

January 2019

Predicting The Structure And Selectivity Of Coiled-Coil Proteins

Mojtaba Jokar

Wayne State University, mojtaba.jokar@wayne.edu

Follow this and additional works at: https://digitalcommons.wayne.edu/oa_dissertations

 Part of the [Biochemistry Commons](#), [Biology Commons](#), and the [Biophysics Commons](#)

Recommended Citation

Jokar, Mojtaba, "Predicting The Structure And Selectivity Of Coiled-Coil Proteins" (2019). *Wayne State University Dissertations*. 2285.

https://digitalcommons.wayne.edu/oa_dissertations/2285

This Open Access Dissertation is brought to you for free and open access by DigitalCommons@WayneState. It has been accepted for inclusion in Wayne State University Dissertations by an authorized administrator of DigitalCommons@WayneState.

PREDICTING THE STRUCTURE AND SELECTIVITY OF COILED-COIL PROTEINS

by

MOJTABA JOKAR

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2019

MAJOR: CHEMICAL ENGINEERING

Approved By:

Advisor

Date

DEDICATION

I would like to dedicate my work to my parents, my sisters, and my brother who provided me with love and support unconditionally and sacrificed a great deal for my education and especially my PhD program.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Korosh Torabi, for his continuous guidance, support and professional academic supervision, which assisted me in completing this PhD dissertation. It was an honor for me to conduct this research under his supervision and guidance. I could not have imagined having a better advisor and I believe the lessons I learned from him would be the invaluable assets for my future career.

I am grateful to Dr. Jeffrey Potoff, and Dr. Zhiqiang Cao from the Department of Chemical Engineering and Materials Science at Wayne State University, and Dr. Takeshi Sakamoto from the Department of Physics and Astronomy at Wayne State University, for their time and effort serving as my dissertation committee members. Besides my dissertation committee members, I would like to thank Dr. Eranda Nikolla for her support, encouragement and insightful guidance, especially during a period that I was serving as her teaching assistant.

My sincere thanks also goes to my undergraduate program mentors, especially Dr. Bahram Dabir, Dr. Hamid Modarress, and Dr. Manouchehr Nikazar, who is no longer with us and may he rest in peace. When I started my PhD program, most of the concepts of Chemical Engineering graduate core courses were already clear to me and I owe it to my professors back home, who were educated in the best universities in the world and now serve at Tehran Polytechnic, which owns one of the best schools of Chemical Engineering in Iran.

Finally, and most importantly, nobody has been more caring to me in the pursuit of my PhD program than the members of my family. I would like to express gratitude to my parents, my sisters, and my brother for their endless love, support, and sacrifices. I am truly thankful for having them in my life.

PREFACE

PREDICTING THE STRUCTURE AND SELECTIVITY OF COILED-COIL PROTEINS

The main objectives of this research are: to develop a model to predict the propensity of a protein sequence to form an isolated coiled-coil structure, and to investigate the selectivity of coiled-coils by studying protein-protein interactions. Possibly, one of the simplest and most studied protein-protein interactions exists in coiled-coil structures. The methods and proposed solutions can substantially reduce the computational effort while maintaining reasonable levels of accuracy.

Keywords: Coiled-coil Proteins; Statistical Mechanics; Molecular Thermodynamics; Protein Structure; Computational Biology

TABLE OF CONTENTS

DEDICATION	ii
ACKNOWLEDGMENTS	iii
PREFACE	iv
LIST OF TABLES	vii
LIST OF FIGURES	viii
CHAPTER 1: Introduction	1
1.1-Statement of the problem	1
1.2- Objectives of the study	1
CHAPTER 2: α -helix structure	3
2.1- Introduction	3
2.2- AGADIR model and its partition function	3
2.3- Energies involved in AGADIR model	5
2.3.1- ΔG_{int} : Intrinsic helical propensity (entropy loss)	6
2.3.2- ΔG_{elec} : Electrostatic interactions	6
2.3.2.1- Ionic strength, pH and temperature dependence	7
2.3.2.2- Calculation of distances between charged groups	7
2.3.2.3- Calculation of electrostatic energy between charged groups	8
2.3.3- ΔG_{dip} : Interaction with helix macro-dipole	9
2.3.3.1- Effect of salt concentration on helix stability through macrodipole	10
2.3.4- ΔG_{HB} : Hydrogen bonding	10
2.3.5- ΔG_{nonH} : Interaction with non-helical residues	10
2.3.6- ΔG_{SD} : Side-chain-side-chain interactions	13

CHAPTER 3: Introduction to coiled-coil protein structures	14
3.1- Introduction	14
3.2- Molecular structure of a coiled-coil protein structures	15
3.3- Developing a statistical mechanical model for the coiled-coil dimerization	16
3.4- Coiled-coil dimerization energies	20
CHAPTER 4: Electrostatic interactions and their effect on coiled-coil specificity	21
4.1- Introduction	21
4.2- Calculating the electrostatic interactions in a coiled-coil structure	27
4.3- Distance between two point-charges	28
4.3.1- Effect of ion concentration on the distances between charged residues	35
4.4- Pairwise electrostatic interactions	36
4.5- Contribution of non- $g_i-e'_{i+5}$ positions to the total electrostatic interaction	40
4.6- Electrostatic interactions and coiled-coil paring specificity	46
CHAPTER 5: Hydrophobic core interactions in coiled-coil protein structures	54
5.1- Introduction	54
5.2- Model for calculating hydrophobic core interactions in a coiled-coil	55
5.3- Contribution of hydrophobic core interactions to stability of the structure	62
5.4- Hydrophobic core interactions and coiled-coil paring specificity	64
CHAPTER 6: Summary, challenges and conclusions	69
CHAPTER 7: Future research directions	73
REFERENCES	75
ABSTRACT	91
AUTOBIOGRAPHICAL STATEMENT	93

LIST OF TABLES

Table 4.1 Distances and coupling energies between charged residues	34
Table 4.2 Coupling energies between charged residues in kcal/mole	38
Table 4.3 Coiled-coil region amino acid sequences used in this work	41
Table 5.1 SASA values for 20 amino acids at different positions of the heptad repeat	60
Table 5.2 SASA values for 20 amino acids when exposed to the solvent	61

LIST OF FIGURES

Figure 3.1 Structure of a coiled-coil dimer	16
Figure 4.1 Change of the distance between D and H at g_i and e'_{i+5} positions	30
Figure 4.2 Change of the distance between E and R at g_i and e'_{i+5} positions	30
Figure 4.3 Change of the distance between E and R at g_i and e'_{i+5} positions	32
Figure 4.4 Change of the distance between E and R at g_i and e'_{i+5} positions	32
Figure 4.5 Charged amino acids and heptad repeat positions along a coiled-coil dimer	33
Figure 4.6 Change of the distances with salt concentration	36
Figure 4.7 Coupling energies between charged residues at g_i - e'_{i+5} positions	39
Figure 4.8 Electrostatic coupling energy of CREB3 with 22 other b-ZIP sequences	42
Figure 4.9 Electrostatic coupling energy of C/EBP α with 22 other b-ZIP sequences	43
Figure 4.10 Electrostatic coupling energy of ATF-6 with 22 other b-ZIP sequences	44
Figure 4.11 Electrostatic coupling energy of ZF with 22 other b-ZIP sequences	45
Figure 4.12 Absolute value of electrostatic energies ($E_{\text{non-}ge'}/E_{ge'}$)	46
Figure 4.13 g_i - e'_{i+5} interaction scores of 22 sequences for charged residues	47
Figure 4.14 g_i - e'_{i+5} interaction scores of 22 sequences for charged and polar residues	48
Figure 4.15 g_i - e'_{i+5} electrostatic interactions of 22 sequences	50
Figure 4.16 Frequency of g_i - e'_{i+5} electrostatic interactions	51
Figure 4.17 Total electrostatic interactions of 22 sequences	52
Figure 4.18 Frequency of total electrostatic interactions	53
Figure 5.1 Side view of a parallel coiled-coil dimer showing a side-chain and a pocket	55
Figure 5.2 Change of SASA for Phe residue at position b	59
Figure 5.3 Change in stability of a leucine-zipper dimer upon mutations at position d	63

Figure 5.4 Change in stability of a leucine-zipper dimer upon mutations at position <i>a</i>	64
Figure 5.5 Absolute value of hydrophobic core over the electrostatic energies	66
Figure 5.6 Summation of packing and electrostatic interactions for 22 sequences	67

1- CHAPTER 1: Introduction

1.1- Statement of the problem

Coiled-coils are protein structural motifs made up of two or more α -helices twisted around one another. Coiled-coils are critical to the function of various motor proteins, cytoskeletal filaments and extra-cellular matrix proteins (Rose et al. 2004; Burkhard et al. 2001).

Dimeric coiled-coils in the form of long rigid rods are responsible for mechanical load transmission and act as lever arms and possibly reversible springs within myosin family of proteins (Taniguchi et al. 2010; Parry et al. 2008; Li et al. 2003; Knight et al. 2005). Mechanical properties of the coiled-coil structures forming the tail domain of myosin II are crucial to its work cycle. Single molecule studies using atomic force microscopy or optical tweezers along with various molecular simulation studies have confirmed the unique mechanical properties of the coiled-coil protein structures (Root et al. 2006; Schwaiger et al. 2002; Kreuzer et al. 2013; Gao et al. 2011). However, single molecule experimental techniques are not capable of probing the mechanical properties of short coiled-coil motifs in their native structural settings and the molecular simulation studies often fail to generate a quantitatively accurate prediction of their response to mechanical load (Torabi and Schatz 2013). At the same time both of these methods are time consuming and expensive (Root et al. 2006; Bornschlöggl and Rief 2008).

1.2- Objectives of the study

In this project, we aim to develop a statistical mechanical model (Jokar and Torabi 2017; Jokar and Torabi 2016) to predict the propensity of forming a given coiled-coil dimer based on available empirical data. Within the proposed model we identify and quantify various energetic and entropic effects, responsible for dimerization of two helical polypeptides into a coiled-coil

structure. We determine our model parameters by examining a relatively large number of solved protein structures that contain coiled-coil motifs. This would allow us to develop a thermodynamic model for predicting the propensity of a given amino acid sequence to form a coiled-coil structure. To do so, we develop a partition function for coiled-coil dimerization. Further incorporation of the above model into the previously developed α -helix tensile mechanics model (Torabi and Schatz 2013) can be a direction for the future research which is predicting the structural response of a given coiled-coil motif to bending and tensile stress. The experimental part of this line of research is done by atomic force microscopy or optical tweezers, which leads to investigating the force-extension of the coiled-coil motifs (Taniguchi et al. 2010; Root et al. 2006; Schwaiger et al. 2002; Gao et al. 2011). Moreover, we can apply this model to various coiled-coil structural motifs within the motor proteins and cytoskeletal filaments for which no theoretical predictions are available (Rose et al. 2004; Li et al. 2003; Knight et al. 2005; Herrmann et al. 2007; Armel et al. 2009; Atzberger et al. 2006; Brown et al. 2011; Blankenfeldt et al. 2006; Tripet et al. 1997).

Since statistical mechanical models are computationally inexpensive and quantitatively reliable, they have been widely used in predicting the secondary structure of DNA molecules (SantaLucia 1998; SantaLucia and Hicks 2004; Huguet et al. 2010) and α -helical polypeptides (Torabi and Schatz 2013; Lacroix et al. 1998). The expected outcome of this work is a novel statistical mechanical model that not only identifies the coiled-coil propensity of a given amino acid sequence but also has the potential to predict its structural response to mechanical tension (Jokar and Torabi 2017; Jokar and Torabi 2018). This predictive tool would be highly valuable to the fields of biomedicine, biomimetics and nanotechnology.

2- CHAPTER 2: α -helix structure

2.1- Introduction

The α -helix is a common secondary structure of proteins. The amino acids in an α -helix are arranged in a right-handed helical structure with 3.6 residues per turn (Eisenberg 2003). Among types of local structure in proteins, the α -helix is the most regular, prevalent, and the most predictable from sequence. Some proteins like keratin and collagen are almost entirely α -helical in structure.

The α -helix is stabilized by hydrogen bonding of the C=O group of the i^{th} amino acid with the N-H group of the $(i+4)^{\text{th}}$ amino acid and this hydrogen bonding is one of the most important characteristics of an α -helix. The residues are conventionally numbered starting from the N-terminus (Torabi and Schatz 2013; Chothia et al. 1981).

2.2- AGADIR model and its partition function

AGADIR is an algorithm which predicts the helical propensity of a given polypeptide sequence. AGADIR statistical mechanical model is based on a large set of experimental data for different polypeptides in aqueous solutions which include entropic and energetic effects. The accuracy of AGADIR model has been tested by circular dichroism (CD) and NMR spectroscopy measurements (Lacroix et al. 1998; Muñoz et al. 1995; Muñoz et al. 1997).

The interactions which affect the helical propensity of a given polypeptide sequence, include intrinsic helical propensity of residues, electrostatic interactions between the charged residues, electrostatic interactions of the helix macrodipole with the charged residues, side-chain-side-chain interactions, the backbone hydrogen bonding, as well as the effect of N and C termini protection. Acetylation or succinylation of the N-terminus or amidation of the C-terminus tends to increase the helicity of a polypeptide by removing the charges that interact unfavorably with

the helix macrodipole and by providing an extra hydrogen bonding site with the helix backbone (Torabi and Schatz 2013; Doig 2008). In the following section of this report, we will explain the experimental data used in AGADIR in detail.

As mentioned before, the aim of AGADIR model is to predict the helical propensity of an amino acid sequence. There could be different helical segments with different probabilities of formation associated with each of them along a polypeptide chain. Therefore, we need to estimate the probability of formation of all the possible helical windows in order to calculate the average helical content of the polypeptide sequence (Torabi and Schatz 2013). AGADIR uses a partition function which is the summation all over the helical conformations along the polypeptide chain:

$$Z = \int e^{-u(r)/k_B T} dr = \int_{rc} e^{-u(r)/k_B T} dr + \sum_{ij} \int_{ij} e^{-u(r)/k_B T} dr \quad (2.1)$$

where Z is the partition function, r is the configuration space vector and $u(r)$ is potential energy as a function of the molecule's configuration, k_B is the Boltzmann's constant and T is the temperature. The partition function is the summation of the partition function of the entirely random-coil (rc) conformation and the summation over partition functions of all possible i,j helical windows. Subscripts i and j indicate the number of the first residue and the length of each helical window, respectively. Equation (2.2) presents a single-segment approximation of the algorithm known as AGADIR-1s. However, we can generalize this equation to formulate conformations which include more than one helical segment formed along the chain (AGADIR-ms model) (Muñoz et al. 1997).

Dividing Z by the partition function of the entirely random-coil conformation is equivalent to shifting the free energy by a constant term, and results in:

$$Z = 1 + \sum_{ij} K_{ij} \quad (2.2)$$

where K_{ij} is the probability of formation of the i,j helical window relative to an entirely random-coil conformation:

$$K_{ij} = \frac{\int_{ij} e^{-\beta u(r)} dr}{\int_{rc} e^{-\beta u(r)} dr} ; (\beta = 1/k_B T) \quad (2.3)$$

Thus, the free energy change of forming a helical window, relative to the entirely random-coil conformation, would be:

$$\Delta G = -k_B T \log K_{ij} \quad (2.4)$$

So far, we discussed how the AGADIR algorithm formulates the partition functions of different helical windows along a given polypeptide chain. In the following sections, we discuss how various free energy terms are estimated from the experimental data and how AGADIR uses several thermodynamic models to make predictions for a large range of pH, ionic strength, and temperature (Muñoz et al. 1995).

2.3- Energies involved in AGADIR model

In AGADIR the free energy of a helical window, contains the following energies: Intrinsic helical propensity (entropy loss), electrostatic interactions, interactions with helix macro-dipole, energy associated with hydrogen bonding, interaction with non-helical residues, and side-chain interactions (Lacroix et al. 1998). We will explain all these energies in this section of the report.

In this report, we use the nomenclature below to refer to different positions in the α -helix:

$N'' \quad N' \quad N_{cap} \quad N1 - N2 - N3 \dots \dots \dots C3 - C2 - C1 - C_{cap} \quad C' \quad C''$

$STC \quad STC \quad STC \quad - He - He - He \dots \dots \dots He - He - He - STC \quad STC \quad STC$

where *STC* (S, strand; T, turn; and C, coil) indicates a non-helical conformation and *He* is a helical residue (Richardson and Richardson 1988).

2.3.1- ΔG_{int} : Intrinsic helical propensity (entropy loss)

This term is explained as the free energy required to fix the dihedral angles of an amino acid residue in α -helical angles which reflects the loss of conformational entropy between the helix and random-coil states. Also other energies, such as the changes in solvation and van der Waals' contacts of the side-chain with the helix should be included here. These components vary according to helix position, as shown theoretically and experimentally for some amino acid residues at the first helical turn of an α -helix. In AGADIR different amino acids can have different intrinsic helical propensities at positions N1, N2, N3 and N4, according to the experimental evidence obtained in poly-alanine-based peptides (Petukhov et al. 1998). AGADIR includes the different helical propensities of either neutral or charged titratable amino acid residues as well (Chakrabartty et al. 1994).

2.3.2- ΔG_{elec} : Electrostatic interactions

The electrostatic model used in AGADIR, includes all electrostatic interactions between two helical, or one helical and one non-helical charged groups, the helix macrodipole and charged helical and non-helical residues. Since electrostatic interactions are distance dependent, determining the distance between charged amino acids is the main step to begin calculating the electrostatic interaction between them (Lacroix et al. 1998). Also, the electrostatic interactions change with the pH of the solution. Depending on the electrostatic environment, the pKa of ionizable groups in a peptide change from their standard values.

We have assumed that there is no energy coupling, other than electrostatic, between residues in the random-coil reference state. However, there are examples to the contrary, e.g. β -turn conformations are found in short peptides in aqueous solutions.

2.3.2.1- Ionic strength, pH and temperature dependence

The model presented for calculating the electrostatic interaction between two charged amino acid residues clearly explains pH, temperature and ionic strength dependence. Electrostatic interactions exponentially decrease with increasing charge-to-charge distances and ionic strength. Different salts at high ionic strength have different effects and the only calibration was done for NaCl (on a neutral peptide) by Baldwin's group (Scholtz et al. 1991). AGADIR's prediction is accurate for all salts at low ionic strength (less than 0.1 M) and for NaCl up to 1 M. Other contributions, such as the effect of salt on hydrophobic interactions, are considered of less importance and are not included.

2.3.2.2- Calculation of distances between charged groups

The database used in this part consists of 279 proteins (Muñoz et al. 1994) and it is included in the program WHATIF (Vriend 1990). It was assumed that the average distance between two charged amino acid residues at different positions of an α -helix in the protein database reflects the average distance found in a helical peptide, and the same applies to the random-coil. In the case of the interaction of a charged group with the helix dipole, they have measured the average distances of charged groups at different positions in the helix and the first four amide groups or the last four carbonyl groups. Previously, they neglected the electrostatic interactions of residues outside the α -helix conformation with the helix macrodipole and/or residues within the helix. However, the maximum distance between an Asp side-chain at position N' and the first four amide groups of the helix is short enough to create electrostatic interactions.

To take into account these interactions they have empirically introduced a linear dependence of the distance with the number of residues separating the nonhelical charged residue and the cap positions. For residues N' and C', the distance is 6 Å°. The separation distance increases by 3 Å° for every extra position after the N' or C' position. In the case of the interaction of charged non-helical residues with charged helical residues, only the caps and residues N' and C' have been considered (Lacroix et al. 1998).

2.3.2.3- Calculation of electrostatic energy between charged groups

The electrostatic contribution of two charged residues on an α -helix is the difference between the electrostatic interaction in the helical and random-coil states:

$$\Delta G_{elect} = G_{Hel} - G_{rc} \quad (2.5)$$

The electrostatic interaction can be obtained using Coulomb's equation:

$$G = \frac{e^2 q_i q_p}{3\pi\epsilon_0\epsilon_r r_{ip}} \times \exp(-K \times r_{ip}) \quad (2.6)$$

where e is the charge of the electron, q_i and q_p are the charges of the two residues, r_{ip} is the distance between the two charges, ϵ_0 is the vacuum permittivity, ϵ_r is the dielectric constant of the medium and K is the Debye-Huckel parameter, which is defined by the following equation:

$$K = \left(\frac{8\pi e^2 N_A I}{1000kT} \right)^{1/2} \quad (2.7)$$

where I is the ionic strength of the solution, N_A is Avogadro's number, k is the Boltzmann's constant and T is the temperature. Therefore, the dependence of electrostatic interaction between two charged residues on ionic strength of the solution and temperature can be explained by Debye-Huckel parameter.

All electrostatic interactions (including charged side-chain groups, free N-terminal and C-terminal main-chain groups, and the succinyl blocking molecule if the peptide is succinylated)

are considered in AGADIR to calculate the electrostatic contribution of the amino acid residues in the random-coil and helical segments, taking into account the ionic strength and assuming full charges ($q_i=q_p=1$ in the Equation 2.6) (Lacroix et al. 1998). Now the pK_a of the residues in the random-coil ($pK_{a(RC)}$), and in the corresponding helical segment ($pK_{a(Hel)}$), are calculated using equations below:

$$pK_{a(RC)} = pK_{a(ref)} + G_{RC}/2.3RT \quad (2.8)$$

$$pK_{a(Hel)} = pK_{a(ref)} + G_{Hel}/2.3RT \quad (2.9)$$

Using the pK_a values, the degrees of ionization are obtained from equation below:

for basic amino acid residues:

$$q = 1/(1 + (10^{pH}/10^{pK_a})) \quad (2.10)$$

for acidic amino acid residues:

$$q = 1/(1 + (10^{pK_a}/10^{pH})) \quad (2.11)$$

2.3.3- ΔG_{dip} : Interaction with helix macro-dipole

All electrostatic interactions between the helix macrodipole or the free N and C termini and groups located in the α -helix are considered in AGADIR as well as the interactions of the helix macrodipole with charged groups located outside the helical segment. Half a charge is assigned to the helix macrodipole (positive at the N terminus and negative at the C terminus). Moreover, there is an effect of ionic strength on helix stability that is due to the helix macrodipole but is independent of the presence of charged amino acids residues. It is suggested that increasing the ionic strength should stabilize the α -helix by shifting the equilibrium between the helical conformation (which should have a large dipole moment) and the random-coil (which has a very small dipole moment, due to the random orientation of the dipoles carried by the

peptide bonds). Experimental characterization of a neutral soluble peptide, under different ionic strength conditions, supported this assumption (Scholtz et al. 1991).

2.3.3.1- Effect of salt concentration on helix stability through macrodipole

The dependence of $\Delta\Delta G$ with salt, is empirically fitted to the following equation (Lacroix et al. 1998; Scholtz et al. 1991):

$$\Delta\Delta G_{hel} = -\alpha(1 - \exp(-\beta I)) \quad (2.12)$$

where $\Delta\Delta G_{hel}$ is the difference in folding free energy of a particular α -helix segment in a solution with ionic strength I and in pure water. The values of α and β are, respectively, 0.15 and 6. This energy has been added to every helical segment in AGADIR. For low ionic strengths (below 0.15 M), the effect on ΔG seems to be linear and similar for different salts. Above this value the dependence follows above equation until approximately 1 M salt (Lacroix et al. 1998).

2.3.4- ΔG_{HB} : Hydrogen bonding

Hydrogen bonds are one of the major structural determinants of an α -helix. In the α -helix structure, the N-H group of an amino acid forms a hydrogen bond with the C=O group of the amino acid at four preceding residues along the structure ($i, i+4$ hydrogen bonds). This repeated hydrogen bonding pattern is the most prominent characteristic of an α -helix. The term ΔG_{HB} reflects the sum of the main-chain-main-chain enthalpic contributions, which include the formation of main-chain-main-chain $i, i+4$ hydrogen bonds (Lacroix et al. 1998).

2.3.5- ΔG_{nonH} : Interaction with non-helical residues

At capping positions, AGADIR distinguishes between the neutral and charged forms of Cys, His and Asp. Neutral Cys is a poor hydrogen-bond donor or acceptor and therefore not a good N-capping residue, but when it is charged, it can interact with the helix macrodipole and

simultaneously make a charged hydrogen bond with the amide group of residue N4. On the other hand, His cannot make a hydrogen bond with the helix N terminus and has a strong repulsion from the helix macrodipole when it is charged (Sancho et al. 1992). From the experimental data for Cys (Kortemme and Creighton 1995) and His (Armstrong and Baldwin 1993), Serrano's group has evaluated that the N-capping contribution of these two residues is 1 kcal/mol more favorable when Cys is charged or His is neutral. When Asp is not charged, it can make a side-chain hydrogen bond to the carbonyl group of residue C3, the same as Asn. Under these conditions, Asp has the C-capping value of Asn (Lacroix et al. 1998).

