Global Symmetry in Protein Sequence and Structure

Daniel Furse and Mikhail Wolfson

(Dated: May 22, 2007)

A protein's symmetry is often essential to its function. Due to the complexity of protein structure, however, even global symmetries are often difficult to spot with the human eye. By combining geometric and informatic techniques, we develop an approximate method for detecting certain types of global rotational symmetry in proteins. We apply the method to several protein structures and show that it can explain many of the pertinent geometric features of the protein structure. We then conclude with remarks on extending this approach to account for local partial symmetries as well.

PACS numbers: 87.15.-v, 87.15.Aa, 87.15.By, 87.15.Cc, 87.15.La

MOTIVATION AND INTRODUCTION

Recognition of tertiary symmetry in proteins has been a guiding principle in the quest to understand overall protein structure and function ever since Linus Pauling's group discovered the alpha helix in 1951[1]. Since that time, research efforts in this area have made tremendous progress, identifying a large class of repeated tertiary structural motifs, now called "folds", which usually involve many more residues than a single alpha helix typically would. Indeed, folds are higher-level structural units than alpha helices or beta sheets, and frequently use these kinds of structures as sub-units. Multiple folds sometimes exist in a single protein, and, reflecting the place in structural hierarchy that folds occupy, such a protein is said to possess quaternary structure. Hemoglobin is a good example of such a protein. In order to be properly called a fold, though, a structural motif should occur in a significant number of proteins. Sometimes, a such a set of proteins mostly perform similar tasks, but in a surprising number of cases a single fold is represented over a huge and diverse array of functions. This fact does not yet have a satisfactory explanation, and neither does the fact that folds should exist at all, for that matter. It is presently very difficult to justify *a priori*, considering the huge number of hypothetically viable structures available, why the full spectrum of protein function should be realized using a comparatively tiny number of motifs.

The puzzle of the existence of repeated and diversely represented tertiary structure motifs is further clouded by the fact that these structures often possess some kind of internal symmetry. This symmetry is usually viewed as aiding in function [2], especially where a protein performs some kind of mechanical function, as well as bolstering a protein against initially folding incorrectly and subsequent denaturation. In the general case, however, manifest tertiary symmetry is only approximate, deviations from exact symmetry being somewhat statistical or driven by need, such as DNA binding or binding to an asymmetric ligand. It is natural to ask, then, why residual symmetry should remain at all in these cases, and, taking a broad view, the rarity of overall asymmetry[2] in proteins is not immediately explicable. In order to address the evolutionary origins of tertiary symmetry in proteins, as well as the broader questions concerning the reasons for the existence of folds, the origins of tertiary symmetry must be better understood. To this end, we may ask a natural first question: Are symmetries in the tertiary structure of a protein reflected in its amino acid sequence?

Of the common symmetric protein folds, the TIM barrel stands out as a candidate for analyzing sequence-structure symmetry correlations. It is one of the most common of all folds[3], as well as one of the most generally symmetric. Being common, it concomitantly has a large number of related variant sequences all representing the same barrel-like fold, allowing for sequence-similarity studies on structurally analogous pieces of different variants.

TIM barrels are comprised of an eight-stranded, finger-trap-like beta barrel, jacketed on the outside by alpha helices. The resulting structure usually exhibits rotational symmetry through $\frac{\pi}{4}$ to a very good approximation. The topology of the resulting structure is quite complicated, and interesting at three distinct scales. Firstly, the whole protein is barrel shaped, gradually winding through a full 2π about the symmetry axis from start to finish. At the next level, the protein strand accumulates 16π radians around a hoop laid inside the rolls of the protein around the central barrel. Finally, the alpha helices on the outside accumulate between two and eight turns before finishing. The resulting structure is symmetric enough to be analyzed by global means, which we pursue, though a sketch of a local algorithm is found in the appendix. The TIM barrel is thus a relevant example of a symmetrical tertiary structure, and, furthermore, symmetric enough to be accessible by simple geometrical methods.

