Published in final edited form as:

Biopolymers. 2021 October ; 112(10): e23411. doi:10.1002/bip.23411.

Dynamic and Conformational Switching in Proteins

H.A. Scheraga^{1,*}, S. Rackovsky^{1,2,†}

¹Department of Chemistry and Chemical Biology, Baker Laboratory, Cornell University Ithaca, NY 14853

²Department of Biochemistry and Biophysics University of Rochester School of Medicine and Dentistry Rochester, NY 14642

Abstract

Using bioinformatic methods for treating protein dynamics, developed in earlier work, we study the relationship between sequence mobility and dynamics in proteins. It is shown that sequence mobility drives a transition between two dynamic regimes in proteins, and that the specific details of this transition differ qualitatively between α -helical proteins and those in other structural classes. We examine the possibility that conformational switching is related to dynamic switching, by considering a specific system of sequences which exhibit the switching phenomenon. It is shown that a relationship between dynamic and conformational switching is entirely plausible.

Introduction

The existence of distinct classes of folded structures has been known since the earliest days of protein science. The early qualitative understanding of this phenomenon was summarized by Richardson, in a comprehensive study¹. A question of central importance arises from this observation- how did diverse fold types arise in the course of biological diversification? Are different folds related to one another, or did they appear as the end result of completely separate development pathways? In the course of the first quantitative classification of protein structures² we suggested, on the basis of structural considerations, that new folds may arise as a result of one, or a very few, strategically located sequence mutations in an existing structure. Experimental studies^{3,4,5,6} have demonstrated the occurrence of this process in a number of systems.

It should be emphasized that the occurrence of fold switching is a significant challenge to conventional methods of protein fold prediction, since it demonstrates that sequences with near-total identity can fold to completely different architectures. We have studied this phenomenon⁷ using Fourier methods, which describe the organization of the physical

[†] srr87@cornell.edu .

^{*} Deceased

Data Availability

Data generated in this study have not been placed in a public repository. They are available from the authors by request.

Conflict of Interest Statement

The authors state that they have no conflicts of interest in connection with this work.

properties of proteins at the whole-sequence level, rather than the organization of evolutionary similarity parameters on a local length scale.

It is well understood that the dynamics of proteins are central to both folding and biological function^{8,9}. It seems clear that changes in dynamic behavior must also be associated with fold switching, and it is easy to imagine that, in some instances, transitions in dynamics may actually drive the fold switching process. The behavior of protein dynamics as a function of sequence properties, a topic of both intrinsic and practical interest, is therefore also of specific importance in connection with the study of large-scale conformational changes in proteins. Until recently, however, technical difficulties have made this area inaccessible. Simulation methods, which are the current standard for the computational study of protein dynamics, are extremely processor-intensive, and it is generally possible, in the course of an investigation, to simulate only one protein, or a small set of proteins of interest. In recent work^{10,11}, we have developed alternative, bioinformatic methods for the study of protein dynamics. These methods make it possible to study the dynamic characteristics of large datasets of proteins simultaneously, and to draw general conclusions based on those studies. The computational demands of the bioinformatic approach are orders of magnitude less than those of simulation, and the results are far more general in character.

We have shown¹¹, by studying the correlation of sequence and dynamic distances between proteins, that those which adopt helical structures behave dynamically in a different manner than members of other structural classes. A significant number of helical proteins seem to be characterized by dynamics associated with the motions of pseudo-one-dimensional local structures, rather than with modes arising from global, 3-dimensional structure. This behavior is much less frequently observed in sheet/barrel and mixed α/β architectures.

In the present work we study the manner in which this correlation is manifested in defined fold groups of differing architecture. We observe several phenomena:

1. In a given fold group, sequence mobility (defined as the sequence average of the residue-specific average B factor, and denoted by B_0) acts as an driving parameter for the transition from one dynamic regime to the other.
2. The relationship between dynamic similarity and sequence mobility differs qualitatively between architecture classes.
3. The characteristics of this relationship shed light on the specific nature of molecular motions in the various protein structural classes.
4. We examine the connection between dynamic properties and fold switching in a specific system, and demonstrate that such a connection is plausible.

