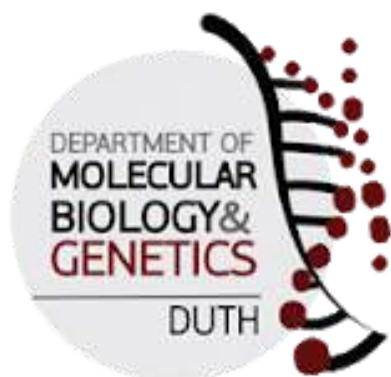




ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ
DEMOCRITUS UNIVERSITY OF THRACE

Democritus University of Thrace



Bachelor Thesis

“Folding simulation of a ftz-derived peptide in a TFE-water solution and NOE analysis”

— by Arvaniti Olympia-Constantina

a project supervised by Dr Nicolaos M. Glykos


Alexandroupolis, 5th July, 2019



Contents

1.	Acknowledgements	2
2.	Abstract	3
3.	Introduction	6
3.1.	The purpose of thesis	6
3.2.	Molecular Dynamics Simulations in biomolecules	6
3.3.	Nuclear Receptors and LxxLL motifs	7
3.4.	FTZ ^{pep} and FTZ-F1 receptor	8
3.5.	Solvents - 2,2,2-Trifluoroethanol (TFE)	10
3.6.	Nuclear Overhauser Effect (NOE)	11
3.7.	Pearson's Coefficient Correlation	11
4.	Methods	13
4.1.	Simulation Protocol - Preparation of Computational System	13
4.2.	Trajectory Analysis	13
4.3.	System Mechanics	14
5.	Results	15
5.1.	Experimental and Simulation Agreement	15
5.2.	Correlation Matrices	20
6.	Conclusion	28
7.	Bibliography	29

Acknowledgments

I would like to thank my supervisor Dr. Nikolaos M. Glykos for helping me with the devising of my thesis concept, his willingful contribution to my academic progression and his insistence on refusing answer all my questions, making the phrase “Go figure.” the echo in every new question that puzzled me during this deed. Moreover, I would like to heartily thank him for his forbearance and sufferance in front of the anxiety and cyclothymia that i showed the past year due to serious personal issues. I thank my family, boyfriend and friends for all the support in every level. Thank you .

Abstract

Key words: Molecular Dynamics, force fields, AMBER99SB-STAR-ILDN, FTZ^{pep}, fushi tarazu, TFE, NOEs, Pearson's correlation coefficient.

The study of coactivator peptide FTZ^{pep} that participates with the orphan nuclear receptor FTZ-F1 in the regulation circuit of the pair-rule gene Fushi tarazu expressed in *Drosophila Melanogaster* embryogenesis is conducted via molecular simulation in order to crosscheck the existent NMR experimental evidence by *Weontae Lee et al.(2012)*, extend the knowledge about peptide spatial conformation and behaviour in a water-TFE (50%) mixture and correlate the components and parameters of the system to extract information about FTZ^{pep} and the puzzling cosolvent TFE. For the comparison between the NMR and molecular simulation findings were calculated and analysed NOEs and for the correlation were created matrices compounded by Pearson's correlation coefficient, r . The results appear to be quite favorable as the simulation data suggest computational and experimental agreement as it provides some crucial notes about system and the preferences of force field. Furthermore, the observation of correlation matrices proved fruitful indicating important correlation between the peptide and TFE contacts with the TFE rich aggregation regions in the solution. Together with previous bibliographic references that prove TFE promotes secondary structure formation, it may be evidence that these aggregation regions are the key in TFE role translated into function.

Περίληψη

Στην παρακείμενη μελέτη πρωταγωνιστεί το συν-ενεργοποιητικό πεπτιδίο FTZ^{pep}, το οποίο συμμετέχει με τον ορφανό πυρηνικό υποδοχέα FTZ-F1 στο κύκλωμα ρύθμισης του pair-rule γονιδίου Fushi tarazu κατά την εμβρυογένεση στη *Drosophila melanogaster*. Αποτελέσματα προερχόμενα από μοριακή προσομοίωση διασταυρώθηκαν με τα υπάρχοντα πειραματικά δεδομένα NMR εντύπως διανεμημένα από τον Weontae Lee (2012), προκειμένου να επεκταθούν οι γνώσεις σχετικά με τη χωρική αναδίπλωση του πεπτιδίου και τη συμπεριφορά του σε ένα μείγμα νερού-TFE (50%) καθώς και να βρεθεί η συσχέτιση των διαφόρων συστατικών και παραμέτρων του συστήματος ώστε να απομυζηθούν όσο το δυνατόν περισσότερες ουσιαστικές πληροφορίες σχετικά με το FTZ^{pep} και τον αιγιματικό συνδιαλύτη TFE. Για τη σύγκριση μεταξύ δεδομένων NMR και μοριακής προσομοίωσης, υπολογίστηκαν και αναλύθηκαν NOEs, ενώ για το συσχετισμό δημιουργήθηκαν πίνακες με τιμές συντελεστή συσχέτισης Pearson, r . Τα αποτελέσματα φαίνεται ότι είναι αρκετά ευνοϊκά, καθώς τα δεδομένα προσομοίωσης υποδηλώνουν υπολογιστική και πειραματική συμφωνία, ενώ παράλληλα, παρέχονται κάποιες κρίσιμες σημειώσεις για το σύστημα και τις προτιμήσεις του πεδίου. Επιπλέον, η παρατήρηση των αποτελεσμάτων οδήγησε σε χρήσιμα συμπεράσματα, υποδεικνύοντας σημαντική συσχέτιση μεταξύ του αριθμού επαφών ανάμεσα σε TFE-πεπτιδίο και τις πλούσιες σε TFE περιοχές συσσωμάτωσης στο διάλυμα. Μαζί με προηγούμενες βιβλιογραφικές αναφορές που υποδεικνύουν ότι το TFE προωθεί τη δημιουργία δευτερογενούς δομής, μπορεί να είναι μια ιδέα ότι αυτές οι περιοχές συσσωμάτωσης είναι το κλειδί στο ρόλο του TFE που μεταφράζεται σε λειτουργία.

Introduction

3.1. The purpose of thesis

The concept of the present thesis originated from the paper of Weontae Lee and collaborators with topic the NMR structural analysis of an LxxLL motif in a coactivator peptide derived by FTZ and interact with FTZ-F1 nuclear receptor¹. The project of the forementioned scientists contained examination of the peptide (FTZ^{pep}) in three different solutions, a water solution, a mixture of water and TFE 50% (v/v) and water solution in complex with the cognate NR FTZ-F1. In ensuing studies by the supervisor of our department and previous fellow undergraduates, a FTZ^{pep} molecular simulation occurred and analyses including calculation of NOEs in comparison with the NMR-derived results. The study resulted in the discovery of evidence that the LxxLL motif persistently preference in helical conformation despite the presence of NR or the solution nature and the helical bias of the whole FTZ^{pep} in a water-TFE solution². A helical secondary structure of this peptide is noticed in complex with FTZ-F1 also. The question comes in mind, associated with the unclear mechanisms that TFE effect peptides, was whether it exists some correlation between TFE and the peptide's formation, and a correlation of other relative parameters as well. The term "relative parameters" is referred to factors such as the individual contacts between: water molecules, TFE molecules, TFE and water molecules, the peptide and water as well peptide with TFE molecules, temperature and HGI%.

3.2. Molecular Dynamics Simulations in biomolecules

The molecular structure and the internal motions cause conformational transitions and thus affect the tertiary structure of biomolecules which is associated directly to their function. The comprehension and prediction of function -especially in proteins- is one of the main problems in study and interpretation of their systemic role in the organism. Molecular dynamics is nowadays a well-known method to approach theoretically the nature of such motions and provide insight to protein models whose tertiary structure and function would be difficult to happen experimentally. Theoretical study of the conformational behaviour of peptides in association with environmental conditions as intramolecular interactions, time and nature of system solvent can be combined with classical experimental data to create a whole picture of the protein structure and function. The essence of theoretical and experimental approach is a useful evaluation technique and a "corrective" mean of computational methods in order to minimize divergents from native protein, especially when some methods have empirical character like force fields.^{3,4}

3.3. Nuclear Receptors and LxxLL motifs

Nuclear receptors are ligand-regulated transcription factors. Non-polar regulatory molecules (like steroid hormones, thyroid hormone, retinoic acid, and oxysterols) are able to diffuse across the plasma membrane and interact with the intracellular nuclear receptor directly by binding to the promoters of target genes and regulating the rate of transcription. Nuclear receptors whose endogenous ligands and function are unknown are termed as *orphan receptors* while orphan receptors with later identification of their corresponding ligand are called *adopted orphan*. In addition, the term *atypical* orphan receptor denotes a more recent superfamily of orphan receptors due to the divergence in conserved domains. This type of nuclear receptors is appeared to be one of the most primitive⁵. Specifically, despite the fact that similarity in many cases is lower, these receptors do not lack the conserved domains (see below) but instead they involve more unconventional versions of them.

