Evolutionary Perspectives on Protein Thermodynamics

Richard A. Goldstein

Division of Mathematical Biology, National Institute for Medical Research, Mill Hill, London, NW7 1AA, UK
richard.goldstein@nimr.mrc.ac.uk

Abstract. While modern evolutionary theory has emphasized the role of neutral evolution, protein biochemistry and biophysics has interpreted the properties of proteins as largely resulting from adaptive evolution. We demonstrate that a number of these properties can be seen as emerging from neutral evolution acting on sequence entropy, that is, the fact that larger numbers of viable sequences have these properties. In this paper, we use a computational model of populations of evolving lattice proteins to describe how the observed marginal stability of proteins as well as their robustness to mutations can result from neutral evolution.

1 Introduction

Imagine a trained physicist from another world, who had never seen a computer before, was given one to analyze. She could disassemble it, making notes regarding the copper wires, the small chips of silicon with particularly-placed impurities, the ferro-magnetic material on a plastic substrate, etc. At some level, however, we would feel that she had somehow missed the essence of the computer, that understanding such an instrument would require a *functional* explanation including the role and purpose of the memory, cpu, and storage devices. This functional explanation is possible only because the computer has gone through a design process which has determined an appropriate form for its intended function. Conversely, an understanding of the functioning of the computer requires knowledge of the properties of doped semiconductors and ferro-magnetic materials, as these properties determined the constraints and opportunities given to the computer designer.

Similarly, an understanding of organisms requires combining a mechanistic description (the physico-chemical properties of organs, cells, proteins, DNA, membranes) with a functional description (the role and "purpose" of the heart, nucleus, histone, enzyme). As in the computer example, a purely mechanistic description would miss the essence of the biological subsystems by neglecting their functional roles, while the functional aspects cannot be understood independently of the constraints and opportunities presented by the mechanistic properties. This duality is again based on history, in this case the process of biological evolution. As in the computer example, evolution has been able to take

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 718–727, 2004.

advantage of the physico-chemical properties of the evolving systems, investing the resulting components with functional roles and purposes. In this instance, however, "design" would not be the appropriate term. Rather we must consider the evolutionary context; by analyzing biological systems in this context, we can work to unify these separate perspectives and understand how evolution utilizes, adopts, and changes the properties of the evolving elements while being constrained by these properties. This particular perspective leads us directly into the heart of what makes biological systems different from non-biological systems, why we can talk about the "purpose" of a lung in a way we cannot talk about the purpose of the electrons in an atom of carbon.

In this paper I focus on the properties of proteins and how they can be understood in an evolutionary context. One of the more interesting aspects of evolution is the separation between genotype and phenotype, between the molecules that are evolving (generally DNA) and the resulting traits that are acted on by evolutionary selection. Proteins can be seen as representing one of the lowest, simplest levels of phenotype, providing an important model for the evolution of higher organizational forms. Proteins are also interesting and important on their own. Various processes that proteins perform, such as folding, are of great theoretical interest – it is difficult to construct theoretical models of protein folding that can explain how such an enormous search problem is solved so quickly. Proteins are also involved in almost all functions in a living system, including respiration, signalling, replication, locomotion, transportation, etc., and are the basis of understanding these processes at a mechanistic and atomistic level. Proteins are the most common targets of pharmaceutical intervention and are thus intrinsically important for biomedical research. In addition, much work is proceeding trying to understand how to engineer proteins with modified or new properties and functions.

Evolution is a complicated procedure, particularly since exceptions to any general principle can lead to an attractive evolutionary niche. One of the more important axes for various models involves the distinction between adaptation and neutral evolution. It is clear that adaptation has had an important role in evolution, in making humans and other biological systems the way that they are. There are, in particular, episodes of adaptive evolution where change is clearly favoured. Some systems, such as pathogens avoiding a host immune response, are likely often undergoing adaptionist evolution. It is also clear that much, if not most, of molecular evolution is neutral in nature, that the vast majority of mutations that occur are either negative, slightly deleterious, or neutral, and that the chance acceptance of neutral or slightly deleterious mutations may often greatly exceed the smaller number of positive changes that might occur. Since the rise of the neutral theory in the late 1960s ([1,2]) much of evolutionary theory has been based on the emphasis of neutral evolution. In contrast, when confronted by the almost miraculous molecular properties of living systems, biochemists have generally thought in terms of adaptation and seen their characteristic traits as having arisen from the "survival of the fittest". Much less work has been performed analyzing specific proteins from a neutralist perspective.