PHYSICAL METHODS

Geometric Model

in order to capture the global symmetries of the structure, we began by discarding side-chain information and focusing only on the structure of the C_α skeleton. Such a model provides a concise structural representation of the protein as a set of vectors that is no longer obfuscated by the side-chain atoms.

the C backbone structure can rotate about its centroid in all directions. a naïve search through all possible rotational angles $(\theta, \varphi, \psi) \in (-\pi, \pi]$ could discover certain triplets that had a large degree of "overlap" with the original structure and thereby address the question of global rotational symmetries. This brute force search, however, would scale with the angular granularity k as $\mathcal{O}\left(k^3\right)$, making accurate conclusions about larger proteins comptutationally unfeasible.

a better approach is to note that any global rotational symmetry posessed by an object will be manifested in its moment of inertia tensor $\hat{\boldsymbol{I}}$. The eigenvectors of $\hat{\boldsymbol{I}}$ (or of its $SO(3)$ cousin, the quadrupole tensor $\hat{\boldsymbol{Q}}$) denote the *principal axes* of the system. Rotation about these axes, rather than some arbitrary set, will directly probe the symmetry of the protein structure. The principal axes approach also reduces the search from $\mathcal{O}\left(k^3\right)$ to $\mathcal{O}\left(k\right)$, lifting the previoiusly imposed size restrictions.

Computational Considerations

the algorithm outlined in the supporting materials takes precisely this approach. After the center of mass of the structure is translated to be the origin, the principal axes are obtained with `CLAPACK` [8] eigensolvers. The eigenvectors $\boldsymbol{\omega}_1, \boldsymbol{\omega}_2$, and $\boldsymbol{\omega}_3$ then become the basis vectors for rotation. The eigenvector $\boldsymbol{\omega}^\star$ that corresponds to the largest eigenvalue is the symmetry axis of the entire structure, and an effective search in symmetry space can be accomplished simply by rotating around this axis in $(-pi, pi]$.

the speedup compared to a brute-force search is significant. Trials for a comparative brute-force method took $\sim$ 1 hour on a Beowulf cluster of AMD Opteron processors. The improved algorithm took $\sim$ 1 second on a laptop processor.

Scoring Function

we now address the issue of an appropriate scoring function to determine the "overlap" between two orientations of a particular protein. a good scoring function of this sort will be insensitive to small fluctuations that are not essential to the structure of the protein, but not introduce any unphysical topological features as artifacts. From these two considerations, we propose the following Mirny scoring function for any two residues i on the original protein and j on the rotated protein

$$m_{ij}\left[\boldsymbol{r}_i, \boldsymbol{r}_j(\theta, \varphi, \psi)\right] = \exp\left(\frac{-\left|\boldsymbol{r}_i - \boldsymbol{r}_j\right|^2}{2d\left|\boldsymbol{r}_i - \boldsymbol{r}_i \cdot \hat{\boldsymbol{z}}\right|}\right) \qquad (1)$$

The Mirny score approaches 1 as the distance between the residues vanishes. It is Gaussian, with a variable parameter d that corresponds to the length scale of the distance distribution. This score has the effect of smearing out every residue with a fuzziness d. As two fuzzy residues map onto one another, the score for the pair smoothly rises to unity.

In the denominator, instead of the usual d^2 Gaussian variance, the right-most term may be thought of as a distance-dependent $d(\boldsymbol{r}_i)$. The distance-dependence compensates for the fact that, due to the cyllindrical radial geometry around the symmetry axis, residues that are farther away from the axis will natrually be farther away from one another.

By rotating about the symmetry axis, the search effectively performs an integral, and we can add the rotations together to produce a position-independent quantity

$$M_{ij} \approx \int_{-\pi}^{\pi} d\theta\, m_{ij}(\theta; \varphi_0, \psi_0) \qquad (2)$$

that is parameterized by the other two orientational angles of the protein. The integrated M_{ij} are the values of the correlation plots discussed in the following sections.

INFORMATIC METHODS

In additino to physical methods, we also developed an informatic, statistical apprach to the problem in order to determine if any of the physical symmetries corresponded to underlying chemical similarities in the sequence.

In order to objectively obtain the sequence information from the correlation plots, we developed a sequence extraction algorithm that located the points on the correlation plot with the most accumulated overlap probability M_{max}. Starting from these points, the algorithm returns neighboring residue pairs whose accumulated probabilities are greater than or equal to M_{max}/M_{cut}, where M_{cut} is an empirically estimated cutoff value.

Running this algorithm on the correlation plots produced several short sequence fragments for each protein. We believed that the best way to account for the chemical similarity of two sequences was first to align them and then compare the alignment scores. This way, we would be comparing optimal similarities across the board, between all sequences.

In order to align the residues by chemical similarity, we created a "chemical similarity scoring matrix," based on the categories defined by "Chemical Group 2" from the CSPAN protein similarity web server [9]. Please see Table I for a listing of the categories.

TABLE I: The categories for amino acids used in the chemical similarity scoring matrix.