Regarding the modular structure of the receptors, sequence and functional analyses indicate a peculiar type of formation with some highly conserved components. In this template are included a variable N-terminal and a C-terminal ligand-binding domain (LBD) interposed from a conserved DNA-binding domain (DBD).⁵ Binding of the ligand to the receptor triggers immediate activation of accessory proteins, the coactivator and the repressors. The first stage of this procedure is the cease of repression that lead in gene silencing; this is accomplished when the ligand binding in LBD changes the conformation of the receptor. This fact permits the release of the corepressor who was bound in the receptor and consequently allows the association of coactivators that lead to the activating path. The interaction of coactivator proteins and the NRs comprises a conserved peptide sequence, a α -helical LxxLL motif which is common and necessary in such protein-protein interactions which mediate the binding of these proteins to the complex of ligand-receptor. The succession of all these interactions regulates the transcription of genes which participate in varying scale of biological processes.^{1, 6, 7}

As previously mentioned, the LxxLL motif has a α -helical structure and appears to be persistent in adopting this conformation even when the cognate nuclear receptor is absent. Two types of interactions are involved in the recognition of LxxLL motif by the NR. Firstly, the hydrogen bonds between the specific charged amino acid residues. In the “charged clamp” of NR’s LBD region are located two highly conserved amino residues that form hydrogen bonds with the hydrogen of the coactivator protein’s LxxLL polypeptide backbone. Secondly, hydrophobic interactions among LxxLL residues and the hydrophobic residues inside the spacious core surface of NR.^{7, 8, 9, 10}

3.4. FTZ^{pep} and FTZ-F1 receptor

Fushi tarazu(ftz) is a pair-rule gene which is necessary for the formation of the boundaries in the alternating parasegments in early *Drosophila melanogaster* embryogenesis. It is expressed in seven stripes and comprises a requirement for the normal development of *Drosophila* since its mutation leads to an embryo with only half the number of segments and for ecdysone-triggered metamorphosis at later stages of development. FTZ-F1 and the gene product FTZ are essential for the regulation of ftz gene.^{11,12}

FTZ-F1 is an orphan nuclear receptor. It acts as a monomer and its structure has also the two conserved regions, DBD and LBD. The DBD consists of two zinc fingers and it does not contact DNA directly, consequently mutation in the respective helix does not result to elimination of DNA binding but disrupt the procedure to some degree that reduced transcriptional activity is presented. LBD is located in the C-terminal. FTZ interacts with FTZ-F1 through an LxxLL motif of its small peptide product, FTZ^{pep} (Figure). In the presence of ligand, performs regulatory functions as coactivator. Evidence that nuclear receptors of FTZ-F1 family have in addition altered phospholipid ligands, lay the idea that non-canonical interactions of FTZ-F1 family with transcription factors may have restrictive role for the cellular specific function of other development nuclear receptors.^{13,14}

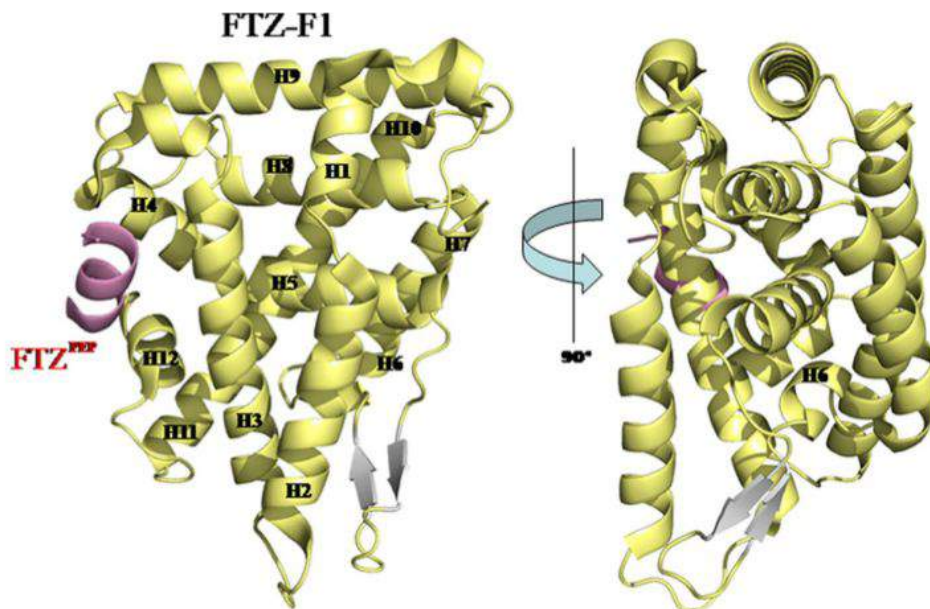


Figure.1 Crystal structure of the FTZ-F1 LBD in complex with its cofactor, FTZ^{pep}. The ribbon representations of the FTZ-F1 LBD in two views are displaced by 90°. FTZ^{pep} is shown in pink.¹⁵

FTZ-F1 is evenly expressed in blastoderm when FTZ is only detected in the stripe of the seven even-numbered parasegments as it is presented in Figure below.

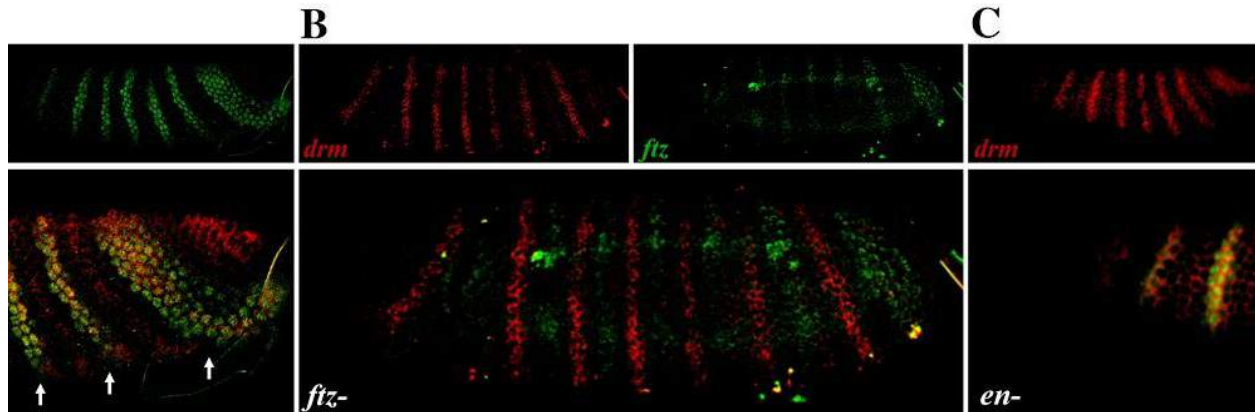


Figure.2 Ftz is required for the expression of primary *drm** stripes independent of *en*. Confocal (A,C) or *ftz* RNA (B). (A) FTZ and *drm* overlap in the primary *drm* stripes.: *drm* RNA (red), FTZ protein (green). (B) Primary *drm* stripes are lost in *ftz* mutants.: *drm* RNA (red), *ftz* RNA (green). The remaining *drm* stripes in the *ftz* mutant embryo are out of register with the *ftz* stripes. (C) Primary *drm* stripes do not require *en*.: *drm* RNA (red), FTZ protein (green). No change in the FTZ-dependent *drm* stripes was observed in *en*. mutants.¹⁶

*drm** is regulated by FTZ and FTZ-F1 and encodes a zinc finger transcription factor involved in differentiation morphogenesis and cell movement during gut morphogenesis

The peptide in this molecular simulation is the nineteen amino-residues segment of FTZ, FTZ^{pep} (P02835 in Uniprot), that corresponds to the 102-120 of the whole protein lengths and respectively has the sequence VEERPST**LRALL**TNPVKKL. The central bolded residues correspond to LxxLL motif and in the following **Figure.3** is used the adjacent numbering.



Figure.3 The LxxLL motif of FTZ^{pep}.

The first leucine in the motif is numbered as +1, the arginine and alanine that constitute the "xx" in the motif are referred as +2 and +3 respectively, when the previous from +1 leucine

residue -in this case threonine- is numbered as -1. In this way the first residue of the sequence is referred as -9 and the last as +12. Number zero is skipped from numbering for convenience.

3.5. Solvents - 2,2,2-trifluoroethanol solvent (TFE)

An essential step in all biomolecular experiments is the devising of an effective solution system. This may be a single solvent system, for example a water solution system, or be composed of a solvent-cosolvent(s) mixture. Due to the presence of cosolvent, the last is prevalent in experimental studies of proteins for it is presenting a variety of effects that are desired. Among these effects are stability increase, increase or decrease of solubility, denaturation and secondary structure formation.^{7, 17-19}

A noted cosolvent is 2,2,2-trifluoroethanol or more casually *TFE* (Fig.1). TFE is categorized in highly fluorinated alcohols and comprises one of the cheapest, most commonly used, and large-scale commercially available solvent of its kind. Fluorinated alcohols in comparison with other non-fluorinated analogue solvents of this category, exhibit higher melting points and lower boiling points• for example TFE has a Mp=-43.5 °C and Bp=74°C.²⁰ The fluorine substituents intensify the inductive effect, increasing the acidity of the hydroxyl proton. In this way, TFE molecule is a superior hydrogen bond donor but poorer as acceptor and as a solvent appears to be highly polar (TFE one of the most polar).²¹ The hydrophobicity that is presented in TFE is indicated also by the CF₃ group; though it is unclear and hard to fully characterize this property of TFE in a water/TFE mixture, according to imaging analyses (NMR spectroscopy and X-rays scattering) are observed rich in TFE regions wherein TFE molecules aggregate, a fact that is reinforced by a highly decreased methane solvation free energy in a TFE mixture than a water solution.

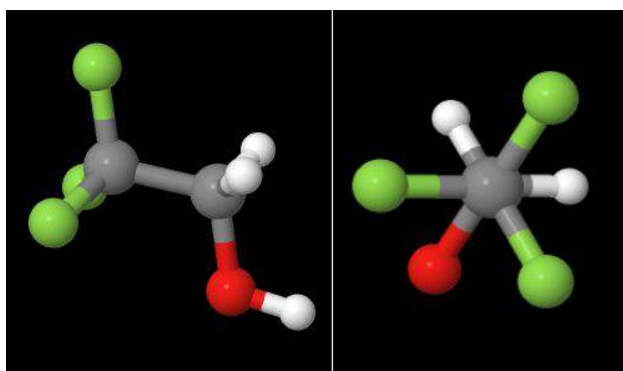


Figure.4 TFE molecule in profile and transversely. It contains a carbon backbone with two carbon atoms, three fluorine substituents, an alcohol molecule and two hydrogen substituents. With grey

colour are presented the atoms of carbon, with red the oxygens, green the fluorines and white the hydrogens.

