Spectral Analysis of a Protein Conformational Switch

S. Rackovsky*

Department of Pharmacology and Systems Therapeutics, Mount Sinai School of Medicine of NYU, One Gustave L. Levy Place, New York, New York 10029
and Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853, USA

(Received 12 January 2011; published 14 June 2011)

The existence of conformational switching in proteins, induced by single amino acid mutations, presents an important challenge to our understanding of the physics of protein folding. Sequence-local methods, commonly used to detect structural homology, are incapable of accounting for this phenomenon. We examine a set of proteins, derived from the G_A and G_B domains of *Streptococcus* protein *G*, which are known to show a dramatic conformational change as a result of single-residue replacement. It is shown that these sequences, which are almost identical locally, can have very different global patterns of physical properties. These differences are consistent with the observed complete change in conformation. These results suggest that sequence-local methods for identifying structural homology can be misleading. They point to the importance of global sequence analysis in understanding sequence-structure relationships.

DOI: 10.1103/PhysRevLett.106.248101 PACS numbers: 87.15.ad, 87.15.Cc

Proteins fold into well-defined three-dimensional structures under the influence of interactions encoded in their amino acid sequences. It is well known that the folds which proteins assume can be divided into discrete, and often dissimilar, classes. The origin of the diversity of these classes is one of the most important problems in current biophysics. We suggested some time ago [1], on purely structural grounds, that, under favorable circumstances, new folds may arise as a result of one or a small number of sequence mutations. This idea poses a clear, important challenge to the conventional understanding of the physics of protein folding, since it requires that sequence pairs with near-total identity fold to completely different architectures.

A number of experimental studies [2–4] have since demonstrated this process by the systematic stepwise mutation of protein sequences. A particularly intriguing conformational switch was recently demonstrated by Alexander *et al.* [5]. This switch occurs between sequence pairs with 77%–98% identity, and provides a well-characterized, experimentally demonstrated set of false positives, for example, for conventional sequence-based structural homology searches. These supplement the well-known, large and very important body of false negatives known collectively as the ''remote homology problem'', which arises from the fact that any reasonably large set of sequences known to fold to a given architecture will contain pairs unrelated by any generally accepted conventional, alignment-based criterion.

Alexander *et al.*, studied a set of 56 residue sequences derived from the G_A and G_B domains of *Streptococcus* protein *G*, and demonstrated that a $4\beta + \alpha$ *IgG*-binding fold is transformed into a 3α albumin-binding fold by a single amino acid mutation. The authors identified a minimal switch pathway for this process, populated by a set of 10 sequences, all of which exhibit at least 77% pairwise identity. It was shown experimentally that five of these sequences assume the $4\beta + \alpha$ fold and five adopt the 3α fold, and that on either side of the single-mutation switch point the native conformation is at least 90% populated. The protein is monomeric, unlike other conformational switches studied experimentally, so that the influence of intermolecular interactions on the switching process can be discounted.

Lattman and Rose [6] have argued on general principles that the signals responsible for protein folding are not local in nature, but rather distributed throughout the entire sequence of the molecule. This suggests that a local, alignment-focused view of folding physics is not always appropriate—a suggestion strongly supported by the existence of the $4\beta + \alpha/3\alpha$ and other conformational switches. In the present work, a quantitative, whole-molecule approach to this important biophysical phenomenon will be presented, and it will be shown that it provides a foundation for understanding the physics of conformational switching.

In a number of recent studies [7–10], we have developed a new approach to the problem of sequence-structure relationships in proteins. This approach is based on the idea that representation of the sequence of a protein by parameters which contain information about the *entire* molecule, rather than parameters which characterize individual amino acids, will make visible properties of the sequence which are relevant to the folding process, and which are not otherwise readily apparent. Following methods set out in previous work, we represent the sequence numerically using 10 property factors [11,12], which form an orthonormal and essentially complete basis set for all the known physical properties of the amino acids, to represent an amino acid as a 10-vector. (The factors are identified in Table I.) A complete protein sequence is thereby transformed into a set of 10 *N*-member numerical strings, each

TABLE I. The Kidera property factors[a]