Category	Residues
small	G, A, S, T
hydrophobic	C, V, I, L, P, F, Y, M, W
polar	N, Q, H
acidic	D, E
basic	K, R

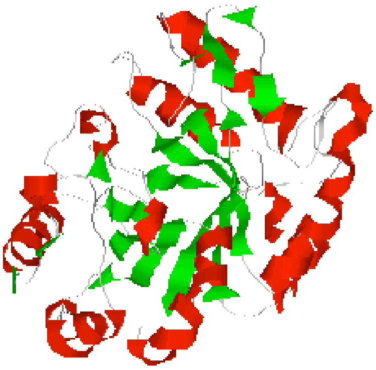

FIG. 1: A RasMac image of hevamine (1HVQ)

The scoring matrix constructed from the similiarity table was like many other match-mismatch matrices. If two amino acids were from the same category, the corresponding entry was 1. In all other cases, the entry was −1. Once the matrix was obtained, each pair of sequences that came from the same protein was globally aligned using an implementation of the Neeldeman-Wunsch algorithm [10].

RESULTS

Approximately Symmetric Example: 1HVQ

The TIM barrel protein 1HVQ, called hevamine, is found in plants and acts as a defense protein[4]. It was selected as an example because of its good $\frac{\pi}{4}$ symmetry and lack of eccentric regions connecting the alpha jacket to the interior barrel. A three-dimensional image of the molecule's structure, made using RasMac, appears in FIG.1.

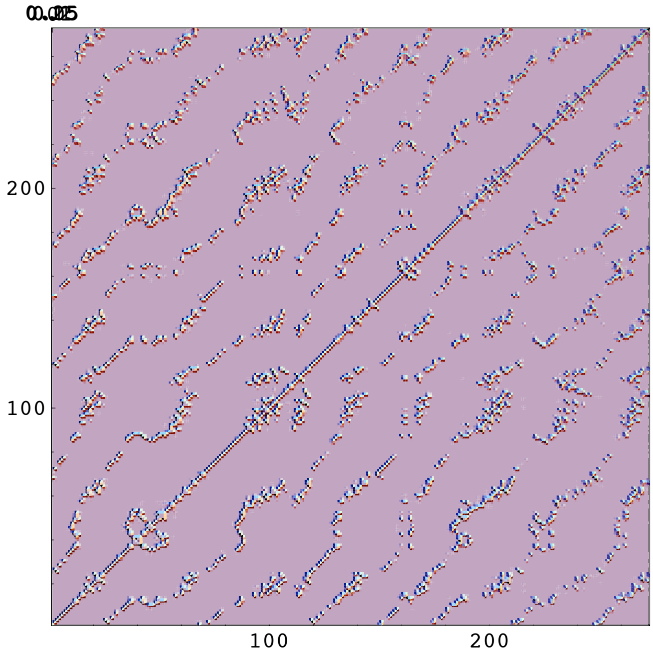

FIG. 2: A rotation-summed correlation plot for 1HVQ

The existence of approximate rotational symmetry in this molecule in immediately apparent upon inspection; one can imagine rotating the molecule against a copy of itself with mean eclipsing taking place every eighth of a turn. By adding up the pairwise correlation scores accumulated through the full rotation of the protein, a map is built up of how strongly each residue is correlated with every other. For a protein exhibiting *perfect* $\frac{\pi}{4}$ symmetry, the resulting correlation plot would contain diagonal rows of unit score, with seven non-zero 'intercepts' spaced exactly evenly along each axis of the plot. The correlation plot so obtained for the approximately symmetric 1HVQ appears in FIG.2.

The resemblance to a correlation plot for an object exhibiting perfect $\frac{\pi}{4}$ symmetry is again readily apparent; one can clearly see seven somewhat jagged "rails" running approximately parallel to the main, straight self-correlation diagonal. Note the symmetry about the main diagonal of the plot.

The deviations from ideal correlation in this case, while showing the overall strength or weakness of the symmetry, also reveal information concerning the small-scale structure of 1HVQ. In following the rail nearest the main diagonal in the plot, the path's behavior can be organized into short, straight segments interspersed with sections that worm around and sometimes develop flanking "dots" shadowing the main path. It turns out that each of these characteristic path behaviors corresponds to a different small-scale structure motif. This interesting feature is entirely due to the fact that 1HVQ is organized in a radial fashion, with closely correlated threads of betas at only relatively small radii, winding threads at intermedi-