3.6. Nuclear Overhauser Effect (NOE)

When the positive charged nuclei of hydrogen are exposed in a static constant magnetic field act like magnetic dipole with their inherent magnetic moment be identified as nuclear spin. Provided that the frequency is appropriate, in other words the oscillation frequency matches the frequency of nuclei, are revealed intermediate higher energy states that the nuclei can reach via energy absorption during excitation. The energy between two rotational states depends on the intensity of the magnetic field and the isotope identity. The nuclear spins polarise and resonate with the magnetic field, vacillates in a lower energy state emitting lower energy photons in the radio wave spectrum. In brief, this is the theory that bases the NMR which is interrelated with NOE.²²

Nuclear Overhauser Effect, that was developed in the namesake of theoretical physicist Albert Overhauser, describes within NMR the transition of a nuclear polarized spin that belong to a population of same spin nuclei to another population via relaxation from the excitation state due to its overpopulation. Qualitatively, this is translated as change the resonance intensity of two neighbour nuclei via coupling and quantitatively as the distance of the nuclei emitting signals which further leads to imaging. The signal intensity is proportional to r^6 that estimates the internuclear distance and consequently the distance of protons. In many experiments of NMR imaging it is common to use r^3 but in smaller peptide it is preferred to calculate the distances with r^6 . Likewise, in the current study was used the r^6 calculation due to the short length of FTZ^{pep}.^{22,23}

3.7. Pearson's coefficient correlation

Pearson's correlation coefficient (PCC) or commonly Pearson's r in statistics is a measure of correlation between two variables, x and y . With the term correlation is meant the association between variables in a statistical relation. In instance of Pearson's correlation coefficient the statistical relation must be linear and the r denotes the strength and direction of this relation between the variables (x & y). PCC can take values in the interval $[-1, +1]$. The strength of correlation depends on the absolute value of r and the direction of correlation on sign. Consequently, the value -1 represents total (or perfect) negative linear correlation, while the value

+1 represents the total positive correlation and 0 is the value that indicates no correlation at least in linear level. The formula for PCC calculation is:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2] - [n\sum y^2 - (\sum y)^2]}}$$

In which n is sample size, x and y are the variables and \sum symbolizes sum mathematically.

Together with Pearson's correlation coefficient there are other similar measures like Spearman correlation, Kendall's t or Cramér's V. Although the title of correlation coefficient characterizes these sizes, the fundamental difference appears to be in the type of variables. Pearson's variable are metric, in other words are variables on which calculations are meaningful, like means or standard deviations and moreover should be continuous, distributed approximately normally and without significant outlier values. In contrast, Spearman correlation, Kendall's t use ordinal variables. Ordinal are the variables that have distinct grades of order but the unit of their measurement cannot be defined, as an example is the gradient of satisfaction in rendering of services: "Good" = +1, "Neutral" = 0 and "Bad" = -1. Finally, Cramér's V bases on nominal variables, i.e. variables without undisputable order. ^{24, 25}

Methods

4.1. Simulation Protocol - Preparation of Computational System

For the dynamic folding simulation of FTZpep was utilized the program NAMD. The optimization of solvation system was followed the LEAP program from AMBER Tool distribution²⁶, the TIP3P water model and the TFE parameters of R.E.D. Library.²⁷⁻³⁰ The simulation was conducted in a cubic unit cell, under AMBER99SB-STAR-ILDN force field³¹⁻³⁴, a solution mixture of 50% TFE- 50% H₂O(v/v)²⁸ and adaptive tempering ranged from 280K 380K featured in NAMD adjustments³⁵. Adaptive tempering is a single-copy replica exchange method in order to update dynamically the simulation temperature, T.³⁶ The variable T changes values according to the simulation step in a random manner which is indicated by the equation $dE/dT = E - E(T) - 1/T + \text{sqrt}(2)T\xi$ (where ξ is Gaussian white noise) and belong to the interval $[T_{min}, T_{max}]$, i.e the range of temperature that in this simulation is [280K, 380K]. The purpose of this method is when the potential energy of the structure is lower than the current system's average energy, the temperature is increased and vice versa. When the energy is higher than the system's average energy, the temperature is decreased. This procedure lasts 1000 steps and leads to energy minimization and consequently to minimum energy structures. After that, starting from 280K and with a 20K temperature step increase the system heats to 380K. The trajectory was compiled by saving point of system coordinates with a 1.000 ps step.

4.2. Trajectory Analysis

For the analysis of the trajectory was mainly used the CARMA³⁷ program and the graphical "user-friendly" interface of CARMA, GRCARMA³⁸. With CARMA were made analyses as the calculation of atomic distances, calculation of the radius of gyration(Rg)^{37, 38}, a dihedral principal component analysis(dPCA)³⁹, the measurement of Q, Qs and q similarity values^{37, 38}. Additional programs as R program - for the Pearson's correlation coefficient matrix and heatmap, 2matrix program - for the creation of Rg-q matrix, noe_averaging_sigma-used in estimating of NOEs values and program r for r⁻⁶ averaging. For secondary structure analyses were used the programs STRIDE⁴⁰ and Weblogo⁴¹. The simulation graphics work, analyzing and processing were made using of programs VMD⁴², plot⁴³, R⁴⁴, GRCARMA, KolourPaint and Krita (free Ubuntu distribution picture editing programs).



4.3. System Mechanics

The simulation was run in a computing cluster name *Beowulf*. Beowulf is our “monster” computer that shelter in its “castle” - very bad joke for the respective tower- 40 CPU cores, unknown memory cards that add 46GB RAM memory on the system, 6 Intel Q6600 Kentsfield general purpose graphics processing unit with 2.4 GHz quad processors shared in 10 nodes. It is connected with an industrial 24G Gigabit internet switch of HP ProCurve 1800 model. This computational masterpiece is used for simulations by NMG group (scientific gaming).

Results

The results are based in two main scientific approaches. The first one is the agreement between experimental NMR evidence of previous work by *Yun et al.* and the simulation derived data at the level of structural comparison, and the second is the creation and study of correlation matrices about molecular contacts in the simulation unit cell when contains FTZ^{pep} in water-TFE 50% mixture (v/v).

5.1. Experimental and Simulation data agreement

The experimental data came from *Yun et al.* printed-only form without the NMR data publicly distributed nor the respective entries in PDB database. Fortunately, the figures of NMR NOEs upper bound were distinct enough to be retrieved ([Figure.5](#)). In the first place, the simulation derived NOEs and then the r^{-3} and r^{-6} average values were calculated, thus only the r^{-6} for better accuracy. Afterwards, a comparison with the experimental upper bounds was made, the violations of upper bound were listed and were calculated the average upper bound violation as show the [Table 2](#) and [Table 3](#) below. The results of the comparison are presented in the upcoming analysis.



[Figure.5](#) Yun et al. 50% TFE/ 40% H₂O/ 10% D₂O mixture at pH 6.5, 25 °C showing the sequential and short-range NOE contacts. The thin lines correspond in weak signal, the lines with interlaced width correspond to medium signal and the thick lines to strong signal. Gaps represent no signal

and grey lines represent over lapses. With α is encoded the H_α hydrogens and respectively, β H_β and N any H, when i represents the number of residues of the peptide.¹

For NMR and consequently for the NOE calculations were examined seven categories of protons, which are:

- $d_{\alpha N}(i, i+1)$: Observation of H_α from residue i to $H_N i+1$
- $d_{NN}(i, i+1)$: Observation of H_N from residue i to $H_N i+2$
- $d_{\alpha N}(i, i+2)$: Observation of H_α from residue i to $H_N i+2$
- $d_{\alpha N}(i, i+3)$: Observation of H_α from residue i to $H_N i+3$
- $d_{\alpha N}(i, i+4)$: Observation of H_α from residue i to $H_N i+4$
- $d_{\alpha\beta}(i, i+3)$: Observation of H_α from residue i to $H_\beta i+3$
- $d_{\beta N}(i, i+1)$: Observation of H_β from residue i to $H_N i+1$

The NMR derived signals, as seen in [Table 1.](#), appertain to three categories based on the signal strength.

Table 1.




Signal Strength	Distance interval in Å	Upper bound	Graphic representation
Strong	1.8 - 2.7	2.7	
Medium	2.7 - 3.3	3.3	
Weak	3.3 - 5.0	5.0	

Table 2.: NOE upper bound violations, TFE/Water simulation

NOE number	Proton Pairs	Residue number & Upper Bound		r^6
1	daN(i,i+1)	-6E ↔ -5E	5.0	2.411
2		-5E ↔ -4R	3.3	2.690
3		-3P ↔ -2S	3.3	2.415
4		-2S ↔ -1T	5.0	2.972
5		-1T ↔ +1L	5.0	3.360
6		+1L ↔ +2R	3.3	3.172
7		+2R ↔ +3A	3.3	3.006
8		+3A ↔ +4L	3.3	2.960
9		+4L ↔ +5L	3.3	2.938
10		+5L ↔ +6T	5.0	2.837
11		+6T ↔ +7N	5.0	2.674
12		+8P ↔ +9V	2.7	2.779
13		+9V ↔ +10K	3.3	2.722
14		+10K ↔ +11K	2.7	2.385
15		+11K ↔ +12L	2.7	2.322
16	dNN(i,i+1)	-5E ↔ -4R	5.0	2.285
17		-1T ↔ +1L	5.0	2.619
18		+1L ↔ +2R	3.3	2.488
19		+2R ↔ +3A	3.3	2.693
20		+3A ↔ +4L	3.3	2.709
21		+4L ↔ +5L	3.3	2.584
22		+5L ↔ +6T	3.3	2.661
23		+9V ↔ +10K	3.3	2.284
24		+10K ↔ +11K	3.3	2.853

25	daN(i,i+2)	-2S ↔ +1L	5.0	4.102
26		+7N ↔ +9V	3.3	4.241
27		+8P ↔ +10K	3.3	4.429
28		+9V ↔ +11K	3.3	4.752
29	daN(i,i+3)	+2R ↔ +5L	3.3	3.600
30		+3A ↔ +6T	3.3	3.784
31		+4L ↔ +7N	5.0	4.001
32		+8P ↔ +11K	5.0	5.338
33	daN(i,i+4)	-3P ↔ +2R	3.3	4.670
34		-2S ↔ +3A	5.0	4.391
35		+1L ↔ +5L	5.0	4.280
36		+2R ↔ +6T	5.0	4.437
37		+3A ↔ +7N	5.0	4.291
38	daβ(i,i+3)	-3P ↔ +1L	5.0	3.424
39		-2S ↔ +2R	3.3	2.947
40		-1T ↔ +3A	5.0	3.296
41		+1L ↔ +4L	5.0	3.060
42		+2R ↔ +5L	5.0	3.288
43		+4L ↔ +7N	5.0	3.585
44	dβ,N(i,i+1)	-5E ↔ -4R	5.0	3.529
45		-3P ↔ -2S	5.0	3.244
46		-2S ↔ -1T	5.0	3.377
47		-1T ↔ +1L	5.0	3.107
48		+1L ↔ +2R	5.0	3.123
49		+2R ↔ +3A	5.0	3.107
50		+3A ↔ +4L	5.0	3.172
51		+4L ↔ +5L	5.0	3.009
52		+5L ↔ +6T	5.0	3.292
53		+6T ↔ +7N	5.0	3.190

54		+8P ↔ +9V	3.3	3.555
55		+9V ↔ +10K	3.3	3.151
56		+11K ↔ +12L	5.0	2.954

Table 3.