The contribution of local sequence motifs involving the interaction of a residue outside the α -helix with a helical residue is added to ΔG_{nonH} in AGADIR. First, a free energy term adds up to the energy of helical segments that contain a hydrophobic staple (Muñoz et al. 1995) or a Schellman motif (Schellman 1980; Aurora et al. 1994; Viguera and Serrano 1995). Second, the capping property of the N-cap is modified if a "capping box" (Harper and Rose 1993) is present; that of the C-cap if there is a Pro-capping motif (Prieto and Serrano 1997). In addition, the combination of free N terminus, capping box motif and an Asp or a Glu at position N4 (Petukhov et al. 1996), contributes -1 kcal/mol to the stability of a helical segment. The stabilization is due to a strong interaction between residue N4, the N-capping residue and the charged N-terminal group.

Using the program WHATIF, strong position dependence for pairs of residues were found (Vriend 1990). They could reflect the formation of local motifs that were not previously described. There are three positive cases found in AGADIR peptide database (Lacroix et al. 1998):

(1) A variant of the capping box motif in which the side-chain of a Thr at position N3 can make a hydrogen bond to the side-chain of Asp, Ser or Asn at position N-cap. The abundance of these pairs is twice what is expected from a random distribution (Prieto and Serrano 1997). The experimental analysis of a protein fragment corresponding to the α -helix of the B1 domain of protein G, with and without this motif, is used to assign the value of a hydrogen bond energetic contribution to this interaction (Blanco et al. 1997). Similar side-chain-side-chain interactions could be established with the side-chain of a Ser at position N3. The expected numbers of cases from a random distribution correspond, however, to the observed numbers and there is no experimental evidence for a stabilizing effect (Lacroix et al. 1998).

(2) A local interaction, similar to the hydrophobic staple motif but involving a charged residue, takes place between a Lys or Arg at position N4 and the main-chain carbonyl group of residue N', when there is a Ser or a Thr at the N-cap. Due to the capping by Ser or Thr, the carbonyl group of the preceding residue points towards residue N4, and is at the right distance to form a hydrogen bond with the side-chain of the basic residues. A value of -0.3 kcal/mol is assigned to the interaction based on the experimental analysis of a peptide series (Lacroix et al. 1998).

(3) Histidine residues at position C1 or the C-cap are frequently guarded by an aromatic residue at C5 (11 cases out of 33) or C4 (12 cases out of 56), respectively, while in the helix center only eight histidine residues out of 165 cases are paired with an aromatic residue at position i-4. The optimal packing of Phe or Tyr side-chains with a histidine residue at position i+4 requires that the residue at position i adopts a rotamer of approximately 180° , and the i+4 residue of approximately -60° . This corresponds to the C5/C1 and C4/C-cap cases (in the helix center the favorite rotamer for His is approximately 180°). Using the analyzed peptides

(Armstrong and Baldwin 1993), in which an $i, i+4$ Phe-His pair is placed at different positions of a polyalanine helix, it is found that the aromatic residue-His pair is three times stronger when the histidine residue is the C-cap or C1 residue (Lacroix et al. 1998).

2.3.6- ΔG_{SD} : Side – chain – side – chain interactions

This term includes side-chain-side-chain interactions between non-charged residues as well as the $i, i+1$, $i, i+3$ and $i, i+4$ electrostatic interactions. In AGADIR, only side-chain-side-chain interactions inside the α -helix are considered as ΔG_{SD} , excluding the electrostatic component between charged groups ($\Delta G_{\text{electrostatic}}$). In some cases, when there is a hydrogen bond between two side-chains that involves a titratable amino acid residue, i.e., Gln-Asp, Gln-Glu (Huyghues-Despointes et al. 1995; Huyghues-Despointes et al. 1997; Smith and Scholtz 1998), the strength of the interaction is pH dependent but independent of salt concentration. In those cases a constant term is added to the side-chain-side-chain energy when the corresponding residue is charged and this term is taken into consideration when calculating its pKa. The values of some side-chain-side-chain interactions were modified, according to some peptide studies by different groups (Huyghues-Despointes et al. 1995; Huyghues-Despointes et al. 1997; Smith and Scholtz 1998; Padmanabhan and Baldwin 1994; Viguera and Serrano 1995; Stapley et al. 1995; Stapley and Doig 1997). Some interactions that were assumed to be attractive based on chemical similarity between amino acid residues, are now repulsive, i.e. the $i, i + 4$ interactions between aromatic side-chains (including His) and β -branched aliphatic residues (Val and Ile). The repulsion comes from the steric incompatibility between these side-chains in the middle of an α -helix (Creamer and Rose 1995). In AGADIR the interaction between an aromatic group and a Cys side-chain is considered as well (Viguera and Serrano 1995).

3- CHAPTER 3: Introduction to coiled-coil protein structures

3.1- Introduction

Coiled-coils are structural motifs in which two or more α -helices come together and make the structure of a dimer, trimer, etc. depending on the number of helices. Coiled-coils contribute to a wide range of cellular functions in both fibrous and globular proteins (Rose et al. 2004). In fact, many structural proteins both inside and outside of cells (keratins, tropomyosin, laminin etc.) that have to bear considerable stress contain a coiled-coil domain. Coiled-coil stabilization is mostly driven by the tendency to bury the hydrophobic side chains in between the two strands so that for the most part, they are not accessed by polar water molecules. Therefore; stability of coiled-coils greatly depends on hydrophobic interactions (Cohen and Parry 1990; Lupas et al. 1991; Crick 1953; Pauling and Corey 1953). In the next sections, we will explain the structure of a coiled-coil in more detail.

Short coiled-coil domains (containing 30 to 50 residues) function as selective dimerization units called leucine-zippers within the broad family of transcription factors (Newman and Keating 2003). Coiled-coil dimerization plays regulatory functions for sensing environmental signals such as temperature, pH and solute concentration. Long rod-shape coiled-coil domains function as shafts or lever arms for force transduction (Blankenfeldt et al. 2006; Kohori et al. 2011), and provide flexibility and extension to cytoskeletal scaffolds and extracellular matrices (Kammerer 1997; Adams et al. 2008).

Coiled-coil domains are also found in the extracellular matrices (Kammerer 1997) where their mechanical properties play an important role in cellular processes such as migration (Sheetz et al. 1998) and differentiation (Engler et al. 2006). Long dimeric coiled-coils transmit mechanical load and possibly act as reversible springs within all three families of motor proteins:

myosin, kinesin and dynein (Taniguchi et al. 2010; Parry et al. 2008; Li et al. 2003; Knight et al. 2005). Myosin's tail domain is a long coiled-coil structure that has critical mechanical and regulatory functions in this diverse family of motor proteins.

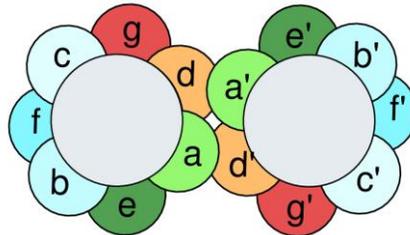
In general sense, protein engineering is expected to create more efficient and practical solutions to treat neurodegenerative diseases such as Alzheimer's disease (AD) and Parkinson's disease (PD), which are increasingly being realized to have common cellular and molecular mechanisms including protein aggregation. For instance, Ferritin, another ubiquitous soluble protein whose function is to store non-heme iron molecules in its sub-units, has been found to play a major role in physiological interactions in the human brain, investigated by several imaging studies (Liu et al. 2016; Ghassaban et al. 2019; Sethi et al. 2019; Chen et al. 2018; Haacke et al. 2018; Wiseman et al. 2019).

3.2- Molecular structure of a coiled-coil protein structure

Coiled-coils usually contain a repeated pattern of seven amino acid residues, referred to as the heptad repeat (Mason and Arndt 2004). A heptad-repeat of amino acids, labeled as *a b c d e f g*, characterizes the sequence of a canonical coiled-coil structure (see Figure 3.1). The residue positions in the heptad repeat are also labeled as “*a b c d e f g*”, where oppositely charged residues frequently found at positions *e* and *g* of the coiled-coil heptad contribute to the stability and specificity of the dimers (Kohn et al. 1995; Monera et al. 1994; Krylov et al. 1994; Krylov et al. 1998; Greenfield 2006). Also, *a* and *d*, which are known as hydrophobic positions, are often occupied by isoleucine, leucine, or valine. The most favorable way for two helices to arrange themselves in an aqueous environment is to pack the hydrophobic side-chains against each other sandwiched between the hydrophilic amino acids. Thus, as mentioned above, it is the burial of hydrophobic surfaces that provides the major thermodynamic driving force for the coiled-coil

dimerization (Acharya et al. 2006; Monera et al. 1995; Litowski and Hodges 2002; Wagschal et al. 1999; Tripet et al. 2000; Acharya et al. 2002; Moitra et al. 1997). We will explain these interactions in more detail in the following sections.

A



B

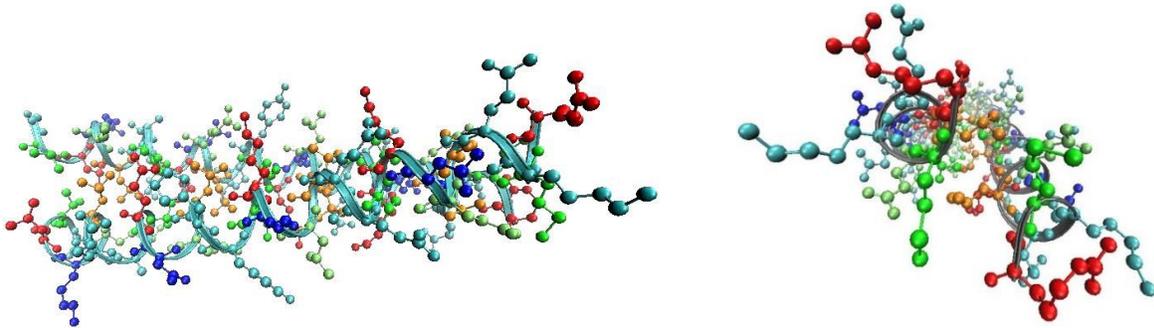


Figure 3.1 Structure of a coiled-coil dimer (A) helical wheel diagram of the heptad repeat in a parallel coiled-coil dimer (B) side view of a coiled-coil dimer. Typically, charged amino acids are located at positions *g* (colored red) and *e* (colored dark green) which interact electrostatically.

Side-chains at positions *a* and *d* form the hydrophobic core of a coiled-coil dimer.

3.3- Developing a statistical mechanical model for coiled-coil dimerization

As mentioned before, the main objective of this work is to develop a theoretical model that can predict the propensity of a given protein sequence to form dimeric coiled-coil structures,

coiled-coil specificity, and also the structural response of a coiled-coil forming domain to environmental signals of temperature, pH, salt concentration.

Available coiled-coil predictive tools are mainly based on some form of a sequence scoring rubric trained against a large data set of known coiled-coil structures (Fong et al. 2004; Gruber et al. 2006). In general, sequence scoring techniques do not capture the effects of solution conditions such as temperature, salt concentration and pH (Pauling and Corey 1953; Fong et al. 2004; Gruber et al. 2006; Lupas and Gruber 2005). We will discuss an example of this technique in the next sections.

Besides sequence scoring methods, molecular dynamics (MD) simulation can also be used to study the structure of α -helical and coiled-coil proteins. However, in order to make quantitatively reliable predictions of polypeptide mechanical properties, improved force fields and solvation models still need to be developed (Best and Hummer 2009). At the same time, equilibrium sampling using MD simulation of a polypeptide pulling experiment is a computationally expensive task. Therefore, MD simulation is not the best approach in developing a predictive tool for tensile mechanics of protein structures that would find a widespread application among a broad range of experimental and theoretical fields.

Statistical mechanical approaches have been highly successful in predicting the secondary structure of RNA and DNA duplexes (SantaLucia and Hicks 2004; Dimitrov and Zuker 2004; Zuker et al. 1999; Reuter and Mathews 2010). These models are based on a unified set of thermodynamic (SantaLucia 1998) parameters and are able to capture the effect of environmental conditions such temperature and salt concentration of the nucleic acid structures (SantaLucia and Hicks 2004; Tan and Chen 2006; Wu et al. 2002). Statistical mechanical models, combined with a proper tensile mechanics model have been used to interpret AFM and

optical tweezers single molecule forced extension experiments of nucleic acid molecules (Huguet et al. 2010; Rouzina and Bloomfield 2001; Williams et al. 2002).

Also, as mentioned before, the AGADIR (Lacroix et al. 1998) algorithm makes quantitatively accurate predictions of a polypeptide sequence considering the effect of a large range of temperature, pH and ionic strength and is based on several thermodynamic models and a large set of experimental data for different polypeptides in aqueous solutions (Muñoz and Serrano 1995). Therefore, the above-mentioned facts support our statistical mechanical model for predicting coiled-coil dimerization, specificity, and tensile mechanics.

To develop our model, we need to consider two different states:

a) First we assume the two strands of interest are not interacting with each other and there is no overlapping between any sections of the strands. Therefore; the coiled-coil is not formed in this state.

We should find the microstates based on the helical windows (Torabi and Schatz 2013), which is the same approach as in AGADIR model. Each strand has a certain number of heptad repeats and residues. Therefore; we can write the probability of formation of i,j and i',j' helical windows for the first and second strands, respectively, as:

$$K_{ij} = \frac{\int_{ij} e^{-\beta u(r)} dr}{\int_{rc} e^{-\beta u(r)} dr} \quad (3.1)$$

$$K_{i'j'} = \frac{\int_{i'j'} e^{-\beta u(r)} dr}{\int_{rc} e^{-\beta u(r)} dr} \quad (3.2)$$

Where β is defined as $1/k_B T$, k_B is the Boltzmann's constant and T is the temperature, r is the configuration space vector and $u(r)$ is the potential energy as a function of the molecule's configuration with respect to each i,j and i',j' helical window. We should notice that these probabilities are relative to random-coil conformation.

Because there is no interaction between two strands and the probabilities of formation of i,j and i',j' helical windows are independent of each other, we can write the partition function as follows:

$$Z = (1 + \sum_{i,j} K_{ij}) (1 + \sum_{i',j'} K_{i'j'}) \quad (3.3)$$

b) Now we assume that two strands are interacting and wherever the two helical windows are overlapping, there is a possibility of coiled-coil formation, considering the interactions between residues in the overlapping segments. Since we know the amino acid sequences of both strands, we can distinguish the heptad repeats in the overlapping part(s) and therefore we can determine the interactions and energy contributions.

In this case, since there are interactions between helices we should add an extra term to our partition function which denotes these interactions:

$$Z_{cc} = \sum_k (1 + \sum_{i,j} K_{ij}) (1 + \sum_{i',j'} K_{i'j'}) e^{-\beta \Delta G_{cc,k}} \quad (3.4)$$

The term $\Delta G_{cc,k}$ is the free energy change of a given coiled-coil structure formed between the two strands in state k . ΔG_{cc} includes various free energy terms associated with hydrophobic, electrostatic, and entropy effects of coiled-coil formation. The first summation term applies to the state in which there is no interaction between the helices, and the second summation term is when there is overlapping between the helices and that is the reason we have included the coiled-coil free energy term in the above equation.

We can find K_{ij} and $K_{i'j'}$ from previously introduced model, AGADIR, and once we derive the data forming the term ΔG_{cc} for different sequences of amino acids, we can obtain the partition function. After we derive the partition function, we can use our model to monitor the behavior of a coiled-coil structure under given solvent conditions.

Here are some assumptions we have made to develop our model: we assume there is only one helical window in each strand. It is a fine assumption especially if each polypeptide has less than 56 residues. In this case even multi-helical segments have almost the same result as single-helical segments, with an error of nearly 0.3 percent (Torabi and Schatz 2013). Therefore; the possibility of having more than one helical segment in each strand is very low. Another point is that in our model, we have neglected mismatches, meaning that we assume there is no possibility that strands shift and form another coiled-coil with another section, and only corresponding regions of strands can dimerize. Also we should know the sequences of both chains, temperature, pH, and ionic strength as inputs of the model in order to predict the structure of the coiled-coil, the same as in AGADIR model.

3.4- Coiled-coil dimerization energies

In the previous section, we derived the partition function for the coiled-coil dimerization model. The most important term in Equation 3.4 is the coiled-coil dimerization free energy (ΔG_{cc}). The majority of coiled-coil inter-chain energetic effects are of the following two types: the electrostatic interactions of the charged residues across the two strands, and the change in the solvation free energy of the residues upon dimerization (hydrophobic effect).

Both interactions appear in the context of a characteristic heptad-repeat sequence which was explained in the previous sections. After deriving these two major energies, we add them to the energies for a single helix, which were used in AGADIR model, and therefore calculate the total dimerization energy of a coiled-coil dimer. The next sections of this report include detailed information on these two types of energies.

4- CHAPTER 4: Electrostatic Interactions and their effect on coiled-coil specificity

4.1- Introduction

Regulating the protein-protein interaction specificity has a wide range of applications in synthetic biology such as protein labeling and purification in high-specificity affinity-tags or cognate-pairs, drugs and toxin delivery and disease modulation (Crooks et al. 2017). In naturally occurring proteins, specificity is achieved via a complex balance of various molecular-level energetic and entropic interactions. Such complexity makes any specificity prediction from the primary sequence data an extremely complicated task.

In a coiled-coil dimer, g and e' pairs occupied by oppositely charged residues provide an additional driving force to formation of parallel coiled-coils (Burkhard et al. 2002; Zhou et al. 1994). The indices indicate the location of the residue along each strand and prime distinguishes the two opposite helical chains. Electrostatic interactions among $g-g'$ and $e-e'$ pairs play the same role in antiparallel coiled-coils. Numerous mutational studies have been carried out to measure the electrostatic interactions within coiled-coil dimers. Replacing attractive electrostatic interactions at g and e' pairs (e.g. Glu-Lys) with a repulsive ones (e.g. Glu-Glu or Lys-Lys) compromises the coiled-coil stability of a parallel coiled-coil dimer (Zhou et al. 1994). Electrostatic interactions have been used to create desired orientation (i.e. parallel vs. antiparallel) in synthetic coiled-coil structures (Monera et al. 1994). Using b-ZIP protein VBP, alanine double-mutant thermodynamic cycles have been carried out to measure the electrostatic interaction of various charged residues at $g_i-e'_{i+5}$ positions within parallel heterodimers (Krylov et al. 1994). At $g_i-e'_{i+5}$ positions, Arg-Glu is shown to be the most favorable electrostatic interaction followed by Lys-Glu, Glu-Arg and Glu-Lys (Krylov et al. 1998). The asymmetry of the electrostatic interactions at $g_i-e'_{i+5}$ positions (e.g. Glu-Arg vs. Arg-Glu) is justified based on

directional asymmetry of the right-handed α -helical structure (Zhou et al. 1994). Repulsive electrostatic interactions at $g_i-e'_{i+5}$ positions are less residue dependent (Kohn et al. 1995). However, a steady loss of stability has been observed upon increasing $g_i-e'_{i+5}$ electrostatic repulsions (Kohn et al. 1995).

In violation of the heptad-repeat pattern, 8% of d positions and 11% of a positions of known coiled-coils are occupied by charged residues: Asp, Glu, Lys and Arg (Akey et al. 2001). Also, 6% of a positions and 11% of d positions are occupied by polar residues: Ser, Thr, Asn and Gln (Akey et al. 2001). In general, buried charged or polar residues undermine the stability of protein structures. Similar effect has been observed in short coiled-coil structures (Chao et al. 1998). Nonetheless, buried polar and charged residues greatly influence the oligomerization state (i.e. dimer vs. trimer vs. tetramer) and helix orientation (i.e. parallel vs. antiparallel) of the coiled-coil structures (McClain et al. 2002; Zeng et al. 2002). A well-known case is the Asn-Asn interaction at $a-a'$ and $a-d'$ positions that promote dimerization of parallel and antiparallel coiled-coils, respectively (Schneider et al. 1997; Lavigne et al. 1995). Furthermore, interhelical salt-bridges formed by buried residues at positions a and d with surface residues at positions e' and g' often occur in natural coiled-coils (Schneider et al. 1997). For example, $d-g'$ salt-bridges occur in Escherichia coli seryl tRNA synthetase and the effector domain of the protein kinase PKN (both antiparallel coiled-coils) (McClain et al. 2002). Similar $a-e'$ electrostatic interactions are seen in the transcript cleavage factor GreA, PKN, and in the transcriptional activator protein, MtaN (McClain et al. 2002). Also, $d-e'$ and $a-g'$ salt-bridges are seen in parallel dimers such as b-ZIP transcription factors (Zeng et al. 2002; Reinke et al. 2010; Reinke et al. 2013) and the rod domain of cortexillin I (Burkhard et al. 2000). While electrostatic interactions between surface residues (e.g. $g_i-e'_{i+5}$ in a parallel dimer) undoubtedly influence the coiled-coil coupling,

frequency of salt-bridges formed between buried and surface residues points to their possible contribution in stability and specificity of the coiled-coil structures.

With regards to the pairing specificity, the leucine-zipper domain of the b-ZIP class of transcription factors in eukaryotic cells is, by far, the most studied coiled-coil dimer (Newman and Keating 2003). b-ZIP dimerization provides a perfect example of how a great degree of protein-protein interaction specificity among a large family of structurally similar proteins can be encoded in the primary sequence of relatively short (about 30 residues long) helical segments (Fong et al. 2004). The first extensive study of leucine-zipper pairing specificity was carried out in protein microarrays evaluating 49×49 pairings among a nearly complete set of human b-ZIP transcription factors across sixteen different families (Newman and Keating 2003). The data set has been further expanded to include 48 synthetic coiled-coils (Reinke et al. 2010) along with b-ZIP transcription factors of four other metazoan species of sea squirt, fruit fly, nematode and sea anemone along with two single-cell organisms, choanoflagellate and yeast (Reinke et al. 2013). Despite their homologous primary structure, the above studies provided an unambiguous evidence of a highly selective coiled-coil dimerization with almost no interactions across different families of b-ZIP transcription factors.

Stability of an isolated coiled-coil and the pairing specificity of a given sequence are the results of a complicated interplay among various energetic and entropic effects. Relative to the random-coil state, free-energy of each helical strand consists of the helical propensities (entropy loss of fixing backbone dihedral angles in the α -helical state) of all the residues, backbone and side-chain hydrogen bonding, electrostatic interaction among polar and charged residues and interaction of the charged residues with the helix macrodipole (Lacroix et al. 1998). The inter-helical free-energy terms include the change in the solvation free-energies upon formation of the

hydrophobic core, van der Waals and steric interactions at the coiled-coil interface, entropy loss due to limited rotamer states of the side-chains at the interface, inter-helical hydrogen bonding, and inter-helical electrostatic interactions among charged residues.

The earliest attempts to predict the specificity of b-ZIP dimerization were based on $g_i-e'_{i+5}$ electrostatic complementarities (Vinson et al. 1993; Parry et al. 1977; McLachlan and Stewart 1975). Vinson et al. (1993) suggested a simple specificity score as the sum of all the attractive minus repulsive $g_i-e'_{i+5}$ pairs. Such specificity score was successful when applied to a limited number of coiled-coil sequences (Vinson et al. 1993). However, when applied to a more comprehensive human b-ZIP data set, the above scoring scheme is shown to be of little utility. In an attempt to improve the above scoring rule, Newman et al. (2003) suggested adding +0.5 for $g_i-e'_{i+5}$ electrostatic attraction involving Asp and the interaction of the polar residue Gln with other charged residues. The new rule correctly assigns a positive score to almost all the 80 strongly interacting b-ZIP sequences examined. However, about 60% of the non-interacting pairs are also predicted to have a positive score. Therefore, the above scoring rule does not distinguish the strongly interacting pairs from the noninteracting ones. Using the experimentally determined (Moll et al. 2002) electrostatic interactions of $g_i-e'_{i+5}$ pairs (instead of only ± 1) does not significantly improve the predictions of the above scoring rule (Newman et al. 2003). Seemingly, non- $g_i-e'_{i+5}$ electrostatic interactions as well as compatibility of hydrophobic packing, van der Waals attractions and steric clashes at the coiled-coil interface play an important role in dimerization specificity of coiled-coils.

In an attempt to improve the coiled-coil specificity predictions, Fong et al. suggested a scoring rule based on seven different inter-helical interactions among the residues at positions a , d , e and g (Fong et al. 2004). Pair-interaction scores were determined based on an optimization

problem with constraints defined according to a training dataset of known coiled-coil interaction in myosin, tropomyosin, cortexillin, types III and V intermediate filament proteins and coiled-coil domains in keratin proteins. The optimized scores perform significantly better than simple $g_i \cdot e'_{i+5}$ scoring rules when applied to a dataset of 80 strongly interacting and 849 non-interacting pairs of human and yeast b-ZIP coiled-coils. Along the same lines, a more accurate specificity predication has been achieved by a recent scoring model optimized according a machine-learning technique that includes both pair and triplet interactions (Potapov et al. 2015).

