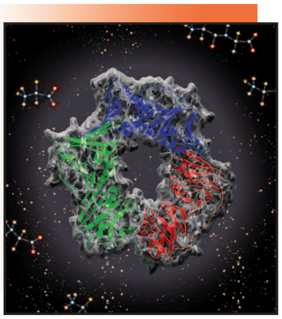


BACKGROUND © PHOTODISC, FOREGROUND IMAGE: U.S. DEPARTMENT OF ENERGY GENOMICS: GTL PROGRAM, HTTP://WWW.ORNL.GOV.HGMIS

Proteins and Their Shape Strings

An Exemplary Computer Representation of Protein Structure

BY ROGER E. ISON,
SVEN HOVMÖLLER,
AND ROBERT H. KRETSINGER

Molecular biologists need to describe, compare, and search the three-dimensional (3-D) structures of known proteins and predict the structure of novel proteins from their amino acid sequences. Such tasks depend on choosing an appropriate representation for the computation to be performed. The obvious representation of a protein as an array of its atom coordinates in 3-D is not the most appropriate choice for many purposes. Protein molecules are nonbranching chains of hundreds of amino acids, but the backbone trajectory, the protein's thread that constrains how the molecule can fold, is not explicit in a constellation of atoms. Moreover, atom coordinates are an analog or continuous representation of structure (floating-point numbers); however, many of the most useful structure comparison techniques, such as graph theory algorithms and character string pattern matching, are discrete methods formulated in terms of explicit relationships among a collection of discrete data objects.

From this perspective, a self-relative tracing of the protein's backbone thread is a more natural representation. Proteins fold by rotations around two chemical bonds that are present in every link of the chain. These two dihedral angles Φ and Ψ, associated with each link, provide enough information to follow the trajectory of the backbone and locate the backbone atoms. Dihedral angles are continuous quantities, but the angles actually adopted during folding are approximately discrete. That is, when one plots on a Φ,Ψ plane the dihedral angles found in protein structures determined to high-resolution, the observations cluster primarily in seven regions of the graph (Figure 1). Consequently, a protein's structure can be approximately represented by a sequence of discrete shape symbols that correspond to the high-occupancy Φ,Ψ regions. In this article, we call such a sequence of symbols a shape string. The reason shape-string representations work is the constancy of bond lengths and angles among the twenty genetically encoded amino acids.

However, "approximately discrete" is a serious qualification. It means that if we fix the location of the first residue of the chain exactly, we can still only estimate the location of the next residue, and from there the third residue is placed even less accurately. Errors accumulate rapidly if we try to reconstruct the entire molecule this way. Except for the highly repetitive α-helix and β-strand secondary structures, the reconstruction is not very usable for segments longer than eight or nine amino acids; it can go wildly wrong after just four or five steps. If the structural alphabet is small (seven or eight symbols), backbone reconstructions are inaccurate. Adding a few more symbols, each again representing a single, canonical Φ,Ψ pair, does not improve the results very much. The advantage of the representation for applications like structure prediction may be lost if the symbol alphabet Σ becomes large because the size of the conformation space to explore for a protein n amino acids long grows as $|\Sigma|^n$.

Structure prediction applications try to predict the locations of a protein's atoms in 3-D space from its amino acid sequence. It is relatively easy to determine the amino acid sequence of a novel protein, either from genomic data or by mass spectroscopy. However, determining the protein's 3-D structure is a much more difficult, expensive, and time-consuming exercise in X-ray crystallography or nuclear magnetic resonance. Certain proteins will not yield to even the most sophisticated laboratory methods of structure determination. Since microbes, plants, and animals have many hundreds of thousands of proteins whose structures we wish to know, there is great interest in developing computational methods to predict protein structure, but it is perhaps the hardest problem in bioinformatics.

Many structure prediction methods produce shape strings, which must then be translated and refined into a more accurate and complete list of atom locations. But since it is difficult to infer accurate atom locations directly from a shape string, additional complex processing is necessary. If the starting structure derived from the shape string is not close to the final structure, the refinement may fail altogether or require human intervention to succeed. Naturalness and convenience notwithstanding, this has limited the appeal and usefulness of shape strings as a structure representation. The ideal would be a representation with a small symbol alphabet not tied to a single point approximation for each region; conformation space could be described and explored with a small number of discrete states, yet backbones could still be reconstructed accurately. These two objectives seem contradictory, but we will describe a new way to reconcile them.

0739-5175/05/$20.00©2005IEEE

From an engineering perspective, the problem of accurately reconstructing a protein from its shape string is similar to digital compression and information recovery. Some compression algorithms derive an expansion table of patterns that recur in the input data; those patterns are replaced in the compressed form by pointers into the expansion table. Shape strings achieve compression by collapsing an amino acid's Φ,Ψ dihedral angles, which together span a 360° × 360° range, into just seven or eight discrete symbols; however, the expansion table needs to be more sophisticated than simply choosing a representative Φ,Ψ pair for each shape symbol. We present a method based on the fact that neighboring amino acids influence each other's Φ,Ψ angles. Instead of an expansion table whose entries are indexed by single shape symbols, ours are indexed by contiguous fragments of the shape string, and one expansion table works for all proteins.

The next section presents background knowledge about proteins necessary for understanding our method and results. Then the method is described, including some summary statistics characterizing its accuracy. Finally, we discuss using shape strings to compare protein structures and indicate how our method can be applied to that problem.

Essentials of Protein Structure

Protein molecules are linear polymers of hundreds of amino acids. Twenty different amino acids are encoded by DNA; they are conventionally represented by three-letter abbreviations or by twenty letters (Table 1). A linear string of letters names the amino acid sequence in order, so this representation is sometimes called the protein's one-dimensional (1-D) structure. Proteins can be compared without any reference to their actual 3-D structures by aligning their sequence strings to see how well the amino acids match up. Modern alignment algorithms are very fast and highly refined, taking into account substitutions of amino acids as they are observed to occur in related molecules and insertions and deletions of amino acids in the sequences. Because we know something about the likelihood of mutations being accepted and surviving in an organism, aligned sequences can be compared to estimate the probability that two proteins are homologous (related by evolutionary descent). Homologous proteins usually have a similar structure and function.