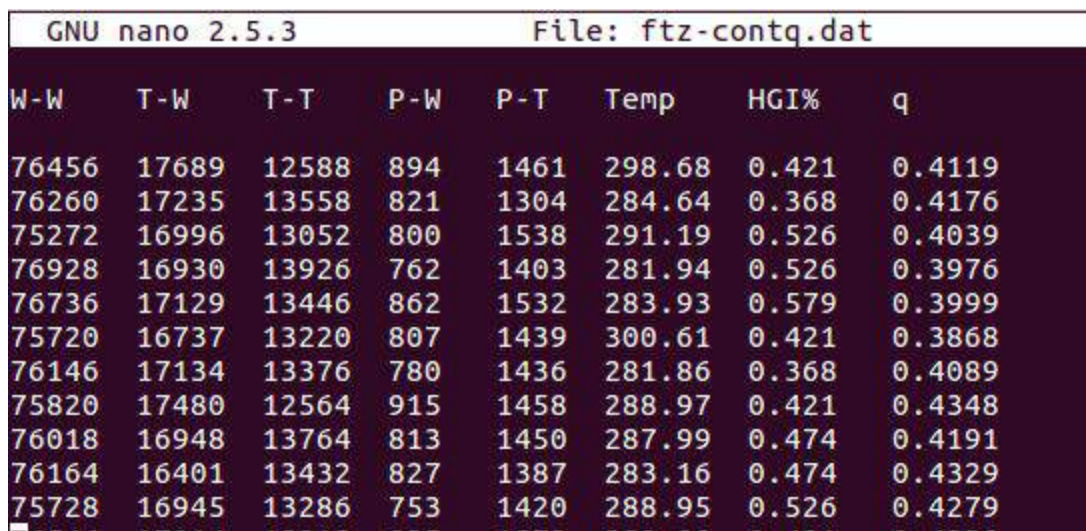
Average violation (r^6) (300K)	Number of violations	Number of protons
0.113 Å	9	56
0.054 Å [excluding daN(i,i+2)]	6	56

The NOE average upper bound violation as seen in Table 3. is 0.109 Å when in total were examined 56 NOEs and noticed 9 violations (Table 2.). This number indicates the TFE contribution in peptide's stabilization. The part of the average violation that is mostly responsible for the total increase belongs to daN(i, i+2) NOEs and specifically to the signal from residue pairs [+7N ↔ +9V], [+8P ↔ +10K], [+9V ↔ +11K] that raise the average violation number to 0.109 Å. When excluding these violations the average is only 0.054 Å. In order to perceive the significance of this variation, in water simulation with a smaller number of recorded NOEs than in TFE/water mixture, the average violation was 0.050 Å with 5 out of 46 NOEs to violate the upper bound. In TFE/water mixture, without the exclusion of daN(i, i+2) the average violation is approximately double(0.109 Å) in comparison with the respective value in water mixture, while in the excluded version is equivalent. The exclusion did not happen arbitrarily, but was led by appearance of reverse turn that promote the related residues (+7,+11), a fact that may indicate a bias of the force-field against reverse turns. This conclusion, however, is dealt with reservations due to limited data sources, yet simulation and NMR data are in agreement.

5.2. Correlation matrix

In order to examine the relation of the components both among them and to other prevailing parameters of the simulation, a study of the simulation contacts was conducted. In this study, the contacts file of the whole simulation was taken. The file contains the contacts of components in the simulation i.e. water, TFE and FTZ^{pep}, the current temperature, the current stride sequence and % content. The data are organised in lines so that each line respectively matches to a simulation frame and in columns by type of the contact pair or the calculated parameter that respects to the frame(see [Figure.6](#)). In order to be precise, an explanation of the shorthands that are used is made. These shorthands are going to be used ahead for brevity reasons.

- WW : Number of contacts between water molecules.
- TT : Number of contacts between TFE molecules.
- WT : Number of contacts between water and TFE molecules.
- PW : Number of contacts between water and FTZ^{pep} molecules.
- PT : Number of contacts between TFE and FTZ^{pep} molecules.
- Temp : Temperature.
- HGI% : Helicity content percentage.
- q : A similarity measure between a comparison and reference structure. (*In comparison with the discrete HGI% values, q values are continuous and provide an easier observation of trends and correlations. q's base structure is in frame 8483221.*)



W-W	T-W	T-T	P-W	P-T	Temp	HGI%	q
76456	17689	12588	894	1461	298.68	0.421	0.4119
76260	17235	13558	821	1304	284.64	0.368	0.4176
75272	16996	13052	800	1538	291.19	0.526	0.4039
76928	16930	13926	762	1403	281.94	0.526	0.3976
76736	17129	13446	862	1532	283.93	0.579	0.3999
75720	16737	13220	807	1439	300.61	0.421	0.3868
76146	17134	13376	780	1436	281.86	0.368	0.4089
75820	17480	12564	915	1458	288.97	0.421	0.4348
76018	16948	13764	813	1450	287.99	0.474	0.4191
76164	16401	13432	827	1387	283.16	0.474	0.4329
75728	16945	13286	753	1420	288.95	0.526	0.4279

[Figure.6](#) The first ten lines from contact file of simulation.

In the analysis of the matrices some crucial observations were noted about the correlations among the contact pairs. These observations are listed and expounded below and can be noticed in [Figure.7](#). The matrix of [Figure.7](#) was created including all the temperature of contact

file, ranged from 280K to 380K. All the values ranged from [-0.20, +0.20] considered as not of an importance correlation but are minded in order to observe the trends and the deviations.

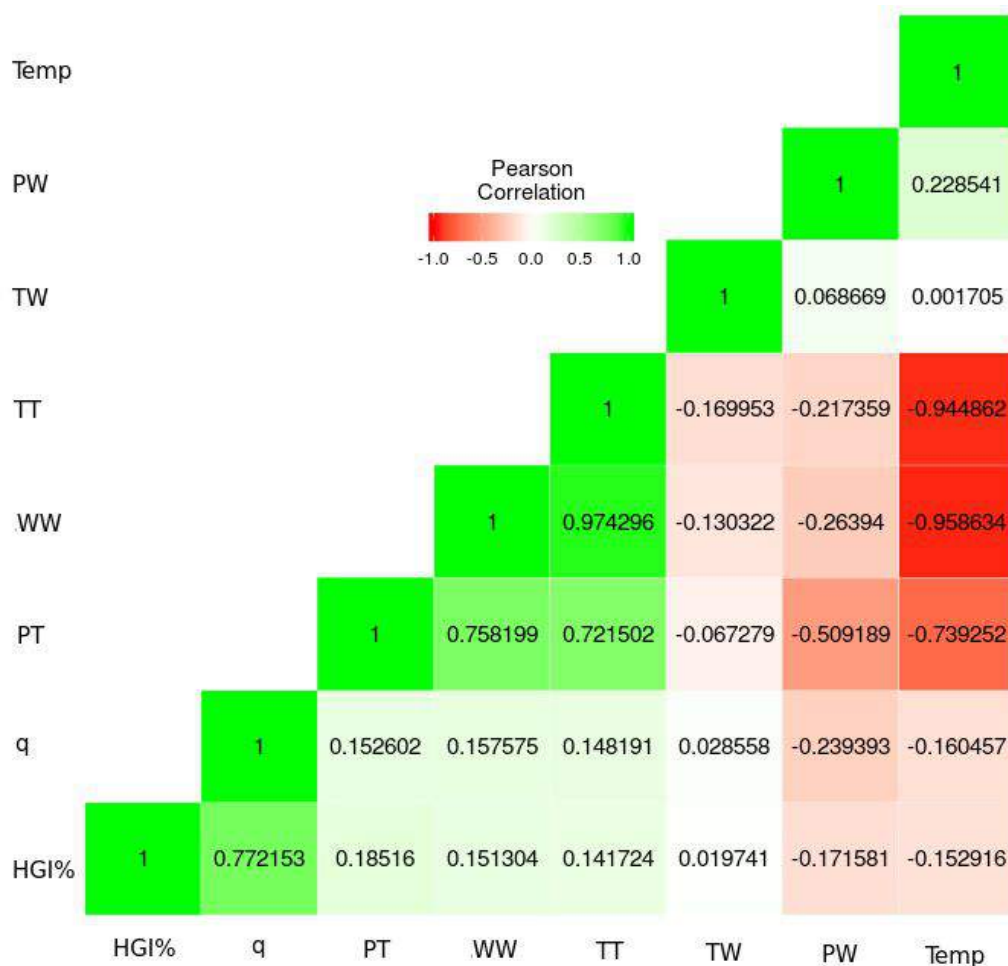


Figure.7 Presentation of the correlation among contact pairs of the simulation in a data heatmap matrix calculated based on Pearson’s correlation coefficient factor r including all temperatures of file [280K-380K]. The matrix and calculations were made with R program using the libraries *reshape2* and *ggplot2*^{44, 45}. (see Annex, pg. 34)

Firstly, a parallel increase of WW contacts is observed respectively with a TT contact increase as shown by the high positive correlation in matrix. In general, when the contacts between a pair of molecules increases the contacts of this molecules with other molecules besides those in pair decrease. This is explained naturally due to a degree of “occupation” of the molecule by the specific pairwise molecule. For better comprehension, in WW-TW-TT example, when contacts between water molecules increase, the contacts between TFE and water molecules decrease. The contacts between self TFE molecules increase also. This is explained because when TFE molecules interact more with self molecules the degree of interaction with

other types of molecules, like water, decreases. In this way, there are more “water-free” interactions among TFE molecules, and the contacts TFE-TFE are increased. This finding reinforces the existent knowledge about TFE rich regions in water-TFE solutions due to TFE trend for aggregation.

