Published in final edited form as:

Proteins. 2015 November ; 83(11): 1923–1928. doi:10.1002/prot.24916.

Nonlinearities in protein space limit the utility of informatics in protein biophysics

S. Rackovsky^{1,2,*}

¹Dept. of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853

²Dept. of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, New York, NY 10029

Abstract

We examine the utility of informatic-based methods in computational protein biophysics. To do so, we use newly developed metric functions to define completely independent sequence and structure spaces for a large database of proteins. By investigating the relationship between these spaces, we demonstrate quantitatively the limits of knowledge-based correlation between the sequences and structures of proteins. It is shown that there are well-defined, nonlinear regions of protein space in which dissimilar structures map onto similar sequences (the conformational switch), and dissimilar sequences map onto similar structures (remote homology). These nonlinearities are shown to be quite common- almost half the proteins in our database fall into one or the other of these two regions. They are not anomalies, but rather intrinsic properties of structural encoding in amino acid sequences. It follows that extreme care must be exercised in using bioinformatic data as a basis for computational structure prediction. The implications of these results for protein evolution are examined.

Keywords

Sequence space; Structure space; Fourier analysis; Conformational switches; Distant homology

Introduction

The protein folding problem is one of the outstanding problems in computational biophysics. Determination of the manner in which the three-dimensional structure of a protein is encoded in its one-dimensional amino acid sequence would lead to rapid and accurate computational prediction of protein structure. The structure is clearly encoded in interactions between atoms in the amino acids of the sequence, and between those atoms and those of the solvent or other environment in which a protein functions. The ultimate goal of computational protein folding, therefore, is to predict structure by minimizing the global free energy of the protein-solvent system as a function of protein conformational variables.

*srr87@cornell.edu.

Attainment of this goal has remained elusive, because of the high dimension and extreme complexity of the conformational energy surface. In an effort to simplify the search for a global minimum in this surface, information-based methods have been developed for analyzing the sequence-structure relationship, and hybrid structure prediction methods have been utilized which combine information- and physics-based approaches. While prediction using hybrid methods has been reasonably successful in many specific instances, two persistent phenomena suggest the existence of an underlying uncertainty as to the precise informatic determinants of protein architecture.

1. The well-known^{1, 2} “Remote Homolog Problem”: the fact that any reasonably large group of sequences which fold to a specified architecture will contain pairs of sequences which are not related by any known criterion.
2. The existence of “conformational switch” sequences^{3–5}, in which a single mutation leads to a change in the fold of the protein. These have also been extensively studied. A database of protein pairs with similar sequences and different structures has been compiled⁶, and the distributions of multiple conformational states of proteins examined⁷. Bioinformatic sequence comparison methods in current use predict, incorrectly, that these sequence pairs should have identical structures.

It is not known quantitatively how prevalent these two phenomena are, and whether they are artifacts arising from shortcomings in the conventional methods of sequence and structure comparison, or whether they are intrinsic features of proteins, which place fundamental limitations on the utility of informatic methods as an aid to physical approaches in structure prediction. The present work is designed to address these questions.

Methods

Our approach rests on the comparison of two completely independent protein spaces- a space of protein sequences and one of protein structures. The concept of a quantitatively defined sequence or structure space is not new^{8–20}. Exhaustive comparison of completely independent spaces, however, has only recently become possible²¹, with the development of new approaches to both sequence and structure comparison. A space is defined by its metric- the function which defines distances between points in the space. It is the near-universal practice in protein studies to measure intersequence distances using alignment-based methods, and interstructure distances using some form of superposition. In previous work²¹ we have discussed the basic assumptions and limitations of these viewpoints. Recent work²² has raised further questions as to the need for some of the evolutionary assumptions underlying current sequence comparison algorithms. It is a particular concern that neither alignment nor superposition can be used to quantitatively compare arbitrary protein pairs. Both depend on an ability to identify equivalent atoms or residues in two (or more) proteins of interest, which is not possible when the molecules differ substantially in chain length or sequence. As a result, not all pairwise distances can be calculated deterministically in a physically meaningful way. The metrics used herein do not suffer from this drawback. Sequence comparison is carried out using a Fourier based, Euclidean metric^{23–28}, acting on a complete, orthonormal numerical representation of amino acid physical properties^{29, 30}. An important characteristic of this metric is that, unlike alignment-based metrics, it is

