Published in final edited form as:

Proteins. 2009 June ; 75(4): 870–884. doi:10.1002/prot.22296.

Local descriptors of protein structure: A systematic analysis of the sequence-structure relationship in proteins using short- and long-range interactions

Torgeir R. Hvidsten^{1,2,*}, Andriy Kryshchak^{3,*}, and Krzysztof Fidelis^{3,**}

¹Linnaeus Centre for Bioinformatics, Uppsala University, Uppsala, Sweden

²Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, Umeå, Sweden

³Genome Center, UC Davis, Davis, CA, USA

Abstract

Local protein structure representations that incorporate long-range contacts between residues are often considered in protein structure comparison but have found relatively little use in structure prediction where assembly from single backbone fragments dominates. Here, we introduce the concept of local descriptors of protein structure to characterize local neighborhoods of amino acids including short- and long-range interactions. We build a library of recurring local descriptors and show that this library is general enough to allow assembly of unseen protein structures. The library could on average re-assemble 83% of 119 unseen structures, and showed little or no performance decrease between homologous targets and targets with folds not presented among domains used to build it. We then systematically evaluate the descriptor library to establish the level of the sequence signal in sets of protein fragments of similar geometrical conformation. In particular, we test whether that signal is strong enough to facilitate correct assignment and alignment of these local geometries to new sequences. We use the signal to assign descriptors to a test set of 479 sequences with less than 40% sequence identity to any domain used to build the library, and show that on average more than 50% of the backbone fragments constituting descriptors can be correctly aligned. We also use the assigned descriptors to infer SCOP folds, and show that correct predictions can be made in many of the 151 cases where PSI-BLAST was unable to detect significant sequence similarity to proteins in the library. Although the combinatorial problem of simultaneously aligning several fragments to sequence is a major bottleneck compared to single fragment methods, the advantage of the current approach is that correct alignments imply correct long range distance constraints. The lack of these constraints is most likely the major reason why structure prediction methods fail to consistently produce adequate models when good templates are unavailable or undetectable. Thus we believe that the current study offers new and valuable insight into the prediction of sequence-structure relationships in proteins.

^{**}Corresponding author: Krzysztof Fidelis, Address: UC Davis Genome Center, 451 East Health Sciences Drive, Davis, CA 95616-8816, USA. Tel: +01 925 980 0870, Fax: +01 530 754 9658, kfidelis@ucdavis.edu.

^{*}Contributed equally

Keywords

protein structure prediction; protein structure comparison; fragment-based methods; local protein substructures; sequence patterns; long-range interactions

Introduction

Functional properties of proteins depend on structural conformations, and the lack of structural information is therefore a major limitation in functional genomics. Although structural genomics projects aim at amending this situation by systematically solving structures experimentally to ensure that most sequences have detectable sequence similarity to at least one known structure^{1,2}, experimentally derived structures still cover only a minute fraction of the proteins that have been sequenced. Therefore, protein structure prediction is a challenge also of great practical interest.

Computational methods can systematically produce models comparable in quality to the structures solved by low resolution X-ray crystallography when sequence identity to structurally solved homologues is at least 25%³. For lower sequence similarity levels (the so-called twilight zone of around 25% sequence identity and the “midnight zone” of around 10% sequence identity⁴), the modeling techniques are less reliable although clearly evolving and improving^{3,5,6}. Even in the most challenging cases, where neither significant sequence similarity nor structural similarity can be detected between the target and solved proteins, structural blocks sufficient to build the target structure might still exist in databases. In fact, a series of CASP experiments have shown that the so-called fragment-based methods that use recurring structural components currently represent one of the most promising approaches to crack nature’s algorithm for protein folding⁷⁻¹¹.

The fragment based methods rely heavily on the protein structure databases having reached a certain “critical mass”, when they contain enough examples to allow a particular level of generalization. Early crystallographic data allowed for the observation of basic secondary structure types and recognition of somewhat larger super-secondary folding units¹²⁻¹⁴. As the databases grew, fragment based techniques became increasingly relevant as demonstrated in a large number of structural analyses and classifications using this approach¹⁵⁻²⁸. Of particular importance is a set of analyses leading to the identification of local structure motifs at either the protein backbone or backbone/side-chain levels^{27,29-39}. Finally, with their complexity ranging from large super-secondary structure blocks to individual secondary structure elements and further to very short segments and discrete residues, fragment based methods have been extensively employed in structure modeling and prediction^{8,9,40-55}. These methods mainly account for short-range residue interactions by typically considering only one continuous backbone fragment at a time. To predict a tertiary structure of a protein, fragments are typically clustered into classes based on sequence and structure similarity, and assigned to the target sequence based on the sequence signal of the cluster (see, e.g.^{56,57}).

We introduce a novel fragment-based method that makes use of combinations of short continuous backbone fragments that are close in three dimensions, but not necessarily along

the sequence (we call these combinations *local descriptors*). The method is designed to reveal local similarities between proteins: in fact, we identify all local similarities (as defined by local descriptors) between all proteins in the ASTRAL database we use in this study⁵⁸, and organize them into a library of recurring local substructures. In terms of size, local descriptors consisting of short backbone segments of typically five to seven residues are small enough to abundantly re-occur in proteins that are not related by sequence homology. Thus the method identifies, in a systematic way, the local building blocks that are common to many proteins with otherwise unrelated global structures (folds). At the same time, local descriptors are large enough to allow identifying meaningful sequence similarity between examples found in the database. The approach may therefore give new insight into the sequence-structure relationships within the folding space⁵⁹⁻⁶⁵.

The descriptor building block approach has many possible applications, including protein structure analysis, classification, alignment, identification of structure domains, and structure and function prediction. Indeed, the approach has already found successful practical applications in protein structure comparison^{66,67}, structure alignment⁶⁸, prediction of protein-ligand interactions⁶⁹, fold recognition⁷⁰, generation of fold specific consensus sequences⁷¹, residue-residue contact prediction⁷², *ab initio* protein structure assembly⁷³, and prediction of function from structure⁷⁴. Here, we describe for the first time the details of how local descriptors are built and how they can be used to construct a manageable library of building blocks for protein structure. Since some practical applications of this methodology have already been explored, we limit our attention to several basic issues critical to the nature of this formalism. We believe that utilizing sequence and structure information from several fragments at a time, including long-range side-chain interactions between these fragments, may offer a valuable advantage over single-fragment methods in predicting tertiary structure. Indeed, the lack of good long-range distance constraints is a major reason why *ab initio* folding methods often fail.

In the light of the above discussion we here set out to answer the following questions: (Q1) Can we build a manageable library of recurring multi-fragment motifs that is sufficiently powerful to represent unseen protein structures? (Q2) Is it possible to recognize sequence patterns (i.e. *sequence signal*) in the recurring structure motifs? (Q3) Are the motif sequence patterns sufficiently strong to allow assigning structure motifs to the correct structure domains? (Q4) Is it possible to use the sequence patterns to align these motifs to the correct positions along the main chain? (Q5) And finally, is it possible to infer the protein SCOP⁷⁵ fold of the target protein, especially in cases of low sequence identity to the protein domains used in building the library?

Materials and Methods

Data

Structures from the ASTRAL 40% sequence identity list (release 1.57, 4013 domains) were used to construct the descriptor database. This pre-CASP5 release made it possible to test our library on proteins from subsequent CASP experiments as well as on a large number of proteins with low sequence similarity to release 1.57. The test set used for fold assignment included 479 protein domains from ASTRAL releases 1.59 and 1.61 having less than 40%

sequence identity to domains in ASTRAL release 1.57 and belonging to one of the 135 folds with at least 5 fold-specific descriptor groups.

