Published in final edited form as:

Methods Mol Biol. 2013 ; 932: 141–158. doi:10.1007/978-1-62703-065-6_9.

A modular perspective of protein structures; application to fragment based loop modeling

Narcis Fernandez-Fuentes¹ and Andras Fiser²

¹Leeds Institute of Molecular Medicine, Section of Experimental Therapeutics, University of Leeds, St. James's University Hospital, Leeds LS9 7TF, UK

²Department of Systems and Computational Biology, Department of Biochemistry Albert Einstein College of Medicine, 1301 Morris Park Ave, Bronx, NY 10461, USA

Summary

Proteins can be decomposed into supersecondary structure modules. We used a generic definition of supersecondary structure elements, so-called *Smotifs*, which are composed of two flanking regular secondary structures connected by a loop, to explore the evolution and current variety of structure building blocks. Here, we discuss recent observations about the saturation of Smotif geometries in protein structures and how it opens new avenues in protein structure modeling and design. As a first application of these observations we describe our loop conformation modeling algorithm, ArchPred that takes advantage of Smotifs classification. In this application, instead of focusing on specific loop properties the method narrows down possible template conformations in other, often not homologous structures, by identifying the most likely supersecondary structure environment that cradles the loop. Beyond identifying the correct starting supersecondary structure geometry, it takes into account information of fit of anchor residues, sterical clashes, match of predicted and observed dihedral angle preferences, and local sequence signal.

Keywords

Secondary structure; Supersecondary Structure; Smotif; Loop modeling; Protein Structure Evolution; Protein Structure Modeling; Protein Structure Design

1. Introduction

1.1. Recent advances in the analysis and definition of supersecondary structure motifs

Protein structures can be decomposed into folds (1) that are determined by the number, arrangement, and connectivity (topology) of secondary structure elements (2). An intermediate structural level between folds and secondary structures are the supersecondary structure elements that are composed of a number of regular secondary structure elements linked by loops (e.g. Rossmann, helix-turn-helix, four strand Greek key, β -meander motifs). Folds can be perceived as an overlapping combination of various supersecondary elements (Fig.1). These supersecondary structure elements are sometimes shared among different folds and sometimes highly repeated within the same one. Some of these supersecondary structure elements are frequently observed because these are either connected to a general functional role (e.g. DNA or co-factor binding) or they present a thermodynamically advantageous structural arrangement, which might have emerged several times during evolution. This latter observation prompted the theory of relic peptide world (3), which proposes that modern, thermodynamically stable proteins are the results of duplication, mutation, shuffling and fusion of a limited set of relic peptides.

Several reports in the recent literature have explored the idea of describing protein structures as superstructures formed by the assembly of a limited number of supersecondary elements. A recent update on the idea of ancient relic peptides found that protein fragments of up-to 20 to 40 residues long can co-occur in different structural contexts suggesting that these to be an ancestral pool of peptide modules (4). Various efforts tried to explore possible toolsets of supersecondary elements, such as antiparallel β -sheets (5), $\alpha\beta\beta$ and $\beta\beta\alpha$ motifs (6), $\alpha\alpha$ -turn motifs (7), or four helix bundles (8). Other studies identified similar structural elements as possible building blocks of structural hierarchy using different approaches. The so-called *Closed Loops* were identified by their close C_{α} - C_{α} contacts from solution structures and found to have a nearly standard size (27 residues ± 5) in agreement with the theoretical optimal size for loop closure derived from polymer statistics (9–11). In another approach, dynamic Monte Carlo simulations of a C_{α} lattice model was used to identify the nearest neighbor residues, labeled as *most interacting residues*, which serve as anchors for protein folding (12). It was found that anchor residues are conserved hydrophobic clusters that keep together the so-called *Tightened End Fragments*, which essentially correspond to the Closed Loop definition. A related, more generic analysis identified fragments that are shared by different folds using an arbitrary length of 5,10,15,20 residues (13) and those were used to establish structural and functional relationships among folds. Voigt et al. introduced peptide “schemas” as subunits of the protein structure that interact the least with their environment. It was assumed that these schemas can be easily swapped among proteins, as these will have the least impact in disturbing the interaction network within a fold (14). Peptide schemas are composed of 20–30 residues and typically are bundled α -helices, α -helices combined with β -strands and β -strands connected by a hairpin turn. The computationally predicted schemas were explored experimentally in five proteins and were found to be viable building units for recombinant hybrid proteins. Nussinov and co-workers defined possible building blocks to better explain the hierarchical nature of folding process (15, 16). The building blocks were determined computationally by a stepwise dissection procedures and a scoring function that was assessing compactness, isolation and hydrophobicity. Building blocks were found sometimes to be as small as a single secondary structure or more complex supersecondary structures and in general these were assumed to be the most highly populated conformations in solution.