Such a neutralist perspective is, however, important in understanding proteins for a number of reasons. Often it is possible to show that neutralism is sufficient to explain the observed properties. Because neutral evolution is always occurring, neutrality represents, when possible, the most parsimonious explanation. For this reason, adaptation is a reasonable explanation only when neutral evolution can be eliminated. Conversely, attempts to explain features based on neutral evolution can highlight when neutrality is *not* an adequate explanation, where an adaptive mechanism might be required. In addition, neutral evolution protects us against the so-called "Panglossian Paradigm" [3] where the current role of a feature is used to explain how and why the feature emerged in the way that Pangloss in Voltaire's Candide explains how we ended up with noses in order to support spectacles and legs in order to fit into trousers. Features that evolved based on one dynamic may end up being used for a completely different function. Finally, neutralist perspectives make us remember that evolution is decidedly *not* an optimization process. There are well-defined stochastic equations regarding the origin, fixation, and elimination of genetic variation, involving the fixation of deleterious mutations as well as the elimination of favourable mutations. Simplistic images that have been imposed on evolution, such as the afore-mentioned "survival of the fittest", may possibly represent more the projection of our psychological need to imagine ourselves at *some* peak of perfection rather than an inherent characteristic of the evolutionary process.

Another important aspect of the evolutionary process is the fact that evolution occurs in finite populations. The stochastic nature of the process results directly from this aspect. It also is important in elimination and fixation, as subpopulations that fall from one to zero can never recover. The stochastic nature of the evolutionary process does not mean that we cannot come up with general principles. Just as the thermodynamic notions of pressure, temperature, and heat represent the random motions of a large ensemble of particles, so we can generate principles based on the dynamics of populations. One important distinction, however, is the size of the populations. There are no populations close to Avogadro's number. For this reason the stochastic element can never be completely averaged away, and we are often left with tendencies and probabilities rather than fixed rules and laws.

The complicated nature of biomolecular evolution involving specifics of protein structure, function, thermodynamics, as well as population and evolutionary dynamics, makes this area attractive for theoretical simulations. The simulations, however, have to make numerous approximations and simplifications. In a number of publications we have described a simplified, lattice protein model which, combined with simulations of population evolution, have provided some interesting perspectives on why proteins are the way that they are. In this paper, we summarize and advance these perspectives, focusing on the thermodynamic and mutational properties of observed proteins including making connection with relevant experiments.

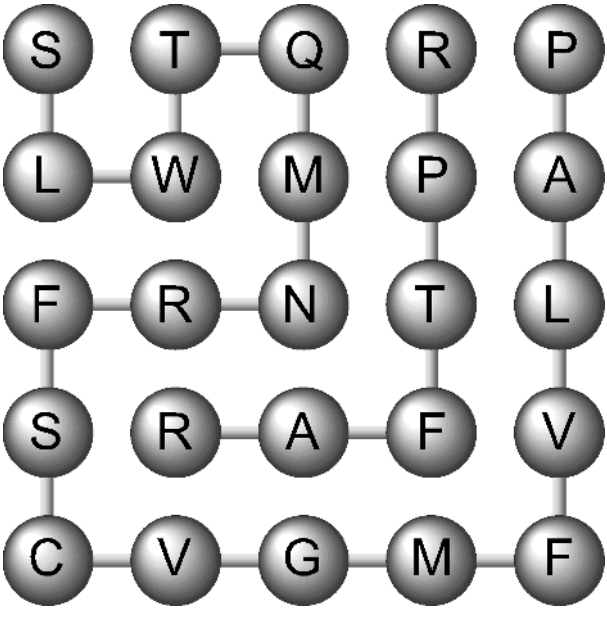

Fig. 1. Model of lattice protein

2 Models

2.1 Modelling the Evolving Proteins

The model of proteins is shown in Fig. 1. Proteins are represented as 16-monomer polypeptides forming a self-avoiding walk on a 5×5 square lattice. Each amino acid occupies exactly one lattice point. There are exactly 1081 possible conformations neglecting structures related by rotation, reflection, or inversion. While a two-dimensional lattice is highly inappropriate for folding simulations – the space of possible conformations is non-ergodic [4] – such a model allows us to have a reasonable ratio of buried to exposed residues with a moderately-sized protein.