Proteins are not extended and loose; they normally exist folded into highly ordered, tightly packed structures whose 3-D shape is determined by the unique amino acid sequence of each protein type. In principle, each protein should fold into the shape that minimizes its free energy, but the folding process is not fully understood. Current models only approximate the relevant energy functions, and computationally minimized structures do not exactly match experimentally determined structures, even if they begin with a structure already near the minimum. Experimentally determined structures become worse, not better, when processed through current energy minimization algorithms.

A protein's fold determines which amino acids are spatially near each other, what hydrogen bonds can form to hold the molecule in shape, which amino acids are exposed on the surface and oriented to interact with other molecules, and how the molecule's physical conformation may flex and change as it interacts with ligands and other macromolecules. These fac-

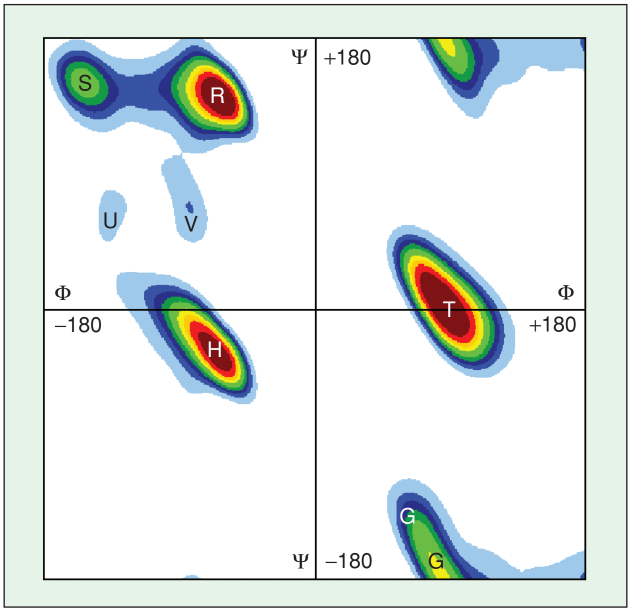


Fig. 1. The approximate locations of seven shape symbols on a composite of the plots of (3). Region details differ for each amino acid. Note that R and H have about the same Φ values but different Ψ. Region G is primarily occupied by Gly, peaking where indicated by the black letter G. Rarely, it is also occupied by other amino acids with a peak indicated by the white letter G. This conformation of non-Gly residues occurs in an uncommon form of turn that connects sequentially adjacent β-strands. (This figure is reprinted from (4) with permission of the International Union of Cyrstallography.)

Table 1. The 20 amino acids encoded by DNA, with their standard abbreviations and one-letter symbols.

Alanine (Ala, A)
Arginine (Arg, R)
Asparagine (Asn, N)
Aspartic acid (Asp, D)
Cysteine (Cys, C)
Glutamic acid (Glu, E)
Glutamine (Gln, Q)
Glycine (Gly, G)
Histidine (His, H)
Isoleucine (Ile, I)
Leucine (Leu, L)
Lysine (Lys, K)
Methionine (Met, M)
Phenylalanine (Phe, F)
Proline (Pro, P)
Serine (Ser, S)
Threonine (Thr, T)
Tryptophan (Trp, W)
Tyrosine (Tyr, Y)
Valine (Val, V)

tors govern how proteins regulate gene expression; catalyze reactions; transport ions and other molecules; form physical structures; identify cell types; transmit information; and generally conduct the business of life at the cellular level.

Each link of the chain (Figure 2, from [1]) consists of a nitrogen atom N, followed by a carbon atom Cα, then another carbon C that bonds to N of the next link. An amino acid side chain R projects sideways from Cα (like a necklace pendant), an oxygen atom O projects from the carbonyl carbon atom, and a hydrogen projects from the nitrogen. The R units are also called *side chains* or *residues*. The backbone bonds are repetitive; only the identity of its side chain R distinguishes one link from another.

The bond lengths are fixed at about 1.5 Å. Proteins fold by rotating around these bonds, with very little stretching or distortion of the bonds. The N-C bond between links of the chain is a partial double-bond that does not rotate freely. This dihedral angle, denoted Ω, is almost always 180°. Successive amino acids would therefore be on opposite sides of the chain if it extended linearly, but it does not; the four bonding electron orbitals of carbon have tetrahedral geometry. These geometric constraints also dictate that the six atoms of each $C\alpha_i$,C,O,N,H,$C\alpha_{i+1}$ group are coplanar. The whole chain can thus be described as a series of linked rectangles, as illustrated in Figure 2; each Cα is associated with two such rectangles.

The flexibility of the protein chain comes from the fact that the rectangles on each side of Cα can rotate. The N-Cα rotation angle is denoted as Φ, and the Cα-C rotation as Ψ. The important point is that since the bond lengths, orbital geometries, and Ω are determinate, the trajectory of the protein's backbone through 3-D space can be described very accurately by a list of Φ,Ψ angle pairs, one pair for each residue. From these angles, one can infer the locations of the N, H, Cα, C, and O atoms; the location of the first carbon of each amino acid residue R; and various hydrogen atoms that satisfy the other bonds. This compact representation, one Φ,Ψ angle pair associated with each amino acid in the protein sequence, is in a deep sense the natural one that reflects the real mechanics of protein folding and the physical constraints on the movements of atoms in a protein. All but two amino acids have additional rotational angles that determine the locations of the atoms of the side chain.

Secondary Structures

The vast majority of folded proteins are composed of semirigid secondary structures connected by turns. Two very common secondary structures that account for 55% of the protein structure other than turns are the α-helix and β-strand (Figure 3).