2. Materials

2.1. Smotif, a general supersecondary structure building block

In order to systematically explore modularity in all known protein structures, we used a general, supersecondary structure classification (17). In this approach a basic supersecondary motif, which we will refer to as *Smotif*, is composed of two regular secondary structure elements linked by a loop. These motifs may or may not serve as possible units for structural evolution; however the definition carries the advantage that one can employ automated algorithms to systematically explore the shape and occurrence of these motifs in all known folds (Fig.1).

Smotifs are characterized in protein structures by the types of sequential secondary structures and the geometry of the orientation of the secondary structures with respect to each other, as described by four internal coordinates, introduced by Oliva et al. (17, 18). Therefore a protein structure can be expressed as a string of overlapping Smotifs. If we used a definition that has higher number of connected secondary structures e.g. 3 or more, the number of combinations would be very large and would prevent us from a systematic classification. Meanwhile, any combination of Smotifs is possible, both sequentially and spatially, which can recapitulate earlier described, more complicated supersecondary structure motifs.

The geometry of Smotifs captures the local structural arrangement of the two flanking secondary structures SS1 and SS2 and is defined by four internal coordinates: a distance and three angles (Fig. 2) (17, 18). D is a distance, expressed in Angstroms, and the typical D values range from 2 and 16 Angstroms. As expected, D mainly depends on the number of residues contained in the loop region. The hoist (δ), pack (θ) and rho (ρ) angles range from 0 to 180, 180 and 360 degrees, respectively.

The geometrical values are distributed in a continuous space, e.g. the hoist angle ranges between 0 to 180 degrees. In order to compare Smotif geometries, the parameter space of geometrical values were binned, where the four internal coordinates define each bin. A range of binning sizes and parameter intervals were explored for the four variables in order to get the sharpest partitioning of the geometrical space with the smallest number of possible bins. The quality of the binning was assessed by calculating the Root Mean Square Deviation (RMSD) and the LGA scores (19) upon structural superposition for all Smotifs that were classified in the same or different geometrical bins. The optimal bin partitioning for each parameter was obtained by studying the distribution of distance and angle values of Smotifs in SCOP 1.71 database proteins (20) and resulted in only 324 types of Smotif definitions using the following binning values: 4 Å bins for distance, 60 degree bins for δ and θ starting at 0 degree, and 60 degree bins for ρ , starting at 30 degree. At this level of bin resolution the RMSD upon structural superposition of more than 75% of Smotifs that belong to the same geometrical bin falls below 1 Å.

Smotifs extracted from protein structures are organized in a library using a two-level hierarchical classification. In the first hierarchical level, Smotif are grouped by type, i.e. Smotifs are identified according to the type of bracing secondary structures: $\alpha\alpha\alpha\beta\beta\alpha$ and $\beta\beta$ according to the definition of secondary structure by the DSSP program (21). In the second hierarchical level, Smotifs are grouped according to their geometry bin, as described above. We have used this library of Smotifs to explore the modular nature of fold organization, both in terms over evolutionary time and the current fold space varieties. Subsequently, we utilized the spatial restraints enforced by the bracing secondary structures within a Smotif we set up a prediction algorithm to model the conformation of the connecting loop segments.

3. Methods

3.1. Saturation of Smotifs in known folds

As we described above, folds can be dissected into their Smotif building blocks, and one can express protein folds as a unique string of overlapping Smotifs (Fig. 1). The frequency of occurrence of Smotifs in all known protein folds was explored, and in addition we studied the increase of coverage of Smotifs in Protein Data Bank (PDB) over time (Fig. 3). Briefly, all protein structures that were identified as “new folds” from SCOP (20) releases 1.73 and 1.75 and CASP 3–6 meetings (22) were decomposed into Smotifs and these were compared to libraries of Smotifs extracted from backdated versions of the PDB. Within the pairs of datasets we evaluated the existence of identical Smotifs in the novel folds and the previously defined folds. The first comparison was based on the type of secondary structures and the geometry (D , hoist, packing, and meridian) of Smotifs. In a second, stricter comparison, the lengths of the flanking secondary elements (SS1 and SS2) were also compared. If these lengths differed by more than 2 or 4 residues in the case of strands or helices, respectively, the Smotifs were considered different.