completely free of structural information incorporated during an initial parametrization, which guarantees that the intersequence distances it measures are indeed structure-independent. The interstructure distance metric, which is also Euclidean, and based on the comparison of distributions of structure fragments of specified size, was used to carry out the first complete quantitative classification of protein structures and characterization of structure space^{8–9,10}. It was later independently shown³¹ to perform comparably to trusted, “gold standard” superposition-based structure comparison algorithms^{32, 33}, in instances where a superposition can be meaningfully defined, while being computationally far less demanding. The central points for the present work are that both the intersequence and interstructure metrics give physically meaningful comparisons between arbitrarily different molecules, and that they are completely independent.

The construction and characterization of these metrics have been discussed previously²¹. It was shown in that work that there is very high correlation ($R \sim 0.8$) between distances calculated using the two independent metrics, so that the distance $d(P, Q)$ between the *sequences* of two proteins P and Q is, in general, reflective of the distance $\delta(P, Q)$ between their *structures*. We ask here what the limits of this relationship are. We investigate this question by examining the environments of the proteins in a large database, in both sequence and structure space.

A more complete outline of the Fourier formalism, and methodological details of the present work, are given in the Supporting Material³⁴.

Results

It is desired to find a measure of the frequency with which either of two “anomalous” situations occurs:

1. Proteins with different sequences exhibit similar structures; or
2. Proteins with similar sequences exhibit different structures.

To this end, we examine the sequence environment of each protein in a very large database used in previous studies²¹. This database, which contains 12,011 proteins, is described in the Supplementary Material³⁴. The results of this calculation are shown in Figure 1. Each point in this graph describes the immediate environment in both sequence and structure space of one protein in the database. It is constructed by identifying the set $\{Y_i(P) | i=1, 2, \dots, M\}$ of M nearest neighbors in sequence space of each molecule P in the database, and calculating average sequence and structure distances between P and the members of $\{Y_i(P)\}$. We also calculate average sequence and structure distances between members of $\{Y_i(P)\}$. In Figure 1, we plot the averaged structure distance between P and its $M=20$ nearest sequence neighbors against the corresponding averaged sequence distance. Both variables are plotted in centered form (i.e., the variable X is rewritten as $X - \bar{X}$, where the overbar denotes a *global* average over the entire database), so that positive values are greater and negative values less than average. The two phenomena we are interested in can then be mapped to well-defined regions of the plot. The first (the remote homolog problem) is represented by points in the fourth (lower right) quadrant of the plot, in which nearest sequence neighbors of P have high averaged sequence distances $\langle SED(P) \rangle$ from P, but averaged structure

distances $\langle \text{STD}(\text{P}) \rangle$ are low. The second (the conformational switch) is represented by points in the second (upper left) quadrant, in which nearest sequence neighbors have low $\langle \text{SED}(\text{P}) \rangle$ values, but $\langle \text{STD}(\text{P}) \rangle$ values are high. The third (lower left) and first (upper right) quadrants together comprise a region of “normal”, linear behavior, where averaged sequence and structure distances are essentially monotonically related.

The plot scale in Figure 1 is highly compressed, because the points represent average values, acting as proxies for the complete set of M actual distances of each type associated with each protein. Averages vary more slowly than the individual points in the ensembles with which they are associated. The first question which must therefore be addressed is whether the points in this plot represent distinguishably different behaviors of sequences and their associated structures. In order to establish the statistical significance of the plot, an ANOVA (Analysis of Variance) study was carried out, which demonstrated that the points in the four quadrants of Figure 1 are drawn from different distributions with respect to both $\langle \text{SED}(\text{P}) \rangle$ and $\langle \text{STD}(\text{P}) \rangle$, with $p < 0.001$.