The backbone of an α-helix forms a right-handed spiral with 3.6 residues (links of the chain) per turn. The side chains, not illustrated, extend radially out from the helix. It is held in shape by hydrogen bonds between each backbone NH and the CO group four residues farther along the chain toward the N-terminus. These H-bonds are formed by overlap of the electron cloud of the hydrogen of NH and the electron cloud of the oxygen. H-bonds are weaker than covalent bonds, and they have some flexibility with respect to length, indicated by yellow and green lines in Figure 3. An α-helix is not entirely rigid, but it has considerable structural strength and imparts stability to the protein. Once formed, it is not easily broken because every link in the helix is bonded to two others, each four residues away.

A β-strand is an extended, ribbonlike structure with its side chains and its CO and NH groups alternating on either side of the strand. This alternation is due first of all to the 180° Ω angle mentioned previously. However, an extended strand by itself does not remain in exactly this configuration; to hold this shape, it usually is next to another β-strand in order to form strand-to-strand H-bonds. Figure 3 shows that every other link forms a hydrogen bond between its O and an N from another

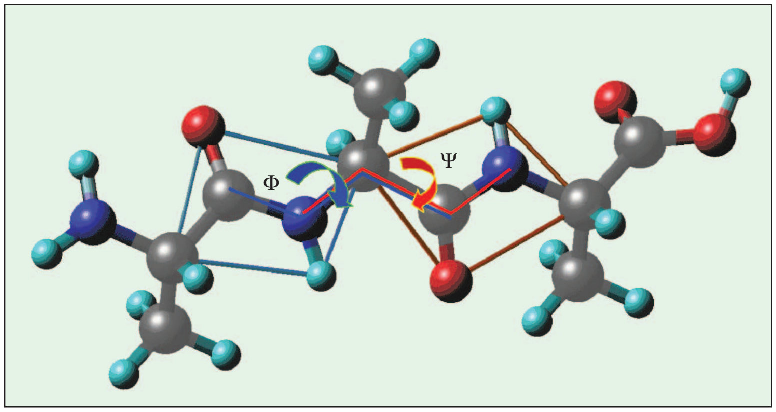


Fig. 2. A repetitive backbone of the protein chain. The Φ dihedral angle of the Cα atom is indicated by the blue arrow and Ψ by the red arrow. In the drawing, $\Phi = \Psi = 180°$; Ψ increases with NH fixed, Cα rotating clockwise, and Ψ increases with Cα fixed, CO rotating clockwise. Here, the amino acid side chain, projecting upward from Cα, is the methyl group of alanine. The backbone atoms—$C\alpha_i$, C_i, oxygen O_i (red), nitrogen N_{i+1} (dark blue) and its associated hydrogen H_{i+1} (pale blue), $C\alpha_{i+1}$—are coplanar. A hydrogen bond may form between the NH group and an O of a different residue in the protein chain. (This figure is reprinted from (1) with permission from Elsevier, 2004).

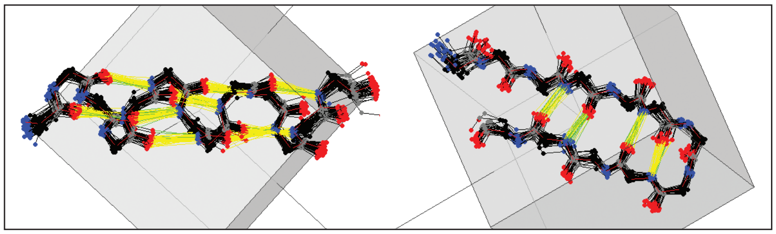


Fig. 3. The α-helix (left) is a right-handed coil with 3.6 residues per turn of the backbone. The amino acids, not shown, extend out radially from the helix. This illustration superimposes a number of α-helices that have been aligned in 3-D. The helix is held in shape by hydrogen bonds between the O of one link and an NH group four residues farther down the chain. Unlike covalent bonds, H-bonds have some freedom of orientation and length, indicated by the green and yellow lines. β-strands (right) are ribbonlike structures with consecutive residues and O and NH groups alternating on each side of the ribbon. Here two β-strands, connected by a short turn, are hydrogen-bonded into a β-sheet.

β-strand. The paired strands may be nearby in sequence and connected by a short turn as illustrated, or they may be widely separated by other segments of the amino acid sequence and come together as a result of the global structure of the protein's fold. Two or more strands can bond in this way to form a wide, warped surface appropriately called a β-sheet.

These secondary structures occur so frequently that protein structures are commonly described by assigning a symbol, H (helix), S (strand), or C (random coil), to each amino acid. Coil really signifies "everything else." HSC notation indicates the structural elements in a protein but is of limited value because C symbols, in contrast to H and S symbols, do not specify the actual trajectory of the backbone. The secondary structures are identified in order but are not placed in three dimensions.

The Ramachandran Plot and Structural Alphabets

In 1963, Ramachandran et al. [2] analyzed the relationship between Φ and Ψ by calculating whether the atoms of a dipeptide (two amino acids of a protein chain, including the backbone and a short side chain) would come into contact at various Φ,Ψ conformations. They mapped a plane encompassing the full ±180° range of each angle and found that certain regions should be accessible while other regions would be excluded by physical (van der Waals) contacts between various atoms. Such maps are now known as Ramachandran plots.