Secondly, the observation with the most scientific interest possibly is the behaviour of contacts between the peptide and TFE as well as the contacts of TT. In the first observation, the occupation scenario explained the decrease of TW contacts in a WW contact increase. The same scenario is followed by PW contacts that are reduced in a WW increase. If the correlation between our components was based in occupation only (*i.e. the availability of a molecule in the mixture in order to form contacts with other types of molecules*), then the contacts between TFE and the peptide would decrease in the TT contact increase, this does not happen, however an expected outcome considering the tendency of TFE to aggregate. Between TT and PT calculation is observed one of the most positive correlations in matrix (actually, the 4th from top equal to +0.721502). Moreover, The conclusion of this finding is possibly that these TFE rich regions (mentioned in paragraph 3.5.) favour the contacts among TFE and peptide molecules.

For the correlation of temperature with other parameters were created two submatrices based on temperature values. The first matrix, contains values range from 280K to 320K, and the second matrix from 320K to 380K as it is presented in [Figure.8.1](#) and [Figure.8.2](#) respectively. Even in a brief examination of the two matrices it is obvious that only four elements have results in the interval of significance and there are the contact pairs WW, TT, PT and TW. The first three contact pairs correlated with temperature resulted highly negative values of r . A fluctuation is noted, where these values become more negative in the matrix of higher temperature values. A simple explanation behind this finding is the raise of temperature is translated in accumulation of molecular kinetic energy and consequently molecular mobility that obstruct bonding. Additionally, considering the boiling points of pure water and TFE (which are 373K and approximately at 347K respectively, see paragraph 3.5. *Solvents - 2,2,2-trifluoroethanol solvent (TFE)*) a hypothesis is born that suggests TFE to be affected more by temperature rise due to a lower boiling point. This hypothesis reinforces the pair TW-Temp, which suggests a different behaviour on each submatrix. In matrix with lower temperatures TW contacts indicate positive correlation with temperature (+0.244229), but in the matrix with the higher temperature this value is inverted to negative correlation (-0.327632). By noticing the PW correlated with Temp in both submatrices, it is observed that despite the fact that the r value is not of importance, it is positive in each of them, in spite of reaching water’s boiling point too, but decreases dramatically. This note may betoken a more independent correlation of PW with temperature due to existent behaviour of peptide to be more disordered in water.

The correlation between q and HGI% is obviously anticipated to be positive because both are sizes that describe the structure conformation. The structure in which bases q calculations presents high helicity.

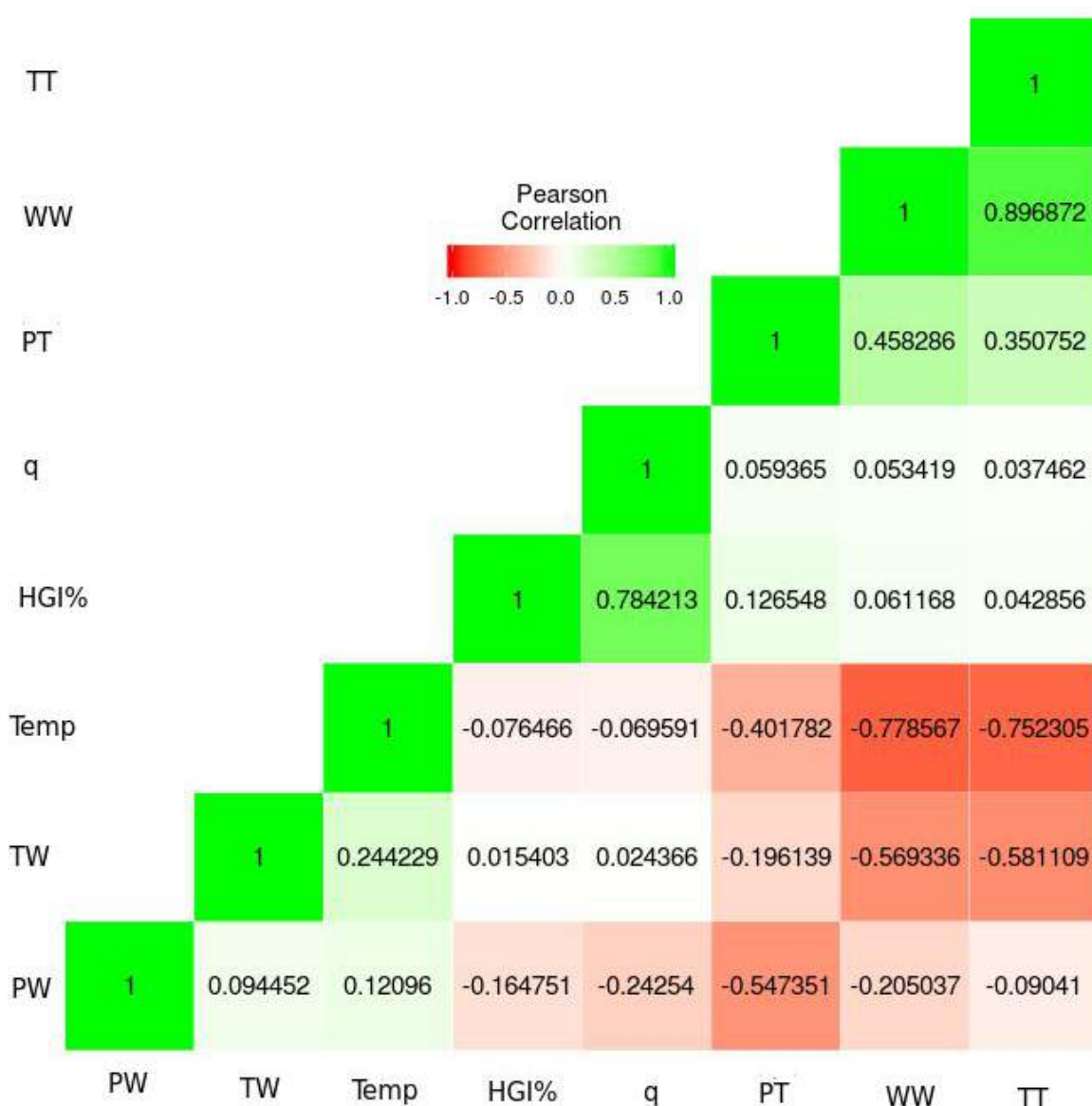


Figure.8.1 Presentation of the correlation among contact pairs of the simulation in a data heatmap matrix calculated based on Pearson's correlation coefficient factor r including only the temperatures [280K-320K] of file. The matrix and calculations were made with R program using the libraries *reshape2* and *ggplot2*. (see Annex, pg. 34)

