca<-read.table("File25_PC.dat")
library(cluster)
library(factoextra)
pdf("File25.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()

pca<-read.table("File26_PC.dat")
library(cluster)
library(factoextra)
pdf("File26.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()
```

```
pca<-read.table("File27_PC.dat")
library(cluster)
library(factoextra)
pdf("File27.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()

pca<-read.table("File28_PC.dat")
library(cluster)
library(factoextra)
pdf("File28.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()

pca<-read.table("File29_PC.dat")
library(cluster)
library(factoextra)
pdf("File29.pdf")
wss<-(nrow(pca)-1)*sum(apply(pca,2,var))
for(i in 2:15) wss[i]<-sum(kmeans(pca,centers=i)\$withinss)
plot(1:15,wss,type="b",xlab="Number of clusters",ylab="Within groups sum of squares")
fviz_nbclust(pca,kmeans,method="silhouette")
fviz_nbclust(pca,pam,method="silhouette")
dev.off()
q()
n
RSCRIPT
```

Script 11: plot_dcd.sh

```
#!/bin/bash
#!/usr/bin/Rscript

R --no-save << RSCRIPT

pca<-read.table("File1_PC.dat ")
d<-dist(pca)
fit<-hclust(d)
png("File1.png")
plot(fit)
rect.hclust(fit,k=7)
dev.off()
groups<-cutree(fit,k=7)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File1.dat",sep=" ")

pca<-read.table("File2_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File2.png")
plot(fit)
rect.hclust(fit,k=8)
dev.off()
groups<-cutree(fit,k=8)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File2.dat",sep=" ")

pca<-read.table("File3_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File3.png")
plot(fit)
rect.hclust(fit,k=6)
dev.off()
groups<-cutree(fit,k=6)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File3.dat",sep=" ")

pca<-read.table("File4_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File4.png")
```

```
plot(fit)
rect.hclust(fit,k=6)
dev.off()
groups<-cutree(fit,k=6)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File4.dat",sep=" ")

pca<-read.table("File5_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File5.png")
plot(fit)
rect.hclust(fit,k=6)
dev.off()
groups<-cutree(fit,k=6)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File5.dat",sep=" ")

pca<-read.table("File6_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File6.png")
plot(fit)
rect.hclust(fit,k=7)
dev.off()
groups<-cutree(fit,k=7)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File6.dat",sep=" ")

pca<-read.table("File7_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File7.png")
plot(fit)
rect.hclust(fit,k=7)
dev.off()
groups<-cutree(fit,k=7)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File7.dat",sep=" ")

pca<-read.table("File8_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File8.png")
```

```
plot(fit)
rect.hclust(fit,k=6)
dev.off()
groups<-cutree(fit,k=6)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File8.dat",sep=" ")

pca<-read.table("File9_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File9.png")
plot(fit)
rect.hclust(fit,k=6)
dev.off()
groups<-cutree(fit,k=6)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File9.dat",sep=" ")

pca<-read.table("File10_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File10.png")
plot(fit)
rect.hclust(fit,k=5)
dev.off()
groups<-cutree(fit,k=5)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File10.dat",sep=" ")

pca<-read.table("File11_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File11.png")
plot(fit)
rect.hclust(fit,k=5)
dev.off()
groups<-cutree(fit,k=5)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File11.dat",sep=" ")

pca<-read.table("File12_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File12.png")
```

```
plot(fit)
rect.hclust(fit,k=5)
dev.off()
groups<-cutree(fit,k=5)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File12.dat",sep=" ")

pca<-read.table("File13_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File13.png")
plot(fit)
rect.hclust(fit,k=5)
dev.off()
groups<-cutree(fit,k=5)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File13.dat",sep=" ")

pca<-read.table("File14_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File14.png")
plot(fit)
rect.hclust(fit,k=4)
dev.off()
groups<-cutree(fit,k=4)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File14.dat",sep=" ")

pca<-read.table("File15_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File15.png")
plot(fit)
rect.hclust(fit,k=4)
dev.off()
groups<-cutree(fit,k=4)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File15.dat",sep=" ")

pca<-read.table("File16_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File16.png")
```