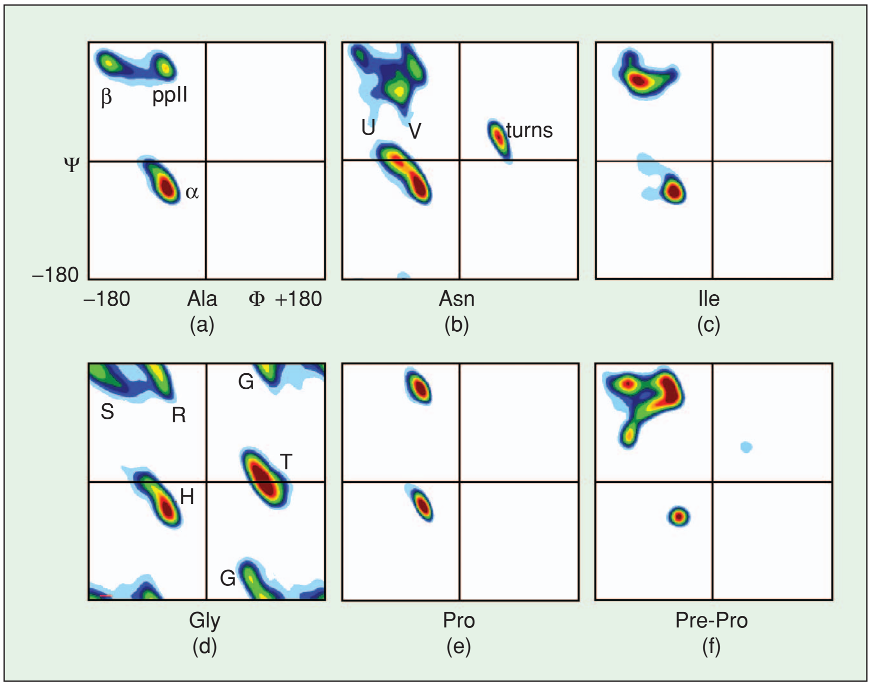


Fig. 4. The Ramachandran plots of amino acids have distinctive fingerprints. (a) Ala is most commonly found in an α-helix, a β-strand, or a polyproline type-II rod. (b) Asn is hydrophilic, so it is commonly found on the surface of proteins where it is exposed to water; therefore, it tends to participate in turns and has a more complex plot than Ala. (c) Ile rarely appears in the T conformation. (d) The side chain of Gly consists of only a hydrogen atom; its small size allows much more conformational freedom than is available to other amino acids. It is also hydrophilic, forms turns, and often ends an α-helix. (e) Pro occupies only two regions of the Ramachandran plot because of the constraint from an extra bond that it forms. (f) Residues immediately preceding Pro have an unusual distribution of conformations. Rarely, such residues may have an Ω dihedral angle of 0° instead of 180°. (This figure is reproduced from (3), with the permission of the International Union of Crystallography, http://journals.iucr. org.)

Very little experimental protein structure data existed in 1963. Today, the Protein Data Bank (PDB) [3], the most widely used public structure repository, contains more than 25,000 structures determined by X-ray crystallography or nuclear magnetic resonance. Hovmöller et al. [4] compared Ramachandran's predictions to the conformations observed for 237,384 amino acids from 1,042 high-resolution ($\leq$2.0 Å) protein structures. Ramachandran's predictions were broadly correct but off the mark in their details. The Ramachandran plots of most amino acids are similar to that of alanine (Ala), but each one has a distinctive "fingerprint." Figure 4, based on data from [4], illustrates some of these. The complete set can be found in [4].

Rooman et al. [5] assigned symbols to label seven high-occupancy regions of the Ramachandran plot. A sequence of such shape symbols, one per residue, is an approximate representation of the backbone conformation. They assigned symbols by reference to a single Φ,Ψ plot assumed to be the same for all amino acids, including the regions occupied by glycine (Gly), the smallest amino acid, which has only a hydrogen atom for its side chain. They estimated the probabilities that specific amino acid sequences will adopt conformations corresponding to sequences of shape symbols and performed a local energy calculation to predict a backbone shape string. Their procedure then used a single, representative Φ,Ψ point within each of the seven regions to construct an approximation of the predicted backbone. They noted that their seven-symbol representation carries much more information than the traditional helix, strand, and coil secondary structure designations, and that it "yields a global and complete description of 3-D structures" [5].

Figure 1 lays out shape symbols on a composite map of Ala and Gly. These do not correspond exactly to [5] but are essentially similar. The labeled regions correspond to conformations with specific physical meanings. The S region usually is occupied by a residue that is part of a β-strand, and H is an α-helical conformation. Region R corresponds to the polyproline type-II helix, a long, flexible rod in which typically every fourth or fifth residue is proline (Pro). U and V are found where β-strands transition into turns. Region G is usually occupied by Gly residues in some sort of turn, although other types of residue occasionally take this conformation. Residues in T conformation are usually found in turns. It is important to note that although these structural contexts are the most common occurrences of each conformation, all of them may also be found in the short turns and more extended strands (random coils) found in some proteins. Several consecutive S or H symbols define a β-strand or an α-helix.

The amino acids within α-helices have a very narrow distribution of

Best exemplars offer a neat solution to the difficult challenge of accurately reconstructing turns and random coil regions from their shape strings.

Φ,Ψ angles in the lower right part of the H region, indicated by very dark red shading in Figure 4(a). The upper-right part of the elongated H region comes mainly from amino acids at both ends of α-helices and in turns where just one or two consecutive amino acids have H conformation. To allow more detailed annotation, our software can split the H region into two parts: A for the pure α-helical conformation (bottom right) and K for the upper-left part of the H region. About 37% of all amino acids are in the A region and 14% in K. Unless otherwise indicated, we used the setting where A and K are merged into the one symbol H.

This original idea of shape symbols is a proper intellectual ancestor of many more recent methods and techniques that use structural alphabets, particularly structure prediction methods. But regardless of the method that assigns a discrete conformation symbol to each residue, any reconstruction that projects a backbone trajectory using one canonical Φ,Ψ point approximation for each region of the Ramachandran plot can produce backbones no more accurate than those approximations allow. Such projections are very good in α-helices and good in β-strands, but some turns will be reconstructed quite inaccurately. As a result, the global structure of the molecule will be grossly incorrect, reflecting the product of cumulative errors.