Defining local descriptors of protein structure

A local descriptor of protein structure is associated with a specific protein and a particular residue in the sequence of that protein (the *central* residue). To build a descriptor we identify all amino acids that are within a certain distance from the central amino acid, and add these “close” residues together with their four sequence neighbors, two on each side, to the descriptors. These continuous five residue-long backbone fragments are called *elements*, and the element including the central residue is called the *zero element*. If some elements overlap they are merged into *segments*; the segment including the zero element is called the *main segment*.

In order to identify “close” amino acids we first represent each residue as a point in space and then calculate the distance between them. We have experimented with several different choices for residue centers including the C_{α} atom, the C_{β} extended point and the geometrical center of the side chain. All the results in this paper were obtained using the so-called C_{β} -extended point as a residue center. This virtual atom, denoted as $C_{\beta x}$, is a point on the continuation of the $C_{\alpha} \rightarrow C_{\beta}$ vector, 2.5 Å away from the C_{α} atom (for glycine this point is built based on a typical configuration of the backbone and the C_{β} atom in other residues). This method for residue representation simplifies several issues. Firstly, we implicitly introduce the presence of side chains in descriptors (something that is not achieved using C_{α} atoms) and, secondly, we use the same position in space (relative to the backbone) to represent geometrically different residues (this is not the case when using the geometrical center of side chains).

We define that two residues y and z are “close” to each other in space if the Euclidian distance between their $C_{\beta x}$ “atoms” satisfies any of the two following conditions:

$$|C_{\beta x}^{(y)} - C_{\beta x}^{(z)}| < 6.5 \text{ Å} \text{ or } 6.5 \text{ Å} < |C_{\beta x}^{(y)} - C_{\beta x}^{(z)}| < 8 \text{ Å} \text{ and } |C_{\beta x}^{(y)} - C_{\beta x}^{(z)}| < |C_{\alpha}^{(y)} - C_{\alpha}^{(z)}| - 0.75 \text{ Å}.$$

These parameters were chosen empirically by making a trade-off between the desire to keep descriptors universal (i.e., small and typical for many different global conformations), and the desire to increase their size in order to enrich the structural context, increase their specificity and include more long-range interactions (see Figure 1 for examples).

Structural comparison of local descriptors

Two local descriptors are considered structurally similar if they comply with a number of requirements computed in a specific order to reduce the computational complexity.

Firstly, the number of elements $NELEM^{(i)}$ and segments $NSEGM^{(i)}$ in two compared descriptors $i = 1, 2$ are allowed to differ no more than

$$1/T_1 < NELEM^{(1)} / NELEM^{(2)} < T_1, \text{ and} \\ 1/T_2 < NSEGM^{(1)} / NSEGM^{(2)} < T_2,$$

where $T_1, T_2 > 1$ are constants.

Secondly, root mean square deviation scores are used to measure the geometrical similarity of the elements of the descriptors. Each element is represented by its C_α coordinates, and a combination of the Kabsch and McLachlan techniques^{76,77} is used to optimally superimpose pairs of elements. The resulting RMSD must satisfy the condition

$$RMSD(ELEM_0^{(1)}, ELEM_0^{(2)}) < T_3,$$

where $ELEM_0^{(i)}$, $i=1, 2$, denotes the zero element of the i^{th} descriptor and T_3 is a RMSD cutoff.

Thirdly, we require that the segments in the two descriptors have similar relative location and orientation in 3D space. This introduces a significant combinatorial problem in that we need to determine which element in one descriptor corresponds to which element in the other descriptor. To reduce the search space we only superimpose pairs of elements of the two descriptors (these so-called duplexes consisting of the zero element and one other element) and count the number of duplexes that fulfill the requirement

$$RMSD(DUPLEX_{n1}^{(1)}, DUPLEX_{n2}^{(2)}) < T_4,$$

where $ni = 1, \dots, NELEM^{(i)} - 1$, $i=1, 2$. Assuming that the first descriptor is the central descriptor of the growing group (i.e. the seed descriptor), the second descriptor is accepted as a preliminary member of this group if

$$p_f / p_t > T_5,$$

where p_f is the number of element complexes in the seed descriptor that are superimposable with a duplex in the other descriptor and p_t is the total number of duplexes in the seed descriptor.

The fourth step is carried out in the context of the group resulting from the seed descriptor being compared to all other descriptors as described in steps one through three. For each element n in the group, we calculate the ratio between the number of descriptors that contain this particular element g_n and the total number of descriptors in the group g_t . Elements that do not satisfy the condition

$$g_n/g_t > T_6 \quad \text{or} \quad g_n > T_7$$

are removed from the group. After this clean up, we also remove all descriptors that do not have the same number of segments as the seed descriptor. Note that this means that a descriptor may be represented by different segments in different groups.

Finally, we calculate the overall RMSD between the seed descriptor and each member of the group, and remove all descriptors with an RMSD higher than 1.5 times the average RMSD score calculated from the best 50% of the descriptors. This removes outliers that were falsely included in the group due to the fact that we only compared pairs of elements.

The descriptor group library discussed in this paper was calculated for the following constants: $T_1=2$, $T_2=1.5$, $T_3=1.3$ Å, $T_4=1.5$ Å, $T_5=2/3$, $T_6=1/2$, $T_7=200$. These values were determined empirically, with the goal of obtaining groups of visually similar descriptors and a library of a reasonable size (see Figure 2 and Table I for examples).

Selecting a group library

We divided the groups into classes based on the secondary structures of their segments (using the DSSP⁷⁸ program). A segment was classified as a helix if more than 50% of the residues were helices and less than 20% were identified as β -strands, and visa versa. In all other cases the segment was classified as a coil. A class consists of all descriptor groups with the same number of segments, the same secondary structure for the main segment and the same number of helical and β -stranded segments. E.g. the class 6h23 includes groups with six segments where the main segment is a helix (h) and where there overall are two helical segments, three β -stranded segments and one coil (6 - (2+3)).

Redundancy in the group library was removed by iteratively taking the largest remaining group in each class and adding/aligning new descriptors from all smaller overlapping groups in that class (if they encompass all the necessary segments and the larger group includes fewer than T_7 descriptors– see above). Since some classes might include very similar descriptor groups (due to the definition of classes), we also repeated this elimination procedure over all groups independently of class.

Defining significant signals

The probability of observing the same amino acid x times in a specific position (column) in a multiple alignment of n sequences is given by the Binomial distribution (see e.g. Wadpole & Meyers⁷⁹)

$$p(x;n,F)=\binom{n}{x}F^x(1-F)^{n-x},$$

where F is the *a priori* frequency of that particular amino acid estimated from the protein database (ASTRAL). To evaluate whether a certain residue was significantly under/over-

represented in a specific position of the descriptor-based alignment, we calculated two p-

values $p_u = \sum_{i=0}^x p(i; n, F)$ and $p_o = \sum_{i=x}^n p(i; n, F)$ describing the probability of obtaining at most or at least the observed number x of appearances of this residue simply from the data background. P-values p_u and p_o below 0.05 are associated with a significant under- /over-representation respectively.

The formula for the variance of x in the Binomial distribution is $\sigma_x^2 = nF(1 - F)$. Furthermore, the variance of the frequency of x (i.e. x/n) is given by $\sigma^2 = F(1 - F)/n$ (where σ is the standard deviation used in Signal 1 defined in the next section).