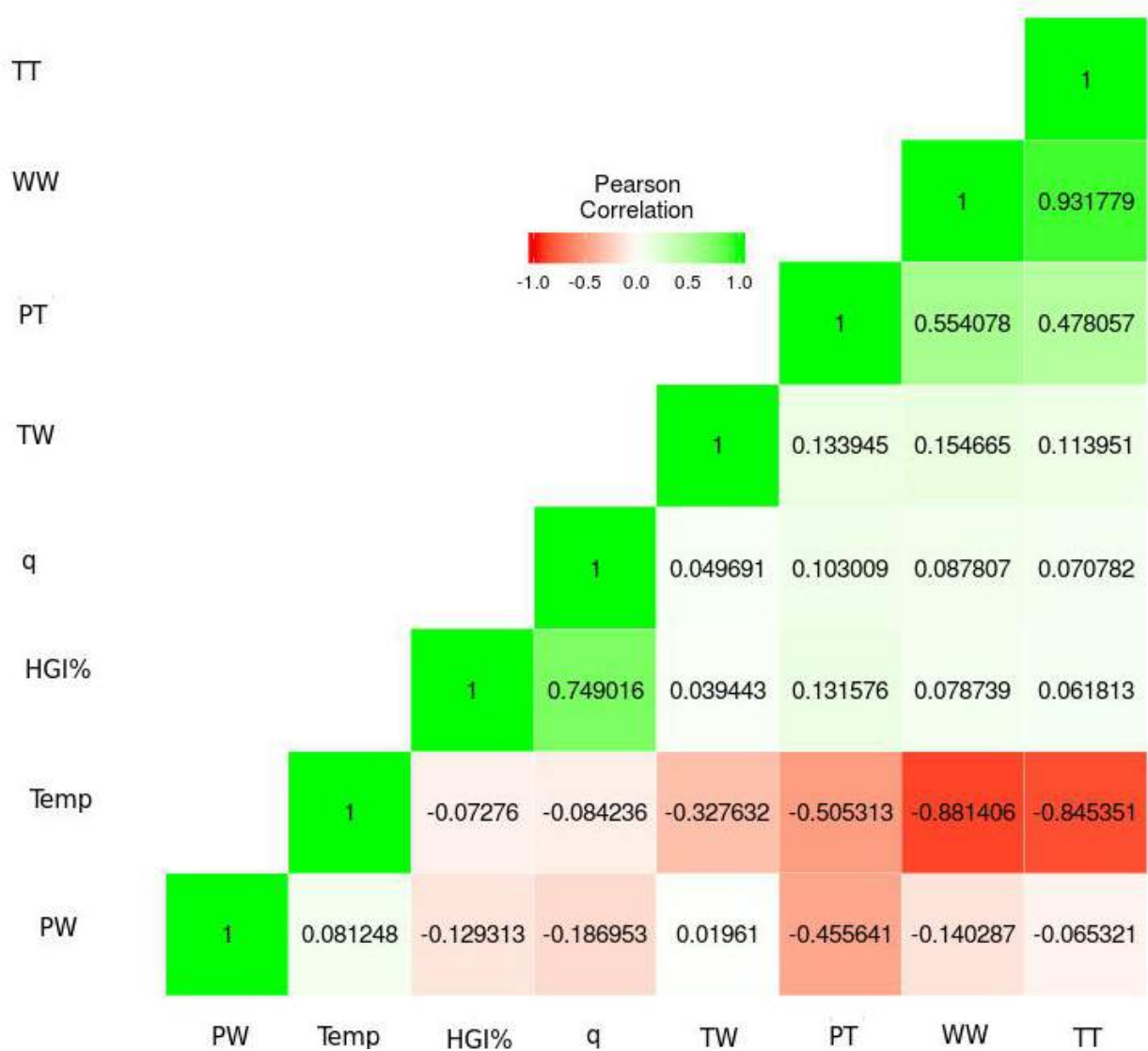
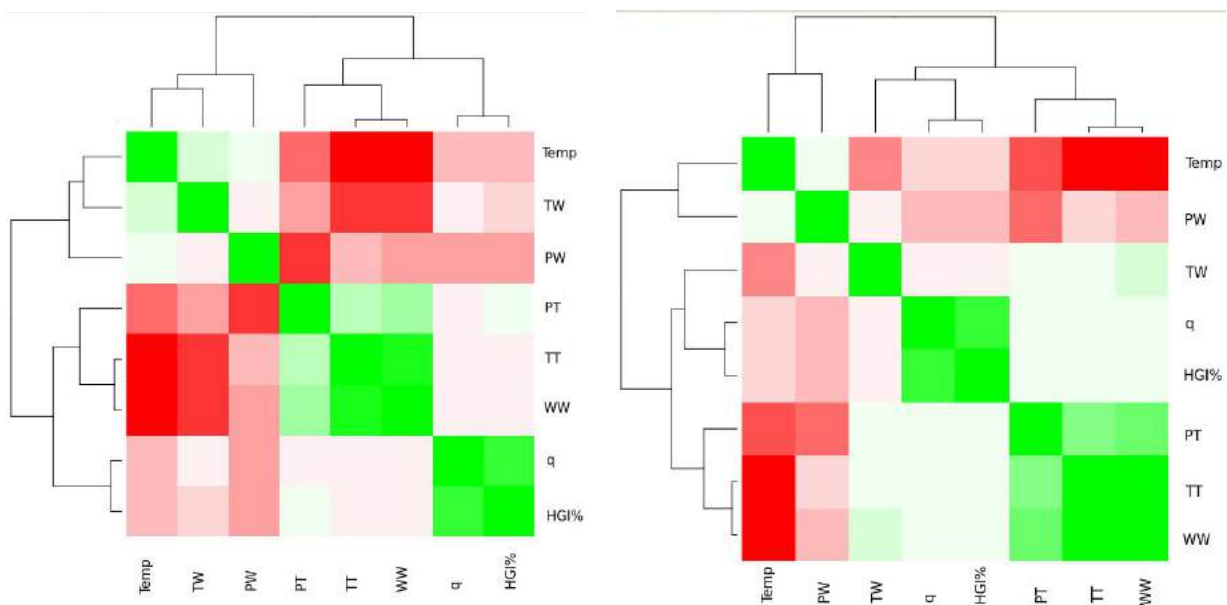


Figure.8.2 Presentation of the correlation among contact pairs of the simulation in a data heatmap matrix calculated based on Pearson's correlation coefficient factor r including all temperatures of file [320K-380K]. The matrix and calculations were made with R program using the libraries *reshape2* and *ggplot2*⁴⁵. (see Annex, pg. 34)

Additionally with the matrices of [Figure.7](#), [Figure.8.1](#) and [Figure.8.2](#) it considered important the creation of a matrix that presents the correlation “affinity” of the individual parameters that were examined. In [Figure.9](#) there are two matrices that respect to [Figure.8.1](#) (left) and [Figure.8.2](#) (right). In the left matrix, like in [Figure.8.1](#) the temperatures are between 280K-320K and in the right matrix correspond the temperature between 320K-380K. The branches represent correlation relationships. Like a phylogenetic tree, the distance from base, i.e. the matrix, corresponds to the degree of correlation of the parameter based on r value. Parameters that are connected near the base are more correlated in a manner that is indicated by the respective colour of matrix. The length of branches represent the calculated their r distance. For example the branch that connects the q and HGI% matrix squares is near the base because their correlation is highly positive, but the adjacent branch which connects TT with WW has smaller length, is nearer in base indicating a closer correlation as shown in the [Figure.7](#).

In a closer observation, there seem to be formed to main clusters similar in both matrices. However, they differ in the branch of TW that from the left to the right matrix alters cluster, indicating the sensitivity of TW in the two different temperature conditions. Moreover, it is noted that PT and TT are in the same cluster and are correlated in a resent level, when WW and PW are very hierarchically far and are localized in different clusters.



[Figure.9](#) Hierarchical tree representation of heatmap. Values with higher correlation appear earlier connection branches. The mirror display have been retained in these heatmaps.

Finally, as a logical continuation of the previous observations were made some scatter plot density matrices via *plot* program. These are scatter plots made by part of the simulation contact file and x and y variables are elements of contact arrays of the file. For example, a subfile of contact file containing the columns of WW and TW contacts was made and used as an input for plot program. Basically, the upcoming figures are screenshots from plot program and present the density of a parameters chosen pair. The colour representation that is used by plot indicates dark blue colorization for minimums and thought yellow, dark red for maximums.

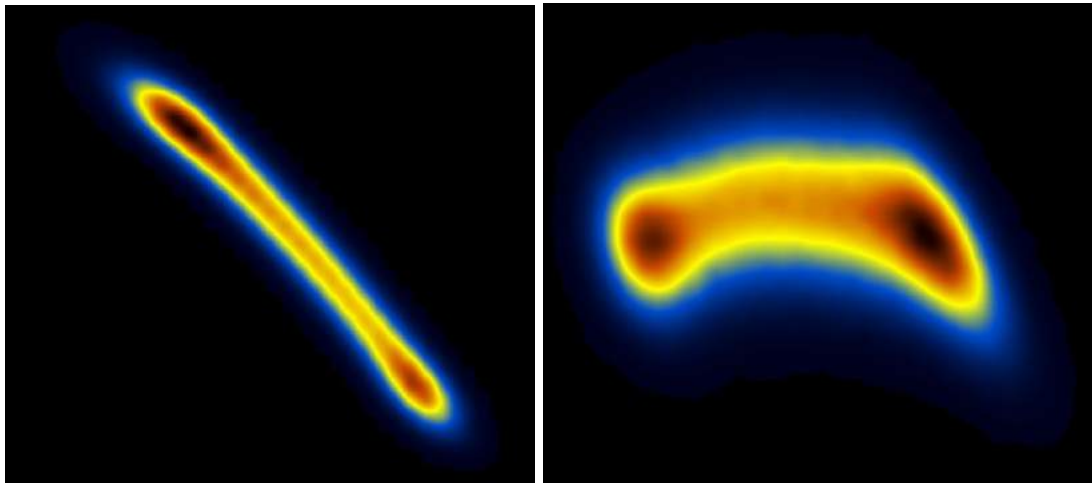


Figure.9.1 Density matrices of contacts. The left plot represents the contact density of WW and TT pair. The right plot represent the contact density of WW and TW pair.

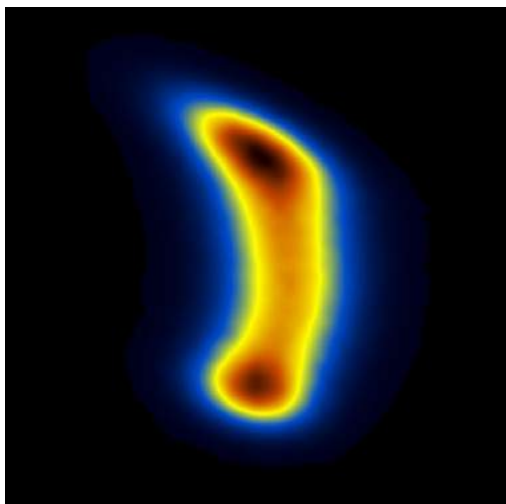


Figure.9.2 The plot represents the contact density of TT and TW pair.

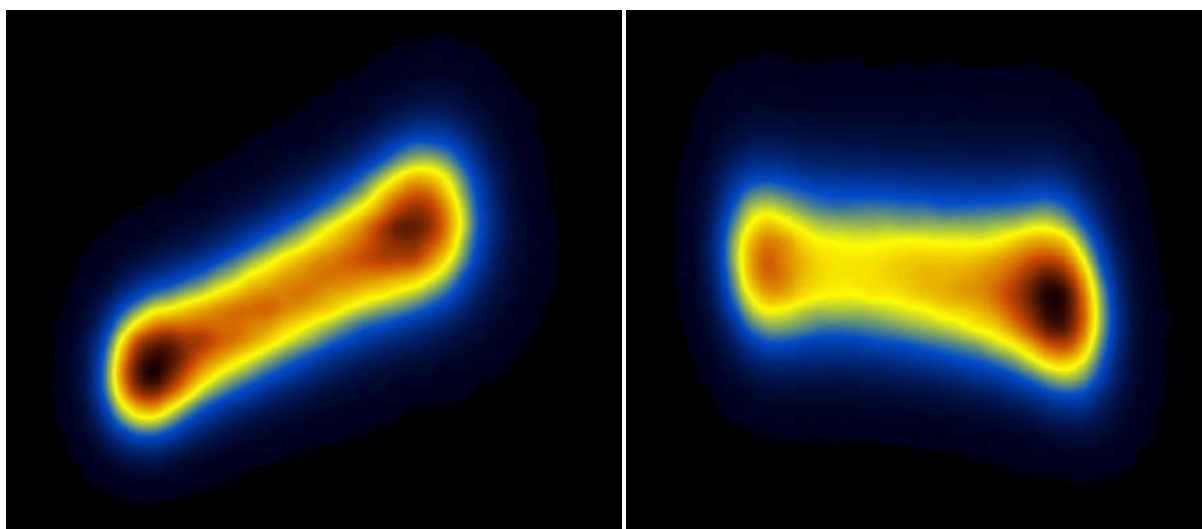


Figure.9.3 Density matrices of contacts. The left plot represents the contact density of TFE and TFE-peptide contacts (TT-PT). The right plot represents the contact density of water contacts and the water-peptide (WW & PW).