1. Helix/bend preference	6. Partial specific volume
2. Side-chain size	7. Flat extended preference
3. Extended structure preference	8. Occurrence in alpha region
4. Hydrophobicity	9. *pK-C*
5. Double-bend preference	10. Surrounding hydrophobicity

[a]The first four factors are essentially pure physical properties; the remaining six factors are superpositions of several physical properties, and are labeled for convenience by the name of the most heavily weighted component.

of which records the course of one property factor along the N-residue sequence. These strings can be Fourier transformed, which results in a representation of the sequence by a set of sine and cosine Fourier coefficients. Each of these coefficients, which is labeled by a wave number k and a property identifier l, encodes information about the entire sequence of the protein. The information encoded at particular k and l values is linearly independent of that encoded at all other values, by construction [7,8,11], since the Fourier decomposition of a sequence is complete and orthonormal with respect to both physical properties and wave number. Furthermore, the Fourier components are determined by information associated with different intrinsic length scales in the sequence- coefficients with wave number k contain information about structural features of size $\sim N/k$. This approach provides a method for systematically studying the presence in each property factor of features on specific scales, and for doing so in a uniform manner in sequences of varying lengths. It shares a global view of sequence properties with Fourier and other periodicity-based approaches previously proposed by a number of workers [13–21]. It differs, however, in that most of those studies have used those tools to examine the role of hydrophobicity and the implications of hydrophobicity patterns in protein sequences. We examine the roles of all sequence properties simultaneously.

In recent work we have shown [9] that the $k = 0$ Fourier coefficients contain sequence information which correctly encodes the *structural* relationships between proteins. We have also demonstrated [10] that different physical properties have distinctly different behaviors as a function of k, and that these differences in k dependence likely induce different folding mechanisms.

We examine the global properties of sequence sets shown by Alexander *et al.* [5] to lie on the key mutation pathway between the 3α and $4\beta + \alpha$ conformations. This includes 5 sequences which adopt the 3α conformation (which we denote as set A) and 5 which adopt the $4\beta + \alpha$ conformation (set B). (All sequences considered in this work are given in Tables 1 and 2 of the supplemental material [22].) Our purpose is to determine whether sequence characteristics can be identified which correlate with the remarkable difference in fold between the two sets of near-identical sequences.

The procedure is straightforward. There are 5 sequences in set A, and 5 in set B. We represent those sequences using the property factors, and Fourier transform the 10 resulting numerical strings. We are interested in the magnitudes of the Fourier coefficients, so we consider the sine and cosine Fourier power spectra [7,8], whose elements are given by the squares of the corresponding Fourier coefficients. In each case we need to determine whether the observed magnitude of the power spectral element is significantly different from that which would be observed on a purely random basis. We therefore compare the computed spectral elements to the value which would result from averaging over all possible sequence permutations, normalized by the corresponding standard deviation. This, of course, is the standard Z function

$$Z([a_k^l]^2) = \frac{[a_k^l]^2 - \langle [a_k^l]^2 \rangle_N}{\sigma([a_k^l]^2)}. \quad (1)$$

Here $a_k^{(l)}$ is the sine or cosine Fourier coefficient with wave number k for property l (where $1 \leq l \leq 10$), the subscripted brackets denote an average over all possible permutations of the sequence, and $\boldsymbol{\sigma}$ denotes the standard deviation of the power spectral element over the ensemble of all possible sequence permutations. By measuring the value of the power spectral element relative to the expected value over this ensemble, we determine the contribution of the specific sequence to the power spectrum, beyond that provided by sequence amino acid composition. We are greatly aided in this undertaking by the fact [8] that the average and standard deviation in this equation can be calculated analytically and exactly. We define a *signal* in the power spectrum by the equation $Z([a_k^l]^2) \geq 2.0$. This is the standard criterion for a power spectral element to be larger than average at the 5% confidence level.