The above specificity scoring algorithms are all developed based on empirical data of coiled-coil dimerization without resorting to atomic-resolution protein structures. Grigoryan, et al. (2006) examined multiple variations of structure-based models to predict b-ZIP specificity based on interaction energies estimated by various protein force-fields. They showed structure-based methods do not capture the core interaction accurately resulting in poor specificity predictions when compared to experimental data. They confirmed that a hybrid-approach of replacing certain interactions with machine-learning weights considerably improves the performance of a purely physical model.

A comprehensive list of various specificity prediction methods and a detailed evaluation of their performance is presented in reference (Potapov et al. 2015). Therein, it is demonstrated that the empirical data-driven scoring algorithms outperform both structure-based and hybrid techniques, when benchmarked against the experimental data (Potapov et al. 2015). This is especially true when the coiled-coil sequences in the testing dataset are similar to those in the training dataset. However, the downside of all scoring approaches is that they provide little to no physical interpretation of various inter-helical interactions (hydrophobic, electrostatic, van der Waals and steric) and their significance to coiled-coil specificity.

What we investigate in this section is the role played by the electrostatic interactions in pairing specificity of the coiled-coils. Here, we only studied a set of sequences that naturally form stable coiled-coils (b-ZIP transcription factors), at least with one other partnering sequence. Since the intra-helical interactions are independent of the opposite strand's sequence, we assume that inter-helical interactions determine the pairing specificity of a coiled-coil forming sequence. As mentioned above, in de novo designed coiled-coil dimers with a regulated interface (e.g. Leu at all *a* and *d* positions) and charged residues only at positions *e* and *g*, it has been demonstrated that the electrostatic interactions at $g_i-e'_{i+5}$ pairs positions determine the pairing specificity (Chao et al. 1998; O'Shea et al. 1993). However, within more irregular sequences, it has been observed that repulsive $g_i-e'_{i+5}$ electrostatic interactions may very well be tolerated in a coiled-coil structure and such simplistic view of electrostatic interaction is not sufficient for pairing specificity predictions (Arndt et al. 2002). Even so, most naturally occurring coiled-coils (e.g. b-ZIP class of transcription factors) require a high degree of specificity that can only be achieved via a more complicated network of interaction beyond simple $g_i-e'_{i+5}$ electrostatic complementarities. Such observations suggest an incomplete picture of the contribution of $g_i-e'_{i+5}$ electrostatic interactions to coiled-coil dimerization specificity (Mason and Arndt 2004; Newman and Keating 2003). In this section of this report, we first demonstrate that the electrostatic interactions among charged residues other than $g(i)-e'(i+5)$ pairs are not negligible within b-ZIP transcription factors. Subsequently, we investigate whether the complete inter-helical electrostatic interaction including both the $g_i-e'_{i+5}$ and non- $g_i-e'_{i+5}$ pairs is capable of making specificity predictions in coiled-coils. If not, the evidence points towards the critical role played by the van der Waals and steric interactions (geometric compatibility at the protein-

protein interfaces) in highly selective partnering behavior of the coiled-coil forming sequences and perhaps more generally in other protein-protein interactions.

Previously, we explained how AGADIR derives the electrostatic interactions between the charged residues of a single polypeptide chain. We are using a similar model to obtain the electrostatic interactions between charged residues of a coiled-coil located on the opposite strands.

4.2- Calculating the electrostatic interactions in a coiled-coil structure

We used Equation (4.1), the linear solution of the Poisson-Boltzmann model (Neves-Petersen et al. 2003) for a pair of point-charges, which is the same approach as shown in Equation (2.6), to calculate the electrostatic interaction between titratable residues within the coiled-coil dimers.

$$G = \frac{e^2 q_1 q_2}{3\pi\epsilon_0\epsilon_r r_{12}} \times \exp(-K \times r_{12}) \quad (4.1)$$

where e is the charge of an electron, q_1 and q_2 are the partial charges of the two residues, r_{12} is the distance between the two point-charges, ϵ_0 is the vacuum permittivity, ϵ_r is the dielectric constant of the medium and K is the Debye-Huckel parameter. The Debye-Huckel parameter is calculated as:

$$K = \left(\frac{8\pi e^2 N_A I}{1000 k_B T}\right)^{1/2} \quad (4.2)$$

where I is the ionic strength (salt concentration) of the solution, N_A is the Avogadro's number, k_B is the Boltzmann's constant and T is the temperature. This model is shown to accurately estimate the electrostatic interactions among solvent-exposed charged residues within isolated α -helices (Lacroix et al. 1998; Sitkoff et al. 1994).

We estimate the partial charges q_1 and q_2 according to acid dissociation constants that are modified with respect to the electrostatic environment of each residue. The acid dissociation constant of a given residue within the coiled-coil structure of known sequence is estimated as:

$$pK_a = pK_a^o + E_{CC}/2.3RT \quad (4.3)$$

where pK_a^o is the reference acid dissociation constant of the amino-acid and E_{CC} is sum of pair-wise electrostatic interactions (calculated by Equation (4.1) using full charges $q_1 = q_2 = 1$) of that residue with all other charged residues of the coiled-coil. E_{CC} includes the electrostatic interactions with other charged residues located along the same strand as well as the opposite strand. We used AGADIR's (Lacroix et al. 1998) database for the distances between residues of the same strand. For two residues located on opposite strands, distances are estimated based on molecular dynamic simulations detailed in the next section.

Having calculated the modified acid dissociation constants within the coiled-coil conformation, we estimate the partial charge of each residue according the following two equations (McClain et al. 2002):

Basic amino-acid residues:

$$q = 1/(1 + (10^{pH}/10^{pK_a})) \quad (4.4)$$

Acidic amino-acid residues:

$$q = 1/(1 + (10^{pK_a}/10^{pH})) \quad (4.5)$$

4.3- Distance between two point-charges

The distances between amino acids along one strand of a peptide are already available in AGADIR (Lacroix et al. 1998) database. To calculate a pair-wise electrostatic interaction using Equation (4.1), first we need to estimate the distance between the point-charges. For all possible combination of two charged residues at all possible heptad positions, we performed an implicit-

solvent molecular dynamics (MD) simulation at temperature 300 K and salt concentration 0.15 M and calculated the average distance over all possible side-chain rotamer states as observed during the MD run. We confirmed that the average distances are not sensitive to neither temperature nor salt concentration.

For the initial structure of each simulation, we used 2ZTA leucine-zipper structure (O'Shea et al. 1991) and mutated it with the two desired residues at desired heptad-repeat positions, using VMD Mutator Plugin (Humphrey et al. 1996). After mutating the structure with residues of interest, we used NAMD (Phillips et al. 2005) to optimize the structure, using CHARMM force-field and continued our MD simulations long enough for the average distance to converge. For charged residues with smaller side chains (e.g. aspartic acid) a 40 ns MD run is sufficient for the average distance to converge (Figure 4.1). For charged residues with larger side-chains (e.g. arginine and glutamic acid) and thus more possible rotamer states, distances are averaged over 80ns or more MD runs.

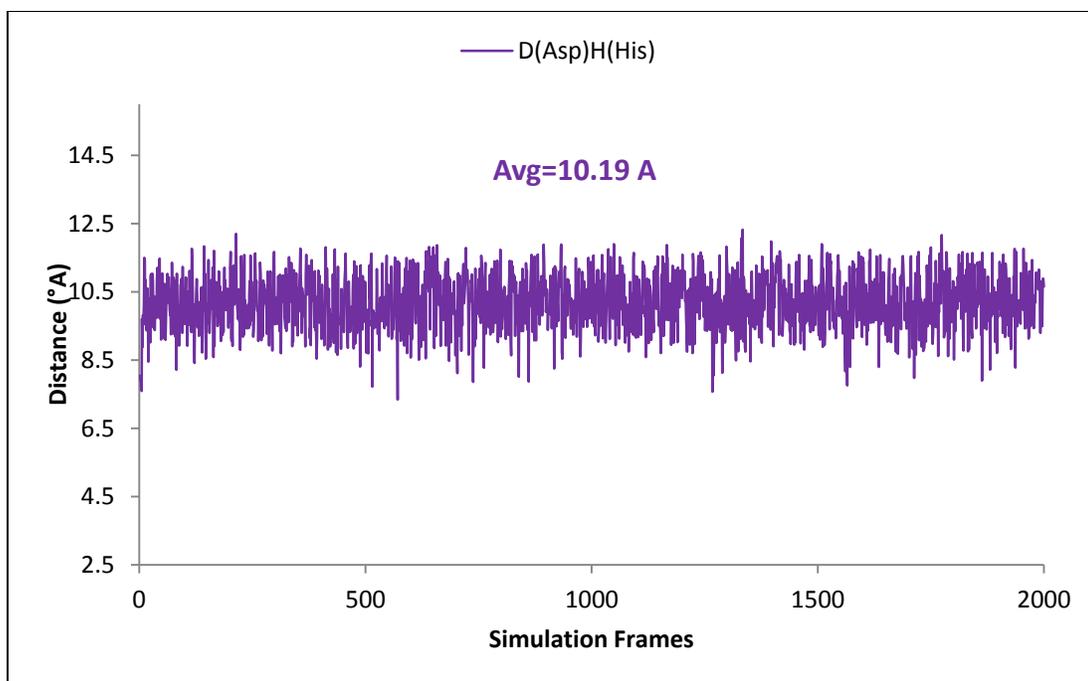


Figure 4.1 Change of the distance between D and H at g_i and e'_{i+5} positions (40ns MD simulation)

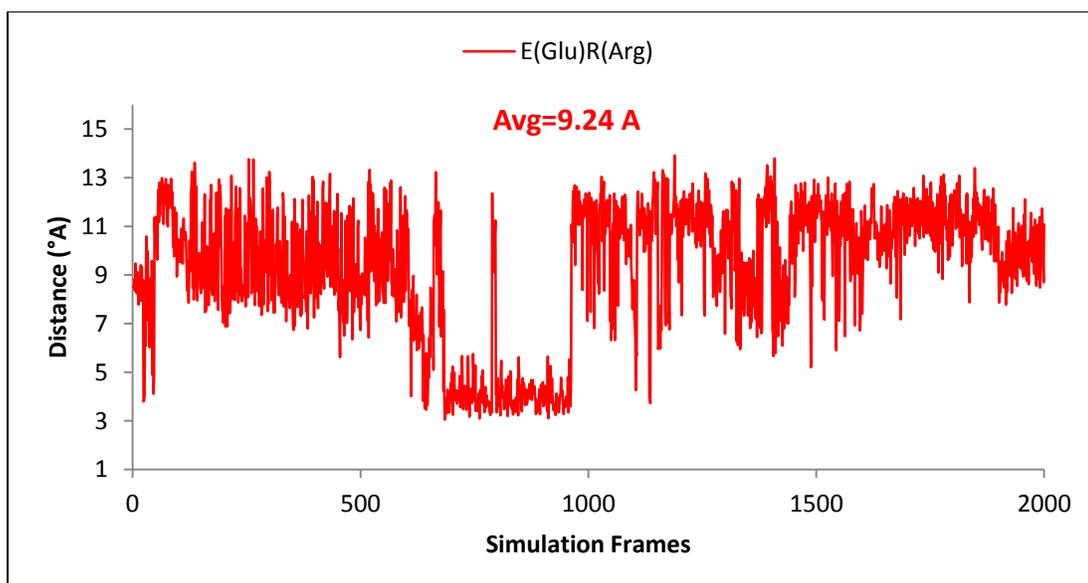


Figure 4.2 Change of the distance between E and R at g_i and e'_{i+5} positions (40ns MD simulation)

For example, Figure 4.2 shows the change of the distance between the charged atoms of glutamic acid (E) and arginine (R) at g_i and e'_{i+5} positions, respectively. We repeated the simulation with exactly the same conditions and plotted the results again, as depicted in Figure 4.3. We ran both simulations for 40 ns. The average distance between these two residues is 9.24 Å and 5.52 Å for the first and second MD simulation, respectively. The reason for such a noticeable difference is that glutamic acid and arginine have longer chains than aspartic acid and histidine. As a result, glutamic acid and arginine can sample multiple rotamer states. Repeating the simulation and having a significantly different average distance means that not all the optimum states had been sampled during the first simulation. Consequently, for these cases, we should run MD for a longer time to assure that we have sampled all the optimum states in order to have a more accurate average value of the distances between charged residues. Figure 4.4 shows the change of distance between charged atoms of glutamic acid and arginine at the same positions of the heptad repeat when we ran MD for 80 ns. We can see that more states are sampled and fluctuations are around considerably different values from the two previous MD runs. In this case the average is 8.08 Å, which is more accurate than what we obtained from 40 ns simulations.

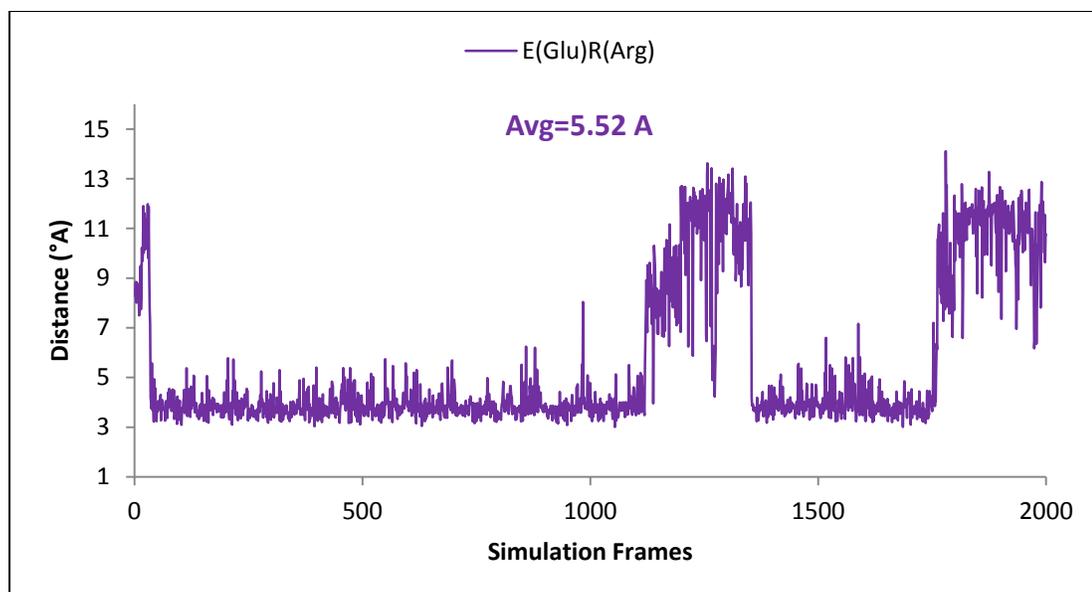


Figure 4.3 Change of the distance between E and R at g_i and e'_{i+5} positions (40ns MD simulation)

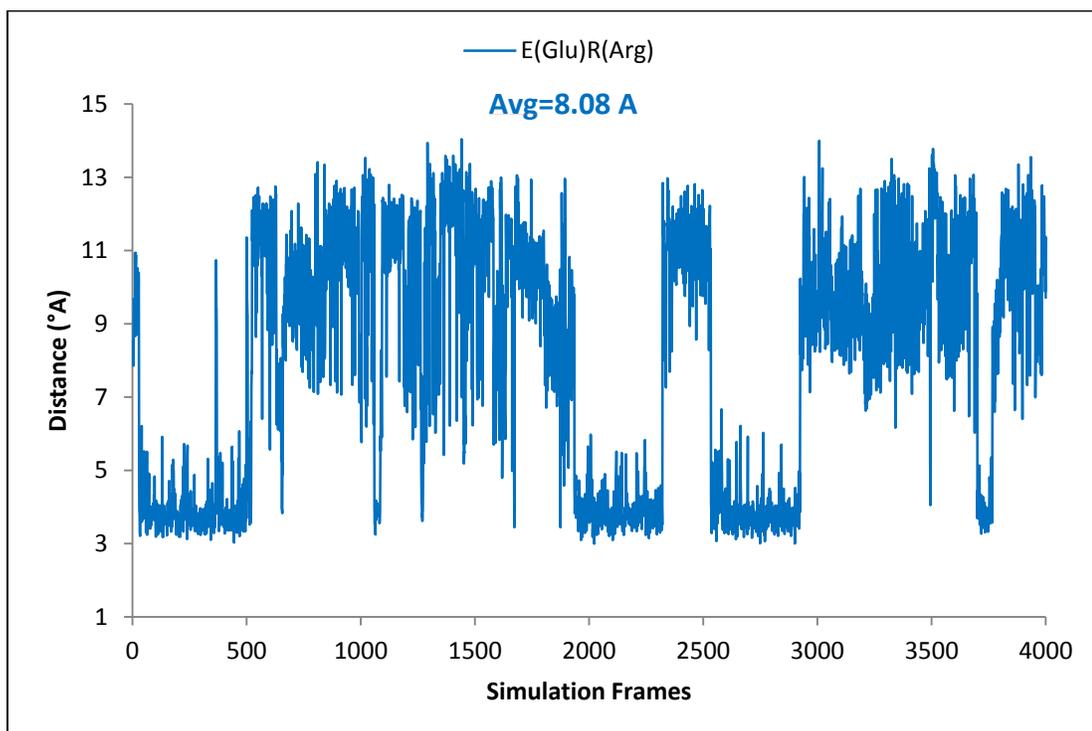


Figure 4.4 Change of the distance between E and R at g_i and e'_{i+5} positions (80ns MD simulation)

All simulations are performed with NAMD's built in Generalized Born Implicit Solvent model with surface tension value of $0.0072 \text{ kcal.mol}^{-1}.\text{\AA}^{-2}$. All simulations are performed while the protein backbone is constrained via a harmonic potential. The values of force constant and the exponent used in the harmonic constraint energy function are both set to $2 \text{ kcal.mol}^{-1}.\text{\AA}^{-2}$. Electrostatic coupling energies are calculated based on charged residues at 8 different heptad position pairs: $g_i e'_{i+5}$, $e_i g'_{i+2}$, $a_i a'_i$, $a_i d'_{i+3}$, $a_i d'_{i-4}$, $a_i g'_{i-1}$, $d_i d'_i$, $d_i e'_{i+1}$, where indices indicate the location of the residue along each strand and prime distinguishes the opposite helical chain. We have confirmed that the distance between charged residues located at any heptad position pairs other than the above 8 are relatively too large to result in any significant electrostatic interaction. Each heptad position in the above 8 pairs may be occupied by any of the 5 charged residues Arg, His, Lys, Asp and Glu, which results in 8×25 pairs (see Figure 4.5). Distance between point charges for all the 200 possible combinations are reported in Table 4.1.

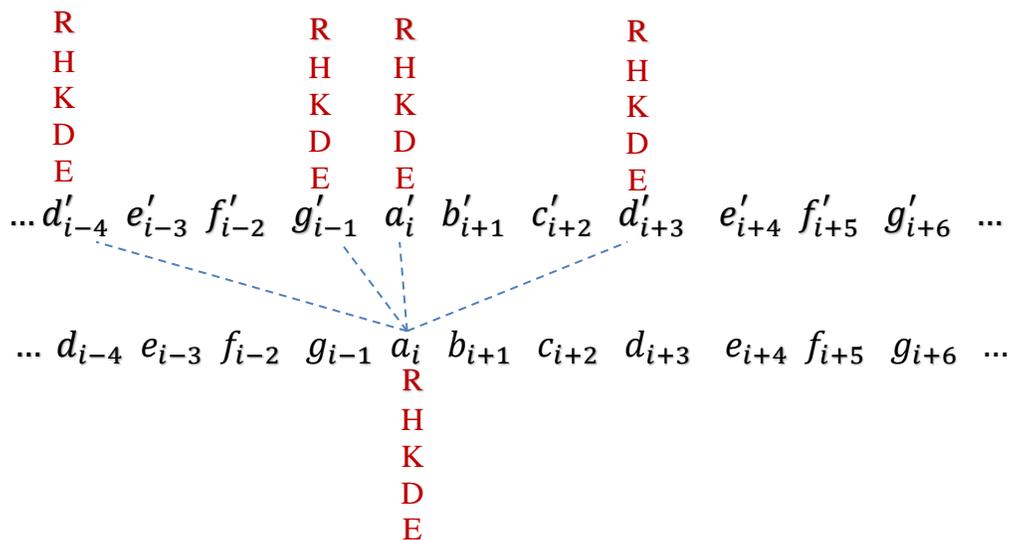


Figure 4.5 Charged amino acids and heptad repeat positions along a coiled-coil dimer

Table 4.1 Distances and coupling energies between charged residues. The coupling energy data are presented at temperature 21°C, pH 7.4 and salt concentration 0.1 M.

positio ns→	g(i),e'(i+5)		e(i),g'(i+2)		a(i),a'(i)		a(i),d'(i+3)	
	distance (Å ²)	Coupling Energy (kcal/mole)						
RE	4.7	-1.098	10.84	-0.257	11.9	-0.105	5.74	-0.503
KE	8.2	-0.443	10.9	-0.254	11.83	-0.106	5.25	-0.598
DR	3.68	-1.553	12.1	-0.203	10.48	-0.138	6.4	-0.267
EE	9.1	0.364	10.34	0.283	9.64	0.163	6.27	1.11
RR	10.63	0.268	10.04	0.301	15	0.061	10.17	0.437
RK	10.14	0.295	14.79	0.127	13.62	0.077	7.85	0.654
RH	9.77	0.027	11.96	0.019	11.47	0.01	7.13	0.103
ED	9.26	0.352	10.2	0.291	8.7	0.198	6.21	1.26
EH	9.97	-0.042	10.69	-0.036	8.57	-0.031	5.48	-0.018
KK	8.35	0.428	12.21	0.199	13.02	0.086	8.37	0.626
KD	9.1	-0.365	10.95	-0.252	11.02	-0.124	4.25	-0.8
KH	10.7	0.024	12.54	0.018	11.56	0.01	7.35	0.127
DD	9.09	0.365	10.4	0.28	6.51	0.33	6.1	1.449
DH	10.19	-0.04	10.36	-0.039	7.34	-0.045	5.98	0.184
HH	10.63	0.002	10.48	0.002	8.82	0.001	4.08	0.004
ER	8.08	-0.562	9.09	-0.365	11.38	-0.1162	7.08	-0.329
EK	7.71	-0.495	11.08	-0.245	11.19	-0.12	4.28	-0.98
RD	8.57	-0.409	13.76	-0.151	10.68	-0.132	3.97	-0.953
KR	9.92	0.308	12.79	0.179	13.99	0.072	8.47	0.585
HR	9.18	0.03	11.79	0.02	11.15	0.011	5.56	0.132
DE	9.39	0.343	11.39	0.231	7.61	0.253	6.23	1.285
HE	9.5	-0.048	10.26	-0.04	9.02	-0.027	5.93	0.096
DK	9.75	-0.319	10.64	-0.267	9.61	-0.164	3.78	-1.052
HK	9.75	0.027	10.72	0.024	10.4	0.012	6.1	0.161
HD	10.15	-0.041	10.5	-0.037	8.2	-0.034	4.87	0.139

positio ns→	d(i),a'(i+4)		g(i),a'(i+1)		d(i),d'(i)		d(i),e'(i+1)	
	distance (Å°)	Coupling Energy (kcal/mole)						
RE	6.32	-0.449	4.11	-1.169	10.6	-0.134	3.92	-1.341
KE	3.7	-1.264	6.62	-0.48	9.9	-0.154	4.5	-1.055
DR	6.02	-0.306	4.95	-0.955	9.69	-0.161	5.92	-0.388
EE	5.69	1.228	5.58	1.007	7.74	0.245	5.77	1.047
RR	6.92	0.741	7.29	0.608	13.3	0.082	5.54	0.938
RK	6.89	0.769	7.12	0.654	12.83	0.089	8.39	0.507
RH	5.94	0.13	6.04	0.047	9.93	0.013	5.3	0.131
ED	5.54	1.431	6.17	1.051	6.86	0.302	6	0.999
EH	5.9	0.094	5.26	-0.201	7.33	-0.045	5.37	0.053
KK	7.25	0.75	7.16	0.648	11.48	0.113	7.46	0.638
KD	4.11	-0.887	6.11	-0.402	8.75	-0.196	4.37	-1.106
KH	5.83	0.162	6.91	0.043	9.39	0.014	6.14	0.159
DD	5.98	1.472	6.13	1.059	5.94	0.383	6.09	1.12
DH	6.54	0.265	4.82	-0.25	6.12	-0.069	6.21	0.248
HH	5.76	0.005	6.48	0.004	6.81	0.001	5.74	0.004
ER	5.85	-0.479	3.79	-1.43	10.6	-0.134	6.61	-0.399
EK	4.13	-0.991	4.45	-1.102	9.63	-0.163	4.92	-0.78
RD	5.43	-0.471	3.43	-1.378	9.91	-0.154	3.82	-1.393
KR	5.76	0.98	6.52	0.72	12.69	0.091	7.86	0.592
HR	7.92	0.099	5.84	0.108	10.2	0.013	7.01	0.042
DE	6	1.3	5.76	0.967	6.95	0.296	6.78	1
HE	4.75	-0.096	5.11	-0.054	7.01	-0.05	6.39	-0.124
DK	4.11	-0.862	4.26	-1.179	8.97	-0.187	4.73	-0.704
HK	7	0.129	5.7	0.134	9.08	0.015	7.2	0.041
HD	5.3	0.133	6.29	0.199	6.11	-0.07	5.84	-0.155

4.3.1- Effect of ion concentration on the distances between charged residues

To understand the effect of different ion concentrations of the solvent on the distances between charged residues, we did MD simulations at different ion concentrations (from 0.1 to 0.99 M) while keeping other parameters fixed. As an example, Figure 4.6 shows the change of the distances between DD and HH pairs (at $g_i-e'_{i+5}$ positions) with salt concentration. According to this figure, we can confirm that as we increase the ion concentration of the solvent, the distances between charged pairs do not change significantly. This fact also proves that the distances between charged residues in AGADIR database are valid, since they are not salt concentration dependent. However, the electrostatic energy decreases while we increase salt concentration, due to shielding effect of the charges.