Matching multiple alignments in groups to target sequences using sequence signals

Each position in a multiple alignment may be described by a set of observed frequencies f_{ij} , where j is a sequence specific position in the alignment and i is an amino acid (or, possibly, amino acid substitution group). Based on these frequencies we may then construct a sequence signal of the form

$$s_{ij} = \frac{f_{ij} - F_i}{\sigma_i}, \quad (\text{Signal 1})$$

where F_i is the *a priori* frequency of amino acid i and σ_i is the standard deviation of the frequency of amino acid i estimated using the Binomial distribution. Notice that the signal s_{ij} is constructed so that positive values indicate overrepresentation of amino acid i in position j (not necessarily significant according to our earlier definition) and negative values indicate underrepresentation. The over/under-representation is measured in standard deviations ensuring that the signal strength is independent of the *a priori* frequency for the amino acid.

Let g refer to the multiple alignment of corresponding segments in a descriptor group (i.e. group alignment) and t refer to an equally long multiple alignment constructed from fragments of the target sequence (to increase the reliability of the predictions we represented each target sequence as a multiple alignment obtained by running PSI-BLAST for this sequence against the non-redundant sequence database of NCBI - <ftp://ftp.ncbi.nlm.nih.gov/>). A match between these two multiple alignments g and t may now be defined as the sum

$$m(g, t) = \sum_{ij} s_{ij}(g) \cdot s_{ij}(t) \cdot \sum_j SS_j(g) \cdot SS_j(t),$$

where s_{ij} is given by equation (Signal 1) and SS_j is the secondary structure in position j . The matching equation is constructed so that identical/opposite signs of the signal in the same position in the two alignments will contribute positively/negatively to the overall score proportionally to the strength of the signal. Also, the score gives preference to the alignments with better secondary structure agreement. Obviously, the secondary structure of

group members is known, while the secondary structure of the target has to be predicted (we used PSIPRED⁸⁰ for this purpose).

For the sake of comparison we also included an additional and commonly used signal definition

$$s_{ij} = \log \left(\frac{Q_{ij}}{F_i} \right), \quad (\text{Signal 2})$$

from the PSI-BLAST package. Here Q_{ij} is the estimated probability of finding amino acid i in position j , Q_{ij} is a weighted average of the observed frequency f_{ij} and prior knowledge of the substitution likelihoods given by a substitution matrix (BLOSUM-62 by default). The larger the multiple alignment is, the more weight is put on the observed frequency f_{ij} .

Local descriptor groups were assigned to target sequences in three steps: First, each segment in the group was scored individually against all fragments of the same length in the target using the scoring scheme described above. Second, target-specific descriptors were assembled from the best scoring target fragments, thereby assigning a local geometrical conformation to a specific region of the target. Given a target sequence of n amino acids and a descriptor group of m segments, there are $O(n^m)$ possible target specific descriptors that can be assembled. Obviously, searching through all these combinations is computationally intractable. We therefore limited the assembly to the 10 target fragments with the best match to each segment alignment. Hence, the complexity of assembling descriptors from a typical target sequence of 300 amino acids using a typical three-segment group was reduced from a maximum of 300^3 combinations to 10^3 combinations. Of these combinations, a considerable number are discarded due to the fact that they include fragments that overlap, have an order along the target sequence that does not occur among any of the descriptors in the group or have completely incompatible secondary structures. Third, an acceptance threshold was inferred for each descriptor group. The threshold was selected to discriminate scores obtained from domains that actually contain the local substructure and scores obtained from domains randomly selected from other folds (one protein from each fold). To avoid overfitting the threshold, scores for proteins containing the local substructures were calculated after preliminary removing their descriptors from the group (i.e. leave-one-out cross-validation).

Assignment of fold from sequence

PSI-BLAST was run for each test sequence against the sequences in the training set and the non-redundant database from NCBI, and the SCOP folds of the nearest neighbors in terms of E-score in the training set were used as fold assignments (PSI-BLAST⁸¹ was run with 10 iterations, an E-score threshold of 0.005 for including at most 10 000 sequences in the model).

For the local descriptor approach, fold assignment was conducted by matching each group in the library to the target and assigning the ones with a matching score higher than the

acceptance threshold. For each fold with assigned groups, the probability of assigning at least that many groups from the fold by chance (P-value) was computed using the hypergeometric distribution

$$p(x;N, n, k) = \frac{\binom{k}{x} \binom{N-k}{n-x}}{\binom{N}{n}},$$

where N is the number of groups in the library, n is the number of groups assigned, k is the number of groups from the fold in the library and x is the number of assigned groups from

the fold. P-values were then computed as the sum $\sum_{i=x}^{\min(n,k)} p(i;N, n, k)$.

Since the acceptance thresholds for each group were computed by matching the sequence signal to sequence domains in the training set (containing sequences with less than 40% sequence identity to each other), they may be too high for more difficult targets with little or no detectable sequence identity to any of the domains in the training set. On the other hand, the sequence length varies dramatically even for sequences from the same SCOP fold, meaning that these same thresholds may be too low for some of the longest sequences. To address these problems, we calculated P-values for all folds given the subset of groups assigned at different weights on the acceptance threshold (i.e. *threshold = original threshold * weight*). We stepwise reduced the weight from 2 to 0.5 or until at least 100 groups were assigned, and recorded the fold with the lowest P-value for the most weights. The weight giving the lowest P-value for that particular fold was used for the final assignments. This approach resulted in better performance for a number of sequences compared to only using the original acceptance thresholds. This is particularly true for difficult targets where PSI-BLAST fails to provide any fold associations. For these targets we see a substantial drop in the weight used due to a weaker sequence signal (see supplementary material).

Results and Discussion

We introduce the concept of local descriptor of protein structure to systematically describe local 3D geometries (or neighborhoods) common to different proteins. In the following sections we provide results to answer the five questions (Q1-Q5) posed at the end of the Introduction. The first three sections describe how we define local descriptors, how we compare and organize recurring descriptors into a manageable library, and whether this library is powerful enough to describe unseen proteins (Q1). The remaining sections then investigate the level of sequence signal in recurring local descriptors (Q2-Q5). Here we argue that to answer questions Q2-Q5 the most objective approach is to assess the ability to predict the structure-sequence relationships using well established techniques from machine learning to modeling and evaluation. Furthermore, we argue that a relevant and sufficiently complex prediction problem is that of assigning sequence domains to correct SCOP folds, in particular in the case of low sequence similarity to the training set. Thus, we answer

questions Q2-Q5 in the context of a reduced problem domain where similar descriptors in the library are only found in the same fold. We introduce this fold-specific library, before presenting more general statistics on the sequence signal in this library (Q2). We then devise a strategy for assigning local descriptors to domain sequences, and assess the success rates of both assignment and alignment on a set of unseen test sequences (Q3 and Q4). Finally, we assess whether we can infer the correct SCOP folds for the test sequences using the assigned local descriptors (Q5), and compare the performance of this approach to the common approach of aligning the test sequences to the training set sequences using PSI-BLAST⁸¹.

Local descriptors of protein structure

In contrast to supersecondary structures that usually encompass entire secondary structure elements, local descriptors have smaller granularity and characterize only the local structural neighborhood of particular amino acid residues. A descriptor is built of short fragments of protein backbone that are close to the selected central residue. Fragments are at least five residues long, continuous and non-overlapping. The number and length of fragments in a descriptor are determined by the side chain packing in the described neighborhood and are implicitly regulated through the distance parameters that define descriptors. This is accomplished by taking into account not only the position of the protein backbone but also the orientation of the side chains (see Figure 1 and Materials and Methods for details).

