A theoretical study on nonnative contacts in protein folding by a modified HP model

Chong Chen¹,² and Ran Huang¹,³,⁴

¹Department of Materials Technology and Engineering, Research Institute of Zhejiang University-Taizhou, Taizhou, Zhejiang 318000, China
²Ningbo Hongzheng Engineering Consulting, Inc., Ningbo, Zhejiang 315000, China
³State Key Laboratory of Microbial Metabolism and School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China
⁴Email: ranhuang@sjtu.edu.cn

Abstract. We have investigated the nonnative contacts for proteins in a modified HP model by generating all possible conformations (enumeration) to carry out exact calculation for the nonnative contact density \( n_\varepsilon(e) \) as a function of the energy \( e \), as well as the thermodynamically average nonnative contacts \( \bar{n}_\varepsilon(T) \) as a function of temperature of \( T \). A series of abstract coarse-grained chains on standard, weekly and strongly interacting models were studied to estimate the properties of \( n_\varepsilon(e) \) and \( \bar{n}_\varepsilon(T) \) to understand the role for nonnative contacts playing in the protein folding process.

1. Introduction

The exploration on protein folding has been a long and tough journey in the last half century, numerous efforts have been devoted into this extraordinary problem, nevertheless, the exact mechanism is still unknown to us. Such folding behavior that a particular linear arrangement of amino acids can automatically achieve its unique 3D structure is one of the key issues in the research on protein [1]. Meanwhile, the study of protein folding problem becomes more and more important in various fields such as biophysics, cell biology medicine, biotechnology, and even materials science, which require an urgent knowledge on the nature of protein folding [2].

Along with experimental techniques such as X-ray crystallography and multi-dimensional NMR being widely used to study the structure as well as the dynamics of proteins [3, 4], theoretical models and computational methods can also help in the structure prediction [5], and usually coarse-grained bead chains are employed for convenient treatment in physics [6, 7]. Out of a number of coarse-grained models, the HP model is a well-developed methodology which puts the protein on a lattice, and presents the amino acids as sites with chemical bonds as edges, with all unoccupied site filled by water molecules [8]. All the monomers are classified into two types: H (hydrophobic species) and P (hydrophilic species) linked in a 2D self-avoiding walk chain [9]. The model simplifies the driving force for folding to be only hydrophobic force, which minimizes the number of hydrophobic residues exposed to water, with nearest-neighbor HH contact interaction being assigned the energy \( \varepsilon < 0 \).

Although being simple and difficult to correlate an experimental evidence, as an exact model, the HP captures the underlying thermodynamics in predicting protein structure from amino acid sequence [10]. The disadvantage is that the model lacks details. The resolution of protein structures and energetic are not accurately represented. In this work, we extend the classical HP model with a number of nonnative
contacts other than the HH pair, in order to make the model more capable to describe the complex reality. We have studied three sets of parameters for interactions presenting different energetic parts between chemically unbounded residues from the conformation and the solvent [11]. The enumeration allows us to investigate exactly the effects of energetic on the native state(s), the effect of small size on protein thermodynamics and the difference between conformations.

2. Model and methods

We treat a protein as a coarse-grained semiflexible self-avoiding chain on a square lattice to study its folding behavior, with different energy penalties assigned to different structures in the conformations generated by program. The semiflexibility of the protein gives rise to a crystalline phase, which represents the native state of the protein at low temperatures.

Figure 1 shows a protein in its compact form so that all the solvent molecules (W) are expelled from inside and surround the protein. The red beads denote hydrophobic residues (H) and blue beads represent hydrophilic residues (P), and this model does not allow voids on the lattice. The nearest-neighbor distinct pairs PP, HH, HP, PW and HW between the residues and the water are also shown. Since only three out of these six contacts are independent on the lattice, we choose the HH, HW, and HP pairs to be our variables.

Figure 1. A compact HP model of protein with water surrounded on a square lattice. (The blue beads are hydrophilic residues, red are hydrophobic, and green beads are water molecules)

2.1. Protein as a small system

We consider a finite protein with M residues in a given sequence $\chi$ of H and P associated with the residues on a square lattice. All the conformations are generated on infinite lattice so that the system is ergodic and the boundary of the volume does not affect the behavior of the system. As we only deal with a single protein, the behavior of one polypeptide chain which consists of $M << \infty$ residues is governed by small system thermodynamics. While the predictions of different ensembles describing a macroscopic system are the same, for a small system, different statistical ensembles differ from each other.

2.2. Interaction energies

The interaction energies are restricted between chemically unbounded residues H and P, and water molecules W that are nearest neighbors of each other. For only three independent interactions out of six pairs of species, we choose the three van der Waals energies $e_{HH}$, $e_{HW}$, and $e_{PH}$ between the unboned pairs HH, HW, and PH.