The energy $E(S, C)$ of a given sequence S in any particular conformation C is a pairwise contact energy equal to

$$E(S, C) = \sum_{<i,j>} \gamma(A_i, A_j) u_{i,j} \tag{1}$$

where $\gamma(A_i, A_j)$ is the interaction energy between the amino acid at locations i and j in the sequence (such as between the serine and threonine in the upper-left corner of the protein in Fig. 1) and $u_{i,j}$ is equal to 1 if residues i and j are in contact (that is, are not covalently connected but are on neighbouring lattice points) and zero otherwise. The values of $\gamma(A_i, A_j)$ are taken from the contact energies derived by Miyazawa and Jernigan based on a statistical analysis

of the available protein database [5]. Because of the nature of the derivation, interactions with solvent, including entropic terms, are explicitly included in the contact energies.

We can assume that the conformation of lowest energy is the native-state C_{ns}, and compute the probability that a protein at equilibrium would be in this state

$$P_{\mathrm{ns}} = \frac{\exp(E(S, C_{\mathrm{ns}}/kT)}{\sum_C \exp(E(S, C)/kT)} \tag{2}$$

which allows us to compute $\Delta G_{\mathrm{folding}}$

$$\Delta G_{\mathrm{folding}} = -kT \log\left(\frac{P_{\mathrm{ns}}}{1 - P_{\mathrm{ns}}}\right) \tag{3}$$

2.2 Modelling Population Evolution

As mentioned before, the population dynamics are essential to any form of evolutionary simulation. In general, we start with a given population of protein sequences, initially identical. For each sequence we calculate the various thermodynamic quantities described above. We then choose a given number of sequence locations, chosen from an appropriate Poisson distribution, to change to another random amino acid. The thermodynamic quantities for the resulting sequences are again calculated. We then apply truncation selection where we decide on which sequences are to be considered viable based on whatever criteria we choose. Non-viable sequences are eliminated from the population, and the remaining sequences are chosen at random, with replacement, to form the next generation with the same population size.

3 Results

3.1 Protein Thermostability

It has been long observed that proteins are marginally stable with typical stabilities (-$\Delta G_{\mathrm{folding}}$) of approximately 10 kcal/mol, equal to a few hydrogen bonds. Two different classes of theories have been advanced for why this occurs, both adaptionist in nature. The first theory is that there is a fitness advantage to marginal stability. This might be for a number of different reasons. Protein functionality might require flexibility, which might be more common in marginally-stable proteins [6,7]. Marginal stability would weaken binding with ligands by requiring an entropy loss upon binding. This might make it easier to modulate binding affinities through mutation or post-translational modification [8,9,10]. Finally, there may be advantages to marginal stability in ensuring sufficiently rapid protein degradation. The second class of explanation revolves around the need for the protein to fulfil multiple selective criteria including functionality, stability, rigidity, etc. There would naturally be trade-offs in these criteria, so

that proteins can optimize stability given the constraints imposed by the other types of selective pressure.

Given the above discussion, it becomes important to investigate whether the observed marginal stability in proteins can result from neutral evolution. We modelled the population dynamics of 3000 sequences, allowing the dynamics to equilibrate for 30,000 generations and gathering data for an additional 30,000 generations [11]. At each generation, 0.2% of the sequences were mutated to an alternative residue. Proteins were considered viable if they were "adequately" stable, that is, with $\Delta G_{\text{folding}}$ less than some "critical" ΔG_{crit}. The result of these simulations were populations of proteins that were marginally stable, with $\Delta G_{\text{folding}} \approx \Delta G_{\text{crit}}$.