We found that proteins that were considered novel folds at CASP 3–6 meetings (years 1998–2004) and in SCOP 1.73, 1.75 (years 2007–2009) do not have any novel Smotif geometries that were not present in previously solved structures. In other words, none of the

Smotifs of novel folds have a unique geometry. For instance, as early as the third round of CASP Meetings in 1998, all of the targets identified as novel folds by the experts could have been reconstructed using Smotifs from known protein structures. In an even stricter comparison, i.e. requiring not just a match in the geometry but also identical lengths of the flanking secondary structures, still less than 6% of the Smotifs in novel folds at CASP meetings would not have a match in already known structures. Similarly, we checked the motif composition of new folds from the archives of SCOP in the 1.73 (2007 November) and 1.75 (2009 June) releases. These contain a total of 233 new folds from 1140 proteins. Similar to the CASP targets, none of these novel folds had a Smotif whose geometry was not already observed in a previously known fold and only 1% of Smotifs has a flanking secondary structure with unique length. Initially, we found 47 Smotifs (out of the 8056 analyzed) that appeared to be new. However, after manual inspection, it turned out that these were artifacts originating from replacing obsolete PDB entries with newer ones (and thus have a newer deposition date assigned).

The above observations suggest that recently solved new folds do not imply the emergence of new Smotifs, and that a protein with a novel fold can be reconstructed using Smotifs from already existing protein folds. Approximately 10 years ago all categories of Smotifs were already represented by at least one example in a known fold. For instance, T0181 (PDB code: 1nyn), a new fold submitted to CASP5, at the time could have been constructed from 7 overlapping Smotifs, all of which can be located in previously solved structures of other proteins representing a variety of different folds (Fig. 4).

3.2. What makes a protein fold new?

Since the repertoire of Smotifs seems to have come close to saturation (Fig. 3) (23), this prompts the question of what is really unique about a fold structure when it is identified as “novel”. When we explored the frequency of occurrence of Smotifs in the non-redundant set of known folds, we observed that novel folds have a larger fraction of Smotifs that have a low frequency of occurrence in the PDB. On the other hand, superfolds (24), those that are adopted by many different sequences (TIM barrels, OB fold, IG superfamily etc) often with different functions, are built by Smotifs that occur with medium or high frequencies in existing folds. This implies that novel folds are composed of a new permutation of existing Smotifs and, specifically, a structure will have a greater likelihood of being “novel” if the structure is enriched with rarely occurring Smotifs.

Another plausible explanation would be to combine, otherwise common Smotifs, in an unusual sequence, and thus resulting in a new topology. To explore this, we calculated a Novelty Z-score for each protein (25), which was obtained from the product of individual Smotif frequencies. The hypothesis is that if the Novelty Z-score of some novel folds is similar to that of known folds, then the novelty for these cases must be a consequence of a never before seen combination of otherwise common Smotifs rather than a result of being constructed from rare Smotifs. Most novel folds are indistinguishable from already known structures in terms of their overall Novelty Z-scores, which indicates that these structures may indeed be just a new topological arrangement of common Smotifs. This means that although novel folds are often built using a higher proportion of rare Smotifs, in many cases these folds are novel because their Smotifs are assembled in an unusual sequence.

These observations open up new venues for structure modeling and design. Fragment-based approaches in structure prediction are becoming the most successful and preferred approaches, especially in the case of new folds (26, 27). These approaches rely on a library of short fragments, which are not necessarily biologically meaningful (usually too short). Using Smotifs for this purpose would drastically decrease the degree of freedom that needs to be explored in the sampling procedure. The remaining bottleneck in all these approaches

is to establish a detectable sequence signal, which can relate the sequence of Smotifs to their conformation. Another venue where Smotifs may prove to be useful is protein design. If one assumes that all building blocks of known and to be discovered folds are already known and present in the current Smotif library, then the remaining challenge is to identify those new combinations of Smotifs which can result in a stable fold for a protein.

3.3. Saturation of loops in databanks

We also approached the previous question on completeness from the perspective of loop fragments, based on the hypothesis that similar structural environments within a Smotif probably also enforce similar connecting segments, i.e. loop conformations. We explored the fraction of loops that is extracted from all known protein sequences, called the *Sequence Space*, that are covered by loops extracted from all known protein structures, called the *Structure Space*. This will estimate the current structural coverage of short segments in the Sequence Space, i.e. in the entire set of known sequences.