The number of bends $N_b$, pairs of parallel bonds $N_p$, and hairpin turns $N_{hp}$ are used to characterize the semiflexibility. A bend is where the protein deviates from its collinear path. Each hairpin turn contains two consecutive bends in the same direction (clockwise or counterclockwise), and a "helical turn" $N_{hl}$ is interpreted by two consecutive hairpin turns in opposite directions (Figure 2):
Figure 2. An illustration of two hairpin turns of clockwise or counter-clockwise respectively. (The two hairpin turns also interpret a helical turn)

Let $e$ denote the set containing all $\{e_i\}$, and $N$ denotes the set containing all $\{N_i\}$, where $i$ stands for b, p, hp, hl, HH, HW, and HP, then we have

$e ≡ \{e_b, e_p, e_{hp}, e_{hl}, e_{HH}, e_{HW}, e_{PH}\}$,

$N ≡ \{N_b, N_p, N_{hp}, N_{hl}, N_{HH}, N_{HW}, N_{PH}\}$.

Let $W(N)$ denotes the number of protein configurations. The energy of particular configuration $\Gamma$ corresponding to the set $N$ is given by

$$E(N) = e \cdot N = \sum_i e_i N_i$$

The entropy function corresponding to configurations with a given $N$ is defined as

$$S(N) = k \ln W(N)$$

Here the Boltzmann constant $k$ is normalized to be 1, and we have $\beta = 1/T$ in the following discussion. As mentioned in the introduction, the parameters are artificial and we simply focus on a general observation on the difference in orders of magnitude between parameter setups. We have tested a series of parameters and found the results to be consistent, therefore, the three submodels presented in this work are for a general illustration. The parameters are shown in Table 1.

### Table 1. Models and their Parameters.

|                  | Standard (A) | Weakly (B) | Strongly (C) |
|------------------|--------------|------------|--------------|
| Bend             | 0            | 1/50       | 1/3          |
| Parallel         | 0            | -1/50      | -1/3         |
| Hairpin          | 0            | -2/50      | -1/3         |
| Helix            | 0            | -2/50      | -1/3         |
| HH               | -1           | -50/50     | -3/3         |
| HW               | 0            | 20/50      | 2/3          |
| PH               | 0            | 5/50       | 1/3          |

2.2.1. The standard model (A). In the standard model, only the interactions between hydrophobic residues are considered, which are always set to be $e_{HH} = -1$. The density $n_{HH} = N_{HH}/M$ is going to be a discrete quantity, and the energy density is presented as $e = E/M = n_{HH}$.

2.2.2. Weakly perturbed model (B). In this model, other energies are non-zero but still small compared to $e_{HH} = -1$. The numerator of various energies are integers and are used to determine the energy $E$ as an integer, which makes it easy to classify energy levels in groups of a given energy. Due to the strong repulsion between H and W, the only energy close to $|e_{HH}|$ is $e_{HW}$.

2.2.3. Strongly perturbed model (C). Other energies are set comparable to $e_{HH} = -1$ as shown in the table. Comparing to model B, this model is more derived from A. Despite this, we will see that both B and C behave very different from A. It should be noted that $W$ does not depend on the model; it is its partition into $W(E)$ that depends on the model.
2.3. Nonnative contacts

The native state of a protein is a compact, lowest free energy state in which a protein is able to perform its biological functions. The term “contact” is defined as any pair of these residues occupying nearest neighbour sites that are not connected by covalent bonds. Nonnative contacts refer to the contacts appearing in the non-native conformations that do not exist in the native conformation.

Figure 3 shows an example of a compact native and nonnative conformation of a protein. While chemical bonds are represented by bold lines, the dash lines indicate contacts. In the nonnative state of a higher-energy denatured protein conformation, nonnative contacts are marked by adding stars on the dash lines. As an illustration, residue 4 here has now contacts with residue 35 and 29, which are not found in the native conformation, so these two contacts are now nonnative contacts as we defined before. However, in the native state, residue 8 and 13 are in contact, while in the denatured state, they are still in contact. By our definition, the pairing 8-13 is not a nonnative contact.

Figure 3. A schematic plot for native and nonnative state for a protein with 36 residues.

Nonnative contact are believed to play a vital role in different stages of protein folding: firstly, the nonnative contacts may affect protein folding rates; secondly, nonnative contacts make decent contributions to the unfolding pathways [12]; and lastly, the nonnative contacts provide rich folding intermediates [13].

Our program runs in a way that given a particular sequence, all possible conformations are generated and examined in turns. The lowest energy conformation is selected as the native conformation. With the native state information stored, the program runs again to collect all nonnative contacts for each conformation by comparing with the native conformation, and captures the nonnative contacts density as a function of the energy, which is defined as

$$n_c(e) = \frac{N_c(e)}{W(e) \times M}$$

where $W(e)$ is the number of conformations that has the same e, and $N_c(e)$ is the sum of nonnative contacts in all the conformations that have the same e. The $n_c(e)$ describes the average number of nonnative contacts for certain energy density e per residue.