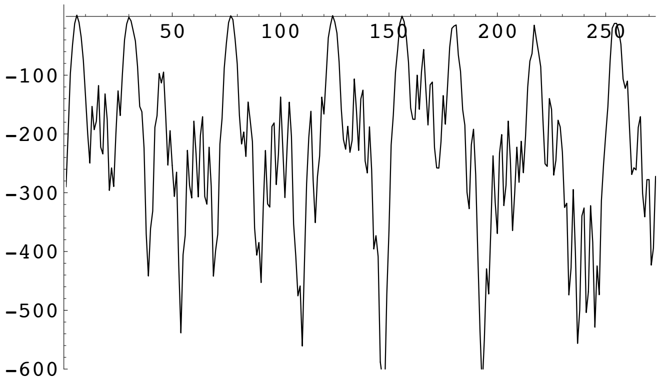

FIG. 3: Correlation strength with first beta-chain residue in 1HVQ

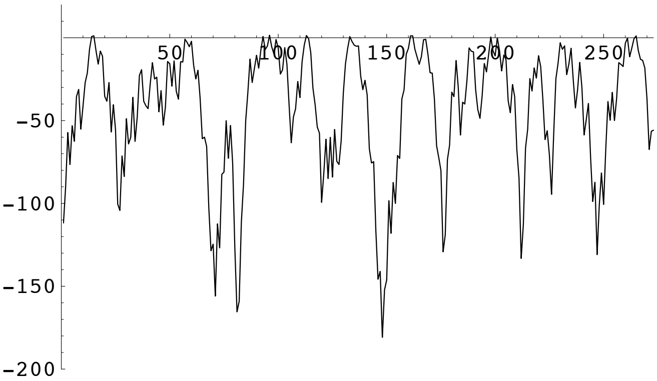

FIG. 4: Correlation strength at $n \sim 100$ in 1HVQ

ate radii, and alpha helices at only relatively large radii. This organization means that when rotating about the symmetry axis, beta sheets will only eclipse other beta sheets, threads will only eclipse threads, and alpha helices will principally eclipse other alpha helices.

Thus, we can immediately classify the behavior: The short straight segments correspond to the interior beta sheets. This region is the most symmetric and thus produces the straightest, cleanest kind of line. The power of the correlations of the beta regions to one and other is apparent in FIG.3, formed by cutting the correlation plot on the first beta region and taking the logarithm of the resulting scores. Almost identical figures are obtained by cutting on the peak positions that occur in this figure.

The regions with flanking 'dots', characterized best by what occurs near residue 100, correspond to alpha helices. The reason for the dots is the almost inevitable difference in orientation of the eclipsing helices, and is best understood in terms of the worst-case scenario. If one imagines exterior helices that are parallel and perpendicular, respectively, to the rotation axis passing through an eclipse, it's easy to visualize the helix perpendicular to the axis 'bowing' (as in the manner of a violin bow) the parallel helix, the turns of the bow periodically approaching and retreating from the outermost turns of the string, generating the dots in the correlation diagram as score is accumulated. This jaggedness, like the beta sheet strength, can be directly viewed by cutting on an alpha helix; cutting around 100 yields FIG.4

How, though, can one be certain that the correlation rails seen actually are accumulated rail-by-rail as the molecule is rotated? This is immediately resolved by cutting at a particular angle rather than summing; an example of this is shown in FIG.5, where a cut is taken at the angle $\frac{\pi}{4}$. Clearly the main diagonal is greatly diminished, and the first rail is the only prominent feature at this angle. This behavior continues as angle is increased: rails appear then disappear in order as eclipses take place.