Smotifs from Structure Space were structurally clustered after an all-to-all comparison and sequence identity cut-offs that ensured structural similarities were identified for each loop length. In the range of 42 to 55% pair-wise sequence identity we found a sharp transition between high to low RMSD values at all lengths (4–14 residue long loops were investigated), suggesting that a 50% sequence identity generally guarantees structural similarity (Fig. 5). Next, all possible loops from the Sequence Space were matched with the sequences from Structure Space, and the coverage assessed. In year 2005, below 40% sequence identity only 20 and 10% of long loops (of length 13 and 14) from Sequence Space could not be matched to at least one loop from Structure Space, while all other loop lengths matched at 100%. Meanwhile all loops (100%) of length 8 from Sequence Space 2005 have at least one loop in Structure Space at 50% or larger sequence identity. The percentages of coverage at 50% sequence cut-off dropped to 96, 94, 68, 53, 33 and 13% for loops of length 9, 10, 11, 12, 13, and 14, respectively. There were no loops at any length up to 14 residues that did not match with a conformational segment that shared at least 20% of sequence identity (Fig. 5 inset).

Finally, we also investigated the growth and change in the databases by repeating these exhaustive comparisons between sequence and structure databases that were available in 2001 and later. We focused our analysis on "medium" and "long" loops that were in the range of 8 to 14 residues. Following the same approach described above, we compared Sequence Space 2005 against Structure Space 2001 and we found that while sequence databases kept growing at an exponential rate, there was almost no unique conformational segments deposited up to 12 residues long fragments during the last 5 years.

3.4. Smotifs and loop modeling

In the absence of an experimentally described structure, computational methods such as comparative modeling, threading or *ab initio* can be used to provide a useful 3D model and fill the gap between the number of sequences and structures. Among these, comparative modeling is currently the most accurate but it is applicable only for that part of the target protein where a suitable template is found (28, 29). Most notably, insertions in the target sequence or segments with very different sequences to the template are not possible to model with this technique. Even above 40% sequence identity, where the core of the fold is well preserved and can be aligned accurately, the surface exposed variable loops can vary substantially among the homologues. Meanwhile loops often represent an important part of the protein structure and functional differences between the members of the same protein family and especially among members of superfamilies are usually a consequence of structural differences of exposed loop regions (30). Active and binding site residues are

more likely to be found in loop regions (31). Text book examples of functionally important loops include antibody complementary determining regions (32), ligand binding sites (ATP (33), calcium binding sites (34), NAD (P) (35)), DNA binding (36) or enzyme active sites (e.g. Ser-Thr kinases (37) or serine proteases (38)). Therefore the accuracy of loop conformations is often a key determinant of the usefulness of computational or experimental models.

Many loop-modeling procedures have been described in the past (39). Similarly to the prediction of protein structures there are *ab initio* (conformational search) methods (40–42), and database search (or knowledge-based) methods (43–45). There are also hybrid procedures that combine the two (7, 46, 47). In *ab initio* prediction a conformational search or enumeration of conformations is conducted in a given environment, guided by a scoring or energy function (41, 42). There are many such methods, exploiting different protein representations, sampling methods, energy function terms and optimization or enumeration algorithm (39). Knowledge-based methods (48), (also known as database search methods) are essentially a loop modeling application of comparative modeling that rely on a library of fragments as templates extracted from known protein structures and on a selection algorithm. Usually, many different alternative segments are obtained, and then sorted according to a variety of criteria, such as geometrical fit or sequence similarity between the template and target loop sequences. We took advantage of the geometrical restraints that are imposed by Smotifs to drive the selection of fragments in loop modeling.

3.5. Overview of the ArchPRED loop modeling method

ArchPRED is a knowledge-based approach to loop modeling that relies on a library of Smotifs and an algorithm containing selection, filtering and ranking steps. In a first step, Smotifs are selected from the library using the geometrical restraints imposed by the bracing secondary structures of the missing loop. Loop length and the type of flanking secondary structure (e.g. α - β) are also used during the selection step. In the second step, at the filtering stage, loops are discarded if the RMSD of their stem residues and the interactions between candidate loops and the new protein environment are unfavorable (i.e. sterical clashes). Finally in the third step, at the ranking stage, the remaining candidate loops are sorted by a composite Z-score. The Z-score function combines a sequence score using a conformational similarity weight matrix, and a Φ/ψ main chain dihedral angle propensities score (Fig. 6).