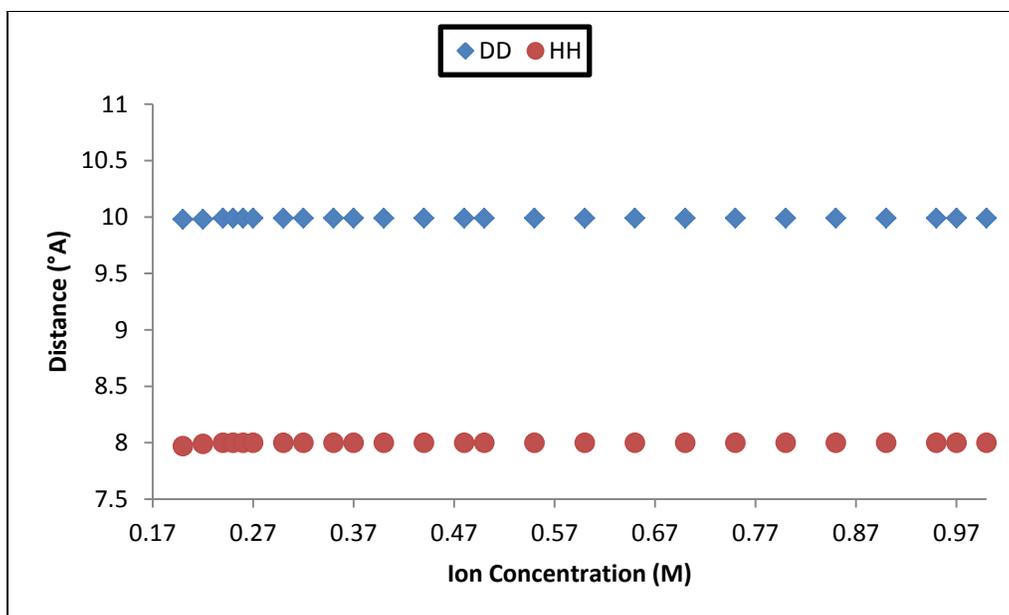


Figure 4.6 Change of the distances with salt concentration

4.4- Pairwise electrostatic interactions

In Table 4.2, we report the model predictions for the electrostatic coupling energies of experimentally available pairs of charged residues at $g_i-e'_{i+5}$ heptad positions. All interactions are calculated at $T=310$ K, $\text{pH}=7.4$ and salt concentration of 0.15 M. The reported distances are average values predicted by the molecular dynamics simulation over all possible side-chain rotamer states. The asymmetry of the distances at $g_i-e'_{i+5}$ positions (e.g. Glu-Arg vs. Arg-Glu) is due to the directional asymmetry of the right-handed α -helical structure (Burkhard et al. 2002). We validated that estimated distances are not sensitive to temperature, salt concentration and pH of the solution. We also confirm that the distances between charged residues at heptad positions b , c and f are too large to result in any significant inter-helical interactions. The average distances for all possible combinations of charged amino acids at heptad positions a , d , e and g are reported in Table 4.1.

As mentioned before, the dielectric constant used in our model is based on the salt concentration in the bulk solution. Therefore, strictly speaking, the model is applicable to solvent-exposed charges and applying the model to buried or semi-buried charges is an approximation. To gain a better confidence on our results, we compared our model predictions to experimentally estimated electrostatic interactions for charged residues at semi-exposed positions g and e . Krylov et al. (1994) have studied the effect of charged and polar amino acids at semi-exposed $g_i-e'_{i+5}$ pairs and estimated the pairwise electrostatic interactions via an alanine double-mutant thermodynamic cycle. Our model predictions for the same interactions and at the same solvent conditions (T=310 K, pH=7.4 and salt concentration=0.15 M) are presented in Table 4.2. As shown in Figure 4.7 our model predictions for the semi-exposed $g_i-e'_{i+5}$ electrostatic interactions are in good agreement with the experimentally estimated data. Nonetheless, the accuracy of our model applied to fully buried side chains at a and d positions remains questionable as using the bulk solution salt concentration for the buried charges may result in underestimation of their contribution to the total electrostatic interactions.

Table 4.2 Coupling energies between charged residues in kcal/mole

Pairs (g,e')	distance (Å)	Electrostatic Interaction Energy (kcal/mole)
EE	9.1	0.3279
ER	8.08	-0.5246
EK	7.71	-0.4572
RE	4.7	-1.0748
RR	10.63	0.2343
RK	10.14	0.2602
KE	8.2	-0.3968
KR	9.92	0.2731
KK	8.35	0.3909

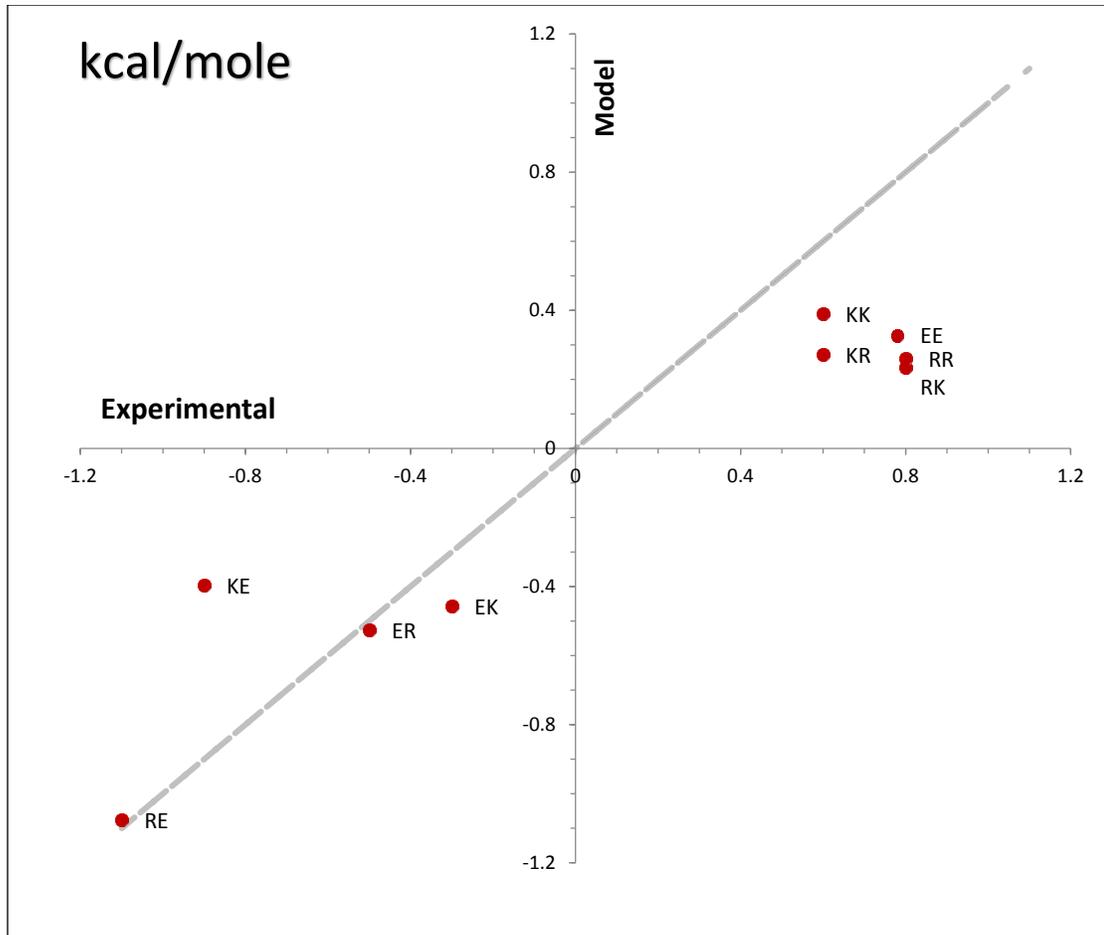


Figure 4.7 Coupling energies between charged residues at g_i - e'_{i+5} positions (experimental data vs. data from our model). Data are presented in kcal/mole at T=310 K, pH=7.4 and salt concentration=0.15 M.

Krylov et al. (1994) investigated the effect of the above-mentioned electrostatic interactions by studying the following sequence in the case of forming a coiled-coil homodimer:

AAFLX'KXNTALX'TXVAELX'KXVGRCX'NI

where X and X' are the mutated residues at g_i and e'_{i+5} positions, respectively, which are located on opposite strands of the coiled-coil structure. We also used our model to calculate the total electrostatic interaction between the two chains at the same solvent conditions as in the experiment, and we obtained results which can be verified by Krylove et al. (1994): among the

pairs of residues that we tested, E-R (at $g_i-e'_{i+5}$) is the most stable pair contributing to electrostatic energy of the coiled-coil. The second most stable pair is E-K. Also, R-R pair is less stable than K-K, meaning that the K-K pair favors coiled-coil dimerization more than R-R at above-mentioned positions.

4.5- Contribution of non- $g_i-e'_{i+5}$ positions to the total electrostatic interaction

As mentioned in the introduction section, in de novo designed coiled-coil dimers with a regulated interface and charged residues only at positions e and g , the net electrostatic interaction of all the $g_i-e'_{i+5}$ pairs is a reasonable predictor of the pairing specificity (Chao et al. 1998). On the other hand, the scoring rules (Vinson et al. 1993; Parry et al. 1977; McLachlan and Stewart 1975) based entirely on the $g_i-e'_{i+5}$ electrostatic interactions fail to predict the pairing specificity among naturally occurring coiled-coils such as human b-ZIP transcription factors (Newman and Keating 2003). One of the main questions we set out to address in this work is whether a more complete estimate of the electrostatic coupling energies does a better job in predicting the pairing specificity. In this section we investigate what fraction of the total electrostatic coupling energies in b-ZIP transcription factors comes from pairs other than $g_i-e'_{i+5}$ and whether they are negligible in comparison to the net electrostatic interaction of the $g_i-e'_{i+5}$ pairs.

The total inter-helical electrostatic interaction of a coiled-coil is calculated as the sum of the pair-wise electrostatic interactions across the coiled-coil interface. We have studied 484 interactions among 22 b-ZIP transcription factor sequences at 21°C, pH=7.4, and ion concentration=0.1 M (Reinke et al. 2013). The details of protein sequences can be found in Table 4.3. These are the same sequences we used in order to investigate the specificity of coiled-coil dimers in the next section of this report.

Table 4.3 Coiled-coil region amino acid sequences used in this work. Sequences are the same as in Newman and Keating (2003), starting from *g* position from left to right.

Family	Name	Coiled-coil Domain Sequence
C/EBP	C/EBP α	RNVETQQKVLELTSDNDRLRKRVEQLSRELDTLRGIFRQL
	C/EBP γ	KAQDTLQRVNQLKEENERLEAKIKLLTKELSVLKDLFLEHAHNLAD
	C/EBP ϵ	RILETQQKVLEMAENERLRSRVEQLTQELDTLRNLFHQI
OASIS	CREB3	YVGGLESRVLYTAQNMELQNKVQLLEEQLNSLLDQLRKLQAMVIEISNKTSS
ATF-6	ATF-6	YMLGLEARLKAALSENEQLKKENGLTKRQLDEVVSENQRLKV
	ZF	YVMGLESRVRGLAAENQELRAENRELGKRVQALQEESSRYLRAVLANETGL
XBP	XBP-1	RMSELEQQVVDLEEENQKLLLENQLLREKTHGLVVENQEL
ATF-2	ATF-2	WVQSLEKKAEDLSSLNGQLQSEVTLRNEVAQLKQLLAHKDC
JUN	JunB	RIARLEDKVKTLKAENAGLSSTAGLLREQVAQLKQKVMTH
FOS	Fos	LTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAHR
ATF-3	ATF-3	KTECLQKESKLESVNAELKAQIEELKNEQHLYMLNLHR
ATF-4	ATF-4	EQEALTGECKELEKKNEALKERADSLAKEIQYLKDLIEEVRKARGKKRV
	ATF-5	EGEALLEGECQGLEARNRELKERAESVEREIQYVKDLLIEVYKARSQ
B-ATF	B-ATF	KADTLHLESEDLEKQNAALRKEIKQLTEELKYFTSVLNSHE
PAR	HLF	KENQJAIASFLEKENSALRQEVADLRKELGKCKNILAKYEARH
smMAF	MafG	QKEELEKQKAELQQEVEKLAENASMKLELDALRSKYEALQTFARTVARS
IgMAF	MafB	QKHHLNEKTQLIQQVEQLKQEVSRRLARERDAYKVKCEKLANSG
CNC	NFE2L1	TILNLERDVEDLQRDKARLLREKVEFLRSRQMKQKVQSLYQEV
	NFE2L2	NIVELEQDLHLKDEKEKLLKEKGENDKSLHLLKKQLSTLYLEV
	NFE2L3	IILNLEDDVCNLQAKKETLKRQAQCENKAINIMKQKLHDL
	BACH1	CIQNLESEIEKLQSEKESLLKERDHILSTLGETKQNLTLGL
	BACH2	CIQNLECEIRKLVCEKEKLLSERNLKACMGEL

As an example, Figure 4.8 illustrates the electrostatic coupling energy of CREB3 (a transcription factor that is a member of the leucine-zipper family) with all the 22 sequences considered in our work, including itself, at the above-mentioned solvent conditions. The sequence that is experimentally (Reinke et al. 2013) shown to bind to CREB3, besides itself, is ATF-4, according to experimental data. The yellow bars show the total electrostatic coupling energies. The red bars show the electrostatic coupling energies calculated based on $g_i-e'_{i+5}$ pairs and the blue bars show the electrostatic coupling energies among charged residues at other positions (non- $g_i-e'_{i+5}$ interactions). An interesting observation is that the sign of the total

electrostatic interaction is predominantly determined by the $g_i-e'_{i+5}$ interactions, which means the $g_i-e'_{i+5}$ interactions are the dominant electrostatic interactions when compared to the non- $g_i-e'_{i+5}$ interactions. However, as seen in Figure 4.8, the contribution of non- $g_i-e'_{i+5}$ interactions to the total electrostatic coupling energy is by no means negligible. We observe a similar trait for several other b-ZIP sequences examined such as C/EBP α , ATF-6 and ZF, as shown in Figures 4.9, 4.10 and 4.11.

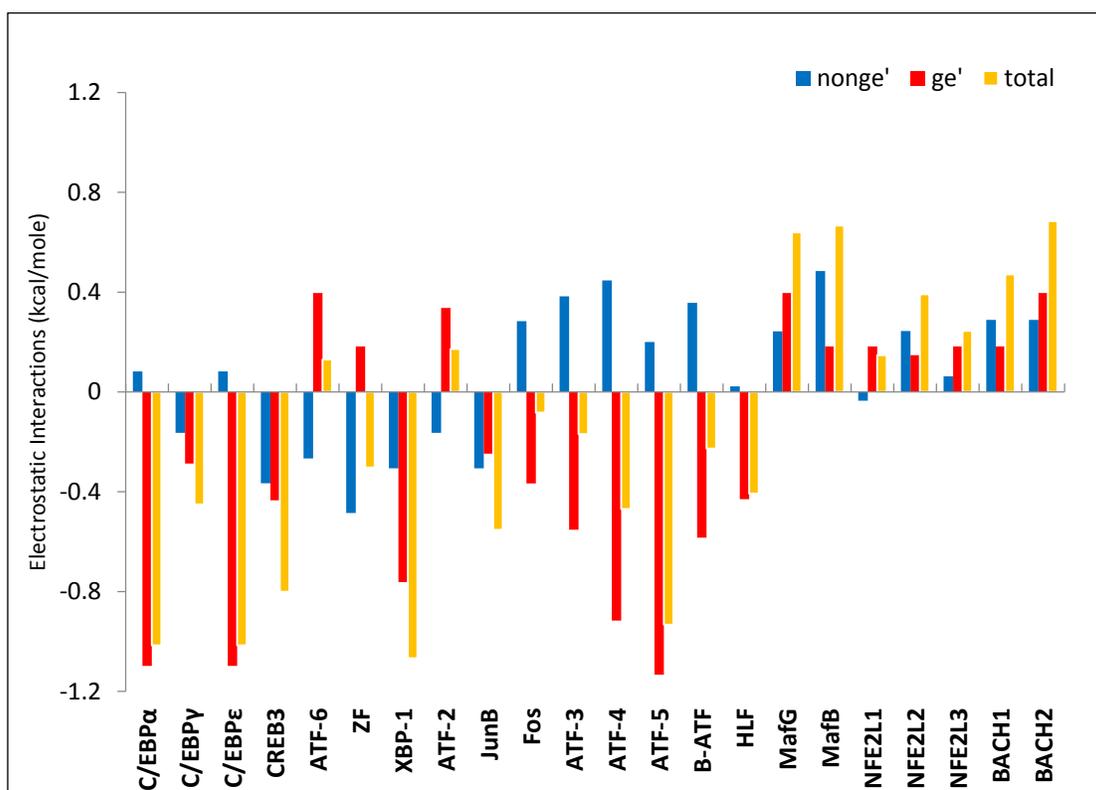


Figure 4.8 Electrostatic coupling energy of CREB3 with 22 other b-ZIP sequences. The yellow, red and blue bars show the electrostatic energies coming from all the positions, from only g_i and e'_{i+5} positions, and from the rest of the positions ($non-g_i-e'_{i+5}$) of the heptad repeat to interact electrostatically, respectively.

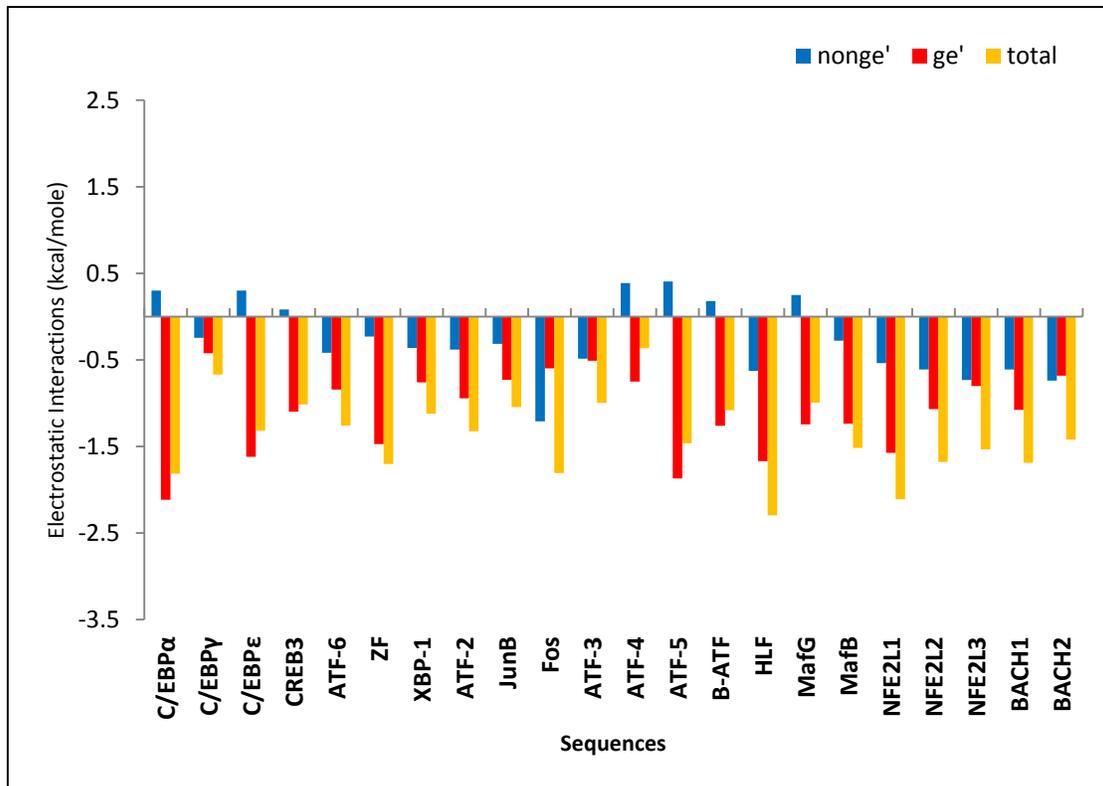


Figure 4.9 Electrostatic coupling energy of C/EBP α with 22 other b-ZIP sequences. The yellow, red and blue bars show the electrostatic energies coming from all interactions, from only $g_i-e'_{i+5}$ interactions, and from the rest of the positions ($non-g_i-e'_{i+5}$) of the heptad repeat to interact electrostatically, respectively.

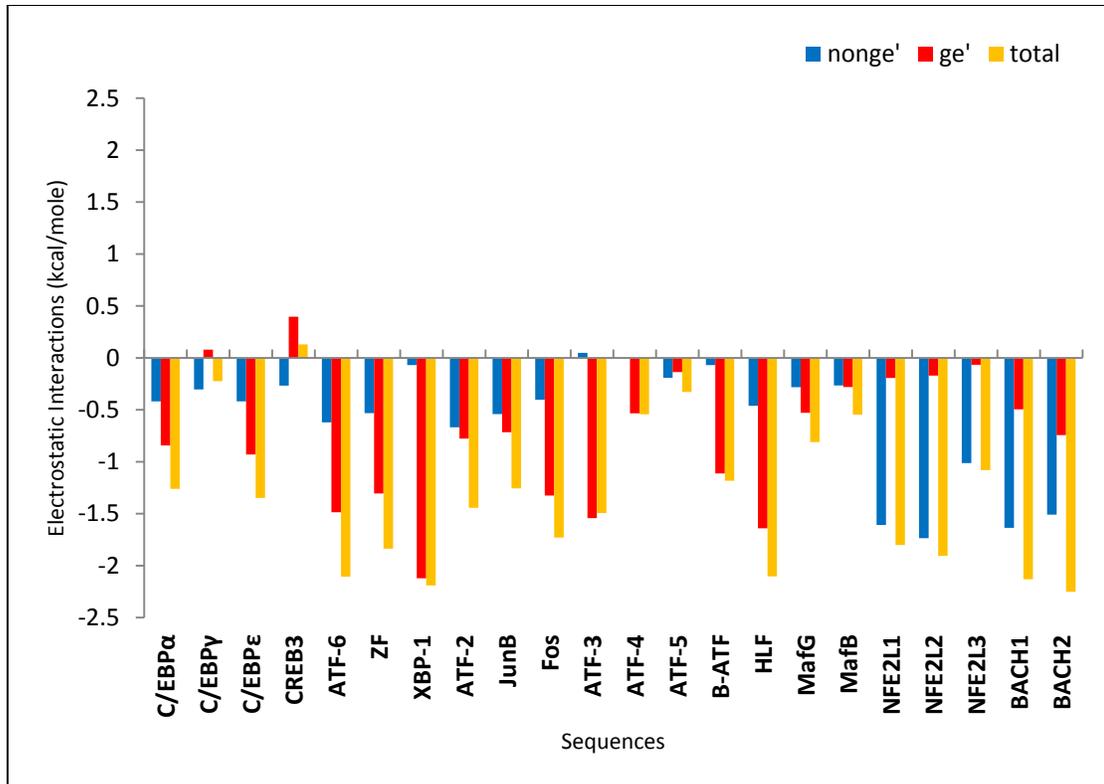


Figure 4.10 Electrostatic coupling energy of ATF-6 with 22 other b-ZIP sequences. The yellow, red and blue bars show the electrostatic energies coming from all interactions, from only $g_i-e'_{i+5}$ interactions, and from the rest of the positions ($non-g_i-e'_{i+5}$) of the heptad repeat to interact electrostatically, respectively.