We found 714,496 descriptors with one or more non-overlapping backbone fragments (*segments*) of five or more residues in 4006 protein domains of known structure with less than 40% sequence identity to each other (ASTRAL). The average segment in these descriptors was 7.3 residues long, with 98% of the descriptors including no more than six segments. Our data show that most residues in proteins have several neighbors that are close in three dimensions but not in sequence. Indeed, only about 20% of the descriptors include only one segment and a little over 25% include two segments (see supplementary material). Since we are interested in multi-fragment descriptors to account for long-range interactions in the sequence-structure relationship, we concentrate our analysis on the 374,558 descriptors of at least three segments. However, we want to mention here that one and two fragment descriptors predominantly are helices or coils with a low percentage of β -strand fragments (on average less than 20%). In fact, it is a general trend in the data that the share of β -strands in descriptors decreases with decreasing number of segments.

A library of local substructures

To find recurring local substructures we performed an all-against-all structural comparison of the descriptors taking into account the following criteria (see Materials and Methods for details):

- number of segments,
- length of segments,
- shapes of individual segments,
- relative orientation of segments and
- overall RMSD (root mean square deviation) between similar descriptors.

For each local descriptor we build a *group* of all structurally similar descriptors (Figure 2 and Table I). We obtain 109,838 descriptor groups containing at least seven descriptors of at least three segments. Thus approximately one in three local descriptors of at least three segments frequently reoccurs in the database.

Obviously, groups centered on all descriptors contain a very large degree of overlap and thus do not constitute an intuitively appealing set of building blocks for protein structure. To deal with this, we preliminarily organized the groups into classes based on the number and secondary structure of their segments, and eliminated overlap in groups, first within classes and then between classes (see supplementary material and Materials and Methods for details). By overlap we here mean groups containing at least one common descriptor. This drastically reduced the size of our library to 4197 non-overlapping groups. From Figure 3 it is evident that less complex geometries are observed more often and in more variants than more complex spatial conformations, i.e. the number of descriptor groups decreases with increasing number of segments. Also, while the mean number of descriptors in groups stays fairly constant over different number of segments, the mean group size is increasingly higher than the median group size for groups with fewer segments. This is because there are a few geometrical conformations of low complexity that are particularly popular. For example, the group 1f97a1#62 (a β -sheet consisting of three strands) contains 12,705 structurally similar descriptors and is the most populated group in our library. None of the descriptors with more than nine segments was common enough to originate a group with more than seven members, and no groups with more than six segments included helical secondary structure.

Structural coverage of the descriptor library

If the approach is to be used for structure classification and modeling it needs to comprehensively describe protein tertiary structures including those structures not used to build the library. To this end, we tested whether the library could describe a set of 119 test proteins and domains that were assessed in CASP experiments. Figure 4 shows the structural coverage of two proteins. It is typical for the entire test set that the regions best represented by matching descriptors correspond to regions with well-determined secondary structure. This can be explained by the fact that descriptors including at least three segments more often originate from conserved protein cores rather than from more variable and less dense surface regions. It is also typical for the data, although not directly visible from Figure 4, that the groups structurally cover regions in the targets that to a large degree overlap with each other (see supplementary material). This indicates the feasibility of structure assembly from assigned descriptors by taking advantage of structural consistency between overlapping groups.

Figure 5 shows the degree of structural coverage over all test proteins, and demonstrates that there is relatively little difference between our ability to describe homology (comparative) modeling targets compared to fold recognition targets or even new fold targets. The graphs show, for example, that 80% of the test domains have more than 80% of their structures covered by descriptor groups in our library. On average, building blocks from our library could reconstruct 83% of a structure in the test set. Even more importantly, the approach covers practically the entire protein core in each case. This clearly demonstrates that the

proposed library of local descriptors constitutes a solid basis for studying the sequence-structure relationship in proteins.

Sequence-derived signals in local substructures

Both topological and evolutionary arguments point to the advantages of using the knowledge of fold in addition to the criterion of local structural similarity when analyzing the sequence signal of local descriptor groups (see supplementary material). Thus, for the sequence signal and SCOP assignment analysis we use a revised library of so-called fold-specific descriptor groups containing local descriptors from the same SCOP fold (3793 descriptor groups from 135 SCOP folds).

Groups of structurally similar descriptors provide links between local protein substructure and sequence fragments. We show here that there is an underlying sequence signal reflecting the structural similarity, and represent this signal in terms of observed frequencies of amino acids in specific positions in the multiple alignments given by the groups. If we possessed a truly representative set of sequence domains, such frequencies would provide the true probabilities for observing each amino acid in a particular structural context. Unfortunately, protein structure databases are quite biased and far from complete. Thus we use both substitution groups and substitution likelihoods to correct this bias in our analysis.

Assuming that corresponding amino acids in aligned fragments from similar local conformations reflect the propensity of proteins to accommodate the topology of that particular substructure, we would expect to see significant differences between amino acid frequencies observed in corresponding positions of these topologically similar regions and the frequencies observed overall in protein databases (henceforth referred to as *a priori* frequencies of amino acids). In order to test this hypothesis, we constructed confidence intervals for observing particular amino acid frequencies in multiple alignments of different size by chance (Figure 6, see Materials and Methods for details). Amino acids with frequencies that are lower/higher than the corresponding lower/upper confidence interval boundaries are considered to be significantly *under/over-represented* ($p < 0.05$). Table II shows that the descriptor groups exhibit larger frequency deviation than what would be expected by chance for several amino acids and all substitution groups. This supports the assumption that the aligned sequence segments in groups reflect some sequence-based similarity (i.e. signal) related to their common geometry. 56%/86% of the positions in the library had at least one amino acid or substitution group significantly under/over-represented, respectively. The higher degree of over- than under- representation, in particular for amino acids, can mostly be attributed to the combination of low *a priori* probabilities and the large number of relatively small groups, leaving no room for observing fewer amino acids than what is expected by chance (see Figure 6). The signal is in general stronger for substitution groups than for single amino acids which might indicate that local substructure similarity depends more on general sequence-related properties than strict sequence identity. However, this is obviously also, in part, due to the higher *a priori* frequencies of substitution groups. Since the descriptor groups represent recurring local substructures of three or more segments they are more likely to be situated in the dense core of the protein. From Table II we can indeed see that hydrophobic amino acids are more often significantly

overrepresented than underrepresented (and visa versa for surface and hydrogen bond amino acids), indicating that most of the positions in the group library are situated in the protein core.

The crucial question of this investigation is whether the sequence signal in our library of descriptor groups is strong enough to correctly assign local substructures to sequences. To this end, we designed a strategy for assigning groups, and thereby local substructures, to target proteins using extracted sequence signals (see Materials and Methods for details). We used the sequence signal in three different forms: two where frequencies of amino acids or amino acid substitution groups in specific positions were normalized by their *a priori* frequencies and standard deviations (denoted as Signal 1a and Signal 1b, respectively) and one where the substitution likelihoods given by a substitution matrix were incorporated (denoted as Signal 2).

Table III shows that the approach is able to discriminate proteins containing a particular substructure from the proteins that do not. It is also clear that target-specific descriptors with the highest scores are often assembled from the correct target fragments. The latter means that the alignment is correct, something that is critical for the assignment of long-range interactions. Amino acids that are close in three dimensions but relatively far apart in sequence are obviously located on different segments in descriptors and therefore we need to align at least two segments correctly in order to correctly assign long-range interactions.

Assigning substructures and SCOP folds to unseen targets

We have shown that it was possible to use the extracted sequence signals to consistently discriminate domains with a particular local substructure from other protein domains. Although these results were obtained using cross validation on domains with less than 40% sequence identity to each other, they were still obtained from domains used to construct the descriptor library. Here we perform a similar analysis on 479 protein domains with less than 40% sequence identity to any domain used to build the groups (Table IV shows some properties of the test set). We match each descriptor group to the test sequence and accept the assignments scoring higher than the group's acceptance threshold. We then check to what degree we were able to assign the correct local substructures and whether we were able to align them correctly.