It is instructive to further characterize the protein environments represented by points in the fourth quadrant of Figure 1. We ask whether the larger average sequence distances here arise from sets of nearest neighbors whose sequences differ substantially from that of the central protein, but which are mutually similar. In order to investigate this point, we examine average sequence distances between members of $\{Y_i(\text{P})\}$. It is found that $\langle \text{SED}(\text{NN}) \rangle$, the averaged sequence distance between nearest sequence neighbors, is, on average, larger in this quadrant than in the third quadrant. ANOVA comparison of the complete distributions of $\langle \text{SED}(\text{NN}) \rangle$ in the two quadrants indicates that they differ from one another with extremely high statistical significance ($p \ll 0.0001$). Nearest sequence neighbors of proteins differ significantly more *from one another* in sequence space in the fourth quadrant than in the third quadrant, but fold to structures which are similar to that of the central protein. This suggests the possibility of degenerate folding mechanisms in the fourth quadrant, in which folding to similar structures proceeds through different pathways, dictated by differing sequences, in the various proteins in $\{Y_i(\text{P})\}$.

We now examine the proteins whose environments are represented by points in the second quadrant of Figure 1. Small sequence differences between these proteins and their nearest neighbors lead to relatively large structure differences. We demonstrate that actual changes in the fold of nearest neighbors are much more likely in this region, as a result of sequence changes, than in the fourth quadrant. In order to do so, we examine the distribution of the match number $v_Y(\text{P})$ - the number of times members of $\{Y_i(\text{P})\}$ fall into the same architectural class (the C Class of the CATH classification, as described in the Supporting Material) as the central protein P . Table I shows the average values of v_Y in the four quadrants. ANOVA analysis of the complete distributions of v_Y in the four quadrants confirms that they differ from one another with $p \ll 0.0001$. We then ask whether these changes in fold are uniform in character among the members of the nearest neighbor set, or whether they result in a range of different structures? To answer this question we examine the values of $\langle \text{STD}(\text{NN}) \rangle$, the averaged structure distance between sequence nearest neighbors. We find that the average of $\langle \text{STD}(\text{NN}) \rangle$ is larger in the second than in the third quadrant, and ANOVA comparison of the complete distributions of $\langle \text{STD}(\text{NN}) \rangle$ in the two

quadrants indicates once again that they differ with extremely high statistical significance ($p \ll 0.0001$). Nearest sequence neighbors differ *structurally* significantly more from one another, and from the central protein, in the second quadrant than in the third. Folding behavior of proteins in the second quadrant is sensitive to small sequence differences, and different folds are accessible as a result of alternative small changes in sequence characteristics. Small changes in sequence lead to large changes in architecture¹⁰. Molecules in this region seem to exhibit chaotic behavior in protein space.

We now ask how often these two phenomena occur. If they were actually anomalous, the second and fourth quadrants of Figure 1 would be sparsely occupied. It can be seen from Table II that this is not the case. Slightly more than half (52%) of the molecules are in the linear region. The distribution can be shown, however, not to be uniform. A chi-square test confirms that the observed distribution differs significantly from one with equal occupancy of the linear and “anomalous” regions, with $p \ll 0.005$. There is high occupancy, and also substantial asymmetry of occurrence, within the “anomalous” regions. There are significantly more proteins in the degenerate region (29% of the total) than in the chaotic region (19%). The observed distribution can be shown, by χ^2 criteria, to differ from uniformity with $p \ll 0.01$. This gives a quantitative estimate of the prevalence of the two phenomena. They are neither rare nor equally probable. The large populations of the second and fourth quadrants suggest that, in fact, they do not represent exceptional behavior at all, but rather reflect intrinsic features of the protein folding code, manifested in varying degrees by a large number of proteins. Some folds are encoded in a range of markedly different sequences, and attainable by multiple pathways, and some sequences are only marginally stable, with respect to fold, under perturbation by mutations.