Having calculated Fourier power spectra for all of the sequences, we seek those values of k and l at which spectral signals are observed with high frequency. We are particularly interested in the differences between the proteins of sets A and B. We therefore calculate values of the difference function $\Upsilon(k, l) = \eta_A(k, l) - \eta_B(k, l)$, where $\eta_X(k, l)$ is the number of signals occurring at wave number k in property l in the sequences of set X. This function can take integer values between -5 and 5, because there are 5 sequences in each set. Not all values of Y correspond to statistically significant differences between the two sets. The statistical significance of Y values can be determined by realizing that the function $\eta_X(k, l)$ describes a Bernoulli process. Standard methods then give a measure of statistical significance for Y, in the form, once again, of a Z function, which we denote $Z(Y)$. This enables us to identify those k and l at which there are statistically significant differences in global sequence properties between set A, which consists of sequences which fold to the 3α conformation, and set B, comprising sequences which adopt the $4\beta + \alpha$ fold.

The results of this calculation are very striking. We find that the very small local differences between the sequences

in set A and those in set B lead to large changes in the global organization of the sequences. This observation is summarized in Fig. 1, in which we plot values of $Z(Y)$ arising from statistically significant differences between the sequences in the two sets. Positive values of $Z(Y)$ occur at values of k and l at which the proteins of set A have significantly more signals than those of set B, and negative values identify values at which this situation is reversed. It will be seen that the proteins of set B- the $4\beta + \alpha$ sequences- are characterized particularly by the presence of signals at low values of k, which are absent in set A. Conversely, the 3α sequences of set A show a rich spectral structure at higher values of k which is nearly absent in set B.

This empirical observation can be made quantitative. We ask whether there is a statistically significant difference between the regions with $k < 20$ and $k \geq 20$ in Fig. 1. Straightforward application of chi-square methods indicates that distributions of Y in the two regions differ with very high significance ($p \ll 0.005$). Considered individually, the $k < 20$ region is statistically indistinguishable from random, while the $k \geq 20$ region differs from a random distribution with $p \ll 0.005$.

The importance of this difference is dramatically emphasized when we characterize $Z(Y)$ not by the wave number k, which is the natural Fourier variable, but rather by the length scale $\xi = N/k$ in sequence space which is interrogated by a signal with wave number k. The most significant single difference arising from a signal unique to set B occurs at $k = 3$ ($\xi \sim 18$ residues), while the most significant difference arising from a signal unique to set A occurs at $k = 20$ ($\xi \sim 2.8$ residues). The unique signals of set A arise from property variations which occur within very short lengths along the chain- on a scale of 2–4 residues. The unique signals of set B, on the other hand, almost all arise from property variations on scales of ~8–56 residues.

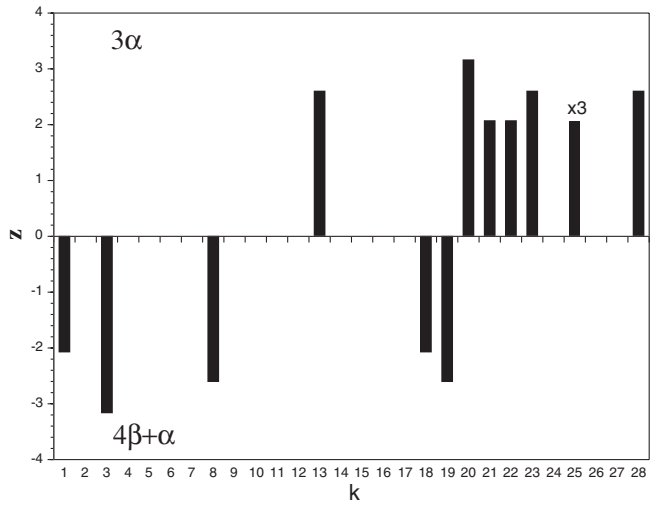

FIG. 1. The function $Z(Y)$ which measures the statistical significance of the spectral difference function Y between sets A and B (see text). Only values which indicate differences significant at the 5% confidence level are plotted. The bar at $k = 25$ is actually a triplet of bars, arising from different property factors.