Table V shows that we are still able to assign local substructures to the correct targets with correct alignment. Although, the performance is expectedly lower than what was achieved within the descriptor library (Table III), on average 56% of the segments in a correctly assigned group are also correctly aligned. Thus proper long-range interactions are identified. It is interesting to notice that there is relatively little difference in our ability to assign groups that structurally match the target (i.e. within the structure similarity requirements used to build descriptor groups, see Materials and Methods) and groups that are from the correct fold (see the two types of sensitivity in Table V). This indicates that local substructures in the targets exhibit similar sequence patterns to those present in the group library even when the structural similarity drops below what is defined as a structural match. From Table V it is also clear, as was the case in Table III, that Signal 1 performs considerably better than Signal

2, however, no immediate gain is observed by adding substitution groups to the signal extraction procedure (Signal 1a versus Signal 1b).

We also tested the ability to infer the correct SCOP fold for test sequences by assigning library groups to the sequences and creating rankings of folds based on those assignments (see Materials and Methods). Thus we test whether there is a bias towards assigning local substructures needed to assemble the actual tertiary structure of the target protein. In many ways, this is less stringent than the preceding test since false positive assignments can be accepted to some degree. Table V shows that the correct fold is our top ranked fold for 70% of the targets and one of our top five folds for 84% of the targets (results using Signal 1b). Obviously, the difficulty of this exercise depends on whether there exist templates of known structure in the training set with at least some sequence similarity to the target. A nearest neighbor approach where PSI-BLAST was used to identify the closest sequence in the descriptor library resulted in fold assignments for 332 of 479 targets of which 329 were correct (Table V: $329/479=0.687$). For the remaining 147 test targets, our approach still had the correct fold as the top ranked fold in 44% of the cases and as one of our top five folds in 64% of the cases. It is interesting to notice that the performance of the descriptor approach correlates much less with the target's sequence identity to the training set than the performance of PSI-BLAST, and that the descriptor approach only correlates modestly with the ability of the local substructure library to describe the target structurally (i.e. coverage) (Table V). Finally, the addition of substitution groups in the signal extraction process (Signal 1b) makes the quality of the fold assignments less dependent on sequence identity and coverage than just using amino acids alone (Signal 1a).

E-score as reported by PSI-BLAST is the likelihood of observing the obtained alignment between the target and a sequence in the database. In Figure 7 the 479 targets are plotted against the E-score from PSI-BLAST and the P-value computed from assigned descriptor groups. It shows that E-scores < 0.05 almost certainly align two sequences of the same SCOP fold. The P-values from our approach are also shown to be a strong indicator of whether the fold assignment is actually correct or not. In addition, the local descriptor approach identifies the correct fold for many of the 151 sequences where PSI-BLAST finds no significant relation to sequences used to build the descriptor library. Thus the two approaches complement each other to cover the entire range of sequence similarity.

Conclusion

We have introduced the concept of local descriptors of protein structure and constructed a library of recurring local substructures based on this principle. This approach has several potential advantages over single-fragment methods currently used in structure prediction. To investigate the properties of the new approach we addressed several questions stated at the end of the Introduction (Q1-Q5). Indeed, the descriptor library exhibits good structural coverage of both seen and unseen structures (Q1), demonstrating that local descriptors can function as building blocks for protein structures.

In general, the level of difficulty of using the descriptor formalism depends on whether the protein structure or the protein sequence is taken as the starting point. It is the latter problem

that poses a greater challenge since it requires identifying the sequence-structure relationships. Here, we present a systematic evaluation of the sequence signal found in the structurally similar descriptors grouped with the objective of fold recognition. We were able to construct descriptor groups for 135 SCOP folds for which a sufficient representation can be assembled from the descriptor library. First, we show that there is a statistically significant, although relatively weak, sequence signal in the descriptor groups in terms of overrepresentation of amino acids/substitution groups at single positions (Q2, Table II, Figure 6). We then show that sequences containing a local substructure can be discriminated from sequences that do not (Q3), and that the corresponding assignments are often correctly aligned (Q4, Table III, Table V). Finally we show that the assignments of local substructures can be used to infer SCOP folds (Q5, Table V, Figure 7).

Comparative modeling methods often rely on (PSI-)BLAST in finding a multiple alignment involving the target and at least one template sequence. To what degree structural information can be carried over from the template to the target depends on the quality of the alignment. In the range of fairly good sequence alignments, PSI-BLAST is a reliable tool for identifying structural similarity. However, when the sequence alignment is poor, no connection to structure can be obtained. In contrast, fragment-based methods might still be able to establish such connection for fragments of the protein and then to assemble a model from these fragments of known structure. Here we show that it is possible to correctly assign multi-fragment local substructures for these so-called difficult targets and subsequently to infer the correct fold.

A natural extension of this approach is *ab initio* assembly of global structural models from the same assigned local substructures. For this to work, we must first and foremost be able to assign the correct local substructures to the correct sequences. However, in order to accomplish correct assembly, the alignment must also be correct, that is, the fragments of the local substructures must be assigned not only to the correct protein, but also to the correct positions along its sequence. This might seem to be a natural consequence of assigning the correct substructures to the correct sequence, but experience shows that the first does not necessarily imply the latter. Global features specific to folds such as for example an overrepresentation of certain amino acids or the number and type of regular secondary structure elements might suffice to facilitate fold recognition even without correct alignment. Our evaluation, however, shows that the alignment often is correct, and we believe that through an assembly procedure correct alignments will emerge in self-consistent global models of the protein structure. Indeed, promising results to this end have already been achieved⁷³. An important effect of correctly aligned fragments from the same descriptor is the automatic inclusion of long-range distance constraints. The local descriptors are defined in such a way that side-chain interactions between distant residues emerge as a side-effect of correct alignment. Naturally, the price to be paid is the addition of a combinatorial problem of assigning many fragments. However, constraints on secondary structure and order along the sequence were shown to be helpful in accomplishing this task. Since lack of good long-range distance constraints is the main challenge in template-free modeling, the concepts introduced in this paper may be of significant importance to the development of new and improved structure prediction methods.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institutes of Health (National Library of Medicine) grant LM07085 (to KF), the Knut and Alice Wallenberg Foundation, and the Swedish Foundation for Strategic Research.