The Smotif library currently classifies 430,000 high quality loop structures. Smotif Library is organized in a three level hierarchy: loops are identified and grouped according to (1) the type of bracing secondary structures; (2) the length of the loop region and (3) four internal coordinates of the bracing secondary structures as defined by a distance vector between the anchor points and three angles: hoist (δ), packing (θ) and meridian (ρ) (18), which bins loops into $20 \times 6 \times 6 \times 8 = 5760$ possible cells or geometrical combinations. Not all cells are equally populated, e.g. short loops cannot have large distance values or $\beta\beta$ -hairpin loops have a restricted geometry in terms of possible angle combinations due to strict hydrogen bond requirements that defines them (49).

3.6 Selecting, filtering, ranking of candidate loops in ArchPRED

The search algorithm is composed of three consecutive steps: (1) selection, (2) filtering and (3) ranking of the suitable segments from the Smotif library. During the (1) selection step, loops queried in a stepwise manner: first, by similar bracing secondary structures, and those having a similar length (± 1 residue) to the query loop. The next selection step in the lookup process involves comparing the four internal coordinates of Smotifs. This initial selection of candidate loops by geometrical requirements quickly narrows the space to be explored by subsequent, more elaborate structural comparisons. For instance, for loops of

lengths 4, 8 and 12, the average number of selected loops by stem residue distances comparison on 50 randomly chosen examples (with a tolerance of 1 Å) is 1534, 683 and 430; while the selected number of loops after geometrical comparison is only 181, 85 and 25, respectively. In the meantime, this filtering step does not eliminate good candidate loops. Comparing the average $\text{RMSD}_{\text{local}}$ of the best fragment between loops that are selected by end point distances and loops selected by geometry, the differences are less than 0.05, 0.09 and 0.11 Å for the test sets of 4, 8 and 12 residue long loops, respectively.

After the initial selection, a two stage, (2) filtering step follows, which checks for the fit of stem residues by superposition of main chain atoms and RMSD calculation and evaluates of steric clashes between the loop and the rest of the protein environment. RMSD cut-offs for superposed stem residues have been applied before in loop structure prediction method either for ranking (50) or filtering (51). RMSD fit of stem residues correlate strongly with the accuracy of prediction of short loops, but this correlation becomes less pronounced for longer loops. The reason is that longer loops (8–14 residues) have more flexibility and their conformations is less restricted by the stem residues than in case of short loops (1–7 residues). Therefore we applied a range of RMSD stem cut-off values as a function of loop length. The second descriptor to filtering of loops explores the conformational fit of candidate loops in the new protein environment. Each candidate loop is structurally fitted in the new protein environment of the query protein and the steric hindrances between the loop and its structural environment are assessed. After these steps the average number of loop candidates in the test sets decreased to 81, 35 and 5 for loops of length 4, 8 and 12, respectively.

In the last, third step (3), the remaining candidate loops are ranked according to sequence similarity and amino acid Φ/Ψ dihedral angle propensities. Sequence and propensity scores have their own range and correlation with prediction accuracy; therefore these scores were converted into Z-scores in order to unify both scores with a comparable and dimensionless criterion. Sequence Z-score gauges the similarity between the sequence of the query and candidate loops and compares it to a reference distribution of randomly selected pairs of loops with similar lengths. A number of different substitution matrices were tested to score sequence similarity and the K3 weight matrix proved to be the most efficient (52) as it was derived from comparisons of Ramachandran maps and was developed to select protein fragments with similar conformations. The second quantitative measure to rank the set of candidate loops is the propensity of amino acids to adopt a specific Φ/Ψ main chain dihedral angle conformation. Propensity is defined as the likelihood that an amino acid residue is found in specific backbone dihedral angles Φ and Ψ . The expected propensity values were obtained from a table that divides the Ramachandran plot into 15 different regions (“p15 propensity” table) (53). The logarithm of the propensity approximates the free energy of a specific residue conformation. The free energy for each position is assumed to be additive, so the score for a sequence fragment is the sum of the log of the propensities at each position (53). The composite Z-score is defined as the sum of the two types of Z-scores.

3.7 Performance of ArchPRED

Benchmarking loop prediction approaches using knowledge-based methods is not straightforward. Some sort of artificially filtered input library needs to be prepared to avoid trivial hits and consequently the overestimation of performance. However, if one overly ambitious in getting rid of all segments in a database that show any level and type of similarity to a query may end up with seriously underestimated method performance. For benchmarking purposed a filtered library was derived by removing any Smotif extracted from protein that share the same SCOP (20) superfamily with the query loop. The performance was compared to a competitive and freely available *ab initio* prediction

method: ModLoop (54) and with the theoretical minimum RMSD, i.e. selecting the best candidate in the library, and thus informing on the practical limits of the method.