In order to investigate the generality of this finding, we carried out a similar analysis of 15 more sequences, also studied experimentally by Alexander *et al.* [5], which differ from some of those in the minimal switch set at only a few sites, and were shown to adopt one or the other of the two conformations, but which are *not* on the minimal switch pathway. Of these, 7 adopt the 3α conformation, and 8 the $4\beta + \alpha$ conformation. We denote the former group of sequences as set C, and the latter as set D. In Fig. 2 we plot the number of statistically significant values of $Z(Y)$ observed, as a function of k. Two facts are apparent from this plot: (i) sets C and D both exhibit significant spectral activity at lower k values. This contrasts with the divergent low-k behaviors of sets A and B. (ii) The behaviors of sets C and D at higher k values are very different, directly paralleling the behavior of the on-pathway sets A and B.

Once again it can be shown that distributions of Y in the regions with $k < 20$ and $k \geq 20$ in Fig. 2 differ with very high statistical significance ($p \ll 0.005$), and that the $k \geq 20$ region differs from a random distribution with $p \ll 0.05$.

These observations suggest that an essential difference between sequences which adopt the 3α structure and those which adopt the $4\beta + \alpha$ structure lies in their spectral properties at higher k values. Sequences which fold to the 3α structure, whether on or off the minimal switch pathway, have a crowded difference spectrum in the interval $20 \leq k \leq 28$, while sequences which fold to the $4\beta + \alpha$ structure have little or no difference-spectral

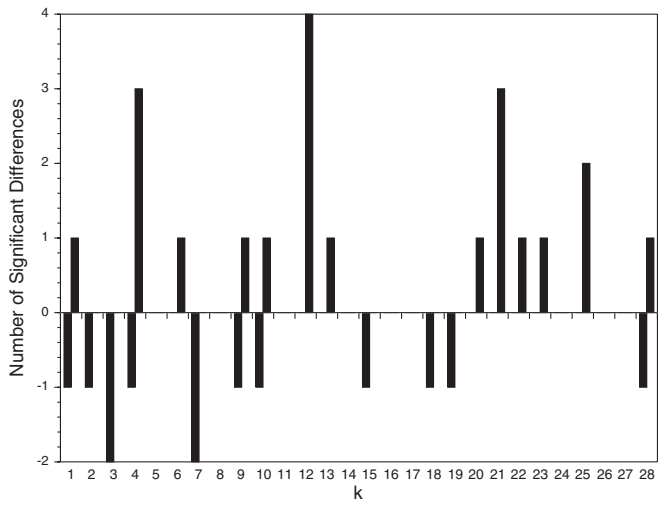

FIG. 2. Spectral difference plot for sets C and D (see text). The number of significant differences between the spectra of the two sets are shown, as a function of k. Positive values are recorded when spectra of set C exhibit significant signals for which there is no equivalent in set D. Negative values indicate the converse situation. The presence of both positive and negative signals at a given k value occurs due to the presence of signals in each set in different property factors.

strength in this range. This indicates the presence in the 3α sequences of a significant number of signals which have no counterpart in the $4\beta + \alpha$ sequences.

A signal in one of the 10 property factors indicates the presence in the sequence of a pattern of strong physical interactions, which is likely to have a determinative effect on the folding pathway. Signals at low k arise from patterns in which relatively large localized clusters of interacting residues are well separated from one another along the sequence. Such interaction patterns are expected to lead to folding by a nucleation mechanism, in which the formation of local folded regions precedes the global self-organization of the molecule. On the other hand, signals at high k arise from patterns in which small interacting regions are closely adjacent over the length of the sequence. These interactions are expected to give rise to nonlocalized collective modes, which lead to folding by ''collapse''-like pathways.

In the present context, it is to be expected that a dramatic change in spectral properties, as the molecule moves along the minimal switch pathway in sequence space, will be accompanied by a change in folding pathway, and that the altered pathway will lead to a different fold. It would appear that the signals observed at lower k values in the off-pathway sequences of set C arise from interactions which stabilize the 3α conformation. As the molecules drift in sequence space toward the switch pathway, those interactions are eliminated, leaving sequences which are marginally stabilized by the interactions which give rise to the high-k spectrum of set A. Further sequence drift eliminates those interactions, causing the conformation to fold, by a different basic mechanism, to the $4\beta + \alpha$ structure. This viewpoint is fully consistent with the stability (T_M) values given by Alexander *et al.* [5].