```
plot(fit)
rect.hclust(fit,k=4)
dev.off()
groups<-cutree(fit,k=4)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File16.dat",sep=" ")

pca<-read.table("File17_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File17.png")
plot(fit)
rect.hclust(fit,k=4)
dev.off()
groups<-cutree(fit,k=4)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File17.dat",sep=" ")

pca<-read.table("File18_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File18.png")
plot(fit)
rect.hclust(fit,k=5)
dev.off()
groups<-cutree(fit,k=5)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File18.dat",sep=" ")

pca<-read.table("File19_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File19.png")
plot(fit)
rect.hclust(fit,k=4)
dev.off()
groups<-cutree(fit,k=4)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File19.dat",sep=" ")

pca<-read.table("File20_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File20.png")
```

```
plot(fit)
rect.hclust(fit,k=5)
dev.off()
groups<-cutree(fit,k=5)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File20.dat",sep=" ")

pca<-read.table("File21_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File21.png")
plot(fit)
rect.hclust(fit,k=5)
dev.off()
groups<-cutree(fit,k=5)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File21.dat",sep=" ")

pca<-read.table("File22_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File22.png")
plot(fit)
rect.hclust(fit,k=5)
dev.off()
groups<-cutree(fit,k=5)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File22.dat",sep=" ")

pca<-read.table("File23_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File23.png")
plot(fit)
rect.hclust(fit,k=5)
dev.off()
groups<-cutree(fit,k=5)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File23.dat",sep=" ")

pca<-read.table("File24_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File24.png")
```

```
plot(fit)
rect.hclust(fit,k=4)
dev.off()
groups<-cutree(fit,k=4)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File24.dat",sep=" ")

pca<-read.table("File25_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File25.png")
plot(fit)
rect.hclust(fit,k=4)
dev.off()
groups<-cutree(fit,k=4)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File25.dat",sep=" ")

pca<-read.table("File26_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File26.png")
plot(fit)
rect.hclust(fit,k=4)
dev.off()
groups<-cutree(fit,k=4)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File26.dat",sep=" ")

pca<-read.table("File27_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File27.png")
plot(fit)
rect.hclust(fit,k=4)
dev.off()
groups<-cutree(fit,k=4)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File27.dat",sep=" ")

pca<-read.table("File28_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File28.png")
```

```
plot(fit)
rect.hclust(fit,k=5)
dev.off()
groups<-cutree(fit,k=5)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File28.dat",sep=" ")

pca<-read.table("File29_PC.dat")
d<-dist(pca)
fit<-hclust(d)
png("File29.png")
plot(fit)
rect.hclust(fit,k=4)
dev.off()
groups<-cutree(fit,k=4)
tab<-data.frame(pca,groups)
write.table(tab,file="Groups_File29.dat",sep=" ")
q()
n
RSCRIPT

for i in {0..28}
do
echo Choose the \file you want to work with.

read pick

awk '{print $1}' $pick.dat > Frames_$pick.dat
awk '{print $7}' Groups_$pick.dat > tmpfile.dat
sed '1d' tmpfile.dat > tmpfile1.dat
mv tmpfile1.dat Clusters_$pick.dat

./converge.pl Frames_$pick.dat Clusters_$pick.dat Frames_Clusters_$pick.dat

echo Write the numbers of clusters \(need eight numbers\). \If there are \less numbers
put zeros.

read p1 p2 p3 p4 p5 p6 p7 p8

if [ $p1 -eq 0 ]
then
   echo There must be at least one cluster\!
        exit 0
else
```

```
   ./Extract_clusters.pl Frames_Clusters_$pick.dat $pick"_"$p1.dat $p1
        carma -v -sort $pick"_"$p1.dat cln025_adapt_STAR_ILDN_COMPLETE.dcd
        mv carma.reordered.dcd carma_$pick"_"1.dcd
fi

if [ $p2 -eq 0 ]
then
   echo There was only one cluster and the job is done\!
        exit 0
else
   ./Extract_clusters.pl Frames_Clusters_$pick.dat $pick"_"$p2.dat $p2
        carma -v -sort $pick"_"$p2.dat cln025_adapt_STAR_ILDN_COMPLETE.dcd
        mv carma.reordered.dcd carma_$pick"_"2.dcd
fi