These results can be made intuitive if we consider the space of all possible sequences. This space is high-dimensional (as many dimensions as the length of the sequence) but extremely sparse (only 20 points along each dimension). The vast bulk of this space consists of proteins that would not fold nor be stable – unviable sequences in our model. There are regions in this space, characterized as hyperspheres, which contains sequences that are viable, that is, will fold into a stable, functional protein. It is a characteristic of high-dimensional spaces that the volume of objects in that space will be dominated by the periphery of that object. (99.95% of the volume of a 150-dimensional sphere is in the outermost 5%.) If the volume of sequence space consists of foldable, stable sequences, while the exterior of the volume consists of unfoldable, unstable sequences, the vast majority of the sequences in this volume will be marginally foldable and marginally stable, purely as a result of the high dimensionality. This will result whenever a) the objects in the space are roughly convex, and b) the fitness criteria are smoothly-varying in the space. If the vast majority of protein sequences are marginally stable, there is no problem explaining the observed marginal stability without resorting to adaptionist arguments, either selective pressure for marginal stability or optimization given constraints. Neutral evolution will be strongly affected by "sequence entropy", the number of sequences consistent with a given property. Sequence entropy will strongly drive protein sequences towards marginal stability.

What of the observation that modifying proteins to increase their thermostability sometimes results in decreased function [12,13,14,15]? Firstly, this is not necessarily always the case [16]. Secondly, if proteins evolve functionality in the context of a natural tendency towards marginal stability, it is not surprising to find mechanisms for functionality that are dependent upon, or at least consistent with, marginal stability. Taverna and Goldstein modelled this behaviour by considering competitive dynamics between three sets of lattice protein models, each with a different mechanism of action [11]. While the exact mechanism was unspecified, one set was modelled as requiring marginal stability, another set requiring moderate stability, while the third required high stability. A member of the first set in a marginally-stable protein had exactly the same fitness as a member of the second set in a moderately-stable form, which had exactly the same fitness as a member of the third set that was highly stable. Conversely,

any member of the first set that was not marginally stable, any member of the second set that was not moderately stable, or any member of the third set with other than high stability, was considered non-viable and was eliminated during the truncation selection. After separate equilibration, the three populations were allowed to compete against each other. In 24 out of the 25 runs, the mechanism consistent with marginal stability became the only form in the population, with the two other mechanisms being completely eliminated. (One run resulted in the domination of the mechanism requiring moderate stability, emphasizing the stochastic nature of the evolutionary process.) With the absolute equivalent fitness of these three populations, the entropic forces again combined with neutral evolution to result in proteins that required marginal stability. In the non-intuitive causality of evolution, marginal stability became required for proteins because they were marginally stable!

This is maybe less non-intuitive than it appears. Globular proteins generally require aqueous environments in order to fold and be stable and functional. One could argue that cells are generally aqueous because this environment is required for the globular proteins. This would correspond to the idea that proteins are marginally stable because this is required for protein function. In reality, the aqueous environment came first, and proteins evolved to function in this milieu. Proteins require aqueous environments to function because they evolved in aqueous environments and adapted themselves to this context. The concept that proteins adapted mechanism consistent to their context, a marginal stability induced by neutral evolution acting through sequence entropy, is not any more surprising.

This is not to say that marginal stability does not result from adaptation, only that the assumption of adaptation is not required to explain the observation. Thus neutral evolution is the most parsimonious explanation, and the observation of marginal stability does not provide any evidence for any selective pressure for marginal stability.

3.2 Proteins and Evolutionary Robustness

Another way to consider the role of sequence entropy is to consider the role of robustness in evolution. Fitness can be defined as the average expected number of viable offspring produced. Generally these offspring will be mutated forms of the parent, and so the fitness of an individual depends upon the fitness of the neighbouring genomes (sequences) in the genome (sequence) space. If mutations are more likely to lead to non-viable offspring, this reduces the fitness of the parent. If a protein depends upon being one of the few sequences with high stability, many more mutants will have reduced stability and thus would be non-viable. Conversely, if a protein has a mechanism consistent with marginal stability, the probability that a mutant would have marginal stability is much higher, ergo a higher fitness.