Methods

We begin by briefly reviewing the methodology on which this work is based. The first step is the definition of a bioinformatic parameter suitable for the description of sequence dynamics. That parameter must be based on actual dynamic data, must be residue-specific, and must distinguish between the 20 amino acids in a statistically significant manner. We

have shown¹⁰ that these requirements are satisfied by the amino-acid-specific average of the crystallographic B factor, which we denote (for amino acid X) by $\langle B(X) \rangle$. The B factor gives a crystallographic estimate of the degree of motion of a specific atom- in this case, the alpha carbon of each amino acid. Values¹¹ of $\langle B(X) \rangle$ were calculated from a very large data set of known structures. In calculating $\langle B(X) \rangle$, care was exercised to exclude structures with missing or manifestly incorrect B factors. The resulting distributions of B factors, which arise from crystal structures determined under differing conditions of temperature, chemical environment, and crystal disorder, encompass the complete range of possible values for each of the amino acids. They therefore contain more information about a specific amino acid than the per-residue mobility arising from simulation of any single protein. It was shown¹⁰ that the distributions of B for the 20 amino acids, and the associated average values, which average over random differences between individual crystal structures, indeed differ from one another in a statistically significant manner.

The physical meaning of the set of parameters $\{\langle B(X) \rangle\}$ can be understood by analogy with hydrophobicity scales, which are widely used in discussing the physical properties of protein sequences. Just as it is not true that, in every case, hydrophilic amino acids are exposed to solvent, residues with high $\langle B \rangle$ values are not always highly mobile. Rather, $\langle B \rangle$ is an intrinsic property which measures the tendency of an amino acid to be in motion. This tendency is particularly significant in the context of the complete, global sequence of the protein, when the longitudinal distribution of $\langle B \rangle$ values encodes important information about the dynamic properties of the molecule. We have developed methods to represent and analyze sequences in terms of global distributions of characteristics^{12,13,14,15,16}.

These methods are based on a signal processing approach, and were originally applied to analyses of the static physical properties of sequences. In recent work^{10,11}, we studied the sequences of proteins written as strings of $\langle B(X) \rangle$ values. (We refer to these as dynamic sequences.) This makes it possible to apply Fourier transforms to analysis of the dynamic properties of sequences. We have shown¹¹ that it is possible to define a distance between sequences based on Fourier coefficients of the longitudinal distribution of $\langle B \rangle$ values. We also constructed a structure-based distance between sequences, based on conformational parameters defined previously². These two distance metrics define two *independent* protein spaces, based on the dynamic and structural properties, respectively, of the same set of proteins. We optimized the match between sequence-sequence distances in these two spaces by varying the dynamic distance function. It was shown that the optimal correspondence between the two sets of distances arises when dynamics-based distances are based only on the $k=0$ Fourier coefficient.

In the present work, we investigate the correspondence between these two sequence spaces- one based on structure, the other on dynamics- in greater detail. We utilize data developed on the basis of our previous work¹¹. The dataset for that study consisted of 5719 proteins taken from the CATH database¹⁷, with no more than 60% sequence identity, and with reliable B factors. We label the members of the data set by their C (class), A (architecture) and T (topology) indices, and identify all (C,A,T) groups with 10 or more members. This gives 29 groups of proteins sorted by structural characteristics, with known dynamic/structural correlation properties. It should be recalled that the index C identifies

the structural class of a protein, and that $C=1$ corresponds to all-helical structures, $C=2$ to sheet/barrel structures, and $C=3$ to mixed α/β structures. Relevant data on the 29 groups is summarized in Table 1.