Turns are often "where the action is" in a protein. They are usually on the surface of the molecule, and biochemically active sites are often located where two turns, though widely separated in the protein sequence, come together spatially in the folded structure. This is another reason why it is important to be able to reconstruct turns accurately and, ultimately, to predict them.

Method and Results: Reconstruction of Turns from Shape Strings

It is well recognized that neighboring residues interact and affect each other's Φ,Ψ conformations. A protein wiggles in constant thermal and vibrational motion, and the molecule will prefer low-energy Φ,Ψ conformations; for example, keeping the amino acids out of each other's way. Consider a set of protein fragments having the same shape string; for example, some specific form of turn. Obviously, they all have approximately the same shape, but we hypothesized that as a group, they would also deviate systematically from the canonical point Φ,Ψ approximations for each shape symbol. Such deviations would be difficult to predict in detail, but nature has already determined the low-energy Φ,Ψ dihedral angles for many turn conformations. Rather than predicting them, it might be possible to simply look them up.

The value of this idea depends on demonstrating that shape strings are effective keys for retrieving protein substructures. This is not a statement about software methods but refers to the fact that the shape symbols used as retrieval keys are discrete, while the Φ,Ψ conformations of protein backbones are continuous quantities. Information is lost when continuous values are represented by discrete symbols; a seven symbol alphabet collapses the entire Φ,Ψ plane into seven points, so it must be demonstrated that protein fragments identified by a specific shape string really are highly congruent, especially when they are found in nonhomologous proteins. One would also hope that many or most turns in proteins come from a reasonably limited list of possibilities so that database lookup gives good coverage of real proteins. We performed computational experiments to assess these questions.

How Shape Symbols Were Assigned

In [4] it was demonstrated that the Ramachandran plots of the 20 amino acids have distinctive "fingerprints" at high resolution and that different regions of certain amino acids may conflict. For example, a Φ,Ψ point corresponding to S (part of a β-strand) for one type of amino acid may be within the R region for another amino acid. Shape symbol assignments should take these differences into account because, ideally, shape strings should retrieve protein fragments that have a similar structural character, not just ones with roughly similar backbone dihedral angles.

Separate plots for each amino acid were prepared from the data of [4], with the boundary of each region outlined manually, and were used to assign seven shape symbols corresponding to Figure 1, tailored explicitly to each amino acid. On certain plots where the boundary between two regions was ambiguous, we followed contours as best as the human eye could allow, but in some cases the R-S, S-U, or R-V boundaries were arbitrary. Ten percent of all residues in our data set lie outside the colored regions of these maps but almost always very close to a boundary. The shape symbol assigned was the one whose boundary is nearest, in Cartesian distance on the Ramachandran plane, to the Φ, Ψ conformation of that residue. We assigned shape strings to every crystal structure entry of the PDB. The shape strings are aligned with the protein sequence, stored so that they can be accessed directly, and linked back to the original PDB atomic position data.

The investigations discussed here are generally restricted to a nonredundant subset of proteins (nrPDB). The PDB contains 25,000+ entries (April 2004) but only has about 3,500 distinct proteins or chains with low sequence similarity. The others are related by homology or laboratory modification; so using them would skew statistical analyses. The nrPDB used in this work includes 2,174 protein subunits having <30% mutual sequence identity, with resolution ≤3.0 Å. It is based on a list generated by PISCES [6]. It includes 546,677 residues, about twice the size of the original data set in [4], which was restricted to structures with <2.0 Å resolution. Structures resolved at 2.0 Å to 3.0 Å are less accurate than those resolved at better than 2.0 Å but are still accurate enough to allow safe assignment of our shape symbols. We used the larger data set to get twice the statistical sample. A shape symbol was assigned to 513,136 residues (94%); the missing ones are at the ends of chains, where Φ or Ψ is undefined, or represent gaps in a PDB entry where no data exists for some parts of the structure.

The "Best Exemplar" of a Shape String

What Φ,Ψ angles most accurately reproduce the actual backbone trajectory of protein fragments that have the same shape string? We wrote software that can retrieve all the instances of any shape string from our nrPDB. The program extracts all the polypeptide fragments having a specified shape string, such as HHHHTRHSSSS. The extracted fragments are all of the same length, but have different amino acid sequences because they came from unrelated proteins. They are represented by the atomic coordinates of their N, Cα, C, O, and Cβ atoms. (Cβ is that carbon atom of the residue R, which is bonded to Cα.)

The software then selects the best exemplar of the set to represent that shape. Each fragment is aligned pairwise in 3-D space with all the other members of the retrieved set, by rotating and translating the fragments so that the RMS distance between corresponding Cα atoms is minimized. This places one fragment on top of the other as well as possible, as measured by the distances between corresponding Cα atoms. The best exemplar is the one with the minimum sum RMS atomic position error, accumulated over its alignments with all the others.

Because it is a fragment of a real protein, the best exemplar has mutually consistent Φ,Ψ angles and atom locations that represent rotational shifts and compensations that have low energy for this particular shape string. We found that the dihedral angles of the best exemplar are consistently superior to an average of corresponding Φ,Ψ dihedral angles over the retrieved set. Averaging washes out the very information we wish to capture: the subtle, interacting compensations that propagate along the chain. The best exemplar method also avoids producing a result that is biased by occasional deviant or extreme cases that might be retrieved by the shape string, reflecting extremes of nature or errors of interpretation in the electron density map.

The software reports statistics characterizing the deviations of atom positions and Φ,Ψ angles from the best exemplar. In interactive mode, the aligned fragments are then displayed as rotatable ball-and-stick figures, along with backbone hydrogen bonds, PDB descriptors, and amino acid sequences. Retrieval, alignment, display, and other computations operate at interactive speeds on an ordinary personal computer.