We find very striking differences in the behavior, in this respect, of proteins of differing structural classes. We label the proteins in our data set by their values of C, the Class index used in the hierarchical structure classification of the CATH database^{35,36}. C=1 denotes all-helical proteins, C=2 denotes sheet/barrel structures, and C=3 denotes mixed helix-sheet/barrel structures. (Note that the three classes constitute 22%, 24% and 54% of the database, respectively.) Each of the three classes is distributed non-uniformly in the environment space of Figure 1, and we find, using χ^2 criteria, that the three distributions differ from one another with very high significance ($p \ll 0.0001$ in all cases). The actual distributions are shown in Tables III–V. A qualitative difference is found between the environments of mixed helix/sheet proteins and those of the other two classes. We observe that 88% of proteins with C=3 are found in the lower half-plane- the region in which $\langle \text{STD}(P) \rangle$ is less than average. The corresponding fractions are 20% for C=1 and 34% for C=2. Mixed helix-sheet/barrel proteins are far more likely to have nearest sequence neighbors which are structurally similar to themselves than are either helical or sheet/barrel structures. The population of the fourth quadrant, for proteins with C=3, is 43% of the total, while that of the second quadrant is 6% of the total. This balance is completely reversed for proteins with C=1 and C=2. The populations of the second quadrant are, respectively, 37% and 33% of the total, while those of the fourth quadrant are 10% and 14%, respectively. Mixed α/β structures are substantially less likely to act as conformational switches than either all- α or all- β proteins.

It should be asked whether the observed prevalence of the phenomena we consider is skewed by the large number of points near the origin of the space (Figure 1), which fall into one or another of the quadrants but might not exhibit “anomalous” properties to a significant degree. We address this question by subdividing the space radially into two regions, in which the distance ρ of the points from the origin is either less than or equal to $(\rho - \bar{\rho})$ or greater than $(\rho - \bar{\rho})$ average. This gives two subregions- inner and outer- within each quadrant, and we can ask whether the proteins which fall within the inner subregions exhibit the same statistically significant differences between quadrants as we observe for the entire dataset.

We have repeated the analysis outlined above on the set of inner subregions, and the results we obtain in this case are entirely consistent with those outlined above. We find that every statistically significant difference observed for the entire dataset is also found between the inner subregions, and remains statistically significant (with $p < 0.005$). We conclude that it is statistically meaningful to consider points within the inner region as contributing to the behavior we have observed, and that the prevalence of the phenomena we have considered is accurately counted.

It should be emphasized that the properties of protein space we observe are calibrated against averages over all permutations of the sequences in our dataset²⁴. They therefore represent intrinsic features of the specific sequences, and effects of database size are automatically accounted for.

These observations suggest that new folds- i.e., structures which differ significantly from their precursors- originate with higher probability from mutations in helical or sheet/barrel structures, whereas evolution of the sequences of mixed helix-sheet/barrel structures is more likely to produce structures of the same type. This picture of relative stability of mixed structures under evolutionary drift is consistent with the suggestion³⁷ that mixed folds are older than other folds, and with the observation³⁸ that they appear to be more stable thermodynamically than other folds.

Discussion

We have demonstrated the following points.

- Three regions of protein space can be defined. In the linear region, the relationship between (P, Q) and $\delta(P, Q)$ is orderly, in the sense that the average structural distance between a protein and its nearest neighbors increases with the average sequence distance between them.
- We also find two significant types of nonlinearity. In the degenerate region, an increase in average sequence distance between a protein and its nearest neighbors is not accompanied by a corresponding increase in the average structure distance.
- In the chaotic region, large average structure distances arise from small values of average sequence distance.

- We find that proteins of different structural classes are distributed very differently among these regions.
- It is likely that new folds originate more often from mutations in purely helical or sheet/barrel proteins than from those in mixed structures.

These observations point to fundamental limitations in knowledge-based approaches to protein structure prediction, and raise fundamental questions. It can not be reliably assumed that structures which are similar arise from sequences which are in any way related, or that even very similar sequences must invariably give rise to similar structures. Behavior which is not consonant with one or the other of those assumptions is manifested to some degree by almost half the proteins in our database. In fact, the resulting informatic uncertainty is implicated in the evolution of new folds, and the stability of others. Taken together with previous results demonstrating limitations of knowledge-based methods^{39–41}, they suggest that considerable care must be exercised in using informatic approaches in computational protein biophysics.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by grant GM-14312 from the National Institutes of Health, and grant MCB-10-19767 from the National Science Foundation.

