Statistical Physics of Biological Information and Complexity

Literature review: Designability of lattice model heteropolymers [1]

Marco V. Bayas **October 15, 2001**

1. Introduction

The folding of proteins is one of the most challenging problems in science. Although, the sequences and the native tridimensional structure of many proteins are known, there is not a general theoretical framework to describe the intermediate steps of folding. On the other hand, experimental research of the intermediates steps is restricted by the range of times involved in the fast phase of folding (~μs), only accessible in few cases [6]. In general, it is known that the tridimensional structure of proteins is determined by its amino acid sequence, but we do not know how this is possible.

One method to analyze protein is the inverse folding strategy [4]. In this scenario the quest is the search for sequences that fold fast into a predefined native conformation. Usually, this is done using the so-called lattice models and Monte Carlo simulations [3,4]. This strategy has been extensively used in the study of protein folding [2,3,4,5] and recently to analyze the concept of designability of protein conformations [1,2]. The designability of a given conformation of lattice model heteropolymers is the number of sequences that fold to that conformation in short time.

The aim of this article is to review the model given by Tiana et al. [1] for the calculation of the designability. Instead of using a 2-letter[1] representation of protein sequences as in the first studies of designability [2] they used a 20-letter representation. The analysis made by Tiana et al. confirmed the relation between stability and designability. According to this, proteins which are stable can tolerate mutations without a lost of activity and consequently without lost of structure. Of course this was known experimentally but there was not a theoretical framework for its interpretation.

2. Lattice Models

In lattice models, a protein is modeled as a string of beads that is arranged on a cubic lattice of step length $\boldsymbol{a}$. In these conditions, the configurational energy of a chain of $\boldsymbol{N}$ monomers is given by:

$$E = \frac{1}{2}\sum_{i,j}^{N} V_{m_i m_j}\,\Delta(\vec{r}_i - \vec{r}_j)$$

where $V_{m_i m_j}$ is the effective interaction potential between monomers m_i and m_j[2]

[1] Polar and Non-polar amino acids.

[2] The values used by Tiana et al. were those provided in the work of Miyazawa et al. [7]. These values correspond to a 20-letter representation of protein sequences.

$\vec{r}_i$ is the position of the *i*th monomer, and

$\Delta(x)$ is the contact function, defined by:

$$\Delta(x) = \begin{cases} 1 & x = a \\ \infty & x = 0 \\ \infty & otherwise \end{cases}$$

In this context, "native" sequences are designed by minimizing, for a fixed amino acid composition, the energy of a target conformation, with respect to the amino acid sequence. The sequences found in this way fold in a short time to the target conformation, which is called "native".

Besides the native sequences, there are other "good folder" sequences [2] which are characterized by a large gap $\delta = E_c - E_n$ (compared to the standard deviation of the contact energies) between the energy E_n in the native conformation and the lowest energy E_c of the conformation structurally dissimilar to the native conformation. E_c is determined by the composition of the protein [4].

3. Designability

The designability of a conformation is the number **n** of sequences that can fold to that conformation. It can be found starting from the minimun energy sequence and counting the number n_m of sets of ***m*** mutations[3] lying within the gap $\delta = E_c - E_n$, that is:

$$n = \sum_{m=1}^{N} n_m$$

Instead of counting the sequences, designability can be evaluated considering the energy distribution probability $P_m(\Delta E)$ associated with m mutations and then evaluating the integral:

$$n_m = n_m^{tot} \int_0^{\delta} P_m(\Delta E)\, d\,\Delta E$$

where n_m^{tot} is the total number of sets of m mutations.

Thus, the problem of determining the designability is reduced to find n_m^{tot} and $P_m(\Delta E)$. It turned out that $n_m^{tot} \approx e^{\alpha m}$, where α is determined by the condition that $n_N^{tot} = N!$ for the case of composition conserving mutations. And, $P_m(\Delta E)$ can be expressed as a convolution of functions $P_2(\Delta E)$ i.e. the energy distribution associated to two mutations.