In all four figures, it is noticeable the distinct contact populations colorized with dark red as a result of different contact densities. Remarkable are the results at [Figure.9.3](#), in which the TT-PT matrix. The subpopulation (who indermediates the two distinct populations) consisted by both TT and PT pairs indicates a connection and a coexistence of TT and PT contacts, reinforcing in this way the hypothesis that TT reach regions hinder PT contact. On the contrary, the right WW-PW plot does not show respective results.



Conclusion

The analyses of peptide NOEs resulted that there is only a narrow distance of the NMR and the simulation derived data. As another evidence of structural agreement the NOE derived distance evaluated in the 0.054 Å after excluding 4 violations from the $d_{\alpha N}(i, i+2)$. The exclusion is based on the existence of reverse turn at that part of the peptide and points a possible disfavour of the force-field for the formation of the specific secondary structure.

After meticulous observation and study of contact correlations in matrices, the findings suggest a strong positive correlation between TT contacts (possibly depict the TFE rich regions of aggregation) with the PT contacts. The results indicate that these regions may have a role in dialysis and comprise structural hindrance for the peptide. Furthermore, strong negative correlation among WW, TT and PT with temperature as well as the alter correlation profile of TW with temperature evinces a shade of sensitivity of TFE in temperature increase, possibly due to lower boiling point than water.

Nevertheless, a lot more steps are needed in the future to fully understand the mechanism under TFE functions in combination with force-fields to promote the secondary structure in peptides, suggesting for the current project a respective study of FTZ^{pep} in complex with the NR FTZ-F1 and in varying proportions in TFE mixtures, e.g. 100% (v/v) TFE.



Bibliography

1. Yun, J.-H.; Lee, C.-J.; Jung, J.-W.; Lee, W. Solution Structure of LXXLL-related cofactor peptide of orphan nuclear receptor FTZ-F1. *Bull. Korean Chem. Soc.* **2012**, *33*, 583–588.
2. Glykos*, N.M.; Adamidou, T.; Arvaniti, O.K.; Folding Simulations of a Nuclear Receptor Box-Containing Peptide Demonstrate the Structural Persistence of the LxxLL Motif Even in the Absence of Its Cognate Receptor. *J. Phys. Chem. B* **2018**, *122*, 106–116.
3. Allen, M.P. Introduction To Molecular Dynamics Simulation. *Computational Soft Matter: From Synthetic Polymers to Proteins, NIC Series*, **2004**, *23*, 1-28.
4. Klepeis, L.J.; Lindorff-Larsen, K.; Dror, O.R.; Shaw, E.D. Long-timescale molecular dynamics simulations of protein structure and function. *Struct. Biol.* **2009**, *19*, 120-127
5. Mangelsdorf, D.J.; Thummel, C.; Beato, M.; Herrlich, P.; Schütz, G.; Umesono, K.; Blumberg, B.; Kastner, P.; Mark, M.; Chambon, P.; Evans, R.M. The nuclear receptor superfamily: the second decade. *Cell.* **1995**, *83*, 835-9.
6. Sever, R.; Glass, C. K. Signaling by Nuclear Receptors. *Cold Spring Harb Perspect Biol.* **2013**, *5*, a016709.
7. Chitra, R.; Smith, P.E.; Properties of 2,2,2-trifluoroethanol and water mixtures. *J. Biochem. Phys.* **2001**, *114*, 426.
8. Enmark, E.; Gustafsson, J.A.; Orphan nuclear receptors--the first eight years. *Mol Endocrinol.* **1996**, *10*, 1293–1307.
9. Savkur, R.S.; Burris, T.P.; The coactivator LXXLL nuclear receptor recognition motif. *J. Peptide Res.*, **2004**, *63*, 207–212.
10. Plevin, M. J.; Mills M. M.; Ikura, M. The LxxLL motif: a multifunctional binding sequence in transcriptional regulation. *Trends Biochem. Sci.* **2005**, *30*, 66-69.
11. Lim, B.; Fukaya, T.; Heist, T.; Levine, M. Temporal dynamics of pair-rule stripes in living *Drosophila* embryos. *Proc. Natl. Acad. Sci.* **2018**, *115*, 8376-8381.
12. Lavorgna, G.; Karim, F.D.; Thummel, C.S.; Wu, C. Potential role for a FTZ-F1 steroid receptor superfamily member in the control of *Drosophila* metamorphosis. *Nat. Ac. Sc.* **1993**, *90*, 3004-3008.

13. Florence, B.; Guichet, A.; Ephrussi, A.; Laughon, A. Ftz-F1 is a cofactor in Ftz activation of the *Drosophila engrailed* gene. *Dev.* **1997**, *124*, 839-847.
14. Yu, Y.; Li, W.; Su, K.; Yussa, M. et al. The nuclear hormone receptor Ftz-F1 is a cofactor for the *Drosophila* homeodomain protein Ftz. *Nat.* **1997**, *385*, 552-555.
15. Yoo, J.; Ko, S.; Kim, H.; Sampson, H.; Yun, J.-H.; Choe, K.-M.; Chang, I.; Arrowsmith, C.H.; Krause, H. M.; Cho, H.-S.; Lee, W. Crystal Structure of Fushi Tarazu Factor 1 Ligand Binding Domain/Fushi Tarazu Peptide Complex Identifies New Class of Nuclear Receptors. *J. Biol. Chem.* **2011**, *286*, 31225–31231. (*Figure retrieved without license.*)
16. Hou, HY.; Heffer, A.; Anderson, W.R.; Liu, J.; Bowler, T.; Pick, L. Stripy Ftz target genes are coordinately regulated by Ftz-F1. *Dev Biol.* **2009**, *335*, 442-53. (*Figure retrieved without license.*)
17. Timasheff, S.N.; Solvent effects on protein stability. *Struct. Biol.* **1992**, *2*, 35-39.
18. Chitra, R.; Smith, E.P. A comparison of the properties of 2,2,2-trifluoroethanol and 2,2,2-trifluoroethanol/water mixtures using different force fields *J. Chem. Phys.* **2001**, *115*, 5521.
19. Franks, F.; Eagland, D.; The role of solvent interactions in protein conformation. *Crit. Rev. Biochem.* **1975**, *3*, 165.
20. Bégué, J.-P.; Bonnet-Delpon, D.; Crousse, B.S.; Shuklov, I. A.; Dubrovina, N. V.; Börner, A. Fluorinated alcohols as solvents, co-solvents and additives in homogeneous catalysis. *Synthesis* **2007**, *19*, 2925-2943.
21. Buck, M.; Radford, S.E.; Dobson, C.M. A partially folded state of hen egg white lysozyme in trifluoroethanol: Structural characterization and implications for protein folding. *Q. Rev. Biophys.* **1998**, *31*, 297.
22. Neuhaus, D. Nuclear Overhauser Effect. John Wiley & Sons, Ltd. published online in **2011**.
23. Zagrovic, B.; van Gunsteren, W.F. Comparing atomistic simulation data with the NMR experiment: how much can NOEs actually tell us? *Proteins.* **2006**, *63*, 210-218.
24. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. Routledge. **1988**.
Found in SPSS Tutorials: Pearson Correlation web library:
<https://libguides.library.kent.edu/SPSS/PearsonCorr>

25. Glen, S. StatisticsHowTo. Powered by: WordPress. **2019**.
<https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/correlation-coefficient-t-formula/>
26. Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R.J. The Amber Biomolecular Simulation Programs J. Comput. Chem. 2005, 26, 1668–1688.
27. Jorgensen W. L.; Chandrasekhar J.; Madura J. D.; Impey R. W.; Klein M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. J. Chem. Phys. 1983, 79, 926-935.
28. Chitra, R.; Smith, P. E. Properties of 2,2,2-trifluoroethanol and water mixtures. J. Chem. Phys. 2001, 114, 426-435.
29. Yu, Y.; Wang, J.; Shao, Q.; Shi, J.; Zhu, W. The effects of organic solvents on the folding pathway and associated thermodynamics of proteins: a microscopic view. Nat. Sci. Rep. 2016, 6, 1-12.
30. Dupradeau, F. Y.; Pigache, A.; Zaffran, T.; Savineau, C.; Lelong, R.; Grivel, N.; Lelong, D.; Rosanski, W.; Cieplak, P. The R.E.D. Tools: Advances in RESP and ESP charge derivation and force field library building. Phys. Chem. Chem. Phys. 2010, 12, 7821–7839.
31. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. Proteins 2006, 65, 712–725.
32. Wickstrom, L.; Okur, A.; Simmerling, C. Evaluating the Performance of the ff99SB Force Field Based on NMR Scalar Coupling Data. Biophys. J. 2009, 97, 853–856.
33. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw D. E. Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field. Proteins 2010, 78, 1950–1958.
34. Best, R. B.; Hummer, G. Optimized Molecular Dynamics Force Fields Applied to the Helix-Coil Transition of Polypeptides. J. Phys. Chem. B 2009, 113, 9004–9015.
35. Kale, L.; Skeel, R.; Bhandarkar, M.; Brunner, R.; Gursoy, A.; Krawetz, N.; Phillips, J.; Shinozaki, A.; Varadarajan, K.; Schulten, K. NAMD2: Greater Scalability for Parallel Molecular Dynamics. J. Comput. Phys. 1999, 151, 283–312.