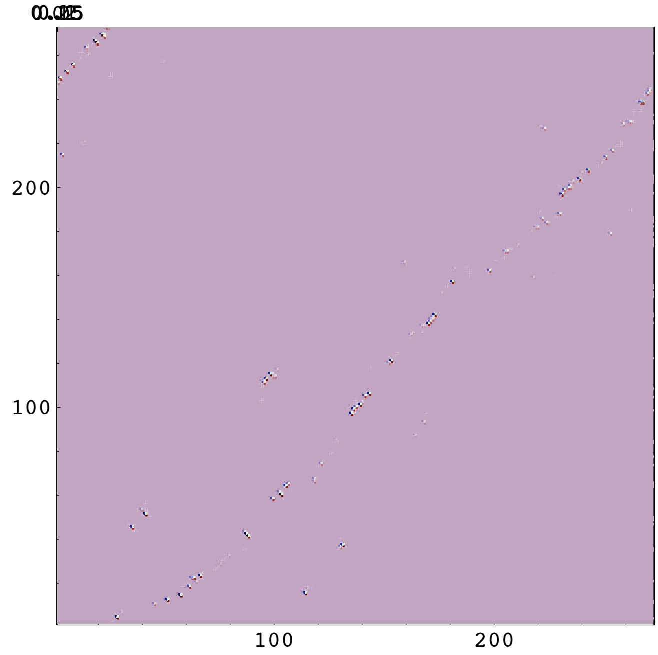

FIG. 5: Correlations at $\theta = \frac{\pi}{4}$ in 1HVQ

Less Symmetric Examples: 2EBN and 2ALR

What occurs then, if this sort of brute-force symmetry axis method is applied to a less symmetric protein? Since the method is somewhat sensitive to internal substructure of a particular fold, it is reasonable to expect that domains corresponding to asymmetric features of the protein should be apparent on the plot. This is indeed the case, and two less symmetric proteins are selected to illustrate what happens when the method is extended somewhat from the approximately symmetric regime.

The first demonstration protein is called Endo-beta-

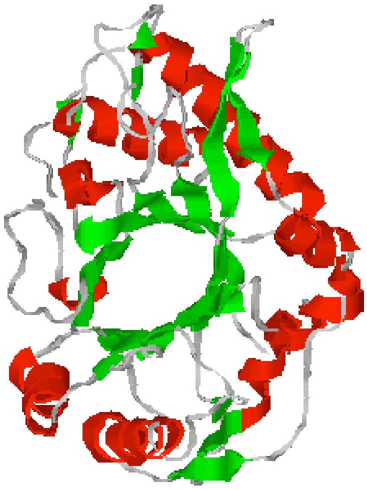

FIG. 6: A RasMac image of endo-beta-n-acetylglucosaminidase (2EBN)

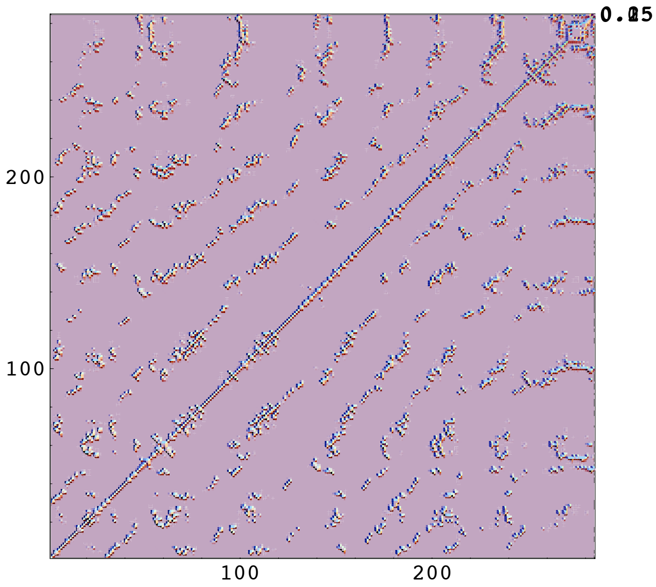

FIG. 7: A rotation-summed correlation plot for 2EBN

N-acetylglucosaminidase[5], or just 2EBN; this protein functions as a chitinase in bacteria. Compared to 1HVQ, it's fairly asymmetric, especially in the central barrel region. 2EBN also bears a superfluous (from a symmetrical perspective) strand at the end of the chain which lies somewhat awkwardly on the top of the protein, as well as having badly formed jacket helices on one side of the protein. Finally, also with regard to the central barrel, the beta-sheet structure 'spills out' of the center over one edge, taking over what would have been 'threads' in 1HVQ. These structural differences with 1HVQ are all visible in Fig.6, a three-dimensional model of 2EBN.

Running the rotation/summing algorithm on 2EBN's raw symmetry axis as defined by the quadrupole tensor yields a correlation plot that looks not entirely dissimilar from 1HVQ, found in Fig.7.

Again, the major features of the correlation plot for 2EBN look similar to those of 1HVQ, but only up to a point. A primary indicator of the breakdown of 2EBN's comparative symmetry is, first of all, the spottiness of the rails in the corners of the plot; in this area even arranging the islands of accumulated score is questionable. The major part of this breakdown is probably due to the fact that alpha helices are conspicuously absent in about half the side of the protein, and rotations near Π bring this region into eclipse with the helix-bearing side. Another offending detail of this plot occurs at the end of the strand where the rails bend up and curve away from the main diagonal path. This feature is entirely due to the free strand mentioned earlier; topologically it does not continue the pattern of toric spiralling, instead just laying on the top of the protein. This clearly distinguishes this region from the rest of the protein, and this pattern breaking is evidenced in the plot.