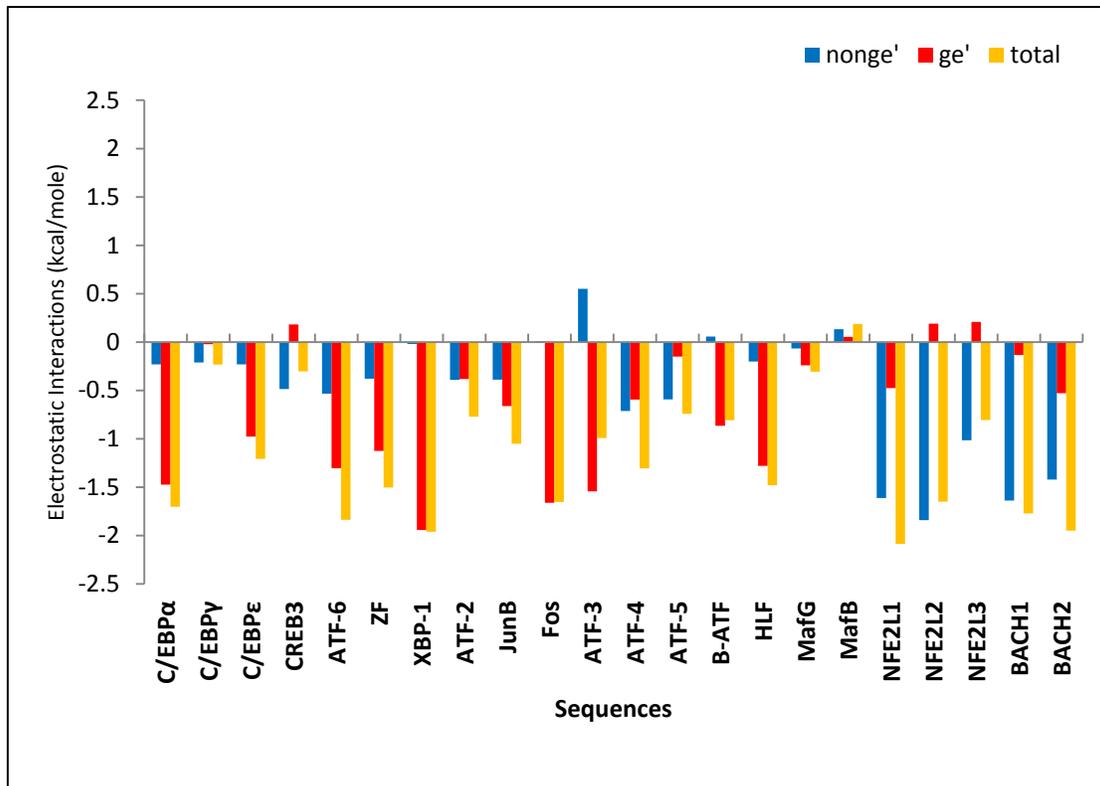


Figure 4.11 Electrostatic coupling energy of ZF with 22 other b-ZIP sequences. The yellow, red and blue bars show the electrostatic energies coming from all interactions, from only $g_i-e'_{i+5}$ interactions, and from the rest of the positions ($non-g_i-e'_{i+5}$) of the heptad repeat to interact electrostatically, respectively.

To further investigate the significance of the non- $g_i-e'_{i+5}$ electrostatic interactions, we calculate the electrostatic coupling energies for all possible pairings of all the 22 b-ZIP sequences. Figure 4.12 illustrates the ratio of the absolute value of the electrostatic coupling energy among the charged residues from positions other than g_i and e'_{i+5} over that of the g_i and e'_{i+5} pairs. As evident in Figure 4, for more than 90% of the pairs (443 out of 484) the ratio is larger than 0.5 and therefore, the non- $g_i-e'_{i+5}$ interactions are not negligible in comparison to the total electrostatic coupling energy. In the next section, we investigate whether we can predict the coiled-coil pairing specificity by taking only $g_i-e'_{i+5}$ electrostatic interactions into account.

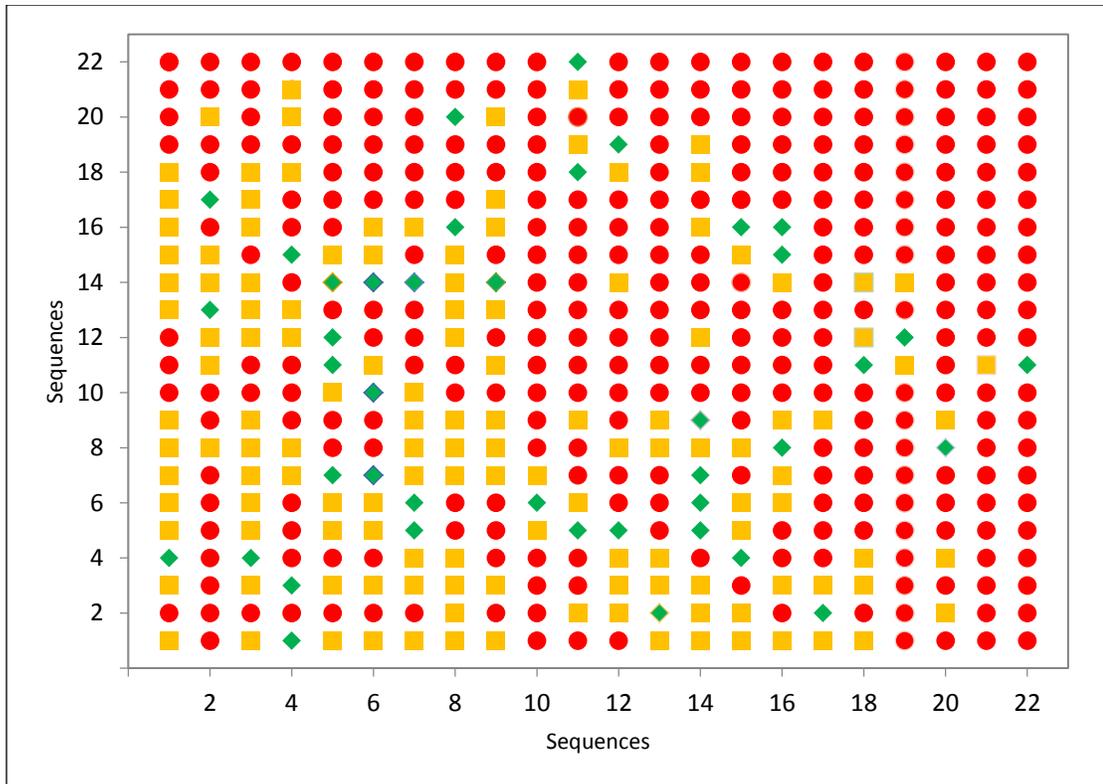


Figure 4.12 Absolute value of electrostatic energies ($E_{non-ge'}/E_{ge'}$). Red: ($E_{non-ge'}/E_{ge'}$) > 0.5, orange: ($0.1 < E_{non-ge'}/E_{ge'}$) < 0.5, green: ($E_{non-ge'}/E_{ge'}$) < 0.1

4.6- Electrostatic interactions and coiled-coil paring specificity

Newman and Keating (2003) used a simple sequence scoring, termed rule 1, to determine the strength of b-ZIP interactions. They simply assigned a score to each b-ZIP pair which is the summation of all the coupling interaction scores taking the weight of each interaction into account. They considered only $\mathbf{g}_i\text{-}\mathbf{e}'_{i+5}$ pairs occupied by charged and polar amino acids. Strong and weak interactions were assigned positive and negative scores, respectively. We used this rule to reproduce the interaction scores for the same 22 sequences in two cases: only for charged residues at \mathbf{g}_i and \mathbf{e}'_{i+5} positions (Figure 4.13) and for charged and two polar residues (glutamine and asparagine) at \mathbf{g}_i and \mathbf{e}'_{i+5} positions (Figure 4.14). The true positive rate for the first case, where we only consider charged residues and not polar ones, is calculated as 88% and for the

second case, where we consider polar residues as well, the true positive rate is 94%. In our electrostatic model, we only consider charged residues and we correct the charged based on the pH of the solution. Here, the reported rates are regarding scoring method, which is not the method we are using in which we calculate electrostatic interactions based solvent conditions.

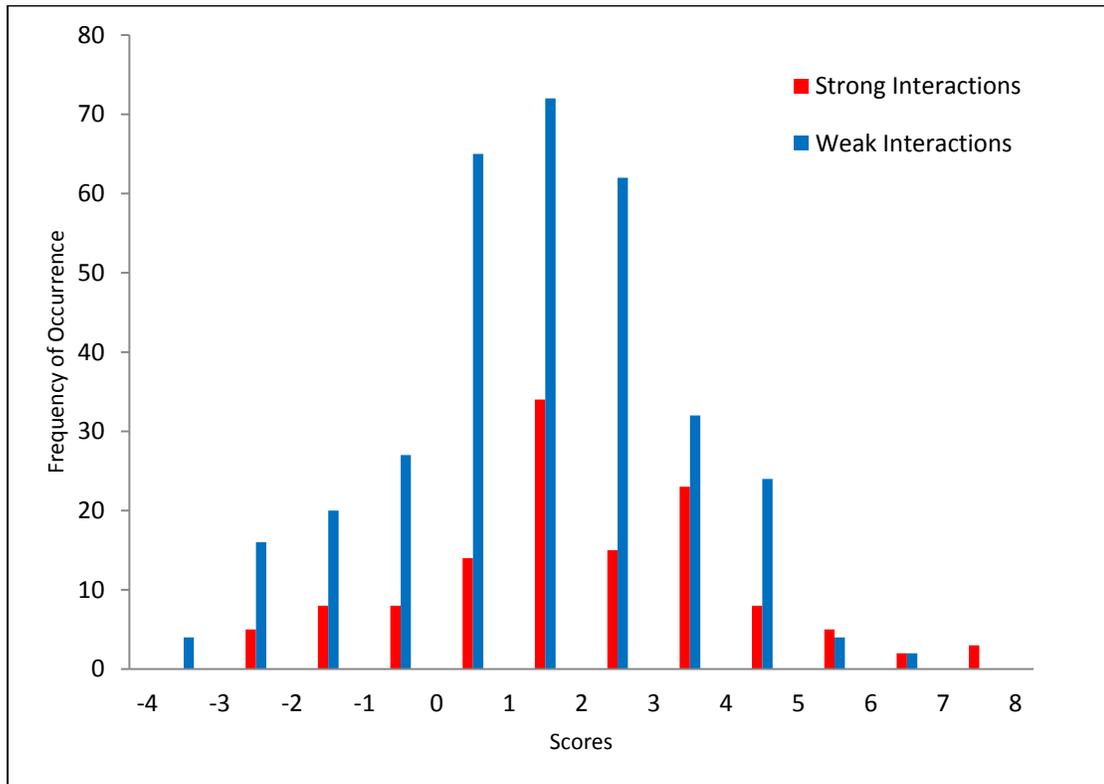


Figure 4.13 $g_i-e'_{i+5}$ interaction scores of 22 sequences for charged residues. Red and blue bars indicate strong and weak interactions, respectively.

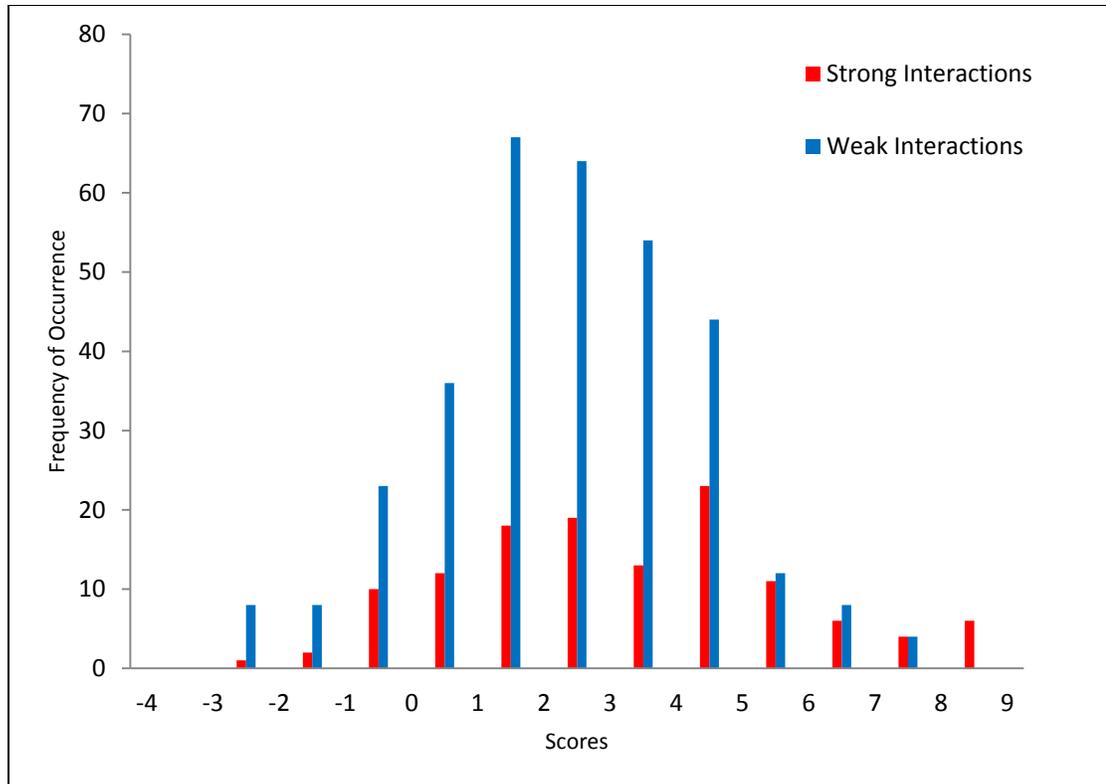


Figure 4.14 $g_i-e'_{i+5}$ interaction scores of 22 sequences for charged and polar residues (Gln and Asn). Red and blue bars indicate strong and weak interactions, respectively.

Figure 4.15 shows the electrostatic coupling energies calculated based on only $g_i-e'_{i+5}$ pair positions at 21°C, pH 7.4 and 0.1 M salt concentration. Following the experimental data reported by Potapov, et al. (2015), we assume pairs with dissociation constants smaller than 250 nM interact strongly and pairs with dissociation constants larger than 5000 nM are considered weak or non-interacting. Coupling state of pairs with intermediate dissociation constants are assumed undetermined and are excluded from our study. Based on their dissociation constants, data points shown in red are strongly interacting pairs and those shown in blue are the non-interacting ones. The y-axis in Figure 4.15 represents the model predictions of only $g_i-e'_{i+5}$ electrostatic coupling energies. Same data are plotted in Figure 4.16 that shows the histogram

distribution of $\mathbf{g}_i\mathbf{e}'_{i+5}$ electrostatic coupling energies. As seen in both Figures 4.15 and 4.16 (399 out of 484) of all the pairs examined in our work have favorable (negative) $\mathbf{g}_i\mathbf{e}'_{i+5}$ electrostatic coupling energies, regardless of being strongly interacting or non-interacting. Using $\mathbf{g}_i\mathbf{e}'_{i+5}$ electrostatic coupling energies to predict coupling specificity results in 83% true-positives predictions (83% of binding couples have negative $\mathbf{g}_i\mathbf{e}'_{i+5}$ electrostatic coupling energies) and 84% false-positives predictions (84% of non-binding couples have negative $\mathbf{g}_i\mathbf{e}'_{i+5}$ electrostatic coupling energies). $\mathbf{g}_i\mathbf{e}'_{i+5}$ electrostatic coupling energies results in 16% true-negative predictions (only 16% of non-binding couples have positive $\mathbf{g}_i\mathbf{e}'_{i+5}$ electrostatic coupling energies). The above results indicate that $\mathbf{g}_i\mathbf{e}'_{i+5}$ electrostatic coupling energies are of little value for specificity predictions.

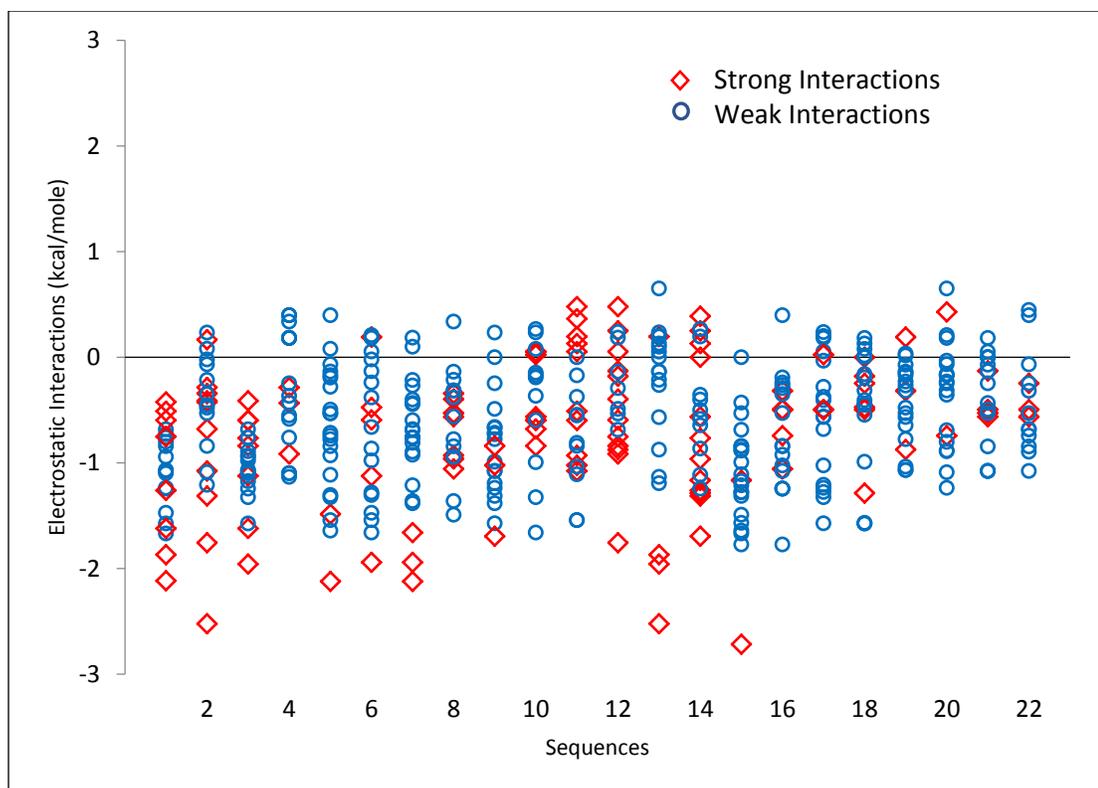


Figure 4.15 $g_i-e'_{i+5}$ electrostatic interactions 22 sequences. Red and blue marks indicate strong and weak interactions, respectively.

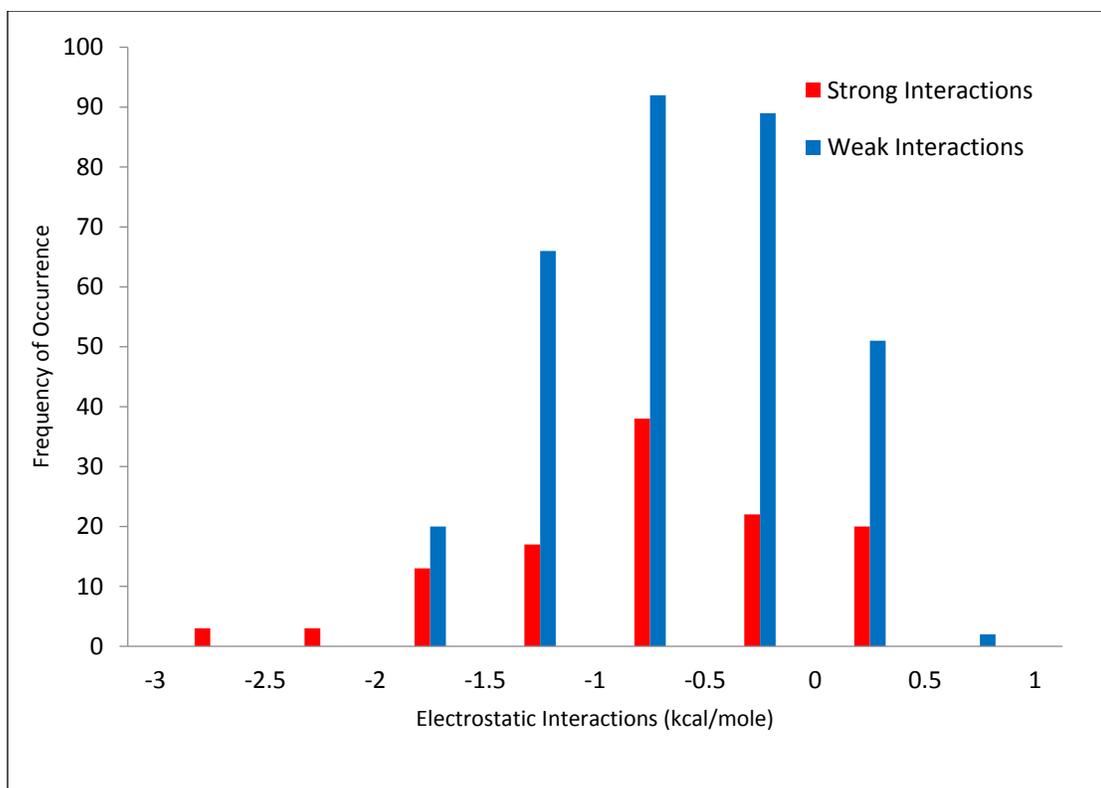


Figure 4.16 Frequency of $g_i-e'_{i+5}$ electrostatic interactions. Red and blue bars indicate strong and weak interactions, respectively.

Figures 4.17 and 4.18 are equivalents of figures 4.15 and 4.16 but based on the entire electrostatic coupling energies. When we consider all the positions, the true positive rate improves to almost 96% (Figure 4.17) while the true negative and false positive rates do not vary significantly.

Figure 4.18 also confirms the fact that when we consider all the electrostatic interactions in addition to $g_i-e'_{i+5}$, we notice more red bars on the negative energies comparing to considering only $g_i-e'_{i+5}$ interactions, meaning the true positive rate in the first case improves. This confirms that the electrostatic interactions coming from positions other than $g_i-e'_{i+5}$ are also important and we can increase the precision in our predictions by taking them into account in calculation of electrostatic interactions (Jokar and Torabi 2019). Also, we see a lot of blue bars, which show weak interactions on negative energies. This means the true negative rate is low and as

mentioned before, although these sequences have favorable electrostatic coupling interactions, some other unfavorable interactions make them not bind to each other. We will discuss this issue in the next section of this report.

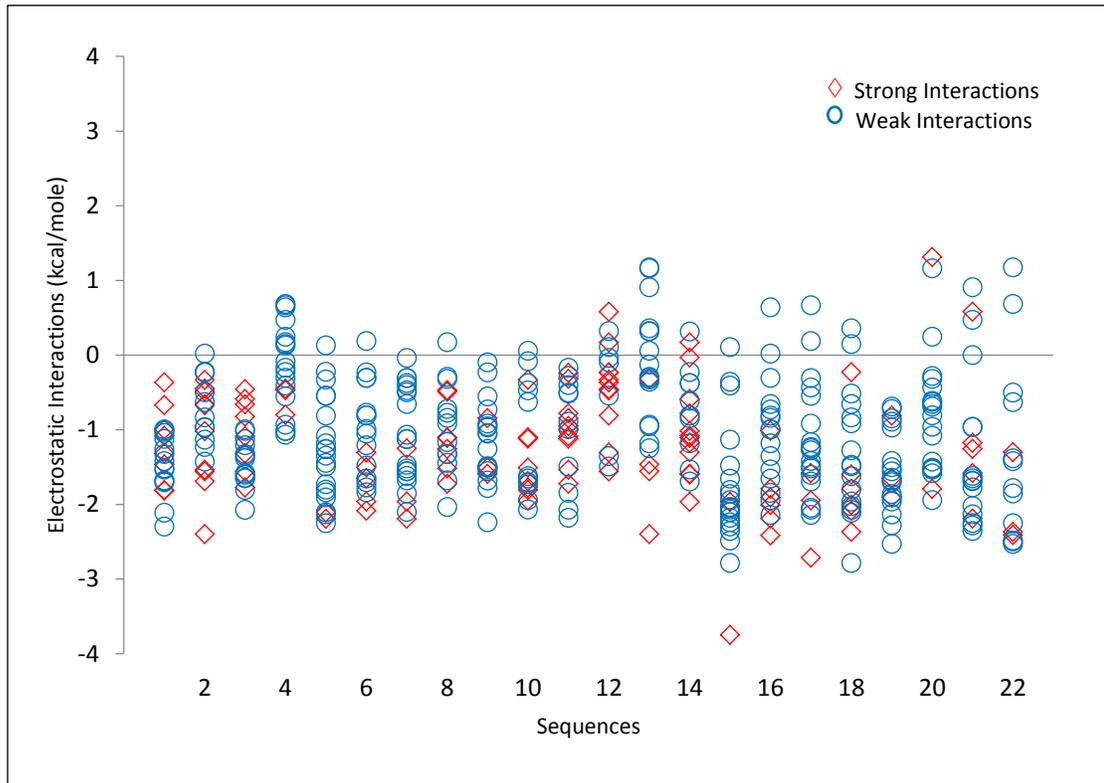


Figure 4.17 Total electrostatic interactions of 22 sequences. Red and blue marks show strong and weak interactions, respectively.