This robustness to mutations can be seen directly, in what has been described as “the survival of the flattest” [17]. It can be observed in the simulation

of the population evolution of lattice proteins, as described above. We can consider the results of the population evolution, and observe the consequence of random mutations. Lattice proteins that evolve based on truncation selection ($\Delta G_{\text{folding}} < \Delta G_{\text{crit}}$) are surprisingly robust to mutations, so that, on average, about half the protein sequences do not have reduced stability upon a random mutation – even when multiple residues are changed [18]. Sequences chosen at random, also subject to the same truncation selection, have almost no probability that a mutation would not be destabilizing. We can have two sequences with the same initial stability, same structure, same observed properties, one derived from population evolution, the other from being selected at random, yet the robustness to mutation is extremely different. In fact, the higher the stability requirement (corresponding to a more-negative ΔG_{crit}), the more likely a mutation will have negligible effect on the stability. Of all of the viable protein sequences, population evolution selects those networks of sequences that have the fewest non-viable neighbours. This perceived robustness of proteins to substitutions has been observed experimentally. For instance, Reddy et al. catalogued a wide range of mutations, observing that approximately 25% actually increased thermal stability [19].

4 Conclusion

Biochemists and molecular biologists have tended to imagine evolution as a constant march to higher and higher fitness levels, while modern evolutionary theory has increasingly emphasized randomness and neutral evolution. When confronted by a characteristic property of living systems, the response of biochemists has been to see the "blind watchmaker" at work, and to inquire how this property increases the fitness of the organisms. In reality, many of these properties can be explained by the process of neutral evolution acting on sequence entropy, taking advantage of the fact that many more sequences have some properties rather than others. In previous work, we have demonstrated how this can result in some structures much more common than others [20,21] (also see [22,23,24]), in proteins that fold into the state of lowest free energy [25], and in work described here, proteins that are marginally stable [11,26] and naturally robust to site mutations [18].

There are direct consequences of this perspective. For instance, the observed robustness of proteins is interpreted as explaining that the mapping of sequence to structure is rather robust and plastic. The alternative viewpoint provided here is that proteins have evolved to be robust to the particular experiments that are being performed. In general, the widespread attempts to understand the relationship between a protein's sequence and its resultant properties through site mutagenesis must take into the fact that proteins have evolved so that these mutations are less likely to change fundamental and important properties. Proteins have, to some extent, prepared for these changes. Finally, robustness is given as evidence why certain properties are not important for the protein. If these properties were under strong selective pressure, it is argued, they should be

"optimized" and therefore highly susceptible to mutational change [27]. Neutral evolution would suggest the opposite conclusion – highly important properties would be "buffered" against change during mutations.

Other conclusions relate to the possibilities and opportunities of protein engineering. If protein sequences are robust to change, this suggests that there are many options to modifying naturally-occurring sequences, in that the sequence can be highly altered while important properties such as stability and foldability will be maintained. Conversely, the evolutionary selection of highly robust sequences suggest that nature finds the flatter peaks in the fitness landscape, even if higher (but narrower) peaks exist elsewhere. This means that it might be possible to design proteins de novo with properties that exceed those found by natural evolution.

Acknowledgements. The work described here was largely performed by Sridhar Govindarajan, Darin Taverna, and Paul Williams. Computer support was provided by Kurt Hillig, Todd Raeker, and Michael Kitson. Funding was provided by NIH grant by NIH grant LM05770.