References

1. Zhang C, Kim SH. Overview of structural genomics: from structure to function. *Curr Opin Chem Biol.* 2003; 7(1):28–32. [PubMed: 12547423]
2. Chandonia JM, Brenner SE. The impact of structural genomics: expectations and outcomes. *Science.* 2006; 311(5759):347–351. [PubMed: 16424331]
3. Tramontano A. Of men and machines. *Nat Struct Biol.* 2003; 10(2):87–90. [PubMed: 12555084]
4. Rost B. Twilight zone of protein sequence alignments. *Protein Eng.* 1999; 12(2):85–94. [PubMed: 10195279]
5. Kryshchukovych A, Venclovas C, Fidelis K, Moult J. Progress over the first decade of CASP experiments. *Proteins.* 2005; 61(Suppl 7):225–236. [PubMed: 16187365]
6. Kryshchukovych A, Fidelis K, Moult J. Progress from CASP6 to CASP7. *Proteins.* 2007; 69(Suppl 8):194–207. [PubMed: 17918728]
7. Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins.* 1999; (3):171–176. [PubMed: 10526365]
8. Bonneau, R., Tsai, J., Ruczinski, I., Chivian, D., Rohl, C., Strauss, CE., Baker, D. *Proteins.* 2001. Rosetta in CASP4: progress in ab initio protein structure prediction; p. 119–126.
9. Jones DT. Predicting novel protein folds by using FRAGFOLD. *Proteins.* 2001; (5):127–132. [PubMed: 11835489]
10. Bradley P, Malmstrom L, Qian B, Schonbrun J, Chivian D, Kim DE, Meiler J, Misura KM, Baker D. Free modeling with Rosetta in CASP6. *Proteins.* 2005; 61(7):128–134. [PubMed: 16187354]
11. Bujnicki JM. Protein-structure prediction by recombination of fragments. *Chembiochem.* 2006; 7(1):19–27. [PubMed: 16317788]
12. Rao ST, Rossmann MG. Comparison of super-secondary structures in proteins. *J Mol Biol.* 1973; 76(2):241–256. [PubMed: 4737475]
13. Levitt M, Chothia C. Structural patterns in globular proteins. *Nature.* 1976; 261(5561):552–558. [PubMed: 934293]
14. Richardson JS. beta-Sheet topology and the relatedness of proteins. *Nature.* 1977; 268(5620):495–500. [PubMed: 329147]
15. Moult J, James MN. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins.* 1986; 1(2):146–163. [PubMed: 3130622]
16. Chothia C, Lesk AM. Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol.* 1987; 196(4):901–917. [PubMed: 3681981]
17. Sibanda BL, Blundell TL, Thornton JM. Conformation of beta-hairpins in protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. *J Mol Biol.* 1989; 206(4):759–777. [PubMed: 2500530]
18. Rooman MJ, Rodriguez J, Wodak SJ. Automatic definition of recurrent local structure motifs in proteins. *J Mol Biol.* 1990; 213(2):327–336. [PubMed: 2342110]
19. Fidelis K, Stern PS, Bacon D, Moult J. Comparison of systematic search and database methods for constructing segments of protein structure. *Protein Eng.* 1994; 7(8):953–960. [PubMed: 7809034]
20. Donate LE, Rufino SD, Canard LH, Blundell TL. Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database for modeling and prediction. *Protein Sci.* 1996; 5(12):2600–2616. [PubMed: 8976569]

21. Geetha V, Munson PJ. Linkers of secondary structures in proteins. *Protein Sci.* 1997; 6(12):2538–2547. [PubMed: 9416603]
22. Oliva B, Bates PA, Querol E, Aviles FX, Sternberg MJ. An automated classification of the structure of protein loops. *J Mol Biol.* 1997; 266(4):814–830. [PubMed: 9102471]
23. van Vlijmen HW, Karplus M. PDB-based protein loop prediction: parameters for selection and methods for optimization. *J Mol Biol.* 1997; 267(4):975–1001. [PubMed: 9135125]
24. Nakashima H, Nishikawa K, Ooi T. Segmentation of a protein into structural elements: analysis and classification of segments. *J Protein Chem.* 1988; 7(4):509–525. [PubMed: 3255375]
25. Micheletti C, Seno F, Maritan A. Recurrent oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. *Proteins.* 2000; 40(4):662–674. [PubMed: 10899788]
26. Kolodny R, Koehl P, Guibas L, Levitt M. Small libraries of protein fragments model native protein structures accurately. *J Mol Biol.* 2002; 323(2):297–307. [PubMed: 12381322]
27. Haspel N, Tsai CJ, Wolfson H, Nussinov R. Hierarchical protein folding pathways: a computational study of protein fragments. *Proteins.* 2003; 51(2):203–215. [PubMed: 12660989]
28. Hunter CG, Subramaniam S. Protein fragment clustering and canonical local shapes. *Proteins.* 2003; 50(4):580–588. [PubMed: 12577264]
29. Alexandrov NN, Takahashi K, Go N. Common spatial arrangements of backbone fragments in homologous and non-homologous proteins. *J Mol Biol.* 1992; 225(1):5–9. [PubMed: 1583693]
30. Orengo CA, Taylor WR. A local alignment method for protein structure motifs. *J Mol Biol.* 1993; 233(3):488–497. [PubMed: 8411157]
31. Alexandrov NN, Go N. Biological meaning, statistical significance, and classification of local spatial similarities in nonhomologous proteins. *Protein Sci.* 1994; 3(6):866–875. [PubMed: 8069217]
32. Artymiuk PJ, Poirrette AR, Grindley HM, Rice DW, Willett P. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J Mol Biol.* 1994; 243(2):327–344. [PubMed: 7932758]
33. Fischer D, Wolfson H, Lin SL, Nussinov R. Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: potential implications to evolution and to protein folding. *Protein Sci.* 1994; 3(5):769–778. [PubMed: 8061606]
34. Wallace AC, Laskowski RA, Thornton JM. Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci.* 1996; 5(6):1001–1013. [PubMed: 8762132]
35. Russell RB. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J Mol Biol.* 1998; 279(5):1211–1227. [PubMed: 9642096]
36. Fetrow JS, Siew N, Skolnick J. Structure-based functional motif identifies a potential disulfide oxidoreductase active site in the serine/threonine protein phosphatase-1 subfamily. *Faseb J.* 1999; 13(13):1866–1874. [PubMed: 10506591]
37. Kleywegt GJ. Recognition of spatial motifs in protein structures. *J Mol Biol.* 1999; 285(4):1887–1897. [PubMed: 9917419]
38. Leibowitz N, Fligelman ZY, Nussinov R, Wolfson HJ. Automated multiple structure alignment and detection of a common substructural motif. *Proteins.* 2001; 43(3):235–245. [PubMed: 11288173]
39. Nakayama S, Willett P. A sphere-based descriptor for matching protein structures. *J Mol Model.* 2002; 8(6):199–207. Online. [PubMed: 12140603]
40. Jones TA, Thirup S. Using known substructures in protein model building and crystallography. *Embo J.* 1986; 5(4):819–822. [PubMed: 3709525]
41. Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol.* 1998; 281(3):565–577. [PubMed: 9698570]
42. Claessens M, Van Cutsem E, Lasters I, Wodak S. Modelling the polypeptide backbone with ‘spare parts’ from known protein structures. *Protein Eng.* 1989; 2(5):335–345. [PubMed: 2928296]
43. Unger R, Harel D, Wherland S, Sussman JL. A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins.* 1989; 5(4):355–373. [PubMed: 2798411]