References

1. Yang Y, Faraggi E, Zhao H, Zhou Y. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics*. 2011; 27:2076–2082. [PubMed: 21666270]
2. Ben-Hur A, Brutlag D. Remote homology detection: a motif based approach. *Bioinformatics*. 2003; 19(Suppl 1):i26–i33. [PubMed: 12855434]
3. Alexander PA, He Y, Chen Y, Orban J, Bryan PN. A minimal sequence code for switching protein structure and function. *Proc Nat Acad Sci USA*. 2009; 106:21149–21154. [PubMed: 19923431]
4. Anderson TA, Cordes MH, Sauer RT. Sequence Determinants of a conformational switch in a protein structure. *Proc Nat Acad Sci USA*. 2005; 102:18344–18349. [PubMed: 16344489]
5. Roessler CG, et al. Transitive homology-guided structural studies lead to discovery of Cro proteins with 40% sequence identity but different folds. *Proc Nat Acad Sci USA*. 2008; 105:2343–2348. [PubMed: 18227506]
6. Kosloff M, Kolodny R. Sequence-similar, structure-dissimilar protein pairs in the PDB. *Proteins: Structure, Function, Bioinformatics*. 2008; 71:891–902.
7. Burra PV, Zhang Y, Godzik A, Stec B. Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure. *Proc Nat Acad Sci USA*. 2009; 106:10505–10510. [PubMed: 19553204]
8. Rackovsky S, Goldstein DA. Protein comparison and classification: A differential geometric approach. *Proc Nat Acad Sci USA*. 1988; 85:777–781. [PubMed: 3422460]
9. Rackovsky S. Quantitative classification of the known protein x-ray structures. *Polymer Preprints*. 1990; 31:205.
10. Rackovsky S. Quantitative organization of the known protein x-ray structures. I. Methods and short-length-scale results. *Proteins: Structure, Function & Genetics*. 1990; 7:378–402.