There are two parameters of importance in the present work. The first is $B_0(P)$, the *mobility* of a given sequence P . This is defined as the average, over the N -amino acid sequence, of the residue-specific average B factor $\langle B(X) \rangle^{11}$:

$$B_0 = \sum_{i=1}^N \langle B(X_i) \rangle \quad (1)$$

(The mobility is, in fact, the $k=0$ cosine Fourier coefficient of the dynamic sequence.) The second is $R(P)$, the Pierson correlation coefficient between structure space and dynamic space distances, between the sequence P and all other sequences in the dataset. This is obtained by calculating the distances in structure space between P and all other members of the dataset, calculating independently the distances in dynamic space between P and all other members of the dataset, and then determining the resulting value of $R(P)$. Distances in both spaces are given by Euclidean functions. The structure-space distance is calculated using a low-resolution representation of structure¹¹, which has been shown to accurately represent the details of architectural relationships. The dynamic-space distance is the result of the optimization¹¹ of a Fourier distance¹⁰ between dynamic sequences.

For each of the 29 (C,A,T) groups, we plot R as a function of B_0 . The results are shown in Figure 1.

Results and Discussion

The significance of the correlation coefficient $R(P)$ was discussed in our previous work¹¹. We demonstrated there that the dataset encompasses proteins which fall into two complementary subsets, which exhibit either normal or anomalous behavior with respect to R . In the normal region, in which $R(P) > 0$, dynamics diverge as structures do, as one would expect. The anomalous region, in which $R(P) < 0$, corresponds to an extended region of structure space in which dynamic behavior remains essentially constant. Physically, $R < 0$ indicates a region in which dynamics are relatively unaffected by structural change- and must therefore be dominated by structural subfeatures common to all the proteins in the region, independent of global architecture. It was shown that almost the entire anomalous region is made up of helical proteins, and it was suggested that, in these molecules, dynamics is dominated by quasi-one-dimensional modes associated with helical segments, rather than by three-dimensional modes determined by the global relationship between helices. It was also noted that not all helical proteins fall into the anomalous region, and that a strong preponderance of sheet/barrel and mixed α/β structures fall in the normal region.

In the present work, we examine this result in greater detail. We wish to characterize the relationship between structure, sequence mobility and dynamics. To this end we examine the relationship between R and B_0 in specific topologies (i.e., specific (C,A,T) groups), with a view to extracting hitherto inaccessible, topology-dependent dynamic information.

It is immediately clear from Figure 1 that the behavior of groups of helical proteins (indexed by $C=1$) is qualitatively different, in fact essentially opposite, to that of sheet/barrel proteins ($C=2$). Whereas, in *all* the $C=1$ groups, $R(P)$ increases with B_0 , in sheet/barrel groups $R(P)$ is a decreasing function of B_0 . Within a given helical (C,A,T) group, the behavior of R as a function of B_0 is essentially sigmoidal, suggesting that mobility serves as a driver of dynamic behavior. As mobility increases, the dynamics of the molecules transition- at first slowly, then more rapidly- from predominantly localized, pseudo-one-dimensional modes to more global, pseudo-three-dimensional behavior. The two large groups of proteins with sheet/barrel structure ($C=2$) exhibit curves which are near-sigmoidal in the opposite sense. As the mobility of these sequences increases, the dynamics transition from three-dimensional to localized behavior. (In the case of sheets and barrels, the localized modes may be pseudo-two-dimensional, reflecting the geometry/topology of the component substructures.) These data also suggest that, in α -helical molecules, the amplitude of motion of local structural blocks is lower than that of larger-scale modes of the structure, and that in all- β proteins the opposite relationship between amplitudes obtains.

It is not surprising to observe that (C,A,T) groups of mixed α/β proteins ($C=3$) exhibit characteristics of both α -helical and all- β proteins. In some groups (for example (3,10,20) and (3,10,100)) the dynamic transition is dominated by β -like behavior, while in others (for example the very large groups (3,40,50) and (3,20,20)) behavior similar to that of helical structures is superimposed on the characteristic sheet/barrel curve. We are currently investigating the factors which determine which behavior is manifested by molecules in a particular topological group.