if [ $p3 -eq 0 ]
then
   echo There were only two clusters and the job is done\!
        exit 0
else
   ./Extract_clusters.pl Frames_Clusters_$pick.dat  $pick"_"$p3.dat $p3
        carma -v -sort $pick"_"$p3.dat cln025_adapt_STAR_ILDN_COMPLETE.dcd
        mv carma.reordered.dcd carma_$pick"_"3.dcd
fi

if [ $p4 -eq 0 ]
then
   echo There were only three clusters and the job is done\!
        exit 0
else
   ./Extract_clusters.pl Frames_Clusters_$pick.dat $pick"_"$p4.dat $p4
        carma -v -sort $pick"_"$p4.dat cln025_adapt_STAR_ILDN_COMPLETE.dcd
        mv carma.reordered.dcd carma_$pick"_"4.dcd
fi

if [ $p5 -eq 0 ]
then
   echo There were only four clusters and the job is done\!
        exit 0
else
   ./Extract_clusters.pl Frames_Clusters_$pick.dat $pick"_"$p5.dat $p5
        carma -v -sort $pick"_"$p5.dat cln025_adapt_STAR_ILDN_COMPLETE.dcd
        mv carma.reordered.dcd carma_$pick"_"5.dcd
fi
```

```
if [ $p6 -eq 0 ]
then
   echo There were only five clusters and the job is done\!
        exit 0
else
   ./Extract_clusters.pl Frames_Clusters_$pick.dat $pick"_"$p6.dat $p6
        carma -v -sort $pick"_"$p6.dat cln025_adapt_STAR_ILDN_COMPLETE.dcd
        mv carma.reordered.dcd carma_$pick"_"6.dcd
fi

if [ $p7 -eq 0 ]
then
   echo There were only six clusters and the job is done\!
        exit 0
else
   ./Extract_clusters.pl Frames_Clusters_$pick.dat $pick"_"$p7.dat $p7
        carma -v –sort $pick"_"$p7.dat cln025_adapt_STAR_ILDN_COMPLETE.dcd
        mv carma.reordered.dcd carma_$pick"_"7.dcd
fi

if [ $p8 -eq 0 ]
then
   echo There were only seven clusters and the job is done\!
        exit 0
else
   ./Extract_clusters.pl Frames_Clusters_$pick.dat $pick"_"$p8.dat $p8
        carma -v -sort $pick"_"$p8.dat cln025_adapt_STAR_ILDN_COMPLETE.dcd
        mv carma.reordered.dcd carma_$pick"_"8.dcd
fi
done
```

Script 12: converge.pl

```
#!/usr/bin/perl -w

( @ARGV==3 ) or die "Usage Take_away.pl file.temps file.dat\n";

open (FILE1 ,$ARGV[0]) or die "Unable to open $ARGV[0]\n";
open (FILE2 , $ARGV[1]) or die "Unable to open $ARGV[1]\n";
open (OUTPUT , ">$ARGV[2]") or die "Unable to open $ARGV[2]\n";

while ($line1 = <FILE1>, $line2 = <FILE2>)
 {
```

```
  chomp($line1);
  print OUTPUT "$line1 $line2";
 }

close(FILE1);
close(FILE2);
close(OUTPUT);
```

Script 13: Extract_clusters.pl

```
#!/usr/bin/perl -w

( @ARGV==3 ) or die "Usage Take_away.pl file.temps file.dat\n";
open (INPUT ,$ARGV[0]) or die "Unable to open $ARGV[0]\n";
open (OUTPUT , ">$ARGV[1]") or die "Unable to open $ARGV[1]\n";

while ($line = <INPUT>)
{
  if ( $line =~ /(\d+)\s(\d)/ )
  {
    if( $2 == $ARGV[2])
    {
      print OUTPUT "$1 \n";
    }
  }
}

close(INPUT);
close(OUTPUT);
```

Script 14: pdb.sh

```
#!/bin/bash

for i in {0..28}
do
echo Choose the cluster you want to work with.

read pick

carma –v –pdb –atmid ALLID cln025_pseudo.psf carma_$pick.pdb
cat carma*.pdb >> $pick.pdb
rm carma*.pdb
```