Finally, we turn our attention to the highly asymmetric protein 2ALR[6]. This protein is an aldehyde reductase from human beings. 2ALR has what might be called a 'decorated' TIM barrel, as two highly deviant substructures hang from the main TIM assembly, loop back, and terminate (in one case) or simply continue the TIM assembly thread. In the case of the terminating loop, the active site of the protein is actually on this terminating thread, thus lending credence to the contention espoused in [2] that sometimes symmetry is broken or altered to meet functional needs. In this case, the active site being in some kind of special, asymmetric position seems reasonable, but the reason for residual symmetry in the reactively inert support structure for this active site is unclear. One might conjecture that the symmetric shape provides physical stability against serious modification of the structure, thus making a TIM barrel a stable substrate on which to attach "arms". This would, however reasonable, be an unsubstantiated conjecture. The protein 2ALR is depicted in Fig.8; again note the appendages.

Clearly one expects the correlation plot for such a structure to be deviant from that of a more agreeably symmetric protein like 1HVQ, and indeed this eventually turns out to be the case. However, obtaining the symmetry axes of 2ALR by diagonalizing the inertia tensor turns out to give a very poor result for the symmetry axis; one cannot even see down the barrel when looking

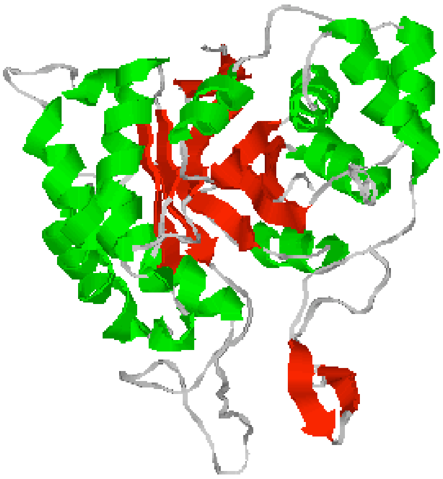

FIG. 8: A RasMac image of a human aldehyde reductase (2ALR)

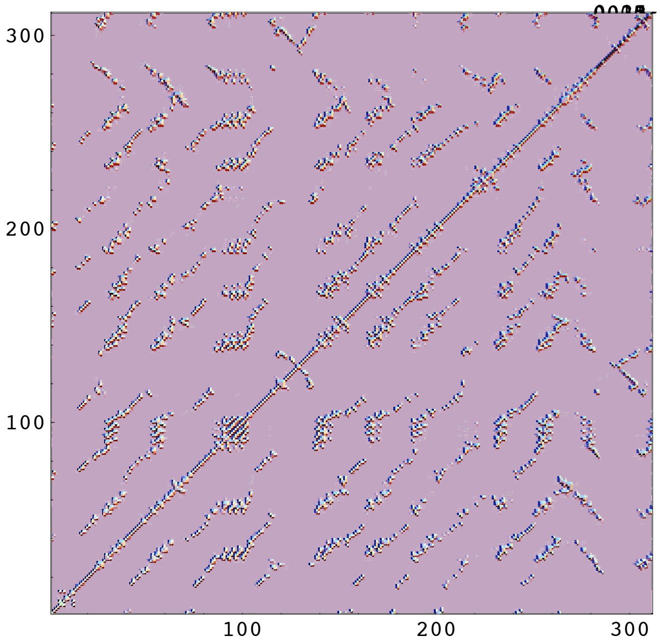

FIG. 9: A rotation-summed correlation plot for 2ALR

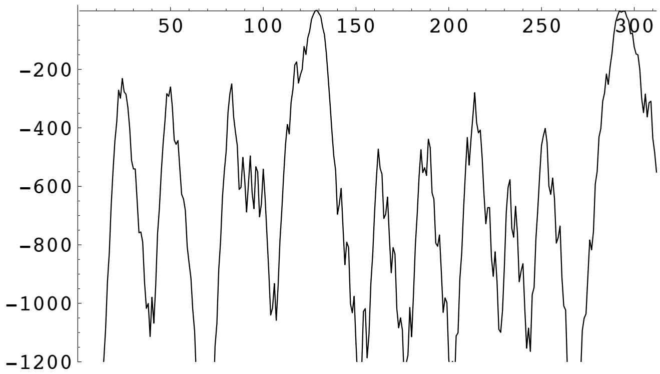

FIG. 10: Correlation strength with farthest excursion of appendages in 2ALR

toward the center along this direction. In order to find the symmetry axis of the TIM barrel part of 2ALR, it was actually necessary to manually remove the structure information corresponding to the appendages from the data and run the algorithm on the symmetrical remainder, then add these pieces back in before applying the rotation and scoring algorithm. This is a very serious drawback to the method, and as noted before an alternative is discussed in the appendix. Having so obtained the principal axes yields a sensible correlation plot, found in FIG9.