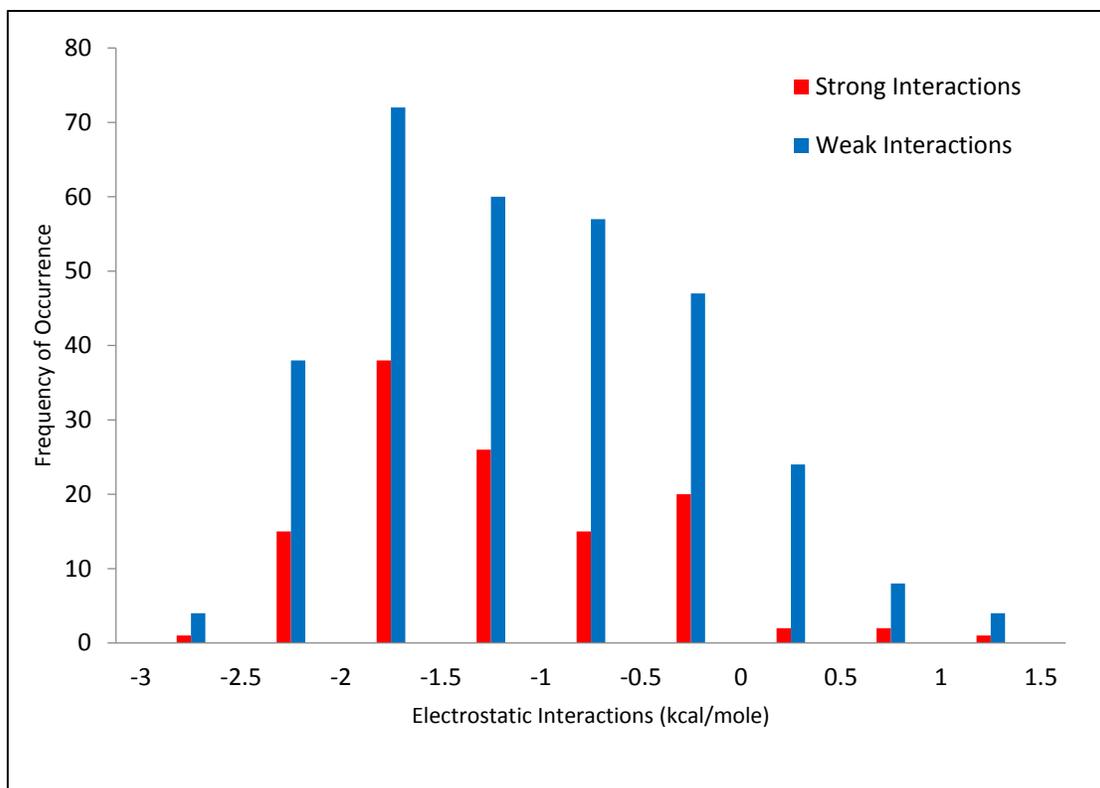


Figure 4.18 Frequency of total electrostatic interactions. Red and blue bars indicate strong and weak interactions, respectively.

5- CHAPTER 5: Hydrophobic core interactions in coiled-coil protein structures

5.1- Introduction

In naturally occurring coiled-coil dimers, hydrophobic residues occupy about 80% (Yu 2002) of *a* and *d* positions creating a hydrophobic core along the coiled-coil axis (O'Shea et al. 1991; Moitra et al. 1997). The hydrophobic core packing in coiled-coil dimers is explained in terms of the knob-and-pocket model (Burkhard et al. 2002). In a parallel dimer (two α -helices running in the same direction), a side-chain at positions *d* packs within the pocket formed by side-chains at positions *a' d' e' a'* (see Figure 5.1) and a side-chain at position *a* packs within the pocket formed by side-chains at positions *d' g' a' d'* (prime indicates the residues on the opposite helical strand). Within a closed-packed interface, core residues at positions *a* and *d* play a key role in coiled-coil stability. Small residues e.g. alanine do not provide enough hydrophobic driving force for dimerization and large (relative to leucine, isoleucine and valine) residues cause steric clashes and packing defects at the interface (Chao et al. 1998). In coiled-coil dimers, leucine is the most commonly observed residue at position *d*, while position *a* favors β -branched amino acids valine and isoleucine (Moitra et al. 1997). Such tendency has also been verified by thermal stability measurements of leucine-zippers point-mutated at positions *a* and *d* (Krylov et al. 1994). The hydrophobic core of a coiled-coil dimer is typically flanked by charged or polar side-chains at positions *g* and *e*. In parallel dimers, side-chains at positions *g_i* of one strand are adjacent to the side-chains at positions *e'_{i+5}* of the opposite strand.

As mentioned before, hydrophobic interactions are one of the main effects in stability of coiled-coil structures. In this part of the report we are going to explain the model that is introduced for helix packing as well as an approach to measure the hydrophobic effect, and

finally the contribution of the interactions in the hydrophobic core to coiled-coil stability and specificity.

5.2- Model for calculating hydrophobic core interactions in a coiled-coil

In order to calculate the solvation energy along a dimeric coiled-coil protein structure, we used the previously mentioned knob and pocket model, with some modifications. We used another pattern for packing of α -helices that has been studied by Joo et al. (2012), which is named knob-socket model. In this model, the socket includes three residues which cover the knob residue and pack it within the socket. In this work, we use this model and the available data, which are basically propensities of different knobs to pack into different sockets, in order to calculate the hydrophobic core interactions along a coiled-coil structure. Figure 5.1 shows an example of a knob and pocket packing in the hydrophobic core, where knob is at position d of the heptad repeat.

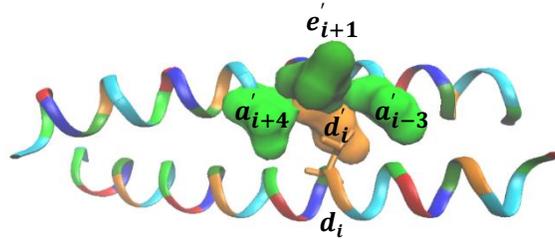


Figure 5.1 Side view of a parallel coiled-coil dimer showing a side-chain and a pocket.

Side chain at positions d_i is packed within the pocket formed by side-chains at positions

$$a'_{i-3} \quad d_i \quad e'_{i+1} \quad a'_{i+4}$$

The following equation shows the change in energy associated with one knob and socket along the coiled-coil dimer:

$$\Delta G_{k\&s} = -k_B T \log \left(\frac{P_{k\&s}}{P_{ref}} \right) \quad (5.1)$$

where $\Delta G_{k\&s}$ is the free energy associated with the knob and socket of interest with respect to a reference, k_B is the Boltzmann's constant, T is the temperature, $P_{k\&s}$ is the probability associated with the knob and socket of interest, and P_{ref} is the probability associated with the reference knob and socket:

$$P_{k\&s} = \frac{F_{k\&s}}{F_{total}} \quad (5.2)$$

$$P_{ref} = \frac{F_{L/LLV}}{F_{total}} \quad (5.3)$$

where $F_{k\&s}$ and F_{total} represent the frequencies of occurrence of the knob and socket of interest and the total frequencies of occurrence of the corresponding socket, respectively. These frequencies are collected by Joo et al. (2012), who studied the propensities of all the twenty naturally occurring amino acids with respect to the sockets in α -helical coiled-coils. $F_{L/LLV}$ is the reference knob-socket, which includes a leucine residue packed into a socket including two leucine residues and a valine, which is one of the most abundant knob-socket patterns in leucine-zippers.

In order to calculate the solvation energy along a coiled-coil structure, we need to do the above calculations for all the possible knob-socket patterns found along the coiled-coil. To do so, we simply sum over all the possible knobs and sockets along the structure (or sequences) of interest. Notice that a lot of the residues overlap with each other and should not be counted more than once. The second term in Equation 5.4 shows the subtraction of these redundant residues from the summation over the knob-socket patterns.

$$\Delta G_{total\ hydrophobic\ core} = \sum_{all\ knobs\ and\ sockets} \Delta G_{k\&s} - \sum_{overlapped\ residues} \Delta G_i \quad (5.4)$$

We should note that the first term is the summation over the knob-socket interactions, which consists of four residues each. The second term, however, is the summation over the single residues that were overlapped and counted more than once. As mentioned above, we can calculate the first term using equations 5.1, 5.2 and 5.3.

In order to calculate the second term, we started with measuring the Solvent Accessible Surface Area (SASA) of the available amino acids at different positions of the heptad repeat from available coiled-coil dimer crystal structures (O'Shea et al. 1991; Lavigne et al. 1998; Gonzalez et al. 1996; Marti et al. 2000; Oshaben et al. 2012). We calculated SASA of different amino acids by using VMD (Humphrey et al. 1996). Obviously not all the combinations of 20 amino acids and 7 positions of the heptad repeat were available in the crystal structures of coiled-coil dimers. In such cases, we had to mutate the crystal structure, which was more challenging comparing to our mutations for calculation of the electrostatic interactions. The reason for such complexity is that hydrophobic core interactions, unlike the electrostatic interactions, are not pair-wise and we should mutate 4 residues instead of two for measuring each knob-socket interaction. Furthermore, unlike electrostatic interactions, we deal with all the 20 naturally occurring amino acids and not only the charged ones in calculation of the hydrophobic core interactions. For these reasons, mutating the crystal structure with all the combinations of amino acids and heptad repeat positions seems like a time-consuming and unreasonable task. Therefore; we used an approximation and classified different positions of the heptad repeat into three groups: exposed (*b*, *c* and *f*), buried in the hydrophobic core (*a* and *d*) and partially exposed (*e* and *g*) positions. Then we mutated 2ZTA crystal structure (O'Shea et al. 1991) using VMD Mutator Plugin (Humphrey et al. 1996), with all the 20 amino acids at the above-mentioned positions along the heptad repeat. After mutating the structure with residues of interest, we ran

molecular dynamics (MD) simulations in order to track the physical movement of the atoms within each structure. All the simulations were performed with NAMD's built in Generalized Born Implicit Solvent model (Phillips et al. 2005) using a surface tension value of $0.0072 \text{ kcal.mol}^{-1}.\text{\AA}^{-2}$. After running MD, we measured SASA using VMD for each amino acid. Here we should note that the residues adjacent to the amino acid of interest affect the value of accessible surface area. Therefore; in our mutations, we also included different neighboring residues with different side-chain sizes which result into different amounts of coverage. However, since we are including single residues here and not knob-socket patterns, classifying the neighboring residue side-chains into categories based on their side-chain size is a reasonable approximation. Another fact is that SASA values do not fluctuate as much as pair-wise distances (which we used in calculation of the electrostatic interactions) and after a certain amount of the simulation time, the average SASA that we obtain seems constant for each structure. Figure 5.2 shows an example of the change in SASA during a simulation which we ran for 40 ns, for Phenylalanine at position **b** of a leucine-zipper coiled-coil.

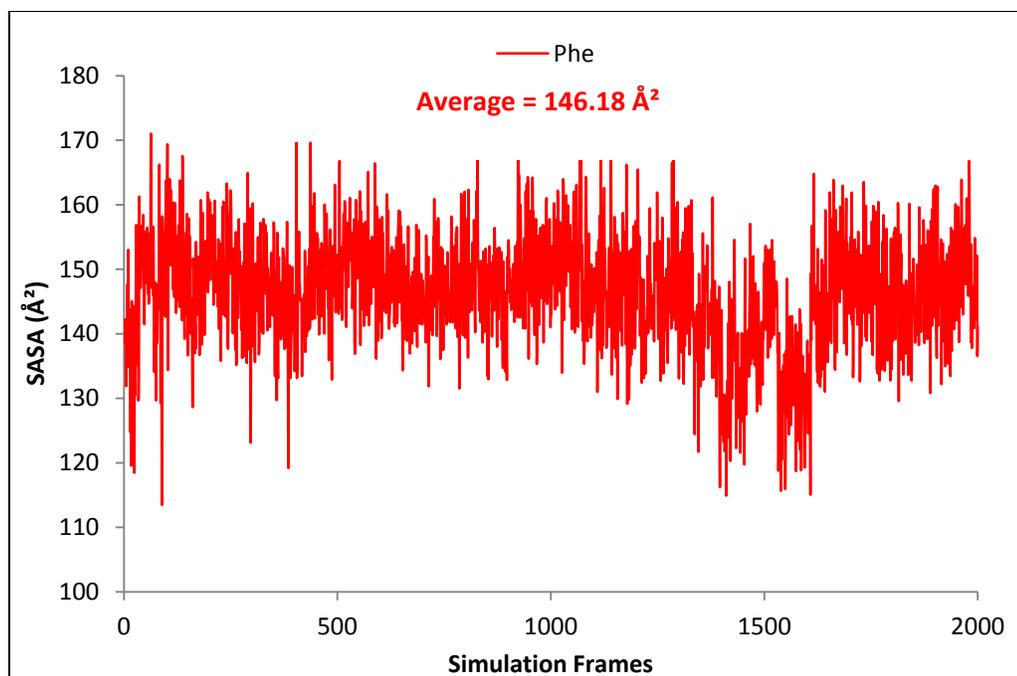


Figure 5.2 Change of SASA for Phe residue at position *b* (40ns MD simulation)

Table 5.1 shows the average SASA of all the 20 amino acids (in Å²) at different positions of a coiled-coil dimer. As mentioned before, some of the data are from the available coiled-coil crystal structures and the rest are averaged values obtained from MD simulations.

Table 5.1 SASA values for 20 amino acids at different positions of the heptad repeat. Data are collected from MD simulations and available crystal structures

amino acids↓/position→	SASA (Å ²) at different heptad repeat positions						
	g	a	b	c	d	e	f
A	31.237	34.46	62.33	58.18	34.46	31.237	54.03
C	62.296	20.08	76.48	76.48	20.08	62.296	76.48
D	60.07	59.2	89.237	89.237	59.2	60.07	89.237
E	49.93	65.78	107.92	82.62	65.78	49.93	95.27
F	123.56	67.019	146.18	146.18	67.019	123.56	146.18
G	19.299	10.762	34.44	34.44	10.762	19.299	34.44
H	110.811	65.51	127.35	109.15	65.51	110.811	145.55
I	84.17	51.897	77.478	77.478	51.897	84.17	77.478
K	117.83	143.6	151.31	151.31	143.6	117.83	151.31
L	62.69	2.16	28.85	28.85	2.16	73.52	28.85
M	101.63	50.895	125.65	125.65	50.895	101.63	125.65
N	65.59	0.68	96.16	96.16	0.68	65.59	96.16
P	57.419	29.093	70.12	70.12	29.093	57.419	70.12
Q	130.76	116.14	94.66	99.79	116.14	130.76	89.53
R	115.59	75.48	101.66	78.62	75.48	115.59	90.14
S	46.31	55.27	68.13	68.13	55.27	46.31	68.13
T	68.827	19.524	75.83	75.83	19.524	68.827	75.83
V	74.5113	21.7	96.138	96.138	21.7	74.5113	96.138
W	137.011	43.98	165.653	165.653	43.98	137.011	165.653
Y	141.218	63.411	119.77	119.77	63.411	141.218	119.77

Now that we have the SASA values for single residues along a coiled-coil dimer, we can calculate the second term in Equation (5.4) which represents the energy associated with the overlapped residues:

$$\Delta G_i = (SASA_{cc,i} - SASA_{exposed,i}) * Surface\ Tension \quad (5.5)$$

where ΔG_i is the free energy change associated with residue i when present in a coiled-coil structure with respect to the state where it is completely exposed to the solvent. $SASA_{cc,i}$ is the solvent accessible surface area of residue i in coiled-coil dimer (from Table 5.1), and $SASA_{exposed,i}$ is the solvent accessible surface area of residue i when it is completely exposed to

the solvent. In order to calculate $SASA_{exposed,i}$ values, we ran MD simulations under the same above-mentioned conditions for 20 amino acids. In this case, mutating the structure with different amino acids was much simpler than what we did to derive $SASA_{cc,i}$ values because the structure here is an α -helix where residues are exposed to the solvent and only their size makes the differences. Also since the structure we used was an α -helix and not a coiled-coil dimer, there is no heptad repeat pattern and the SASA values are not dependent on the positions. The data associated with SASA values for exposed residues are shown in Table 5.2. Notice that since the residues here are more exposed to the solvent than in a coiled-coil structure, the corresponding SASA values are larger.

Table 5.2 SASA values for 20 amino acids when exposed to the solvent. Data are collected from

MD simulations			
amino acids	SASA (\AA^2)	amino acids	SASA (\AA^2)
A	52.86	M	149.52
C	79.01	N	74.75
D	83.58	P	63
E	99.48	Q	111.11
F	138.12	R	153.06
G	30.16	S	63.26
H	120.87	T	72.55
I	97.71	V	75.43
K	152.4	W	173.48
L	106.96	Y	158.92

Once we monitor all the overlaps of residues in the knob-socket patterns along the structure, we should sum over them and subtract them from knob-socket interactions, according to Equation (5.4).

5.3- Contribution of hydrophobic core interactions to stability of the structure

In Figure 5.3, we report the model predictions for the energetic contribution of seven amino acids located at the *d* position of a leucine-zipper dimer and compare them to experimental data. Moitra et al. (1997) have obtained thermal stabilities of mutated leucine-zippers, monitored by CD spectroscopy, and compared the contribution of different amino acids to the stability of the coiled-coil dimer. The *d* position of the heptad repeat is mutated by leucine, methionine, isoleucine, alanine, valine, cysteine and serine, among which none is charged and therefore the contribution does not come from electrostatic interactions. The experimental conditions were 162.5 mM salt concentration, pH 7.4 before thermal melting.

Figure 5.3 shows the stability of the leucine-zipper structure relative to mutated alanine at position *d*, obtained from experiments and the model we described in the previous section of this report. It should be noted that the dimerization free energy has been extrapolated to 37 °C and we calculated all the interactions at the same temperature.

As shown in Figure 5.3 our model prediction for the changes in relative stability upon different mutations are in good agreement with the experimentally estimated data with a R^2 value of 0.83. Specifically, it has been confirmed that leucine is the most stabilizing amino acid at position *d* of a leucine-zipper dimeric coiled-coil and our data perfectly verifies that. Also, in experiments, the order of stability of the above-mentioned amino acids is L, M, I, V, C, A and S. Our model predictions are consistent with this order of stability. However, in experiments, methionine shows more contribution to stability than isoleucine at position *d*, which according to Figure 5.3 is not confirmed by our model. That being said, our model shows a difference of 0.38 kcal.mole⁻¹ between the contributions of these two amino acids to overall stability, which is relatively not a very large number.

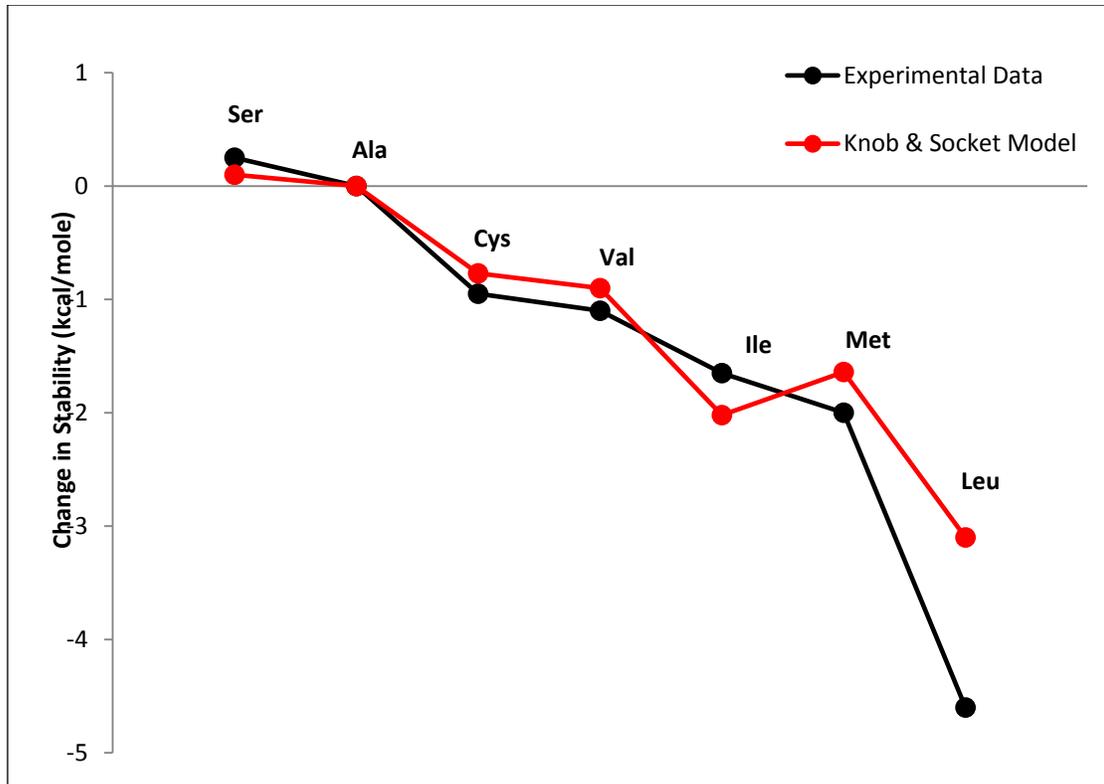


Figure 5.3 Change in stability of a leucine-zipper dimer upon mutations at position *d*

Figure 5.4 shows the model predictions for the energetic contribution of nine amino acids located at the *a* position of the α -helical coiled-coil dimers and the corresponding experimental data. Wagschal et al. (1999) have studied the role of position *a* in the hydrophobic core of a coiled-coil protein in determining coiled-coil stability and oligomerization state. In their structure of interest, position 19 which corresponds to position *a* of the central heptad repeat, is mutated with all the naturally occurring amino acids. However, some of these mutations favor coiled-coil trimers more than dimers and therefore are excluded from the data shown in Figure 5.4. The experimental data result from CD spectroscopy under benign conditions (~50 mM, phosphate, 100 mM KCl, pH 7.0).