Given our observations on the influence of sequence mobility on dynamics, and the postulate that dynamics and fold switching are connected, one can ask whether any explicit relationship can be demonstrated between mobility and fold switching. We have examined this question, although our efforts are necessarily limited by the lack of a substantial database of fold switching sequences. We consider a set of sequences from the G_A/G_B proteins studied by Alexander *et. al.*⁶. Thermal denaturation midpoints are available for these sequences, and we examine the proposition that those midpoints should decrease with increasing B_0 . This is physically reasonable, since sequences with higher intrinsic mobility should require less thermal input to denature. Furthermore, it is known that the two sequences on either side of the fold switch, which differ by one amino acid, have particularly low values of T_M , and that, as one moves through sequence space away from the architectural transition, in either direction, T_M increases, which establishes a relationship between the denaturation midpoint and fold switching. If a connection can be established between T_M and B_0 , a link between dynamic behavior and fold switching becomes a reasonable supposition.

In figure 2 we show a plot of T_M vs. B_0 for 15 sequences from this set- 7 which fold to a 3-helix-bundle structure, and 8 which fold to a $4\beta+\alpha$ architecture. Regression lines are shown separately for each set of sequences. Two results are apparent from this plot.

1. B_0 does indeed tend to decrease with T_M for both sets of sequences. This observation is highly suggestive of a relationship between dynamics and fold switching- one which is not accessible by conventional methods.
2. Values of B_0 are larger for the 3- α proteins than for the mixed α/β sequences. Viewed in light of our previous work¹¹, this is consistent with the possibility that the dynamics of these 3 α proteins are dominated by globally determined, three-dimensional modes, rather than pseudo-one-dimensional helical modes.

We would not be justified in positing a general connection between dynamics and fold switching based on the study of a single system. Nevertheless, the behavior of the G_A/G_B sequence set is supportive of this hypothesis, and points a way for future study. It is to be hoped that a significant number of systems which exhibit sequence-dependent conformational switching, independent of ligand binding or other intermolecular phenomena, will become available in the near future.

Conclusions

We have demonstrated several characteristics of dynamic behavior in the protein universe, and investigated the relationship between dynamics and fold switching in an experimentally well-studied system. We find the following results:

1. A specifically defined measure of sequence mobility drives a transition between dynamic regimes.
2. The way in which this transition is manifested differs qualitatively between structure classes.
3. Detailed examination of the conformational switch in the G_A/G_B system suggests a link between dynamics and fold switching.

Acknowledgement

This work was supported by the National Institutes of Health/NIGMS under grant GM14312.

References

1. Richardson JS, Adv. Prot. Chem 1981, 34,167.
2. Rackovsky S, Proteins 1990, 7(4), 378. [PubMed: 2381907]
3. Ambroggio XI, Kuhlman B, Curr. Opinion Struc. Biol 2006, 16(4), 525.
4. Roessler CG, Hall BM, Anderson WJ, Ingram WM, Roberts SA, Montfort WR, Cordes MHJ, Proc. Nat. Acad. Sci. USA 2008, 105(7), 2343. [PubMed: 18227506]
5. Anderson TA, Cordes MHJ, Sauer RT, Proc. Acad. Sci. USA 2005, 102(51), 18344.
6. Alexander PA, He Y, Chen Y, Orban J, Bryan PN, Proc. Nat. Acad. Sci. USA 2009, 106(50), 21149. [PubMed: 19923431]
7. Rackovsky S, Phys. Rev. Letters 2011, 106, 248101.
8. Ming D, Kong Y, Lambert MA, Huang Z, Ma J, Proc. Nat. Acad. Sci. USA 2002, 99(13), 8620. [PubMed: 12084922]
9. Doruker P, Jernigan RL, Proteins 2003, 53(2), 174. [PubMed: 14517969]
10. Scheraga HA, Rackovsky S, Proteins 2019, 87(10), 799. [PubMed: 31134683]
11. Rackovsky S, Scheraga HA, Proc. Nat. Acad. Sci. USA (2020) doi: 10.1073/pnas.2008873117.