44. Pedersen JT, Moulton J. Ab initio structure prediction for small polypeptides and protein fragments using genetic algorithms. *Proteins*. 1995; 23(3):454–460. [PubMed: 8710838]
45. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol*. 1997; 268(1):209–225. [PubMed: 9149153]
46. Deane CM, Blundell TL. A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins. *Proteins*. 2000; 40(1):135–144. [PubMed: 10813838]
47. de Brevern AG, Etchebest C, Hazout S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins*. 2000; 41(3):271–287. [PubMed: 11025540]
48. Burke DF, Deane CM. Improved protein loop prediction from sequence alone. *Protein Eng*. 2001; 14(7):473–478. [PubMed: 11522920]
49. De Brevern AG, Valadie H, Hazout S, Etchebest C. Extension of a local backbone description using a structural alphabet: A new approach to the sequence-structure relationship. *Protein Sci*. 2002; 11(12):2871–2886. [PubMed: 12441385]
50. Tosatto SC, Bindewald E, Hesser J, Manner R. A divide and conquer approach to fast loop modeling. *Protein Eng*. 2002; 15(4):279–286. [PubMed: 11983928]
51. Hunter CG, Subramaniam S. Protein local structure prediction from sequence. *Proteins*. 2003; 50(4):572–579. [PubMed: 12577263]
52. Kolodny R, Levitt M. Protein decoy assembly using short fragments under geometric constraints. *Biopolymers*. 2003; 68(3):278–285. [PubMed: 12601789]
53. Das R, Qian B, Raman S, Vernon R, Thompson J, Bradley P, Khare S, Tyka MD, Bhat D, Chivian D, Kim DE, Sheffler WH, Malmstrom L, Wollacott AM, Wang C, Andre I, Baker D. Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins*. 2007; 69(8):118–128. [PubMed: 17894356]
54. Zhang Y. Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins*. 2007; 69(8):108–117. [PubMed: 17894355]
55. Zhou H, Pandit SB, Lee SY, Borreguero J, Chen H, Wroblewska L, Skolnick J. Analysis of TASSER-based CASP7 protein structure prediction results. *Proteins*. 2007; 69(8):90–97. [PubMed: 17705276]
56. Han KF, Baker D. Recurring local sequence motifs in proteins. *J Mol Biol*. 1995; 251(1):176–187. [PubMed: 7643386]
57. Han KF, Baker D. Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc Natl Acad Sci U S A*. 1996; 93(12):5814–5818. [PubMed: 8650175]
58. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res*. 2000; 28(1):254–256. [PubMed: 10592239]
59. Shindyalov IN, Bourne PE. An alternative view of protein fold space. *Proteins*. 2000; 38(3):247–260. [PubMed: 10713986]
60. Harrison A, Pearl F, Mott R, Thornton J, Orengo C. Quantifying the similarities within fold space. *J Mol Biol*. 2002; 323(5):909–926. [PubMed: 12417203]
61. Taylor WR. A ‘periodic table’ for protein structures. *Nature*. 2002; 416(6881):657–660. [PubMed: 11948354]
62. Friedberg I, Godzik A. Connecting the protein structure universe by using sparse recurring fragments. *Structure (Camb)*. 2005; 13(8):1213–1224. [PubMed: 16084393]
63. Krishna SS, Grishin NV. Structural drift: a possible path to protein fold change. *Bioinformatics*. 2005; 21(8):1308–1310. [PubMed: 15604105]
64. Lindorff-Larsen K, Rogen P, Paci E, Vendruscolo M, Dobson CM. Protein folding and the organization of the protein topology universe. *Trends Biochem Sci*. 2005; 30(1):13–19. [PubMed: 15653321]
65. Kolodny R, Petrey D, Honig B. Protein structure comparison: implications for the nature of ‘fold space’, and structure and function prediction. *Curr Opin Struct Biol*. 2006; 16(3):393–398. [PubMed: 16678402]

66. Kryshtafovych A, Milostan M, Szajkowski L, Daniluk P, Fidelis K. CASP6 data processing and automatic evaluation at the protein structure prediction center. *Proteins*. 2005; 61(7):19–23. [PubMed: 16187343]
67. Wang G, Jin Y, Dunbrack RL Jr. Assessment of fold recognition predictions in CASP6. *Proteins*. 2005; 61(7):46–66. [PubMed: 16187346]
68. Daniluk P, Kryshtafovych A, Fidelis K. Comparing protein structures using local structure descriptors. 2008 Manuscript.
69. Strombergsson H, Kryshtafovych A, Prusis P, Fidelis K, Wikberg JE, Komorowski J, Hvidsten TR. Generalized modeling of enzyme-ligand interactions using proteochemometrics and local protein substructures. *Proteins*. 2006; 65(3):568–579. [PubMed: 16948162]
70. Hvidsten TR, Kryshtafovych A, Komorowski J, Fidelis K. A novel approach to fold recognition using sequence-derived properties from sets of structurally similar local fragments of proteins. *Bioinformatics*. 2003; 19(2):II81–II91. [PubMed: 14534176]
71. Drabikowski M, Nowakowski S, Tiurny J. Library of local descriptors models the core of proteins accurately. *Proteins*. 2007; 69(3):499–510. [PubMed: 17623841]
72. Björkholm P, Daniluk P, Krystofovich A, Fidelis F, Andersson R, Hvidsten TR. Using multi-data hidden Markov models trained on local neighborhood of protein structure to predict residue-residue contacts. 2008 Manuscript.
73. Drabikowski, M. Assembling proteins structure from local descriptors. Warsaw: University of Warsaw; 2006. Groups and signals analyzes.
74. Hvidsten TR, Lægheid A, Kryshtafovych A, Andersson G, Fidelis K, Komorowski J. High-throughput protein function prediction using local substructures. 2008 Manuscript.
75. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*. 1995; 247(4):536–540. [PubMed: 7723011]
76. Kabsch W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr*. 1976; A32:922–923.
77. McLachlan AD. A mathematical procedure for superimposing atomic coordinates of proteins. *Acta Crystallogr*. 1972; A28:656–657.
78. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 1983; 22(12):2577–2637. [PubMed: 6667333]
79. Wadpole, RE., Meyers, RH. Probability and Statistics for Engineers and Scientists. New York: McMillan; 1985.
80. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*. 1999; 292(2):195–202. [PubMed: 10493868]
81. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997; 25(17):3389–3402. [PubMed: 9254694]

Figure 1.

(a) Three local descriptors centered on residues 79 (bottom), 173 (middle), and 192 (top) are illustrated as protein cartoons on the wireframe structure of the anaerobic cobalt chelatase (PDB code 1qgo, chain A). C_{β} atoms of the central residues of the descriptors are shown as dark balls. (b) The descriptor centered on residue 80 in the same protein as in (a). Because local descriptors incorporate side chain directionality, close residues pointing in opposite directions may result in very different descriptors. This is particularly well illustrated with the central residues 79 and 80 in this example [i.e., (a) bottom vs. (b)].

Figure 2. Optimal structural superpositions of descriptors similar to the (seed) descriptor (a) 1qgoa_#192 (202 descriptors) and (b) 1qgoa_#80 (378 descriptors) (descriptors are named using the following syntax “domain name ‘#’central amino acid”). Note that the seed descriptors of these two groups are also shown in Figure 1 (although slightly rotated). The corresponding sequence alignment for some of the proteins in group 1qgoa_#192 is shown in Table I.

Figure 3.
The number and size of descriptor groups (logarithmic scale) with different number of segments.

Figure 4. Structural coverage of the Nifb protein (PDB code 1O13, CASP comparative modeling target) and the SOR45 protein (PDB code 1TZA, CASP fold recognition target) with local structures from the group library. The bars show how many groups cover the particular residue in the sequence.

Figure 5.

Coverage of CASP targets using the local substructures from the group library. ALL, all targets assessed; CM, comparative modeling targets (87 domains); FR, fold recognition (56 domains); NF, new fold (15 domains). Splines have been fitted through the data for each of the four presented data sets.

Figure 6.

Frequencies for Valine and hydrophobic amino acids as a function of descriptor group size. The solid lines correspond to the upper and lower boundaries for the 90% confidence interval defining significant under- and over-representation. In general, smaller variation around the *a priori* frequencies is required for an observed frequency to be significant as the group size increases and more evidence is available. The small *a priori* frequencies for single amino acids give little room for significant underrepresentation, something that may only be observed for relatively large groups.

Figure 7.