The functions $P_2(\Delta E)$ were found with MC simulations considering specific lattice model heteropolymers. It was found that these functions have two peaks and that the

---

[3] In this context one mutation corresponds to the replacement of one amino acid by another.

resultant distribution can be fitted by the sum of two gaussian distributions. Further analysis shows that the peaks of the distribution are associated to the cold and hot sites[4]. The Gaussian behavior of $P_2(\Delta E)$ has been attributed to the random nature of the interactions associated to the cold sites.

Under these considerations, it was found for the designability that:

$$n \propto e^{\beta\delta}$$

where β depends on the parameters of the gaussian distribution associated to the cold sites. The calculation of the designability in the case of the 36mer[5] gives a value of 6 x 10^{30} which should be compared with $n_{36}^{tot} = 3.72x10^{41}$ for the case of composition conserving mutations.

The fact that the designability is determined only by δ shows that the concepts of designability and foldability are intimately connected. If a protein is stable in its native conformation, such conformation is highly designable and viceversa.

4. Discussion

The result obtained by Tiana et.al allows one to estimate the probability of that a random sequence folds to given conformation. It will be equal to the ratio between the designability and the total number of possible sequences. In the case of the 36mer this number is in the order of~10^{-12}. This number is very small, as it could have been expected from the fact that if one chose a random sequence of amino acids with the same composition of any protein, this will not fold to its native conformation.

Manifestations of the designability can be found in real proteins where mutations in the "cold" sites maintain a functional protein. Site directed mutagenesis experiments have show that this is possible. Of course, the amount of mutations we can make is limited. It is also known that proteins with the same function in different species share the same overall tridimensional structure even though there are several differences in their amino acid sequences. Again, it is also known that those differences are limited to a small number of amino acids. The expression found by Tiana et.al is a mathematical expression of these facts. It is reasonable to think that the designability in real protein is also quantifiable.

The fact that the structure of a protein can be described using a small number of secondary structures such that α-helices and β-strands suggests that at least at this level of organization there are no many restrictions on the sequence of amino acids necessary to have a given secondary structure. So, secondary structures are more designable than tertiary structures and consequently more stable. Experimental evidence support this conclusion [6].

Finally, it is important to mention that the results reported by Tiana et al. rely on the Gaussian character found for the function $P_2(\Delta E)$. This was attributed to the randomness of

[4] A mutation in a cold site does not change the configurational energy significantly where as a mutation in a hot site cause a big change and even denaturation.

[5] The 36mer is chain of 36 amino acids designed to have the lowest conformation on a given cubic lattice [1]

the interactions associated to cod sites are random and independent of the specific amino acid in the site. This is no necessarily the case in real proteins where the difference between amino acids could matter. Moreover, this behavior of the cold sites found using lattice models could be a consequence of the interaction matrix [7] used in the simulations. The result should be generalized by considering a general form of the interaction matrix. This is necessary considering that it has been pointed out that the matrix used by Tiana et al. is not totally accurate in the description of the interactions in proteins [8].

5. References

1. Tiana, G. Broglia, R. A. Provasi, D. 2001. *Designability of lattice model heteropolymers*. Phys. Rev. E. 64: 011904.

2. Melin, R. et al. 1999. *Designability, thermodynamic stability, and dynamics in the protein folding: A lattice model study.* J. Chem. Phys. 110: 1252-62.

3. Tiana, G. et al. 1998. *Folding and misfolding of designed protein like chains with mutations*. J. Chem. Phys. 108: 757-761

4. Broglia, R. A. Tiana, G. 2001. *Hierarchy events in the folding of model proteins*. J. Chem. Phys. 114: 7267-73.

5. Abkevich, V.I. Gutin, A. M. Shakhnovich, E. I. 1994. *Free energy landscape for folding kinetics: Intermediates, traps, and multiple pathways in theory and lattice models simulations.* J. Chem. Phys. 101: 6052-62.

6. Guebele, M. 1999. *The fast protein folding problem*. Annu. Rev. Phys. Chem. 50: 485-516.

7. Miyazawa, S. Jernigan, R.L. 1985. *Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures: Quasi-Chemical Approximation*. Macromolecules. 18: 534-552

8. van Mourik, J. et al. *Determination of interaction potentials of amino acid from native protein structures: Test on simple lattice models.* J. Chem. Phys. 110: 10123-33