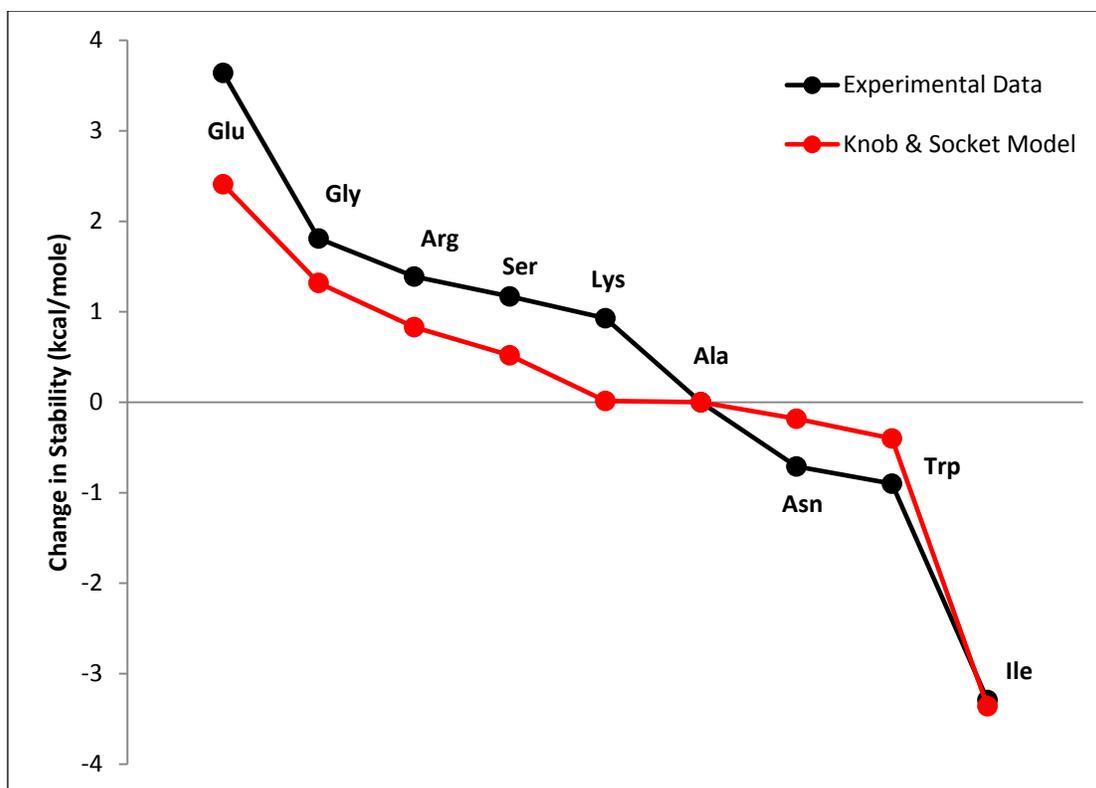


Figure 5.4 Change in stability of a leucine-zipper dimer upon mutations at position *a*

Figure 5.4 shows the change in stability of the coiled-coil dimers relative to mutated alanine at position *a*. We can see that also for this position our model predictions for the changes in relative stability upon different mutations are in good agreement with the experimentally estimated data with a R^2 value of 0.86, as well as the order of stabilities.

5.4- Hydrophobic core interactions and coiled-coil pairing specificity

In the previous chapter, we tried to check if we can predict the specificity of leucine-zipper coiled-coil dimers based on only electrostatic interactions. We calculated the total electrostatic coupling energy between 22 sequences of interest and obtained a true positive rate of 96%. However, the true negative rate was very low, confirming the fact that for the sequences that do not bind to each other, there must be a very unfavorable type of interaction which makes the total interaction between them unfavorable, despite the potential favorable electrostatic

interaction. In this section, we investigate the effect of hydrophobic core interactions in predicting the coiled-coil pairing specificity and compare it to the case where we studied pairing specificity using only electrostatic interactions. Our final goal here is to add up the main dimerization energies and see how much the true positive, true negative and accuracy rates improve.

Here we calculated the hydrophobic interactions along the same set of b-ZIP sequences (see Table 4.3) studied in the previous chapter (Newman and Keating 2003) using the model that was explained above. In order to gain an insight into the values of hydrophobic core interactions comparing to electrostatic energies, we reported the ratio of the absolute values of hydrophobic core over that of electrostatic interactions for all the possible sequence pairs, as shown in Figure 5.5. According to this figure, in almost 60% of the pairwise interactions, hydrophobic core energy is larger than electrostatic (data shown in red and orange) and the rest 40% are the cases where electrostatic dominate the hydrophobic core interactions.

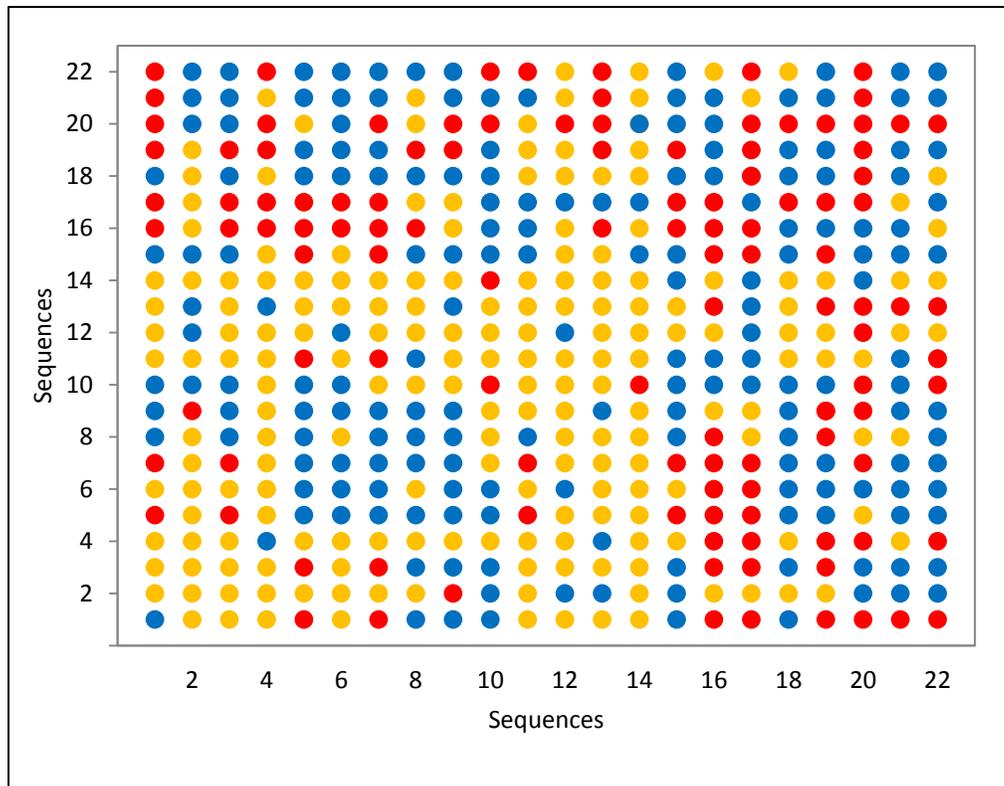


Figure 5.5 Absolute value of hydrophobic core over the electrostatic energies. Red: hydrophobic \gg electrostatic, orange: hydrophobic $>$ electrostatic, blue: hydrophobic $<$ electrostatic

Notice that the data presented in red show the interactions that are highly dominated by mostly unfavorable hydrophobic core interactions, due to very unfavorable and infrequent knob and socket combinations observed in the structure.

Figure 5.6 shows the summation of the hydrophobic and electrostatic interactions based on all the positions of the heptad repeat at 21°C, pH 7.4 and 0.1 M salt concentration. Here we make the same assumption as presented in the previous chapter: pairs with dissociation constants smaller than 250 nM interact strongly and pairs with dissociation constants larger than 5000 nM are considered weak or non-interacting (Potapov et al. 2015). Also pairs with intermediate dissociation constants are assumed undetermined and are excluded. Based on the mentioned

dissociation constant values (Newman and Keating 2003; Potapov et al. 2015) data points shown in red are strongly interacting pairs and those shown in blue are the non-interacting ones.

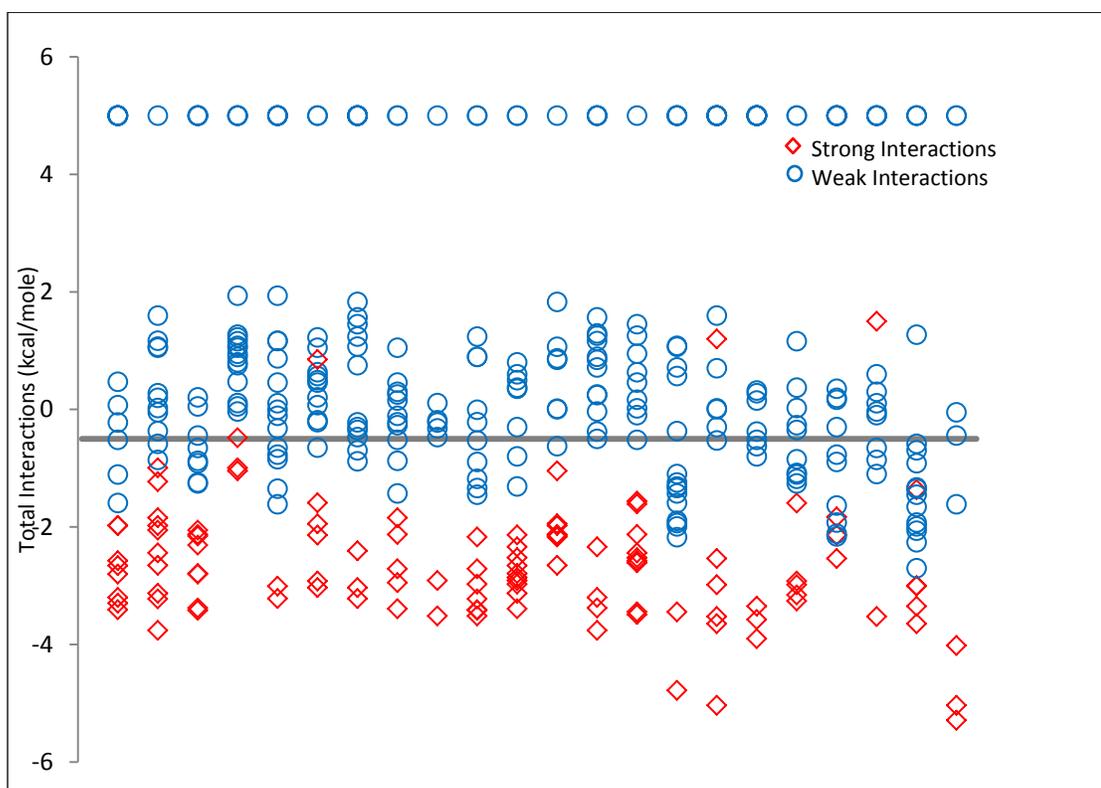


Figure 5.6 Summation of packing and electrostatic interactions for 22 sequences. Red and blue marks indicate strong and weak interactions, respectively.

The y axis in Figure 5.6 represents the model predictions of the summation of hydrophobic and electrostatic coupling energies. Using this set of data, we can predict the specificity of the b-ZIP sequences with a true positive rate of 96% and a true negative rate of 73%. This confirms the importance of the packing of the residues present in the hydrophobic core in predicting coiled-coil pairing specificity. Also almost 90% of all the pairs examined in our work have favorable coupling electrostatic energies which do not guarantee their dimerization. However, by adding the hydrophobic interactions on top of electrostatic coupling energies, we better understand the role of unfavorable knobs and sockets in discouraging the

dimerization of the sequences. This can be confirmed by the improvement in true negative rate, from 16% to 73%. Notice that the data points representing weak interactions that are all aligned on the same energy level in Figure 5.6, have multiple highly unfavorable knob-socket packing patterns in the structures associated with them. To elaborate more, in terms of frequency of occurrence, these packing patterns possess a frequency equal to zero, and we assigned a relatively high energy level to them (5 kcal.mole^{-1}) so that we can visualize them in Figure 5.6. Also the bold data points on any energy level in Figure 5.6 show overlapping data for the same sequences.

The accuracy rate, defined as the summation of the true positive and true negative rates, over the summation of the true positive, true negative, false positive and false negative rates, is 80% in this case. However, by using optimization (Ghouchanian et al. 2017) and regression methods, we can improve the accuracy. For example, if we move the cutoff line that separates the binding and unbinding states from each other to $-1.5 \text{ kcal.mole}^{-1}$ energy level, true negative rate improves to 91% and true positive rate would be 90%. The maximum accuracy can be achieved with a cutoff line at $-1.7 \text{ kcal.mole}^{-1}$ energy level, where we can obtain a true positive, true negative and accuracy of 84%, 94% and 91.1%, respectively.

6- CHAPTER 6: Summary, challenges and conclusions

In this work, we used different computational methods in order to study the structure and selectivity of coiled-coil dimers based on a primary amino acid sequence and solvent conditions. The main objectives of this work are: a) to develop a model to predict the propensity of a protein sequence to form an isolated coiled-coil dimer, and b) to investigate the selectivity of coiled-coils by studying protein-protein interactions. The motivation behind this work was from a simple statistical mechanical model which was used to predict the structure (Lacroix et al. 1998) and mechanical properties (Torabi and Schatz 2013) of an α -helix structure. This model showed promising results when was compared to experimental data. We decided to expand this model to coiled-coil dimers which are abundant structures that are involved in significant biological functions such as the regulation of gene expression, known as transcription factors. Also coiled-coil structures entail unique mechanical properties critical to the function and integrity of various motor proteins, cytoskeletal filaments and extra-cellular matrix proteins (Rose et al. 2004; Burkhard et al. 2001).

We started with studying the structure of coiled-coil dimers and came up with models for calculating the main dimerization energies. We used the same model as implemented in AGADIR (Lacroix et al. 1998) to calculate the electrostatic interactions. However, in the case of coiled-coil, comparing to an α -helix structure, calculating these interactions is more challenging. One of the complexities comes from the distances between the point charges located on opposite strands, which is dependent on the type of the amino acid, the positions of the heptad repeat, and the solvent conditions.

After deriving all the necessary parameters to calculate pairwise electrostatic coupling energies along a coiled-coil dimer, we compared our results to the available experimental data

and observed an acceptable agreement between them, comparing to other computational methods such as sequence scoring.

The next step was investigating the effect of inter-helical electrostatic interactions on pairing specificity of b-ZIP coiled-coils. One finding was that although in most sequences $g_i-e'_{i+5}$ electrostatic interactions are significant; we clearly showed that in calculation of the electrostatic coupling interactions for dimers, one cannot simply ignore the importance of other interactions along the heptad repeat. We proved this by improvement of true positive rate when we considered all other pairwise interactions in our calculations, and also by showing that the ratio of $g_i-e'_{i+5}$ over non- $g_i-e'_{i+5}$ electrostatic energies is not negligible. Besides, non- $g_i-e'_{i+5}$ distances are not significantly larger than those of other pairs of positions, confirming the fact that their electrostatic interactions are not negligible (see Table 4.1). Also, even if we use sequence scoring, in addition to $g_i-e'_{i+5}$ positions, we have to include all other possible positions' scores in order to increase the accuracy of our prediction. However, the scoring method does not consider solvent conditions and also is only based on $g_i-e'_{i+5}$ interactions while the method used in this work calculates the electrostatic interactions based on all the possible positions along the protein structure as well as temperature, pH and ion concentration of the solution.

Furthermore, by considering only electrostatic interactions, we can predict strong interactions between the chains with 96% true positive rate. On the other hand, we cannot predict weak interactions as accurate as strong interactions and our true negative rate is very low. This means that most of the sequences that we know form coiled-coil dimers, have favorable electrostatic interactions with the opposing strand. On the other hand, according to our results, the opposite of this statement is not true. To elaborate more, favorable electrostatic energy does

not always guarantee dimerization and clearly there is another factor that plays an important role in coiled-coil dimerization.

Before studying the specificity problem, we were aware of the fact that hydrophobic interactions are also significant in coiled-coil stability. However, insufficiency of the electrostatic interactions in predicting pairing specificity and the fact that we could not predict weak interactions using solely electrostatic energies, was another motivation for us to study the interactions in the hydrophobic core. The model we used in order to calculate hydrophobic core interactions is based on the patterns that are observed in packing of the α -helices, namely knobs and sockets (Joo et al. 2012). We converted the available frequencies of occurrence of these patterns to probabilities, and calculated free energies based on these values. The challenge here was the subtraction of the single residues that were counted multiple times from the free energy terms, which was explained in detail in the previous chapter. After completing the required datasets, we compared the model predictions to the available experimental data and obtained satisfactory results. Then, same as in the case of studying electrostatic interactions, we moved forward to study the specificity of coiled-coil dimers and the question was if we can predict specificity by summing over the hydrophobic core and electrostatic interactions along a coiled-coil. Comparing to the case of predicting specificity using only electrostatic energies, when we added the hydrophobic core interactions to the model, we observed a significant improvement in true negative rate. This confirms the fact that the model predicts unfavorable coupling energies between sequences very well. In terms of hydrophobic core interactions, this unfavorable effect mostly originates from knob and pocket/socket poor interactions, such as steric effects, clashes, etc. and that is the reason some chains do not dimerize in spite of favorable coupling electrostatic energy. It should be noted that the predicted hydrophobic core interactions from the model were

consistent with sequences that bind to each other as well. This can be confirmed by the fact that the high true positive rate did not decrease after taking hydrophobic interactions into account, and that is the reason we can obtain such high accuracy rate (80%).

7- CHAPTER 7: Future research directions

The models presented in this work are dependent on the solvent conditions and we can always alter them as inputs to the model. In this work, we studied coiled-coil pairing specificity at 21°C, pH 7.4 and 0.1 M salt concentration, which are the same conditions as the experiments. One immediate future research can be investigating the specificity of the same dataset of b-ZIP sequences at different solvent conditions. That being said, insufficiency of experimental data at different conditions may be a challenge.

Another future line of research is using other methods to calculate coiled-coil dimerization energies, especially hydrophobic interactions, and comparing them to the models presented in this work and check if we can obtain more successful results. For the case of hydrophobic interactions, there are already online applications that are able to predict solvent accessible surface area and solvation energies. However, one of the drawbacks of these applications is that they usually do not consider the effect of solvent conditions on the interactions and the results can only be obtained at fixed conditions. Another disadvantage of such models is that usually their input is a crystal structure, and it cannot be any sequence of interest. One of the strengths of our model is that the user can enter any random sequence and if the sequence does not exist naturally or is very infrequent, the results would show low probabilities of occurrence which corresponds to unfavorable energy terms.

Furthermore, investigating other types of interactions besides what we studied in this work can be another area for the future research. Here we only studied two main dimerization energies that are known to make the most contribution to coiled-coil stability. However, in order to make more accurate predictions, one can add salt bridges, entropy effects, etc. on top of

electrostatic and hydrophobic core interactions and modify the dimerization energy terms based on the updated data sets.

Recently, machine learning techniques have been developed in order to predict coiled-coil stability and calculation of different energy terms. Using such methods can be another future research direction for this work. For example, the knob-socket frequencies of occurrence used in calculation of the hydrophobic interactions can be implemented as training data, and the predictions can be made without the computer being explicitly programmed to calculate the interactions.

Finally, similar to the available model for an α -helix structure, the work described in this report can result in a statistical mechanical model that predicts the probability distribution of all possible structures of a coiled-coil dimer forming motif based on its primary sequence at given solvent conditions. Having the partition function, we can also determine the average helical content of different sequences. Also just the same as the model available for an α -helix structure (Torabi and Schatz 2013), another direction of the future research can be studying the tensile mechanics of a coiled-coil dimer by monitoring its force-extension behavior. This is possible by including a force-extension energy term to the partition function of coiled-coil forming motif and deriving the probability terms (see Equation 3.4). Consequently, we can modify the probabilities of all available structures with respect to a given mechanical tension. After formulating our tensile mechanics models, we can validate model predictions through comparison with available single molecule experimental data, which would be computationally inexpensive to run and can be made available as an online application with a user-friendly interface. Such application would be of great utility to a broad range of single molecule experimental research, protein design and engineering, molecular motor research and nanotechnology.

REFERENCES

Acharya, A., Rishi, V., & Vinson, C. (2006). Stability of 100 homo and heterotypic coiled-coil pairs for ten amino acids (A, L, I, V, N, K, S, T, E, and R). *Biochemistry*, *45*(38), 11324-11332.

Acharya, A., Ruvinov, S. B., Gal, J., Moll, J. R., & Vinson, C. (2002). A heterodimerizing leucine zipper coiled coil system for examining the specificity of a position interactions: amino acids I, V, L, N, A, and K. *Biochemistry*, *41*(48), 14122-14131.

Adams, J. C., Bentley, A. A., Kvensakul, M., Hatherley, D., & Hohenester, E. (2008). Extracellular matrix retention of thrombospondin 1 is controlled by its conserved C-terminal region. *J Cell Sci*, *121*(6), 784-795.

Akey, D. L., Malashkevich, V. N., & Kim, P. S. (2001). Buried polar residues in coiled-coil interfaces. *Biochemistry*, *40*(21), 6352-6360.

Armel, T. Z., & Leinwand, L. A. (2009). Mutations in the β -myosin rod cause myosin storage myopathy via multiple mechanisms. *Proceedings of the National Academy of Sciences*, *106*(15), 6291-6296.

Armstrong, K. M., & Baldwin, R. L. (1993). Charged histidine affects alpha-helix stability at all positions in the helix by interacting with the backbone charges. *Proceedings of the National Academy of Sciences*, *90*(23), 11337-11340.

Arndt, K. M., Pelletier, J. N., Müller, K. M., Plückthun, A., & Alber, T. (2002). Comparison of in vivo selection and rational design of heterodimeric coiled coils. *Structure*, *10*(9), 1235-1248.

Atzberger, P. J., & Peskin, C. S. (2006). A Brownian dynamics model of kinesin in three dimensions incorporating the force-extension profile of the coiled-coil cargo tether. *Bulletin of mathematical biology*, 68(1), 131.

Aurora, R., Srinivasan, R., & Rose, G. D. (1994). Rules for alpha-helix termination by glycine. *Science*, 264(5162), 1126-1130.

Best, R. B., & Hummer, G. (2009). Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *The journal of physical chemistry B*, 113(26), 9004-9015.

Blanco, F. J., Ortiz, A. R., & Serrano, L. (1997). Role of a nonnative interaction in the folding of the protein G B1 domain as inferred from the conformational analysis of the α -helix fragment. *Folding and Design*, 2(2), 123-133.

Blankenfeldt, W., Thomä, N. H., Wray, J. S., Gautel, M., & Schlichting, I. (2006). Crystal structures of human cardiac β -myosin II S2- Δ provide insight into the functional role of the S2 subfragment. *Proceedings of the National Academy of Sciences*, 103(47), 17713-17717.

Bornschlöggl, T., & Rief, M. (2008). Single-molecule dynamics of mechanical coiled-coil unzipping. *Langmuir*, 24(4), 1338-1342.

Brown, J. H., Kumar, V. S., O'Neill-Hennessey, E., Reshetnikova, L., Robinson, H., Nguyen-McCarty, M., ... & Cohen, C. (2011). Visualizing key hinges and a potential major source of compliance in the lever arm of myosin. *Proceedings of the National Academy of Sciences*, 108(1), 114-119.

Burkhard, P., Ivaninskii, S., & Lustig, A. (2002). Improving coiled-coil stability by optimizing ionic interactions. *Journal of molecular biology*, 318(3), 901-910.

Burkhard, P., Kammerer, R. A., Steinmetz, M. O., Bourenkov, G. P., & Aepli, U. (2000). The coiled-coil trigger site of the rod domain of cortexillin I unveils a distinct network of interhelical and intrahelical salt bridges. *Structure*, 8(3), 223-230.

Burkhard, P., Stetefeld, J., & Strelkov, S. V. (2001). Coiled coils: a highly versatile protein folding motif. *Trends in cell biology*, 11(2), 82-88.

Chakrabartty, A., Kortemme, T., & Baldwin, R. L. (1994). Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. *Protein Science*, 3(5), 843-852.

Chao, H., Bautista, D. L., Litowski, J., Irvin, R. T., & Hodges, R. S. (1998). Use of a heterodimeric coiled-coil system for biosensor application and affinity purification. *Journal of Chromatography B: Biomedical Sciences and Applications*, 715(1), 307-329.

Chen, B. T., Ghassaban, K., Jin, T., Patel, S. K., Ye, N., Sun, C. L., ... & Saykin, A. J. (2018). Subcortical brain iron deposition and cognitive performance in older women with breast cancer receiving adjuvant chemotherapy: A pilot MRI study. *Magnetic resonance imaging*, 54, 218-224.

Chothia, C., Levitt, M., & Richardson, D. (1981). Helix to helix packing in proteins. *Journal of molecular biology*, 145(1), 215-250.

Cohen, C., & Parry, D. A. (1990). α -Helical coiled coils and bundles: How to design an α -helical protein. *Proteins: Structure, Function, and Bioinformatics*, 7(1), 1-15.

Creamer, T. P., & Rose, G. D. (1995). Interactions between hydrophobic side chains within α -helices. *Protein Science*, 4(7), 1305-1314.

Crick, F. H. (1953). The packing of α -helices: simple coiled-coils. *Acta crystallographica*, 6(8-9), 689-697.

Crooks, R. O., Lathbridge, A., Panek, A. S., & Mason, J. M. (2017). Computational Prediction and Design for Creating Iteratively Larger Heterospecific Coiled Coil Sets. *Biochemistry*, *56*(11), 1573-1584.

Dimitrov, R. A., & Zuker, M. (2004). Prediction of hybridization and melting for double-stranded nucleic acids. *Biophysical Journal*, *87*(1), 215-226.

Doig, A. J. (2008). Stability and design of α -helical peptides. *Progress in molecular biology and translational science*, *83*, 1-52.

Eisenberg, D. (2003). The discovery of the α -helix and β -sheet, the principal structural features of proteins. *Proceedings of the National Academy of Sciences*, *100*(20), 11207-11210.