11. Kolodny R, Pereyaslavets L, Samson AO, Levitt M. On the universe of protein folds. *Ann Rev Biophys.* 2013; 42:559–582. [PubMed: 23527781]
12. Taylor WR. Evolutionary transitions in protein fold space. *Curr Opin Struct Biol.* 2007; 17:354–361. [PubMed: 17580115]
13. Kolodny R, Petrey D, Honig B. Protein structure comparison: Implications for the nature of ‘fold space’ and structure and function prediction. *Curr Opin Struct Biol.* 2006; 16:393–398. [PubMed: 16678402]
14. Pascual-Garcia A, Abia D, Ortiz AR, Bastolla U. Cross-over between discrete and continuous protein structure space: Insights into automatic classification and networks of protein structures. *PLOS Comput Biol.* 2009; 5:e1000331. [PubMed: 19325884]
15. Sadreyev RI, Kim B-H, Grishin N. Discrete-continuous duality of protein structure space. *Curr Opin Struct Biol.* 2009; 19:321–328. [PubMed: 19482467]
16. Sadowski MI, Taylor WR. On the evolutionary origins of “fold-space continuity”: A study of topological convergence and divergence in mixed alpha-beta domains. *J Struct Biol.* 2010; 172:244–252. [PubMed: 20691788]
17. Harrison A, Pearl F, Mott R, Thornton J, Orengo C. Quantifying the similarities within fold space. *J Mol Biol.* 2002; 323:909–926. [PubMed: 12417203]
18. Valas RE, Yang S, Bourne PE. Nothing about protein structure classification makes sense except in the light of evolution. *Curr Opin Struct Biol.* 2009; 19:329–334. [PubMed: 19394812]
19. Nepomnyachiy S, Ben-Tal N, Kolodny R. Global view of the protein universe. *Proc Nat Acad Sci USA.* 2014; 111:11691–11696. [PubMed: 25071170]
20. Hou J, Sims GE, Zhang C, Kim S-H. A global representation of the protein fold space. *Proc Nat Acad Sci USA.* 2003; 100:2386–2390. [PubMed: 12606708]
21. Scheraga HA, Rackovsky S. Homolog detection using global sequence properties suggests an alternate view of structural encoding in protein sequences. *Proc Nat Acad Sci USA.* 2014; 111:5225–5229. [PubMed: 24706836]
22. Skolnick J, Gao M. Interplay of physics and evolution in the likely origin of protein biochemical function. *Proc Nat Acad Sci USA.* 2013; 110:9344–9349. [PubMed: 23690621]
23. Rackovsky S. “Hidden” sequence periodicities and protein architecture. *Proc Nat Acad Sci USA.* 1998; 95:8580–8584. [PubMed: 9671720]
24. Rackovsky S. Characterization of architecture signals in proteins. *J Phys Chem B.* 2006; 110:18771–18778. [PubMed: 16986867]
25. Rackovsky S. Sequence physical properties encode the global organization of protein structure space. *Proc Nat Acad Sci USA.* 2010; 106:14345–14348. [PubMed: 19706520]
26. Rackovsky S. Global characteristics of protein sequences and their implications. *Proc Nat Acad Sci USA.* 2010; 107:8623–8626. [PubMed: 20421501]
27. Rackovsky S. Spectral analysis of a protein conformational switch. *Phys Rev Letters.* 2011; 106:248101.
28. Rackovsky S. Sequence determinants of protein architecture. *Proteins: Structure, Function & Bioinformatics.* 2013; 81:1681–1685.
29. Kidera A, Konishi Y, Oka M, Ooi T, Scheraga HA. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J Prot Chem.* 1985; 4:23–55.
30. Kidera A, Konishi Y, Ooi T, Scheraga HA. Relation between sequence similarity and structural similarity in proteins: role of important properties of amino acids. *J Prot Chem.* 1985; 4:265–297.
31. Budowski-Tal I, Nov Y, Kolodny R. FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately. *Proc Nat Acad Sci USA.* 2010; 107:3481–3486. [PubMed: 20133727]
32. Subbiah S, Laurents DV, Levitt M. Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. *Curr Biol.* 1993; 3:141–148. [PubMed: 15335781]
33. Shindyalov IN, Bourne PE. Protein structure alignment by combinatorial extension (CE) of the optimal path. *Protein Eng.* 1998; 11:739–747. [PubMed: 9796821]
34. See Supplementary Material for a detailed discussion of methods.

35. Orengo CA, et al. CATH- A heirarchic classification of protein domain structures. *Structure*. 1997; 5:1093–1108. [PubMed: 9309224]
36. www.cathdb.info
37. Winstanley HF, Abeln S, Deane CM. How old is your fold? *Bioinformatics*. 2005; 21S1:i449–i458. [PubMed: 15961490]
38. Minary P, Levitt M. Probing protein fold space with a simplified model. *J Mol Biol*. 2008; 375:920–933. [PubMed: 18054792]
39. Rackovsky S. On the nature of the protein folding code. *Proc Nat Acad Sci USA*. 1993; 90:644–648. [PubMed: 8421700]
40. Solis AD, Rackovsky S. On the use of secondary structure in protein structure prediction: a bioinformatic analysis. *Polymer*. 2004; 45:525–546.
41. Solis AD, Rackovsky S. Property-based sequence representations do not adequately encode local protein folding information. *Proteins: Structure, Function & Bioinformatics*. 2007; 67:785–788.

Figure 1.

The environment space for the 12011 proteins in our database. The average structure distance between a protein P and its 20 nearest sequence neighbors is plotted against the corresponding average sequence distance. Variables are shown as centered values, $\bar{X} - \bar{X}$, where the overbar denotes a global average over all the proteins of the dataset. Positive values are therefore greater than average, and negative values are less than average.

Table I

Average values of v_Y in the four quadrants of Figure 1

6.05	6.0
11.2	10.8

Table II

Populations of the four quadrants of Figure 1, all proteins

2321	2482
3765	3443

Table III

Populations of the four quadrants of Figure 1, C=1

993	1138
278	263

Table IV

Populations of the four quadrants of Figure 1, C=2

961	950
579	339

Table V

Populations of the four quadrants of Figure 1, C=3

367	393
2908	2776