The minimum value of RMSD that can be obtained with loops available in the Smotif library (i.e. the loop with the smallest RMSD) are on average 0.25, 0.5 and 1 Å more accurate (for 4, 8 and 12 residues long loops, respectively) than the best results obtained by ModLoop, thus indicating that there are good candidate loops in the library and for all loop lengths (Fig. 7). This observation agrees with our analysis in database completeness (see section 3.3) and also agrees with the conclusions by Du et al (55) and Choi et al. (56), who found that even for long loops there are suitable candidates in the current database of structures with an RMSD of 2 Å or less. Therefore the bottleneck in knowledge-based loop modeling does not appear to be the sampling (completeness of database segments), but the search algorithm and scoring function to locate these segments. ModLoop on average outperformed ArchPRED for all loop lengths if we force the current search algorithm to locate a segment for all possible queries even if these are not very good candidates (i.e. falling below a certain cut-off Z-score). However, averages of both methods fall within the boundaries of 1 Å standard deviation. Finally, the accuracy obtained with ArchPRED is clearly much higher than the accuracy obtained with a random prediction (Fig. 7).

3.8 Coverage versus accuracy: identifying confidence Z-score thresholds

It is important to assign confidence values to a prediction and for that we explored the performance of the method as a function of Z-scores. Z-score cutoffs were defined in such a way that Smotifs selected with more significant Z-scores will have equal or better accuracy than the average accuracy of fragments obtained by ModLoop. As expected, the RMSD values decrease as the composite Z-scores and the accuracy of predictions increase; however the corresponding coverage of the prediction decreases. For loops between lengths of 4–7 residues a Z-score of 1.0 gives an equal or better performance than ModLoop with a corresponding coverage of around 90%. In the case of loops between 8–11 residues a Z-score larger than 2 to 3 is required and the average coverage is around 50% to 60%. For longer loops, beyond 12 residues long the coverage rapidly drops (Table 1).

3.9. ArchPRED web server

ArchPRED is implemented as web server for the modeling of missing protein loops in protein structures. Users are required to provide the query structure in PDB format and define the sequential location of the missing loop. The user can also select whether to query the Smotif database by geometry or by Euclidian distance of the stem residues. If geometry is selected, then the type of bracing regular secondary structural elements (e.g. α – β) have to be defined and if these elements are beta strands then further distinguish between hairpin or link types. Once the prediction is completed, results are sent by email in form of a link pointing to a temporary web page. Optionally, users can select to re-build the side chains of predicted loops and to perform a limited minimization (conjugate gradient minimization) to anneal the stems in the protein framework. The server is accessible at <http://www.fiserlab.org/servers/ArchPRED>. (Fig. 8)

Acknowledgments

This work was supported by NIH grant R01GM096041. This review is partially based on our previous publications of (17, 23, 25, 57). NFF acknowledges support from the Research Councils UK under the RCUK Academic Fellowship scheme.