Engler, A. J., Sen, S., Sweeney, H. L., & Discher, D. E. (2006). Matrix elasticity directs stem cell lineage specification. *Cell*, *126*(4), 677-689.

Fong, J. H., Keating, A. E., & Singh, M. (2004). Predicting specificity in Bzip coiled-coil protein interactions. *Genome biology*, *5*(2), R11.

Gao, Y., Sirinakis, G., & Zhang, Y. (2011). Highly anisotropic stability and folding kinetics of a single coiled coil protein under mechanical tension. *Journal of the American Chemical Society*, *133*(32), 12749-12757.

Ghassaban, K., He, N., Sethi, S. K., Huang, P., Chen, S., Yan, F., & Haacke, E. M. (2019). Regional High Iron in the Substantia Nigra Differentiates Parkinson's Disease Patients From Healthy Controls. *Frontiers in aging neuroscience*, *11*, 106.

Ghassaban, K., Liu, S., Jiang, C., & Haacke, E. M. (2019). Quantifying iron content in magnetic resonance imaging. *Neuroimage*, *187*, 77-92.

Ghouchanian, K., Ghassaban, K., & Jokar, M. (2017). Optimization of the Capsule Production Stochastic Cycle Time. *Research in Economics and Management*, *2*(4), 21.

Gonzalez, L., Brown, R. A., Richardson, D., & Alber, T. (1996). Crystal structures of a single coiled-coil peptide in two oligomeric states reveal the basis for structural polymorphism. *Nature structural biology*, 3(12), 1002.

Greenfield, N. J. (2006). Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions. *Nature protocols*, 1(6), 2527.

Grigoryan, G., & Keating, A. E. (2006). Structure-based prediction of Bzip partnering specificity. *Journal of molecular biology*, 355(5), 1125-1142.

Gruber, M., Söding, J., & Lupas, A. N. (2006). Comparative analysis of coiled-coil prediction methods. *Journal of structural biology*, 155(2), 140-145.

Haacke, E. M., Ghassaban, K., Chen, Y., Liu, S., Buch, S. (2018). T2*: Susceptibility Weighted Imaging and Quantitative Susceptibility Mapping. *Quantitative MRI of the Brain*, 115-128.

Harper, E. T., & Rose, G. D. (1993). Helix stop signals in proteins and peptides: the capping box. *Biochemistry*, 32(30), 7605-7609.

Herrmann, H., Bär, H., Kreplak, L., Strelkov, S. V., & Aebi, U. (2007). Intermediate filaments: from cell architecture to nanomechanics. *Nature reviews Molecular cell biology*, 8(7), 562.

Huguet, J. M., Bizarro, C. V., Forns, N., Smith, S. B., Bustamante, C., & Ritort, F. (2010). Single-molecule derivation of salt dependent base-pair free energies in DNA. *Proceedings of the National Academy of Sciences*, 107(35), 15431-15436.

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: visual molecular dynamics. *Journal of molecular graphics*, 14(1), 33-38.

Huyghues-Despointes, B. M., & Baldwin, R. L. (1997). Ion-pair and charged hydrogen-bond interactions between histidine and aspartate in a peptide helix. *Biochemistry*, *36*(8), 1965-1970.

Huyghues-Despointes, B. M., Klingler, T. M., & Baldwin, R. L. (1995). Measuring the strength of side-chain hydrogen bonds in peptide helices: the Gln. cntdot. Asp (i, i+ 4) interaction. *Biochemistry*, *34*(41), 13267-13271.

Jokar, M., & Torabi, K. (2017). A Statistical Mechanical Model for Coiled-Coil Protein Structure Prediction. *Biophysical Journal*, *112*(3), 193a.

Jokar, M., & Torabi, K. (2017, July). Modeling Coiled-Coil Protein Structures. In EUROPEAN BIOPHYSICS JOURNAL WITH BIOPHYSICS LETTERS (Vol. 46, pp. S346-S346). 233 SPRING ST, NEW YORK, NY 10013 USA: SPRINGER.

Jokar, M., & Torabi, K. (2019). Partnering Specificity and Electrostatic Interactions in Coiled-Coil Protein Domains. *Biochemistry*.

Jokar, M., & Torabi, K. (2016). Tensile Mechanics of Coiled Coil Protein Structures. *Biophysical Journal*, *110*(3), 536a.

Jokar, M., & Torabi, K. (2018). Thermodynamics of a Coiled-Coil Protein Structure. *Biophysical Journal*, *114*(3), 580a.

Joo, H., Chavan, A. G., Phan, J., Day, R., & Tsai, J. (2012). An amino acid packing code for α -helical structure and protein design. *Journal of molecular biology*, *419*(3-4), 234-254.

Kammerer, R. A. (1997). α -Helical coiled-coil oligomerization domains in extracellular proteins. *Matrix biology*, *15*(8-9), 555-565.

Knight, P. J., Thirumurugan, K., Xu, Y., Wang, F., Kalverda, A. P., Stafford, W. F., ... & Peckham, M. (2005). The predicted coiled-coil domain of myosin 10 forms a novel elongated domain that lengthens the head. *Journal of Biological Chemistry*, *280*(41), 34702-34708.

Kohn, W. D., Kay, C. M., & Hodges, R. S. (1995). Protein destabilization by electrostatic repulsions in the two-stranded α -helical coiled-coil/leucine zipper. *Protein Science*, 4(2), 237-250.

Kohn, W. D., Monera, O. D., Kay, C. M., & Hodges, R. S. (1995). The effects of interhelical electrostatic repulsions between glutamic acid residues in controlling the dimerization and stability of two-stranded α -helical coiled-coils. *Journal of Biological Chemistry*, 270(43), 25495-25506.

Kohori, A., Chiwata, R., Hossain, M. D., Furuike, S., Shiroguchi, K., Adachi, K., ... & Kinoshita Jr, K. (2011). Torque generation in F1-ATPase devoid of the entire amino-terminal helix of the rotor that fills half of the stator orifice. *Biophysical journal*, 101(1), 188-195.

Kortemme, T., & Creighton, T. E. (1995). Ionisation of Cysteine Residues at the Termini of Model α -Helical Peptides. Relevance to Unusual Thiol pKaValues in Proteins of the Thioredoxin Family. *Journal of molecular biology*, 253(5), 799-812.

Kreuzer, S. M., & Elber, R. (2013). Coiled-coil response to mechanical force: global stability and local cracking. *Biophysical journal*, 105(4), 951-961.

Krylov, D., Barchi, J., & Vinson, C. (1998). Inter-helical interactions in the leucine zipper coiled coil dimer: pH and salt dependence of coupling energy between charged amino acids. *Journal of molecular biology*, 279(4), 959-972.

Krylov, D., Mikhailenko, I., & Vinson, C. (1994). A thermodynamic scale for leucine zipper stability and dimerization specificity: e and g interhelical interactions. *The EMBO journal*, 13(12), 2849-2861.

Lacroix, E., Viguera, A. R., & Serrano, L. (1998). Elucidating the folding problem of α -helices: local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters. *Journal of molecular biology*, 284(1), 173-191.

Lavigne, P., Crump, M. P., Gagné, S. M., Hodges, R. S., Kay, C. M., & Sykes, B. D. (1998). Insights into the mechanism of heterodimerization from the ^1H -NMR solution structure of the c-Myc-Max heterodimeric leucine zipper. *Journal of molecular biology*, 281(1), 165-181.

Lavigne, P., Kondejewski, L. H., Houston Jr, M. E., Sönnichsen, F. D., Lix, B., Sykes, B. D., ... & Kay, C. M. (1995). Preferential heterodimeric parallel coiled-coil formation by synthetic Max and c-Myc leucine zippers: a description of putative electrostatic interactions responsible for the specificity of heterodimerization. *Journal of molecular biology*, 254(3), 505-520.

Li, Y., Brown, J. H., Reshetnikova, L., Blazsek, A., Farkas, L., Nyitray, L., & Cohen, C. (2003). Visualization of an unstable coiled coil from the scallop myosin rod. *Nature*, 424(6946), 341.

Litowski, J. R., & Hodges, R. S. (2002). Designing Heterodimeric Two-stranded α -Helical Coiled-coils EFFECTS OF HYDROPHOBICITY AND α -HELICAL PROPENSITY ON PROTEIN FOLDING, STABILITY, AND SPECIFICITY. *Journal of Biological Chemistry*, 277(40), 37272-37279.

Liu, M., Liu, S., Ghassaban, K., Zheng, W., Diccio, D., Miao, Y., ... & Haacke, E. M. (2016). Assessing global and regional iron content in deep gray matter as a function of age using susceptibility mapping. *Journal of Magnetic Resonance Imaging*, 44(1), 59-71.

Lupas, A. N., & Gruber, M. (2005). The structure of α -helical coiled coils. In *Advances in protein chemistry* (Vol. 70, pp. 37-38). Academic Press.

Lupas, A., Van Dyke, M., & Stock, J. (1991). Predicting coiled coils from protein sequences. *Science*, 1162-1164.

Marti, D. N., Jelesarov, I., & Bosshard, H. R. (2000). Interhelical ion pairing in coiled coils: solution structure of a heterodimeric leucine zipper and determination of p K a values of Glu side chains. *Biochemistry*, 39(42), 12804-12818.

Mason, J. M., & Arndt, K. M. (2004). Coiled coil domains: stability, specificity, and biological implications. *Chembiochem*, 5(2), 170-176.

McClain, D. L., Gurnon, D. G., & Oakley, M. G. (2002). Importance of potential interhelical salt-bridges involving interior residues for coiled-coil stability and quaternary structure. *Journal of molecular biology*, 324(2), 257-270.

McLachlan, A. D., & Stewart, M. (1975). Tropomyosin coiled-coil interactions: evidence for an unstaggered structure. *Journal of molecular biology*, 98(2), 293-304.

Moitra, J., Szilák, L., Krylov, D., & Vinson, C. (1997). Leucine is the most stabilizing aliphatic amino acid in the d position of a dimeric leucine zipper coiled coil. *Biochemistry*, 36(41), 12567-12573.

Monera, O. D., Kay, C. M., & Hodges, R. S. (1994). Electrostatic Interactions Control the Parallel and Antiparallel Orientation of α -Helical Chains in Two-Stranded α -Helical Coiled-Coils. *Biochemistry*, 33(13), 3862-3871.

Monera, O. D., Kay, C. M., & Hodges, R. S. (1994). Protein denaturation with guanidine hydrochloride or urea provides a different estimate of stability depending on the contributions of electrostatic interactions. *Protein Science*, 3(11), 1984-1991.

Monera, O. D., Sereda, T. J., Zhou, N. E., Kay, C. M., & Hodges, R. S. (1995). Relationship of sidechain hydrophobicity and α -helical propensity on the stability of the single-

stranded amphipathic α -helix. *Journal of peptide science: an official publication of the European Peptide Society*, 1(5), 319-329.

Muñoz, V., Blanco, F. J., & Serrano, L. (1995). The hydrophobic-staple motif and a role for loop-residues in α -helix stability and protein folding. *Nature structural biology*, 2(5), 380.

Muñoz, V., & Serrano, L. (1997). Development of the multiple sequence approximation within the AGADIR model of α -helix formation: Comparison with Zimm-Bragg and Lifson-Roig formalisms. *Biopolymers: Original Research on Biomolecules*, 41(5), 495-509.

Muñoz, V., & Serrano, L. (1995). Elucidating the Folding Problem of Helical Peptides using Empirical Parameters. III> Temperature and pH Dependence. *Journal of molecular biology*, 245(3), 297-308.

Muñoz, V., & Serrano, L. (1994). Intrinsic secondary structure propensities of the amino acids, using statistical ϕ - ψ matrices: comparison with experimental scales. *Proteins: Structure, Function, and Bioinformatics*, 20(4), 301-311.

Neves-Petersen, M. T., & Petersen, S. B. (2003). Protein electrostatics: a review of the equations and methods used to model electrostatic equations in biomolecules—applications in biotechnology. *Biotechnol Annu Rev*, 9(315), 95.

Newman, J. R., & Keating, A. E. (2003). Comprehensive identification of human Bzip interactions with coiled-coil arrays. *Science*, 300(5628), 2097-2101.

O'Shea, E. K., Klemm, J. D., Kim, P. S., & Alber, T. (1991). X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. *Science*, 254(5031), 539-544.

O'Shea, E. K., Lumb, K. J., & Kim, P. S. (1993). Peptide 'Velcro': design of a heterodimeric coiled coil. *Current Biology*, 3(10), 658-667.

Oshaben, K. M., Salari, R., McCaslin, D. R., Chong, L. T., & Horne, W. S. (2012). The native GCN4 leucine-zipper domain does not uniquely specify a dimeric oligomerization state. *Biochemistry*, *51*(47), 9581-9591.

Padmanabhan, S., & Baldwin, R. L. (1994). Helix-stabilizing interaction between tyrosine and leucine or valine when the spacing is $i, i+4$. *Journal of molecular biology*, *241*(5), 706-713.

Padmanabhan, S., & Baldwin, R. L. (1994). Tests for helix-stabilizing interactions between various nonpolar side chains in alanine-based peptides. *Protein Science*, *3*(11), 1992-1997.

Parry, D. A. D., Crewther, W. G., Fraser, R. D. B., & MacRae, T. P. (1977). Structure of α -keratin: structural implication of the amino acid sequences of the type I and type II chain segments. *Journal of molecular biology*, *113*(2), 449-454.

Parry, D. A. D., Fraser, R. B., & Squire, J. M. (2008). Fifty years of coiled-coils and α -helical bundles: a close relationship between sequence and structure. *Journal of structural biology*, *163*(3), 258-269.

Pauling, L., & Corey, R. B. (1953). Compound helical configurations of polypeptide chains: structure of proteins of the α -keratin type. *Nature*, *171*(4341), 59.

Petukhov, M., Muñoz, V., Yumoto, N., Yoshikawa, S., & Serrano, L. (1998). Position dependence of non-polar amino acid intrinsic helical propensities. *Journal of molecular biology*, *278*(1), 279-289.

Petukhov, M., Yumoto, N., Murase, S., Onmura, R., & Yoshikawa, S. (1996). Factors that affect the stabilization of α -helices in short peptides by a capping box. *Biochemistry*, *35*(2), 387-397.

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., ... & Schulten, K. (2005). Scalable molecular dynamics with NAMD. *Journal of computational chemistry*, 26(16), 1781-1802.

Potapov, V., Kaplan, J. B., & Keating, A. E. (2015). Data-driven prediction and design of Bzip coiled-coil interactions. *PLoS computational biology*, 11(2), e1004046.

Prieto, J., & Serrano, L. (1997). C-capping and helix stability: the Pro C-capping motif. *Journal of molecular biology*, 274(2), 276-288.

Reinke, A. W., Baek, J., Ashenberg, O., & Keating, A. E. (2013). Networks of Bzip protein-protein interactions diversified over a billion years of evolution. *Science*, 340(6133), 730-734.

Reinke, A. W., Grant, R. A., & Keating, A. E. (2010). A synthetic coiled-coil interactome provides heterospecific modules for molecular engineering. *Journal of the American Chemical Society*, 132(17), 6025-6031.

Reuter, J. S., & Mathews, D. H. (2010). RNAstructure: software for RNA secondary structure prediction and analysis. *BMC bioinformatics*, 11(1), 129.

Richardson, J. S., & Richardson, D. C. (1988). Amino acid preferences for specific locations at the ends of alpha helices. *Science*, 240(4859), 1648-1652.

Root, D. D., Yadavalli, V. K., Forbes, J. G., & Wang, K. (2006). Coiled-coil nanomechanics and uncoiling and unfolding of the superhelix and α -helices of myosin. *Biophysical journal*, 90(8), 2852-2866.

Rose, A., & Meier, I. (2004). Scaffolds, levers, rods and springs: diverse cellular functions of long coiled-coil proteins. *Cellular and Molecular Life Sciences CMLS*, 61(16), 1996-2009.

Rouzina, I., & Bloomfield, V. A. (2001). Force-induced melting of the DNA double helix
1. Thermodynamic analysis. *Biophysical journal*, 80(2), 882-893.

Sancho, J., Serrano, L., & Fersht, A. R. (1992). Histidine residues at the N-and C-termini
of. alpha.-helices: perturbed pKas and protein stability. *Biochemistry*, 31(8), 2253-2258.

SantaLucia, J. (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA
nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences*, 95(4),
1460-1465.

SantaLucia Jr, J., & Hicks, D. (2004). The thermodynamics of DNA structural motifs.
Annu. Rev. Biophys. Biomol. Struct., 33, 415-440.

Schellman, C. (1980). In Protein Folding (Jaenicke, R., ed.).

Schneider, J. P., Lear, J. D., & DeGrado, W. F. (1997). A designed buried salt bridge in a
heterodimeric coiled coil. *Journal of the American Chemical Society*, 119(24), 5742-5743.

Scholtz, J. M., York, E. J., Stewart, J. M., & Baldwin, R. L. (1991). A neutral, water-
soluble,. alpha-helical peptide: the effect of ionic strength on the helix-coil equilibrium. *Journal
of the American Chemical Society*, 113(13), 5102-5104.

Schwaiger, I., Sattler, C., Hostetter, D. R., & Rief, M. (2002). The myosin coiled-coil is a
truly elastic protein structure. *Nature materials*, 1(4), 232.

Sethi, S. K., Kisch, S. J., Ghassaban, K., Rajput, A., Rajput, A., Babyn, P. S., ... &
Haacke, E. M. (2019). Iron quantification in Parkinson's disease using an age-based threshold on
susceptibility maps: The advantage of local versus entire structure iron content measurements.
Magnetic resonance imaging, 55, 145-152.

Sheetz, M. P., Felsenfeld, D. P., & Galbraith, C. G. (1998). Cell migration: regulation of
force on extracellular-matrix-integrin complexes. *Trends in cell biology*, 8(2), 51-54.

Sitkoff, D., Lockhart, D. J., Sharp, K. A., & Honig, B. (1994). Calculation of electrostatic effects at the amino terminus of an alpha helix. *Biophysical journal*, 67(6), 2251-2260.

Smith, J. S., & Scholtz, J. M. (1998). Energetics of polar side-chain interactions in helical peptides: salt effects on ion pairs and hydrogen bonds. *Biochemistry*, 37(1), 33-40.

Stapley, B. J., & Doig, A. J. (1997). Hydrogen bonding interactions between glutamine and asparagine in α -helical peptides. *Journal of molecular biology*, 272(3), 465-473.

Stapley, B. J., Rohl, C. A., & Doig, A. J. (1995). Addition of side chain interactions to modified Lifson-Roig helix-coil theory: Application to energetics of phenylalanine-methionine interactions. *Protein Science*, 4(11), 2383-2391.

Tan, Z. J., & Chen, S. J. (2006). Nucleic acid helix stability: effects of salt concentration, cation valence and size, and chain length. *Biophysical journal*, 90(4), 1175-1190.

Taniguchi, Y., Khatri, B. S., Brockwell, D. J., Paci, E., & Kawakami, M. (2010). Dynamics of the coiled-coil unfolding transition of myosin rod probed by dissipation force spectrum. *Biophysical journal*, 99(1), 257-262.

Torabi, K., & Schatz, G. C. (2013). Tensile mechanics of α -helical polypeptides. *Macromolecules*, 46(19), 7947-7956.

Tripet, B., Vale, R. D., & Hodges, R. S. (1997). Demonstration of Coiled-Coil Interactions within the Kinesin Neck Region Using Synthetic Peptides IMPLICATIONS FOR MOTOR ACTIVITY. *Journal of Biological Chemistry*, 272(14), 8946-8956.

Tripet, B., Wagschal, K., Lavigne, P., Mant, C. T., & Hodges, R. S. (2000). Effects of side-chain characteristics on stability and oligomerization state of a de novo-designed model coiled-coil: 20 amino acid substitutions in position "d". *Journal of molecular biology*, 300(2), 377-402.

Viguera, A. R., & Serrano, L. (1995). Experimental analysis of the Schellman motif. *Journal of molecular biology*, 251(1), 150-160.

Viguera, A. R., & Serrano, L. (1995). Side-chain interactions between sulfur-containing amino acids and phenylalanine in α -helices. *Biochemistry*, 34(27), 8771-8779.

Vinson, C. R., Hai, T. S. O. N. W. I. N., & Boyd, S. M. (1993). Dimerization specificity of the leucine zipper-containing Bzip motif on DNA binding: prediction and rational design. *Genes & development*, 7(6), 1047-1058.

Vinson, C., Myakishev, M., Acharya, A., Mir, A. A., Moll, J. R., & Bonovich, M. (2002). Classification of human B-ZIP proteins based on dimerization properties. *Molecular and cellular biology*, 22(18), 6321-6335.

Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. *Journal of molecular graphics*, 8(1), 52-56.

Wagschal, K., Tripet, B., Lavigne, P., Mant, C., & Hodges, R. S. (1999). The role of position a in determining the stability and oligomerization state of α -helical coiled coils: 20 amino acid stability coefficients in the hydrophobic core of proteins. *Protein Science*, 8(11), 2312-2329.

Wiseman, N. M., Ghassaban, K., Utriainen, D. T., Buch, S., Kou, Z., & Haacke, E. M. (2019). Perfusion and Susceptibility Weighted Imaging in Traumatic Brain Injury. In *Neurosensory Disorders in Mild Traumatic Brain Injury* (pp. 303-319). Academic Press.

Williams, M. C., Rouzina, I., & Bloomfield, V. A. (2002). Thermodynamics of DNA interactions from single molecule stretching experiments. *Accounts of chemical research*, 35(3), 159-166.

Wu, P., Nakano, S. I., & Sugimoto, N. (2002). Temperature dependence of thermodynamic properties for DNA/DNA and RNA/DNA duplex formation. *European journal of biochemistry*, 269(12), 2821-2830.

Yu, Y. B. (2002). Coiled-coils: stability, specificity, and drug delivery potential. *Advanced drug delivery reviews*, 54(8), 1113-1129.

Zeng, X., Herndon, A. M., & Hu, J. C. (1997). Buried asparagines determine the dimerization specificities of leucine zipper mutants. *Proceedings of the National Academy of Sciences*, 94(8), 3673-3678.

Zhou, N. E., Kay, C. M., & Hodges, R. S. (1994). The role of interhelical ionic interactions in controlling protein folding and stability: de novo designed synthetic two-stranded α -helical coiled-coils. *Journal of molecular biology*, 237(4), 500-512.

Zuker, M., Mathews, D. H., & Turner, D. H. (1999). Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In *RNA biochemistry and biotechnology* (pp. 11-43). Springer, Dordrecht.

ABSTRACT**PREDICTING THE STRUCTURE AND SELECTIVITY OF COILED-COIL PROTEINS**

by

MOJTABA JOKAR**August 2019****Advisor:** Dr. Korosh Torabi**Major:** Chemical Engineering**Degree:** Doctor of Philosophy

A coiled-coil protein structure consists of two (in coiled-coil dimers) or more interacting α -helical strands that together form a left-handed supercoil structure. Many coiled-coil proteins are involved in significant biological functions such as the regulation of gene expression, known as transcription factors. Also coiled-coil structures entail unique mechanical properties critical to the function and integrity of various motor proteins, cytoskeletal filaments and extra-cellular matrix proteins. Engineering these transcription factors is also expected to create more efficient and practical solutions to treat neurodegenerative diseases such as Alzheimer's disease (AD), Parkinson's disease (PD), Huntington's disease (HD), amyotrophic lateral sclerosis (ALS) and prion diseases, which are increasingly being realized to have common cellular and molecular mechanisms including protein aggregation. The main objectives of our work are: a) to develop a model to predict the propensity of a protein sequence to form an isolated coiled-coil structure, and b) to investigate the selectivity of coiled-coils by studying protein-protein interactions. Control over protein-protein interaction specificity has a wide range of applications in synthetic biology such as protein labeling and purification (as high-specificity affinity tags or cognate pairs), drugs and toxin delivery and disease modulation. In naturally occurring proteins,

specificity is achieved via a complex balance of various molecular-level energetic and entropic interactions. Such complexity makes any specificity prediction from the primary sequence data an extremely complicated task. Possibly, one of the simplest and most studied protein-protein interactions exists in coiled-coil structures.

AUTOBIOGRAPHICAL STATEMENT

Mojtaba Jokar joined Wayne State University in January 2015, holding a BSc degree in Chemical Engineering from Amirkabir University of Technology (Tehran Polytechnic). He continued studying towards his doctoral degree in Chemical Engineering at Wayne State University. During his education at Wayne State, he has served as a Graduate Research Assistant and Graduate Teaching Assistant for multiple undergraduate and graduate courses at the Department of Chemical Engineering and Materials Science. His main research interest is computational biology, with a focus on studying neurodegenerative diseases such as Alzheimer's and Parkinson's.