We can ask which specific property factors are responsible for the signals observed in the two sets of sequences. These data are given in Tables 3 and 4 of the supplemental material [22]. In previous work [10] we have shown that the property factors naturally cluster into two classes on the basis of their k-dependent behavior. One set, denoted by C_1, is a set of 5 property factors most associated with helix or bend formation. The complementary set, C_2, includes factors primarily associated with extended structure formation. We find that, in both comparisons described above (set A/set B and set C/set D) there is no statistically significant difference between the properties in which signals occur in the region $k < 20$. However, in both cases, the signals which are observed for $k \geq 20$ are expressed with high statistical significance ($p < 0.05$) by the properties in C_1 rather than C_2. This is entirely consistent with the switch to a 3α conformation, and reinforces the impression that the sequence signals encode the switch mechanism.

The information which gives rise to these observations is not embodied in sequence-local treatments of sequence-structure relationships in proteins [23,24]. The physical processes which direct the folding of proteins are nonlocal in nature, and information about their operation is only accessible through nonlocal analysis.

We have shown that protein sequences which are locally almost identical can have global physical patterns which are very different. This demonstration is made possible by the ability to construct complete, global sequence representations, and to calculate their statistical properties exactly. The data of Alexander *et al.* [5] emphasize the fact that local methods of sequence comparison are not always a reliable guide, even in cases of extreme similarity, to the presence or absence of structural homology. The present results demonstrate that a more global view of sequence characteristics is required for the complete understanding of sequence or structure relationships in proteins.

The author is grateful to Professor Philip Bryan for an important communication.

*Shalom.Rackovsky@mssm.edu

[1] S. Rackovsky, Proteins **7**, 378 (1990).
[2] X. I. Ambroggio and B. Kuhlman, Curr. Opin. Struct. Biol. **16**, 525 (2006).
[3] C. G. Roessler *et al.*, Proc. Natl. Acad. Sci. U.S.A. **105**, 2343 (2008).
[4] T. A. Anderson, M. H. J. Cordes, and R. T. Sauer, Proc. Natl. Acad. Sci. U.S.A. **102**, 18 344 (2005).
[5] P. A. Alexander *et al.*, Proc. Natl. Acad. Sci. U.S.A. **106**, 21 149 (2009).
[6] E. A. Lattman and G. Rose, Proc. Natl. Acad. Sci. U.S.A. **90**, 439 (1993).
[7] S. Rackovsky, Proc. Natl. Acad. Sci. U.S.A. **95**, 8580 (1998).
[8] S. Rackovsky, J. Phys. Chem. B **110**, 18 771 (2006).
[9] S. Rackovsky, Proc. Natl. Acad. Sci. U.S.A. **106**, 14 345 (2009).
[10] S. Rackovsky, Proc. Natl. Acad. Sci. U.S.A. **107**, 8623 (2010).
[11] A. Kidera *et al.*, J. Prot. Chem. **4**, 23 (1985).
[12] A. Kidera *et al.*, J. Prot. Chem. **4**, 265 (1985).
[13] D. Eisenberg *et al.*, Proc. Natl. Acad. Sci. U.S.A. **81**, 140 (1984).
[14] H. Xiong *et al.*, Proc. Natl. Acad. Sci. U.S.A. **92**, 6349 (1995).
[15] M. W. West and M. H. Hecht, Protein Sci. **4**, 2032 (1995).
[16] B. M. Broome and M. H. Hecht, J. Mol. Biol. **296**, 961 (2000).
[17] J. L. Cornette *et al.*, J. Mol. Biol. **195**, 659 (1987).
[18] K. B. Murray, D. Gorse, and J. M. Thornton, J. Mol. Biol. **316**, 341 (2002).
[19] A. Giuliani *et al.*, Chem. Rev. **102**, 1471 (2002).
[20] K. A. Selz *et al.*, Biopolymers **85**, 38 (2007).
[21] M. Colafranceschi *et al.*, OMICS **14**, 275 (2010).
[22] See supplemental material at http://link.aps.org/supplemental/10.1103/PhysRevLett.106.248101.
[23] S. Rackovsky, Proc. Natl. Acad. Sci. U.S.A. **90**, 644 (1993).
[24] A. D. Solis and S. Rackovsky, Polymer **45**, 525 (2004).