References

1. 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; 247:536. [PubMed: 7723011]
2. Hadley C, Jones DT. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure.Fold.Des.* 1999; 7:1099. [PubMed: 10508779]
3. Lupas AN, Ponting CP, Russell RB. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J.Struct.Biol.* 2001; 134:191. [PubMed: 11551179]
4. Alva V, Remmert M, Biegert A, Lupas AN, Soding J. A galaxy of folds. *Protein Sci.* 19:124–130. [PubMed: 19937658]
5. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. *J.Mol.Biol.* 1993; 233:123. [PubMed: 8377180]
6. Boutonnet NS, Kajava AV, Rooman MJ. Structural classification of alphabeta and betabetaalpha supersecondary structure units in proteins. *Proteins.* 1998; 30:193–212. [PubMed: 9489927]
7. Wintjens RT, Rooman MJ, Wodak SJ. Automatic classification and analysis of alpha alpha-turn motifs in proteins. *J Mol Biol.* 1996; 255:235–253. [PubMed: 8568871]
8. Presnell SR, Cohen BI, Cohen FE. A segment-based approach to protein secondary structure prediction. *Biochemistry.* 1992; 31:983. [PubMed: 1734974]
9. Berezovsky IN, Grosberg AY, Trifonov EN. Closed loops of nearly standard size: common basic element of protein structure. *FEBS Lett.* 2000; 466:283–286. [PubMed: 10682844]
10. Trifonov EN, Frenkel ZM. Evolution of protein modularity. *Curr Opin Struct Biol.* 2009; 19:335–340. [PubMed: 19386484]
11. Chintapalli SV, Yew BK, Illingworth CJ, Upton GJ, Reeves PJ, Parkes KE, Snell CR, Reynolds CA. Closed loop folding units from structural alignments: experimental foldons revisited. *J Comput Chem.* 31:2689–2701. [PubMed: 20839296]
12. Papandreou N, Berezovsky IN, Lopes A, Eliopoulos E, Chomilier J. Universal positions in globular proteins. *Eur J Biochem.* 2004; 271:4762–4768. [PubMed: 15606763]
13. Friedberg I, Godzik A. Connecting the protein structure universe by using sparse recurring fragments. *Structure.* 2005; 13:1213–1224. [PubMed: 16084393]
14. Voigt CA, Martinez C, Wang ZG, Mayo SL, Arnold FH. Protein building blocks preserved by recombination. *Nat Struct Biol.* 2002; 9:553–558. [PubMed: 12042875]
15. Tsai CJ, Maizel JV Jr, Nussinov R. Anatomy of protein structures: visualizing how a one-dimensional protein chain folds into a three-dimensional shape. *Proc Natl Acad Sci U S A.* 2000; 97:12038–12043. [PubMed: 11050234]
16. Tsai CJ, Polverino de Laureto P, Fontana A, Nussinov R. Comparison of protein fragments identified by limited proteolysis and by computational cutting of proteins. *Protein Sci.* 2002; 11:1753–1770. [PubMed: 12070328]
17. Fernandez-Fuentes N, Oliva B, Fiser A. A supersecondary structure library and search algorithm for modeling loops in protein structures. *Nucleic Acids Res.* 2006; 34:2085–2097. [PubMed: 16617149]
18. Oliva B, Bates PA, Querol E, Aviles FX, Sternberg MJ. An automated classification of the structure of protein loops. *J.Mol.Biol.* 1997; 266:814. [PubMed: 9102471]
19. Zemla A. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res.* 2003; 31:3370–3374. [PubMed: 12824330]
20. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.* 2008; 36:D419–D425. [PubMed: 18000004]
21. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers.* 1983; 22:2577–2637. [PubMed: 6667333]
22. Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol.* 2005; 15:285–289. [PubMed: 15939584]

23. Fernandez-Fuentes N, Fiser A. Saturating representation of loop conformational fragments in structure databanks. *BMC Struct Biol.* 2006; 6:15. [PubMed: 16820050]
24. Orengo CA, Pearl FM, Bray JE, Todd AE, Martin AC, Lo CL, Thornton JM. The CATH Database provides insights into protein structure/function relationships. *Nucleic Acids Res.* 1999; 27:275. [PubMed: 9847200]
25. Fernandez-Fuentes N, Dybas JM, Fiser A. Structural characteristics of novel protein folds. *PLoS Comput Biol.* 6 e1000750.
26. Zhang Y. Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins.* 2007; 69(Suppl 8):108–117. [PubMed: 17894355]
27. Das R, Baker D. Macromolecular Modeling with Rosetta. *Annu Rev Biochem.* 2008
28. Fiser A, Feig M, III CLB, Sali A. Evolution and Physics in Comparative Protein Structure Modeling. *Acc Chem Res.* 2002; 35:413–421. [PubMed: 12069626]
29. Baker D, Sali A. Protein structure prediction and structural genomics. *Science.* 2001; 294:93–96. [PubMed: 11588250]
30. Blouin C, Butt D, Roger AJ. Rapid evolution in conformational space: A study of loop regions in a ubiquitous GTP binding domain. *Protein Sci.* 2004; 13:608–616. [PubMed: 14978301]
31. Fiser A, Simon I, Barton GJ. Conservation of amino acids in multiple alignments: aspartic acid has unexpected conservation. *FEBS Lett.* 1996; 397:225. [PubMed: 8955352]
32. Kim ST, Shirai H, Nakajima N, Higo J, Nakamura H. Enhanced conformational diversity search of CDR-H3 in antibodies: role of the first CDR-H3 residue. *Proteins.* 1999; 37:683–696. [PubMed: 10651282]
33. Saraste M, Sibbald PR, Wittinghofer A. The P-loop--a common motif in ATP- and GTP-binding proteins. *Trends Biochem Sci.* 1990; 15:430–434. [PubMed: 2126155]
34. Kawasaki H, Kretsinger RH. Calcium-binding proteins 1: EF-hands. *Protein Profile.* 1995; 2:297–490. [PubMed: 7553064]
35. Wierenga RK, Terpstra P, Hol WG. Prediction of the occurrence of the ADP-binding beta alpha beta-fold in proteins, using an amino acid sequence fingerprint. *J Mol Biol.* 1986; 187:101–107. [PubMed: 3959077]
36. Tainer JA, Thayer MM, Cunningham RP. DNA repair proteins. *Curr Opin Struct Biol.* 1995; 5:20–26. [PubMed: 7773744]
37. Johnson LN, Lowe ED, Noble ME, Owen DJ. The Eleventh Datta Lecture. The structural basis for substrate recognition and control by protein kinases. *FEBS Lett.* 1998; 430:1–11. [PubMed: 9678585]
38. Wlodawer A, Miller M, Jaskolski M, Sathyanarayana BK, Baldwin E, Weber IT, Selk LM, Clawson L, Schneider J, Kent, et al. Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease. *Science.* 1989; 245:616–621. [PubMed: 2548279]
39. Fiser A, Do RK, Sali A. Modeling of loops in protein structures. *Protein Sci.* 2000; 9:1753. [PubMed: 11045621]
40. Fine RM, Wang H, Shenkin PS, Yarmush DL, Levinthal C. Predicting antibody hypervariable loop conformations. II: Minimization and molecular dynamics studies of MCPC603 from many randomly generated loop conformations. *Proteins.* 1986; 1:342. [PubMed: 3449860]
41. Moult J, James MN. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins.* 1986; 1:146. [PubMed: 3130622]
42. Brucoleri RE, Karplus M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers.* 1987; 26:137. [PubMed: 3801593]
43. Jones TA, Thirup S. Using known substructures in protein model building and crystallography. *EMBO J.* 1986; 5:819. [PubMed: 3709525]
44. Chothia C, Lesk AM. Canonical structures for the hypervariable regions of immunoglobulins. *J.Mol.Biol.* 1987; 196:901. [PubMed: 3681981]
45. Fidelis K, Stern PS, Bacon D, Moult J. Comparison of systematic search and database methods for constructing segments of protein structure. *Protein Eng.* 1994; 7:953. [PubMed: 7809034]
46. Deane CM, Blundell TL. CODA: a combined algorithm for predicting the structurally variable regions of protein models. *Protein Sci.* 2001; 10:599. [PubMed: 11344328]