- 
36. Zhang, C.; Ma, J. Enhanced Sampling and Applications in Protein Folding in Explicit Solvent. *J. Chem. Phys.* 2010, 132, No. 244101.
 37. Glykos, N. M. CARMA: A Molecular Dynamics Analysis Program. *J. Comput. Chem.* 2006, 27, 1765–1768.
 38. Koukos, P. I.; Glykos, N. M. grcarma: A Fully Automated Task-Oriented Interface for the Analysis of Molecular Dynamics Trajectories, *J. Comput. Chem.* 2013, 34, 2310-2312.
 39. Baltzis, A. S.; Koukos, P. I.; Glykos, N. M. Clustering of Molecular Dynamics Trajectories via Peak-Picking in Multidimensional PCA-derived Distributions. *arXiv*. 2015, 1512.04024 [q-bio.BM].
 40. Frishman, D.; Argos, P. Knowledge-Based Protein Secondary Structure Assignment. *Proteins* 1995, 23, 566–579.
 41. Crooks, GE.; Hon, G.; Chandonia, JM.; Brenner, SE. WebLogo: a sequence logo generator. *Genome Res.* 2004, 14, 1188–1190.
 42. Humphrey, W.; Dalke, A.; Schulten, K. VMD—Visual Molecular Dynamics. *J. Mol. Graphics* 1996, 14, 33–38.
 43. Glykos, M.N. Plot Program. Utopia. <https://utopia.duth.gr/glykos/plot/>
 44. R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>
 45. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.

Annex

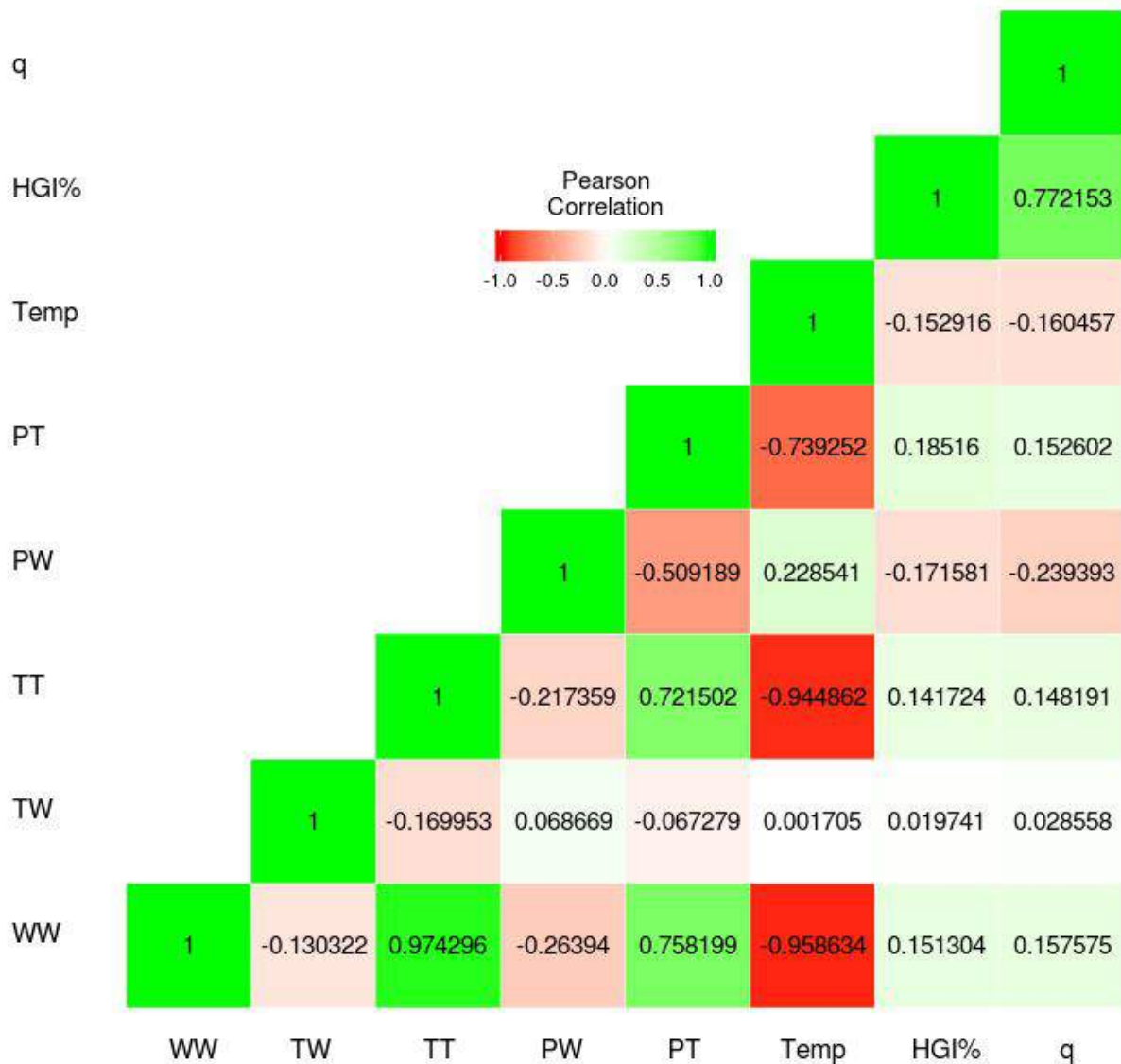


Figure.7 Presentation of the correlation among contact pairs of the simulation in a data heatmap matrix calculated based on Pearson's correlation coefficient factor r including all temperatures of file [280K-380K]. The matrix and calculations were made with R program using the libraries *reshape2* and *ggplot2*^{44, 45}. This figure is another version of [Figure.7](#) in pg. 22 . In this version the correlation matrix was not reorder using the function *hclust*.



Figure.8.1 Presentation of the correlation among contact pairs of the simulation in a data heatmap matrix calculated based on Pearson's correlation coefficient factor r including only the temperatures [280K-320K] of file. The matrix and calculations were made with R program using the libraries *reshape2* and *ggplot2*⁴⁵. This figure is another version of [Figure.8.1](#) in pg. 24 . In this version the correlation matrix was not reorder using the function *hclust*.

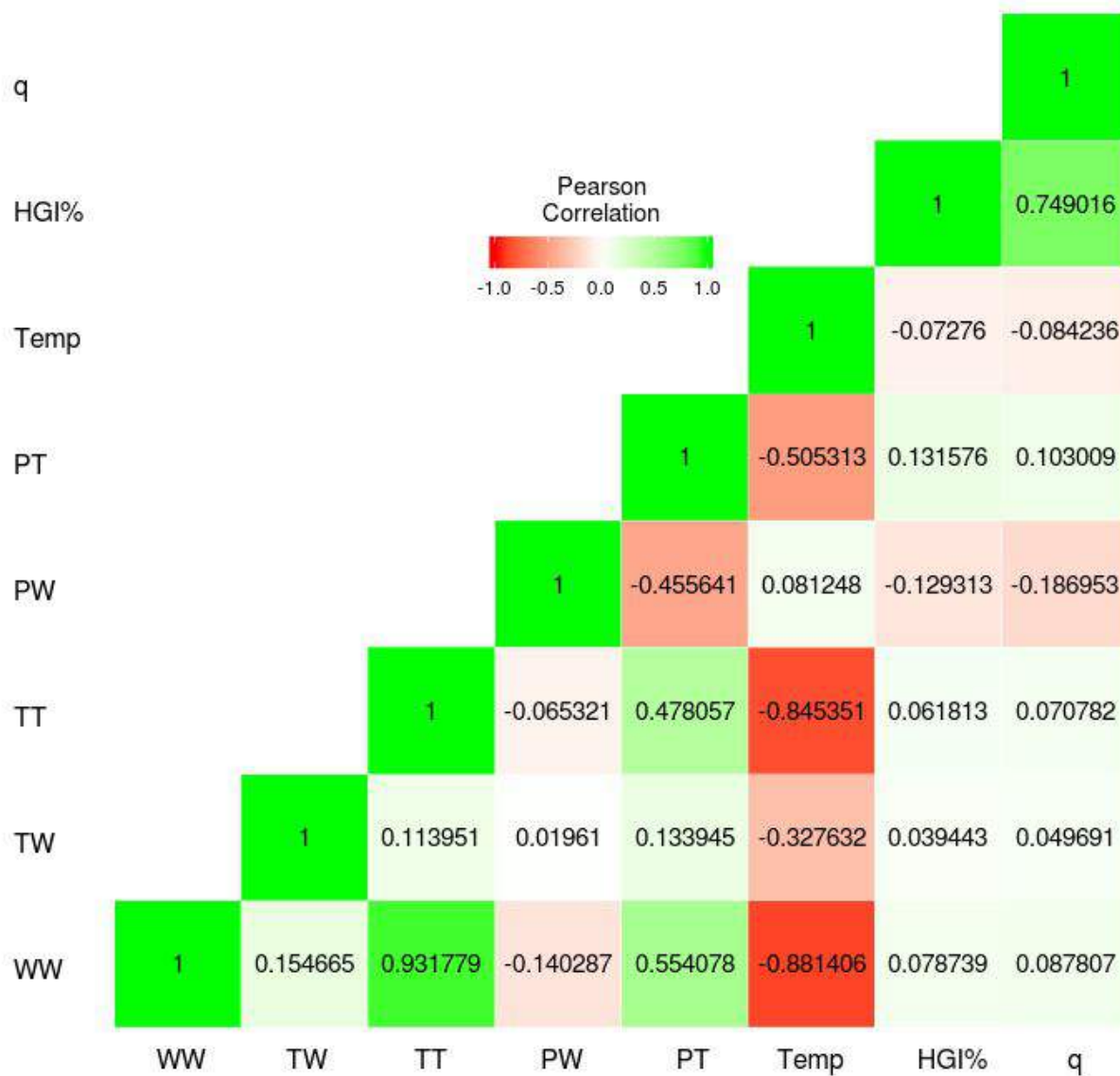


Figure.8.2 Presentation of the correlation among contact pairs of the simulation in a data heatmap matrix calculated based on Pearson's correlation coefficient factor r including all temperatures of file [320K-380K]. The matrix and calculations were made with R program using the libraries *reshape2* and *ggplot2*⁴⁵. This figure is another version of [Figure.8.2](#) in pg. 25 . In this version the correlation matrix was not reorder using the function *hclust*.