The deviation of this plot from the ideal is immediately apparent; huge sections of rails make sharp 90 degree turns before disappearing, and large strips of the plot seem almost wholly bereft of any islands. This paucity of accumulated score is due to the physical separation of the appendages from the barrel; were they to wrap around the barrel or otherwise come in proximity to it the matrix would be affected. Note that the blank appendage regions have a strong diagonal rail where they intersect, this indicates that one appendage does a good job of eclipsing the other. This can be seen by cutting the correlation plot at the farthest residue from the center of the barrel, yielding FIG.10

These three proteins, 1HVQ, 2EBN, and 2ALR characterize the structural spectrum of usefulness of the global symmetry brute-force method for rotationally symmetric proteins. The difficulties encountered in the analysis of 2ALR could be delayed algorithmically with some effort, but this would ultimately prove rather limited, as a meaningful definition of symmetry involves much more than the simple rotational eclipsing found in this way. If we allow symmetry to be a measure of 'self-relatedness', a completely different approach is needed, rooted instead in *local* geometries and groups of transformations.

CONCLUDING REMARKS

Though our intraprotein bioinformatical study of structurally analgous domains showed no correlations, this is not wholly disheartening. We can think of no inherent reason why structural similarities (especially structural repetitiveness) within a single protein should be completely unreflected in the underlying sequence, and believe with a more appropriate metric for chemical similarity, correlations could still be found. We conclude, then, that if correlations of this kind do indeed exist, that the global gapped-alignment method is simply improperly optimized for finding them. Indeed, this is a reasonable stance, as such methods were initially designed for finding hereditary relationships among long sequences and genes, not for producing chemical similarity scores between short pieces of protein of often highly variable length. Also, in the interests of looking at statistically meaningful sets of data, implementation and tuning of the local algorithm would be very helpful; having such a tool at one's disposal opens up many new channels of search, including severely broken TIM barrel symmetry as well as twin-TIM proteins, resembling the isomerases where this fold was originally noticed.

[1] L. Pauling & R. B. Corey, "Atomic Coordinates and Structure Factors for Two Helical Configurations of Polypeptide Chains," PNAS, **37**; 235-240 (1951)

[2] D.S. Goodsell DS, A.J. Olson, "Structural symmetry and protein function," Annual Rev. Biophys. Biomol. Struct. **29**; 105–53 (2000).

[3] M. Gerstein, M. Levitt, "A structural census of the current population of protein sequences," Proc. Natl. Acad. Sci. U S A. **94**,22; 11911–11916 (1997 Oct 28).

[4] A.C. Terwisscha van Scheltinga, K.H. Kalk, J.J. Beintema, B.W. Dijkstra, "Crystal structures of hevamine, a plant defence protein with chitinase and lysozyme activity, and its complex with an inhibitor," Structure, **2**,12; 1181–9 (1994 Dec 15).

[5] P Van Roey, V Rao, T.H. Plummer Jr, A.L. Tarentino, "Crystal structure of endo-beta-N-acetylglucosaminidase F1, an alpha/beta-barrel enzyme adapted for a complex substrate," Biochemistry, **33**,47; 13989-96 (1994 Nov 29).

[6] O. El-Kabbani, N.C. Green, G. Lin, M. Carson, S.V.L. Narayana, K.M. Moore, T.G. Flynn, L.J. Delucas, "Structures of Human and Porcine Aldehyde Reductase: An Enzyme Implicated in Diabetic Complications," Acta Crystallogr. **D**,50; 859 (1994)

[7] N. J. Mitra, L. Guibas, M. Pauly, "Partial and Approximate Symmetry Detection for 3D Geometry ," ACM SIGGRAPH (2006)

[8] CLAPACK

[9] CSPAN

[10] Needleman Wunsch

Appendix: An algorithm for local symmetry detection in proteins

Symmetry detection algorithms in various stages of development already exist[7], and have been shown capable of detecting symmetries in three-dimensional surface data. In such algorithms, calculation of the local curvature tensor and corresponding unit normal are extremely important for assigning direction and character to points. Unfortunately, analogous quantities do not immediately exist for proteins, being sets of points, and the protein backbone is angular enough that assuming some differentiable limit in order to define something like curvature, torsion, and an osculating plane is a poor choice. In order to harness the powerful idea of finding clusters in transformation space, we have to adequately characterize sufficiently local regions of a protein. Our algorithm characterizes, instead of single points, small clusters of points; large enough to characterize but small enough to remain local with respect to the size of a typical symmetry element.