12. Rackovsky S, Proc. Acad. Sci. USA 1998, 95(15), 8580.
13. Rackovsky S, J. Phys. Chem. B 2006, 110(38), 18771. [PubMed: 16986867]
14. Rackovsky S, Proteins 2013, 81(10), 1681. [PubMed: 23720385]
15. Scheraga HA, Rackovsky S, Proc. Acad. Sci. USA 2016, 113(7), 1808.
16. Scheraga HA, Rackovsky S, Proc. Acad. Sci. USA 2014, 111(14), 5225.
17. Orengo CA, Mitchie AD, Jones S, Jones DT, Swindells MB, Thornton JM, Structure 1997, 5(8), 1093. [PubMed: 9309224]

Figure 1:
Plots of R vs. B_0 for 29 (C,A,T) groups (see text).

Figure 2:
Plot of B_0 vs. T_M for proteins in the G_A/G system (see text). Data points for proteins with 3- α -helix structure are shown in red, and data points for those with 4 β + α structure are shown in blue. The two proteins on either side of the conformational switch, which differ in sequence by only one amino acid, are indicated by x symbols.

Table 1.
CAT Groups Used in The Present Work:

C, A, and T are the CATH indices (see text). NLT0 is the number of proteins in the (C,A,T) group with $R < 0$, NGE0 is the number with $R > 0$, and NTOT is the total number of proteins in the group. Z is the Z-function, in units of standard deviation, which measures the significance of the difference between NLT0 and NGE0 (i.e., the truth of the hypothesis that $NLT0 \neq NGE0$).

C	A	T	NLT0	NGE0	NTOT	Z	Description
1	20	5	33	93	126	3.09	Single helices in coiled-coils or other helix-helix interfaces
1	10	287	25	43	68	4.64	Helix Hairpins
1	10	490	45	4	49	14.42	Globin-like
1	10	8	16	28	44	3.67	Helicase, Ruva Protein; domain 3
1	10	760	13	28	41	2.73	Cytochrome Bc1 Complex; Chain D, domain 2
1	10	150	15	26	41	3.58	DNA Polymerase; domain 1
1	20	120	20	20	40	5.85	Four-helix bundle (Hemerythrin (Met), subunit A)
1	20	1050	20	7	27	8.21	Glutathione S-transferase Yfyf (Class Pi); Chain A, domain 2
1	50	10	19	5	24	8.42	Glycosyl transferase
1	20	1260	10	14	24	3.42	Ferritin
1	10	40	9	8	17	4.15	Ribonucleotide Reductase Protein R1; domain 1
1	20	90	10	6	16	5.07	Phospholipase A2
1	20	1250	6	10	16	2.34	Growth Hormone; Chain A
1	10	472	12	3	15	6.75	Cyclin A; domain1
1	10	418	10	5	15	5.35	Actin-binding Protein, T-fimbrin; domain 1
1	20	140	14	0	14	8.56	Butyryl-CoA Dehydrogenase, subunit A; domain 3
1	10	630	6	7	13	2.96	Cytochrome P450
1	10	340	11	0	11	7.59	Endonuclease III; domain 1
1	20	1270	7	3	10	4.65	Substrate Binding Domain of Dnak; Chain A; domain 2
2	60	40	37	404	441	-4.20	Immunoglobulin-like
2	60	120	6	177	183	-4.62	Jelly Rolls
3	40	50	43	661	704	-6.76	Rossmann fold
3	20	20	6	213	219	-5.25	TIM Barrel
3	40	30	1	63	64	-3.13	Glutaredoxin
3	10	20	16	38	54	2.72	Ubiquitin-like (UB roll)
3	10	100	1	40	41	-2.36	Mannose-binding Protein A; chain A
3	40	640	0	37	37	-2.65	Aspartate Aminotransferase; domain 2
3	60	20	0	37	37	-2.65	Glutamine Phosphoribosylpyrophosphate, subunit 1, domain 1
3	40	190	0	35	35	-2.57	D-Maltodextrin-Binding Protein, domain 2