References

1. Kimura, M., Evolutionary rate at the molecular level. Nature (London) **217** (1968) 624–626
2. King, J.L., Jukes, T.H., Non-Darwinian evolution. Science **164** (1969) 788-798
3. Gould, S.J., Lewontin, R.C., The spandrels of San Marco and the Panglossian Paradigm: A critique of the adaptationist programme. Proc. Royal Soc. London, Series B **205** (1979) 581-598
4. Abkevich, V.I., Gutin, A.M., Shakhnovich, E.I., Impact of local and non-local interactions on thermodynamics and kinectics of protein folding. J. Mol. Biol. **252** (1995) 460-471
5. Miyazawa, S., Jernigan, R.L., Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. Macromol. **18** (1985) 534-552
6. Wagner, G., Wuthrich, K., Correlation between the amide proton exchange rates and the denaturation temperatures in globular proteins related to the basic pancreatic trypsin inhibitor. J. Mol. Biol. **130** (1979) 31–37
7. Tang, K.E.S., Dill, K.A., Native protein fluctuations: The conformational-motion temperature and the inverse correlation of protein flexibility with protein stability. J. Biomol. Struct. Dyn. **16** (1998) 397–411
8. Dunker, A.K., et al., Protein disorder and the evolution of molecular recognition: Theory, predictions and observations. Pacific Symp. Biocomputing **3** (1998) 473–484
9. Wright, P.E., Dyson, H.J., Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. J. Mol. Biol. **293** (1999) 321–331
10. Dunker, A.K., Obradovic, Z., The protein trinity–linking function and disorder. Nat. Biotechnol. **19** (2001) 805–806
11. Taverna, D.M., Goldstein, R.A., Why are proteins marginally stable? Proteins: Struct., Funct., Genet. **46** (2002) 105–109

12. Alber, T., Wozniak, J.A., A genetic screen for mutations that increase the thermal-stability of phage-T4 lysozyme. Proc. Natl. Acad. Sci. USA **82** (1985) 747–750
13. Bryan, P.N., et al., Proteases of enhanced stability: Characterization of a thermostable variant of subtilisin. Proteins: Struct. Funct. Genet. **1** (1986) 326–334
14. Liao, H., McKenzie, T., Hageman, R., Isolation of a thermostable enzyme variant by cloning and selection in a thermophile. Proc. Natl. Acad. Sci. USA **83** (1986) 576–580
15. Shoichet, B.K., et al., A relationship between protein stability and protein function. Proc. Nat. Acad. Sci. USA **92** (1995) 452-456
16. Giver, L., et al., Directed evolution of a thermostable esterase. Proc. Nat. Acad. Sci. USA **95** (1998) 12809-12813
17. Nimwegen, E.v., Crutchfield, J.P., Huynes, M., Neutral evolution of mutational robustness. Proc. Nat. Acad. Sci. USA **96** (1999) 9716-9720
18. Taverna, D.M., Goldstein, R.A., Why are proteins so robust to site mutations? J. Mol. Biol. **315** (2002) 479-484
19. Reddy, B.V.B., Datta, S., Tiwari, S., Use of propensities of amino acids to the local structure environment to understand effect of substitution mutations on protein stability. Protein Eng'g **11** (1998) 1137-1145
20. Govindarajan, S., Goldstein, R.A., Searching for foldable protein structures using optimized energy functions. Biopolymers **36** (1995) 43–51
21. Govindarajan, S., Goldstein, R.A., Why are some protein structures so common? Proc. Natl. Acad. Sci. USA **93** (1996) 3341–3345
22. Li, H., et al., Emergence of preferred structures in a simple model of protein folding. Science **273** (1996) 666-669
23. Shakhnovich, E.I., Protein design: a perspective from simple tractable models. Folding & Design **3** (1998) R45-R58
24. Finkelstein, A.V., Ptitsyn, O.B., Why do globular proteins fit the limited set of folding patterns. Prog. Biophys. Mol. Biol. **50** (1987) 171–190
25. Govindarajan, S., Goldstein, R.A., On the thermodynamic hypothesis of protein folding. Proc. Natl. Acad. Sci. USA **95** (1998) 5545–5549
26. Williams, P.D., Pollock, D.D., Goldstein, R.A., Evolution of functionality in lattice proteins. J. Mol. Graphics Modell. **19** (2001) 150–156
27. Kim, D.E., Gu, H., Baker, D., The sequences of small proteins are not extensively optimized For rapid folding by natural selection. Proc. Natl. Acad. Sci. USA **95** (1998) 4982-4986