Connective Shapes (Handles and Turns)

These tools were used to survey the connective shapes joining α-helix and β-strand secondary structures of the proteins in the nrPDB. We identified all the shape strings actually occurring in the nrPDB that consisted of three residues in strand conformation (shape string SSS = S) or four residues in α-helix (HHHH = H) conformation, followed by a turn one to four residues long, then another S or H segment. For example, HHHHURSHHHH is a connective consisting of two α-helical handles joined by the three-residue turn URS. For each such shape string, all its instances were retrieved, the best exemplar was selected, all the other fragments with that shape were aligned to the best exemplar, and goodness-of-fit statistics were computed for the entire retrieved set.

How many distinct connectives occur in the nrPDB? Are some more common than others? A seven-symbol alphabet can combinatorially generate 2,058 different connectives having shape strings of the forms HHHHxHHHH, HHHHxxHHHH, HHHHxxxHHHH, or HHHHxxxxHHHH. The leftmost symbol of the x segment must differ from the left handle, and the rightmost symbol of x must differ from the right handle. Considering all four handle combinations (**H-H, S-S, H-S, S-H**), the total is 8,230 possibilities. Yet, as Table 2 shows, just 39 shapes account for 50% of all 11,967 such connectives in the nrPDB, and 247 shapes account for 80% of them. Proteins are very constrained in how they form short turns that connect secondary structures.

Connectives with the same shape string are highly congruent with their best exemplars. Table 3 shows that the mean RMS deviation of all Cα atom locations from their best exemplars is about 0.60 Å, with a small standard deviation of 0.35 Å. These are excellent results; for comparison, the size of a hydrogen atom is about 1.0 Å. The mean RMS discrepancy of Φ and Ψ angles is 15.8°, with a standard deviation of 15.2°.

Particularly when one or both handles of a connective are α-helices, the atoms at the ends of the α-helices diverge from the best exemplar more than atoms near or in the turn. However, this can be largely remedied by retrieving fragments with longer handles.

Extended and Unusual Shapes

Do the instances of shapes that are not short connections between two secondary structures also align tightly with their best exemplars? This question is harder to address rigorously because of our nrPDB's limited size. We enumerated unusual shapes of length nine or longer, such as those that contain no more than three consecutive H conformations and no more than

Table 2. The number of distinct connective shapes and instances retrieved from the nrPDB. The connections are of lengths 1, 2, 3, or 4, so HHHH-HHHH includes HHHHxHHHH, HHHHxxHHHH, HHHHxxxHHHH, and HHHHxxxxHHHH.

Connective	Number of Instances	Number of Shapes	Number of Shapes Occurring More Than Once	Number of Shapes Including 50% of Instances	Number of Shapes Including 80% of Instances
HHHH-HHHH	3,954	411	216	9	53
HHHH-SSS	2,811	315	172	8	54
SSS-SSS	2,589	320	168	11	58
SSS-HHHH	2,543	370	199	11	82
Total	11,967	1,416	755	39	247

two consecutive S conformations, but sometimes these come from repeated, highly similar structures within the same protein. (We did not remove low-complexity or tandem repeat regions.) Other examples suggest distant homologies because they come from proteins in the same or related structural families, even though they have low sequence identity.

Table 3. The average RMSD error of Cα location and average RMS angular discrepancy between best exemplar and all other instances of the same shape, for the most common shapes accounting for 80% of the four general connective forms.

Connective	Mean RMSd of α-carbon	Standard Deviation of RMSd	Mean RMS Discrepancy of Φ,Ψ	Standard Deviation of Φ,Ψ Discrepancy
HHHH-HHHH	0.63 Å	0.32 Å	14.1°	14.6°
SSS-SSS	0.52 Å	0.37 Å	17.1°	15.7°
SSS-HHHH	0.65 Å	0.37 Å	15.9°	16.2°
HHHH-SSS	0.59 Å	0.32 Å	16.2°	14.2°
Average	0.60 Å	0.35 Å	15.8°	15.2°

Still, having spent many days browsing through such structures, our strong impression is that shape-string retrieval also produces highly congruent sets from clearly unrelated proteins among these unusual shapes, consistent with what is described above. Very good congruence of the retrieved set can be observed even for rather complex shape strings of 16 residues found in nonhomologous proteins.

Best Exemplars Can Be Precomputed

Best exemplars offer a neat solution to the difficult challenge of accurately reconstructing turns and random-coil regions from their shape strings. The most common conformations can be precomputed, so a good answer is returned by in-memory lookup. The best exemplars of unusual shapes that have not been precomputed can be often found by a quick search of the nrPDB. A long shape string that is not found in the nrPDB, for example one produced by a structure prediction algorithm, could be patched together by overlapping shorter segments whose best exemplars are known. There are obvious ways to do this based on congruence of the overlapping regions, but we have not yet determined the best method. There will ordinarily be several combinations of best exemplars that would work, and it may be appropriate to consider both sequence and shape information when choosing the best combination to assemble.

In summary, 55% of a protein structure consists of α-helices and β-strands that can be reconstructed accurately because the backbone hydrogen bonds that hold them in shape are regular and repetitive. The difficulty of translating shape strings into accurate atomic locations primarily results from errors that arise when reconstructing the turns connecting secondary structures. Such errors accumulate and are magnified by the long lever arms of the α-helix or β-strand secondary structures. Most turns are short segments, and these occur in a surprisingly small number of distinct conformations. They can be represented by best exemplars consisting of α-helix or β-strand handles joined by the turning segments. A table of precomputed best exemplars, indexed by their corresponding shape strings, can provide accurate Φ,Ψ dihedral angles from which the backbone atom locations can be inferred. We anticipate that longer, irregular segments can be reconstructed by stringing together overlapping best exemplars.

Method and Results: Aligning Shape Strings

Structure comparisons can be used to organize protein molecules into families and to identify structural similarities between molecules that may reveal something about their biological functions, but proteins are so complex and diverse that no single measure of structural similarity is definitive. Novotny et al. [7] evaluated eleven methods that are available as Web servers and found significant variations in performance. It is beyond the scope of this article to review these methods, but they are generally based on identifying patterns or constellations of interatomic distances.