The 479 test domains plotted against E-score to the closest training domain as given by PSI-BLAST, and P -value for the top fold assignment as given by the descriptor approach (E-scores were not available (NA) for targets where PSI-BLAST could not find any relation to any training domain, and these domains were therefore given random E-scores larger than 10). The test domains are plotted with different colors and shapes to show which method, if any, was able to correctly identify domain folds. Hollow shapes indicate that the descriptor approach did not identify the correct fold as its top assignment, but that the correct fold still was in one of the top five assignments given by this method. The horizontal line indicates the E-score threshold at which predictions from PSI-BLAST may no longer be trusted. The vertical line indicates the corresponding P -value threshold for the descriptor approach. This P -value threshold was selected to maximize the sum of sensitivity (i.e. fraction of correctly predicted test domains with a lower P -value) and specificity (i.e. fraction of incorrectly predicted test domains with a higher P -value). The Receiver Operating Characteristic (ROC) curve shows sensitivity versus specificity for the full spectrum of such P -value thresholds. The selected threshold $3.39\text{E-}14$ corresponds to the sensitivity of 0.84 and the specificity of 0.80.

Table 1

The sequence alignment of a local descriptor group. Sequence alignments resulting from the corresponding structure alignment of members in the descriptor group 1qgoa_#192 (Figure 1a). The complete group contains 202 descriptors from 178 different proteins. Dashes indicate that we could not find a structural fit for some parts of the segment.

Descriptor	Fold	RMS	Segment 1	Segment 2	Segment 3
1qgoa_#192	c92	0.00	DSLRDEGVTVGHL 185-197	AIQSL 137-141	AGIPA 227-231
1doza_#105	c92	1.27	AEMHKDGITEAVS 98-110	KMGLL 5-9	GGLTI 137-141
1hrka_#182	c92	1.16	EEMERDGLERAIA 175-187	KTGIL 68-72	PTMKW 217-221
1thfd_#165	c1	1.04	VEVEKRGAGEILL 158-170	VVVAI 125-129	TTLPI 194-198
1rpxa_#37	c1	1.16	KAIEQAGCDWIHV 30-42	VSPSI 13-17	TDLPL 67-71
1qdlb_#46	c26	1.55	---ERIDPRLII 42-51	LTLII 3-7	KRTPI 76-80
1fpza_#72	c45	0.76	EELKSCGIQDIFV 65-77	KTLIH 135-139	CGIIT 99-103
1b8oa_#108	c56	0.80	RVFRLLGVETLVV 101-113	QVAVI 26-30	CGLRV 231-235
1g2oa_#112	c56	1.04	RAARAAGAQI--- 105-114	DVAVV 29-33	AGAEV 219-223
1emsa2#186	d160	1.22	LWNRKRGAQLLSF 179-191	RLGLS 163-167	NQCYV 215-219

Table II

The sequence signal in the local descriptor library. The table shows the fractions of positions in the group library where the 20 amino acids and 10 additional substitution groups are significantly under/over-represented (a position is a specific column in a multiple alignment of corresponding segments in a descriptor group; there are 102674 such positions in all). At the significance level of 0.05, we would expect an amino acid to be significantly over-/under-represented in 5% of the positions by chance. Underlined numbers in bold indicate that the over-/under-representation occurred in more than 5% of the positions.

Amino acid	Significant under-representation (P<0.05)	Significant over-representation (P<0.05)	Amino acid or Substitution group	Significant under-representation (P<0.05)	Significant over-representation (P<0.05)
Alanine (A)	0.005	<u>0.081</u>	Serine (S)	0.004	0.041
Cysteine (C)	0.000	0.029	Threonine (T)	0.002	0.048
Aspartic acid (D)	0.008	0.045	Valine (V)	0.003	<u>0.126</u>
Glutamic acid (E)	0.010	0.047	Tryptophan (W)	0.000	0.027
Phenylalanine (F)	0.001	<u>0.069</u>	Tyrosine (Y)	0.001	<u>0.055</u>
Glycine (G)	0.011	<u>0.068</u>	Hydrophobic(AGILMFPPVV)	<u>0.139</u>	<u>0.267</u>
Histidine (H)	0.000	0.030	Charged (RDEHK)	<u>0.141</u>	<u>0.107</u>
Isoleucine (I)	0.002	<u>0.101</u>	Large (WQEKRYHF)	<u>0.178</u>	<u>0.119</u>
Lysine (K)	0.007	<u>0.043</u>	Medium (VILPM)	<u>0.111</u>	<u>0.233</u>
Leucine (L)	0.006	<u>0.104</u>	Small (TDNC)	0.040	<u>0.075</u>
Methionine (M)	0.000	0.035	Tiny (AGS)	<u>0.052</u>	<u>0.118</u>
Asparagine (N)	0.003	0.036	Aromatic (FWYH)	0.010	<u>0.091</u>
Proline (P)	0.007	0.035	Surface (RNDEQGHKSTY)	<u>0.298</u>	<u>0.147</u>
Glutamine (Q)	0.002	0.037	Beta branched (ITV)	0.025	<u>0.188</u>
Arginine (R)	0.004	0.047	Hydro bonded (RNDEQHKSTY)	<u>0.289</u>	<u>0.140</u>

Table III

Discriminatory capabilities within the training set. The table shows to which degree the extracted sequence signals from the library of descriptor groups were able to discriminate sequence domains in the training set. Sensitivity/specificity is the fraction of domains that was correctly recognized to contain/not contain the library substructure. Alignment is the highest fraction of correctly aligned fragments for a descriptor having a matching score higher than the acceptance threshold. All statistics were averaged over all descriptor groups.

Method	Sensitivity	Specificity	Alignment
Signal 1a	0.83	0.97	0.73
Signal 1b	0.83	0.97	0.73
Signal 2	0.70	0.96	0.65

Table IV

The test set. The table shows how the test set is distributed over the classes in the SCOP hierarchy. It shows the number of domains from each class and the number of folds from each class covered by these domains. Descriptor coverage is the average fraction of the sequence of a test domain structurally covered by at least one descriptor group.

SCOP class	No. domains	No. folds	Descriptor coverage
All alpha proteins (α)	49	12	0.32
All beta proteins (β)	133	24	0.61
Alpha and beta proteins (α/β)	167	29	0.54
Alpha and beta proteins ($\alpha+\beta$)	93	26	0.47
Multi-domain proteins (alpha and beta)	7	4	0.34
Membrane and cell surface proteins and peptides	10	3	0.28
Small proteins	20	2	0.17
Sum/average	479	100	0.50

Table V

Discriminatory capabilities on the external test set. The table shows the results of assigning local substructures, and inferring SCOP fold, for the 479 test domains. Sensitivity is the fraction of descriptor groups from the correct fold assigned to the target *and* the fraction of structurally matching groups correctly assigned to the target (not all substructures from a fold necessarily match all domains in that fold). Specificity is the fraction of groups from wrong folds not assigned. Alignment is the highest fraction of segments of a correctly assigned group that also were assigned to the correct positions along the target sequence. Identification of SCOP fold: Top1 is the fraction of test domains with the correct fold as the top prediction, while Top5 is the fraction of test domains with the correct fold as one of the top five predictions. Sequence identity correlation shows how the rank of the correct fold correlates with the sequence identity to the closest sequence in the training set. Coverage correlation shows how the rank of the correct fold correlates with target coverage (see Table IV). Notice that substructure sensitivity was calculated for the subset of 438 test domains that were structurally covered by at least one descriptor group, and that alignment was calculated for the subset of *these* groups that actually were assigned to the test domain (with a score higher than the acceptance threshold).

Method	Sensitivity (fold/substruct.)	Specificity	Alignment	Identification of SCOP fold			
				Top1	Top5	Seq. id. correlation	Coverage correlation
Signal 1a	0.46/0.49	0.96	0.56	0.70	0.84	-0.23	-0.14
Signal 1b	0.45/0.48	0.96	0.56	0.70	0.84	-0.19	-0.11
Signal 2	0.36/0.39	0.96	0.53	0.63	0.78	-0.22	-0.16
PSI-BLAST	-	-	-	0.69	0.69	-0.75	-