Note that e is the energy numerically counted for a particular configuration, which is independent to the temperature. To relate the density variables to the temperature, canonical average of $n_c(E)$ by the Boltzmann weights is necessary to characterize the behaviour of nonnative contacts with respect to temperature:

$$\tilde{n}_c(T) = \frac{\sum_e n_c(e) W(e) \exp(-\beta E)}{\sum_e W(e) \exp(-\beta E)} = \frac{1}{M} \frac{\sum_e N_c(e) \exp(-\beta E)}{\sum_e W(e) \exp(-\beta E)}$$

(4)
3. Results and discussion

Figure 4. The behaviour of $n_e(e)$ for: (a) 20 residue protein in the standard model; (b) 20 residue protein in the weakly interacting model; (c) 20 residue protein in the strongly interacting model; (d) 11 residue protein in the strongly interacting model; (e) 14 residue protein in the strongly interacting model.

Figure 4 presents the $n_e(e)$ for various sequence in different models, and all the results are found almost linear to $e$. For a weakly perturbed model protein, $n_e(e)$ distributes itself around the value of $e$ found in for the standard model, because of the slight changes in the energy. The low energy end is erratic and
discrete (Figure 4b). In a strongly perturbed protein (Figure 4c), the energy level is fully spread out, so the linear shape appears smoother. When the number of residue $M$ decreases (Figure 4c, 4d and 4e), the low energy density end appears erratic and not smooth, and $n_c(e)$ is also found to be saturated when $e$ reaches its high end.

For a protein that has only one lowest energy state (native state), notable in weakly perturbed model (Figure 5a) and strongly perturbed model (Figure not shown), $\bar{n}_c(T)$ increases from zero as $T$ increases. Clearly the only possible state for the protein is its unique compact native state when $T = 0$, so there is no nonnative contacts. As the temperature goes up, $\bar{n}_c(T)$ increases dramatically and finally slows down and saturates. However, for a standard model protein (Figure 5b), the lowest energy conformation is not unique, the $\bar{n}_c(T)$ does not start from zero; while the similar behaviour is still obtained as the increasing temperature.

![Figure 5](image1.png)

**Figure 5.** The behaviour of $\bar{n}_c(T)$ for: (a) 14 residue protein in the weakly interacting model; (b) 14 residue protein in the standard model.

As previous literatures indicate that the sequence plays important role in the protein function [14, 15], in order to justify the effect of sequence in our model, we tested several protein sequences. Figure 6 presents two example sequence $\chi_1$ (pphhpphhhhpppphp) (a) and $\chi_2$ (hphphppphpphhppp) (b), which are randomly generated with equal H and P type residues. No significant difference is observed in $n_c(e)$ and $\bar{n}_c(T)$ behaviors.

![Figure 6](image2.png)
Figure 6. The density behaviour for a 20-residue protein in the weakly interacting model with various sequences: (a) $n_c(e)$ of $\chi_1$; (b) $n_c(e)$ of $\chi_2$; (c) $\bar{n}_c(T)$ of $\chi_1$; (d) $\bar{n}_c(T)$ of $\chi_2$.

4. Conclusions
The nonnative contacts were studied on a modified HP model, the density of contacts versus energy density and temperature is captured by exact calculation and enumeration. We have observed that: 1) The density $n_c(e)$ is always monotonically increasing in the standard model, while in the weakly and strongly interacting model, the density $n_c(e)$ is spread out, but still generally increasing; 2) The $\bar{n}_c(T)$ for all cases is generally monotonically increasing; 3) In the present cases, the thermal property is not sequence dependent. Although the abstract observations so far are not well correlated to the experimental reality, we hope this work may serve an introductory reference for the following works on not only the topic of protein folding itself, but also the polymer thermodynamics, the restricted random-walk dynamics, the design of macromolecules like foldamer, etc.

Acknowledgments
This work is financially supported by the Taizhou Municipal Science and Technology Program (2015cls01, 1701gy15 and 1801gy16).

References
[1] Merz Jr K and Le Grand S M 1994 The Protein Folding Problem and Tertiary Structure Prediction (Boston: Birkhauser)
[2] Dill K A and MacCallum J L 2012 Science 338 1042
[3] Gutin A, Abkevich V and Shakhnovich E 1995 Biochemistry 34 3066
[4] Dobson C M 2003 Nature 426 884
[5] Lane1 T J, Shukla D, Beauchamp K A and Pande V S 2013 Curr. Opin. Struc. Biol. 23 58
[6] Ming D and Wall M E 2005 Phys. Rev. Lett. 95 198103
[7] Bereau T and Deserno M 2009 J. Chem. Phys. 130 235106
[8] Berger B and Leighton T 1998 J. Comp. Biol. 5 27
[9] Chikenji G and Kikuchi M 2000 Proc. Natl. Acad. Sci. 97 26
[10] Guo Y, Tao F, Wu Z and Wang Y 2017 BMC Syst. Biol. 11 93
[11] Chikenji G and Kikuchi M 2000 Proc. Natl. Acad. Sci. 97 26
[12] Guo Y, Tao F, Wu Z and Wang Y 2017 BMC Syst. Biol. 11 93
[13] Irback A and Potthast F 1995 J. Chem. Phys. 103 23
[14] Moores S L, Schaber M D, Mosser S D, Rand J, Ohara M B, Garsky V M, Marshall M S, Pompliano D L and Gibbs J B 1991 J. Biol. Chem. 266 22