The basic plan of the algorithm is to assign pairs of clusters to the points in transformation space that connect them, then look for clusters in transformation space. Such clusters should then map back to larger regions of the protein that are somehow similar, or symmetric. Note the *nativeness* of this algorithm; it makes no appeal whatsoever to geometry or external manipulation, it simply looks at how the protein might map back onto itself.

Clusters

In order to discuss this in more detail, we must first define a cluster. A cluster should be adequately represented by four connected backbone carbons. For each cluster, then, define the following two vectors:

- The principal vector $\boldsymbol{P}$ extends from the first to the last carbon
- The normal vector $\boldsymbol{N}$ connects the principal axis to the second carbon

Map from configuration space to transformation space

Now in order to map out of physical configuration space, iteratively step through the protein, taking all $\binom{N}{2}$ pairs of points, assigning them to transformation space in the following way:

- Find the displacement vector $\boldsymbol{d}$ that connects the base of $\boldsymbol{P}$ to $\boldsymbol{P}'$
- Find the smaller of the two axis-angle rotations that takes the vector $\boldsymbol{P}$ to the same direction as $\boldsymbol{P}'$
- Find the smaller of the two angles that takes $\boldsymbol{N}$ to $\boldsymbol{N}'$

The point $\mathcal{P}$ in transformation space mapping $\boldsymbol{P}$ into $\boldsymbol{P}'$ as far as possible without dilation is comprised of the set $\{\boldsymbol{d}, \boldsymbol{\Theta}, \phi\}$, and these points should naturally fall into clusters of symmetrically associated pairs. We have no dilation in this problem because the length from alpha carbon to alpha carbon is for all practical purposes fixed, thus fixing only one length scale. Dilation as part of the transformation might actually have the adverse effect of coarse–graining over weakly associated structures by smearing their transformation points out in yet another dimension of transformation space. Instead of dilation, which in the surface symmetry problem is curvature matching, we can assign a structural similarity score to two clusters. This is done as follows:

- Align the principal and normal vectors of two clusters, as at the end of the transformation procedure
- Pair off the carbons in the aligned clusters in order
- For each carbon, assign a Mirny score with characteristic distance d_M and then sum over the pairs, accumulating a total score

The set $\{\boldsymbol{d}, \boldsymbol{\Theta}, \phi\}$ along with the structure score S now fully characterizes the points in transformation space. Such a score would report excellent agreement between segments of alpha helices or beta sheets, for instance, while reporting a low score in a cross-comparison.

Locating clusters of points in transformation space

In order to define a cluster of points in transformation space, we must construct some sense of distance between points; such a rule is called a metric. The form of the metric in transformation space is not at all clear from the outset, and would most likely need to be optimized, along with the Mirny distance in the structural similarity score, by a neural net or some genetic algorithm. Nonetheless, we can make some blanket statements about what quantities will be more and less important:

- Rotations between principal vectors will be important
- Direction of translation vectors will be important
- Structure score will be important
- Angles between normal vectors will be less important
- Length of translation vectors $\boldsymbol{d}$ will be comparatively unimportant

After having established a sense of distance, we can link up elements of clusters into trees of nearest points. The idea is, for each point, find the nearest point, hop to that point, find that point's nearest point, and continue until we've reached a pair of points that are reciprocally closest neighbors. Do the same for the next free point, finding chains of nearest neighbors, stopping when an established chain is tapped into or when a new reciprocal relationship is found. In this way mark all points. Then look only at the points in reciprocal relationships that terminate the trees of nearest neighbor points. Find the nearest neighbors among this special class of points, and then terminate. If the clusters are reasonably tightly grouped, this will serve to find the vast majority of clusters. One can then recursively peel away clusters from the space by starting at each special point and seeking the ends of the cluster web. Note that an excellent model for the transformation space as filled with points would be a doubly linked list from which points could be popped easily to facilitate the recursive cluster harvesting at the end.

Clusters to symmetrically-related points

With clusters in hand, finding pairs of candidate regions is as easy as reading the parent pair's sequence number information out of each point in the clusters. These clusters, by construction, will map back to regions whose points are mutually associated by similar transformations. The effectiveness of an implementation of this algorithm will be mainly a function of how well the programmer is able to capture the notion of 'similarity' between different transformations in a way that is generally meaningful for proteins—such choices are in this algorithm completely parameterized by the metric.