Can one successfully compare structures by aligning their shape strings in a manner similar to the well-developed methods that compare proteins by aligning their sequence strings? Shape-string alignments would be much faster, and in some ways more flexible, than widely used structure comparison methods that manipulate arrays of atom coordinates. Ye et al. [8] recently described a predictive algorithm of this type that simultaneously aligns the amino acid sequences and short shape strings describing possible, local conformations of the proteins being compared.

We now briefly discuss some problems that arise when aligning shape strings and suggest how the previously discussed methods could resolve them.

Consider a protein, which we can represent as a string of its one-letter amino acid abbreviations. Over evolutionary time, the gene that encodes this protein will mutate and evolve, resulting in some changes in the amino acid sequence. The changes may be substitutions (for example, an alanine amino acid replaced by a lysine). Mutation also may occasionally delete some amino acids from the sequence or insert new ones (indels). However, if the changes are not too extreme, we can detect remaining similarities between the sequences that indicate their evolutionary relationship.

Imagine writing the original and mutated sequences on two lines, one above the other, and sliding them left or right so that letters that are still identical align vertically. The order of the letters cannot be changed, but gaps, representing indels, may be inserted in the strings to achieve the best pairing of letters. Letters may also be aligned that are not identical, but are commonly observed substitutes for one another. (For an illustration, see the following example, in which the shape strings of two proteins are aligned instead of their amino acid sequences.)

Alignment algorithms use a scoring matrix that gives a numerical value to every pairwise combination of symbols. When aligning amino acid sequences, these values represent the probability that one letter may be substituted for another (for example, the probability that alanine will

> **Instead of an expansion table whose entries are indexed by single shape symbols, ours are indexed by contiguous fragments of the shape string, and one expansion table works for all proteins.**

replace lysine in two proteins that are related by evolutionary descent). Identical letters get the highest scores, substitutions get lower values, and gaps are penalized. The best alignment is the one that has the highest total score for its aligned pairs of letters. The "dynamic programming" algorithm is a well-known, efficient way to compute the best alignment given a scoring matrix; there are faster methods that also produce good results such as the widely used BLAST [9] program.

Some caution is appropriate when trying to apply the extensive knowledge of sequence alignment methods directly to shape strings. Sequence alignment is based on the principle that the probabilities of substitutions at different points in a sequence are independent; therefore, an algorithm can find the best alignment between two sequences by scoring each possible pairing of residues independently. Statistically, knowing the identity of one amino acid in a sequence reveals almost nothing about the identities of the preceding and following ones. But shape symbols are not serially independent. Knowing one or several preceding symbols tells a great deal about which shape symbols are most likely to come next. Also, whereas amino acids often can be substituted like construction toy parts, changing a shape symbol changes the meaning of a shape string by altering its backbone trajectory. Some shape-symbol substitutions would not change that trajectory very much, but others would completely redirect it, and groups of substitutions are often compensatory.

Insertions and deletions (indels) of amino acids can also be problematic. Indels in a turn can often be accommodated without dramatically changing the global fold, because turns are usually on the surface of the folded molecule where there is room for rearrangement. In contrast, adding (or deleting) a residue that changes the length of an α-helix will significantly alter the direction of the chain at the end of the helix, which will exit farther (or less far) around the last turn of its spiral axis as a result of the indel. This is like taking the wrong exit out of a subway station; it is not so much that one walks farther but that one walks off in the wrong direction.

Despite all these caveats, the shapes of many structural motifs are so strongly conserved in nonhomologous proteins that they can be identified immediately by simple dynamic programming alignment of the shape strings. This works because when two proteins have significant structural similarity, small differences in local shape are often accommodated by compensating adjustments in the conformations of a few adjacent residues, so the global alignment stays on track.

The example in Figure 5 shows part of a shape-string alignment between PDB entry 1AHR (calmodulin, chicken) and 2SCP (calcium binding protein, sandworm). Calmodulin contains a number of α-helices and four copies of the well-known EF-hand shape motif that is involved in calcium-binding proteins. A standard dynamic programming alignment algorithm was used, along with a scoring system that allowed substitutions of certain shape symbols that are near each other on the Ramachandran plot.

A BLAST amino acid sequence alignment of these two molecules across their entire lengths reports no significant similarity, yet their EF-hand shape strings are almost identical. (Other sequence alignment methods exist that would detect a relationship between these two proteins, for example, by comparing them both to other proteins that are more similar. The example illustrates that although the sequences are very different, their shape is highly conserved.)

In Figure 5, the first and last lines are amino acid sequences; the middle lines show the shape-string alignment. In the middle line, colon (:) characters mark identical shape symbols. Dots (.) indicate allowable shape symbol substitutions from nearby regions of the Ramachandran plane; these commonly occur at the ends of α-helices or β-strands, where the helix or strand is distorted as it transitions into a turn. Dashes (-) designate gaps inserted, where no allowable substitution was found. Eight shape symbols were used in this example; as mentioned previously, the large H region has been partitioned into A and K.

We also aligned the whole 1AHR calmodulin shape string with each protein chain in the PDB. Of the 250 top-scoring entries in the PDB, the 98 best all contain EF-hand structures that aligned with corresponding structures of calmodulin, demonstrating that embedded, similar (but somewhat variable) substructures can be discovered by comparing shape strings. Almost all the rest were oxygen storage or transport molecules that were also formed from a number of α-helices but folded in a different way. These had

```
1AHR: REAFRVFDKDGNGFISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVTM
1AHR: AAAAAAAVKKTKTSRRAAAAAAAAAAKKTSSSRAAAAAAAAKKKKRKKTKTSSRAAAAAAAK
      ::::::::.:::::.::::::::::.::.--::..::::......:::::::::::::.
2SCP: AAAAAAAAAVAKTKTSSRAAAAAAAAAAAAKTR  RAKKAAAAAAAAAVAKTKTSSRAAAAAAAA
2SCP: APLFFRAVDTNEDNNISRDEYGIFFGMLGL  DKTMAPASFDAIDTNNDGLLSLEEFVIA
```

Fig. 5. Sequence and shape strings for proteins 1AHR and 2SCP.

high similarity scores because their long α-helices match up well with those of calmodulin, and they contain some turns that are similar to those in EF-hands.

Thus, one must recognize that the folds are different. This can be done by looking up the best exemplars of turns to determine how consecutive secondary structures are situated with respect to each other in three dimensions. A fully developed shape-string alignment method would be a hybrid procedure, but it should still be very fast and flexible.

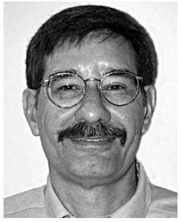
Since shape symbols are not serially independent, one could also train a software learning algorithm, such as a hidden Markov model, to recognize shape strings that characterize a particular fold. Compared to shape-string alignment, such methods have the disadvantage that one must first identify a training set by some independent means. Learning methods can provide powerful classification tools but may be less flexible for identifying embedded similarities, variations not included in the training set, and turns with different shape strings having similar 3-D backbone trajectories.

Conclusion

We have described how the configurations of protein backbone turns can be recovered from shape-string descriptions, which are compressed and approximate, by looking them up in a table. This is more than just a software trick. The underlying scientific observations are that proteins have a surprisingly small number of distinct turn conformations and that instances of such a turn tend to be highly congruent—they have very similar 3-D shapes—even when they come from unrelated proteins with very different amino acid sequences. That is why the best exemplar method works.

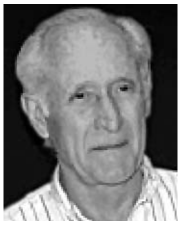
Although they are used internally in many algorithms, shape strings are not generally regarded as an annotation that humans should read. This is unfortunate. Protein structure is much more conserved by evolution than sequence is. Shape strings can be very revealing to the human eye and should identify distant homologues whose sequences have diverged in the fog of time.

A more detailed report is available on request from the corresponding author. A server that produces shape strings for all proteins in the PDB is available at [10]. That Web site also provides a complete set of Ramachandran plots for each amino acid, like those in Figure 4 and [4].



Roger E. Ison received his Ph.D. in computer science from the University of Virginia and then worked at Hewlett-Packard as an engineer and research and development manager. He is owner and cofounder of Mantic Software Corporation, where, since 1992, he has developed specialized modeling software in areas as diverse as financial portfolios and protein structure. He is a visiting research professor in the Computer Science and Engineering Department of the University of Colorado at Denver. His current research focus is protein structure comparison and prediction.

Sven Hovmöller is at the Arrhenius Laboratory at Stockholm University. He has worked for many years on crystal structure determination by electron microscopy but is now turning towards bioinformatics, mainly protein structure description and prediction, using three-dimensional (3-D) structural data from the Protein Data Bank.



Robert H. Kretsinger completed his undergraduate degree in chemistry at the University of Colorado and his Ph.D. in biophysics at Massachusetts Institute of Technology. Following postdoctoral fellowships with John Kendrew at the Medical Research Council Laboratory of Molecular Biology and with Eduard Kellenberger at the Laboratoire de Biologie Moleculaire, Université de Genève, he joined the Department of Biology at the University of Virginia. His research has addressed the structures, functions, and evolution of several protein families, including EF-hand calcium binding proteins, annexins, and $(\beta/\alpha)_8$ proteins, such as DAHP synthases. Much of his recent interest concerns the prediction ab initio of protein structure.

Address for Correspondence: Roger Ison, Department of Computer Science and Engineering, University of Colorado at Denver and Health Sciences Center, Campus Box 109, P.O. Box 173364, Denver, CO 80217-3364 USA. Phone: +1 303 556 5294. Fax: +1 303 556 8369. E-mail: roger.ison@cudenver.edu.

References

[1] R.H. Kretsinger, S. Hovmöller, and R.E. Ison, "Prediction of protein structure," *Meth. Enz.*, vol. 383, pp.1–27, 2004.

[2] G.N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, "Stereochemistry of polypeptide chain configurations," *J. Mol. Biol.*, vol. 7, pp. 95–99, July 1963.

[3] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, "The protein data bank," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–242, 2000.

[4] S. Hovmöller, T. Zhou, and T. Ohlson, "Conformations of amino acids in proteins," *Acta. Crystallogr.*, vol. D58, pp. 768–776, May 2002.

[5] M.J. Rooman, J.-P.A. Kocher, and S. Wodak, "Prediction of protein backbone conformation based on seven structure assignments," *J. Mol. Biol.*, vol. 221, no. 3, pp. 961–979, Oct. 1991.

[6] G. Wang and R.L. Dunbrack (2003) Culling the PDB by resolution and sequence identity [Online]. Available: http://www.fccc.edu/research/labs/dunbrack/pisces/

[7] N. Novotny, D. Madsen, and G. Kleywegt, "Evaluation of protein fold comparison servers," *Proteins*, vol. 54, no. 2, pp. 260–270, Feb. 2004.

[8] Y. Ye, L. Joroszewski, W. Li, and A. Godzik, "A segment alignment approach to protein comparison," *Bioinformatics*, vol. 19, no. 6, pp. 742–749, Apr. 2003.

[9] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman "Basic local alignment search tool," *J. Mol. Biol.,* no. 215, no. 3, pp. 403–410, Oct. 5, 1990.

[10] S. Hovmöller and T. Zhou (2004) Protein shape strings and DNA sequences [Online]. Available: http://www.fos.su.se/~pdbdna/pdb_shape_dna.html.