47. Martin AC, Cheetham JC, Rees AL. Modeling antibody hypervariable loops: A combined algorithm. *PNAS*. 1989; 86:9268–9272. [PubMed: 2594766]
48. Greer J. Comparative model-building of the mammalian serine proteases. *J.Mol.Biol.* 1981; 153:1027. [PubMed: 7045378]
49. Gunasekaran K, Ramakrishnan C, Balaram P. Beta-hairpins in proteins revisited: lessons for de novo design. *Protein Eng.* 1997; 10:1131–1141. [PubMed: 9488138]
50. Michalsky E, Goede A, Preissner R. Loops In Proteins (LIP)--a comprehensive loop database for homology modelling. *Protein Eng.* 2003; 16:979. [PubMed: 14983078]
51. Heuser P, Wohlfahrt G, Schomburg D. Efficient methods for filtering and ranking fragments for the prediction of structurally variable regions in proteins. *Proteins*. 2004; 54:583–595. [PubMed: 14748005]
52. Kolaskar AS, Kulkarni-Kale U. Sequence alignment approach to pick up conformationally similar protein fragments. *J Mol Biol.* 1992; 223:1053–1061. [PubMed: 1538389]
53. Shortle D. Composites of local structure propensities: evidence for local encoding of long-range structure. *Protein Sci.* 2002; 11:18–26. [PubMed: 11742118]
54. Fiser A, Sali A. ModLoop: automated modeling of loops in protein structures. *Bioinformatics*. 2003; 19:2500. [PubMed: 14668246]
55. Du P, Andrec M, Levy RM. Have we seen all structures corresponding to short protein fragments in the Protein Data Bank? An update. *Protein Eng.* 2003; 16:407. [PubMed: 12874373]
56. Choi Y, Deane CM. FREAD revisited: Accurate loop structure prediction using a database search algorithm. *Proteins*. 78:1431–1440. [PubMed: 20034110]
57. Fernandez-Fuentes N, Zhai J, Fiser A. ArchPRED: a template based loop structure prediction server. *Nucleic Acids Res.* 2006; 34:W173–W176. [PubMed: 16844985]

Figure 1.

A protein structure decomposed into a set of overlapping Smotifs. Secondary structure elements at the bottom are shown in cylinders and arrows representing α -helices and β -strands, respectively.

Figure 2.

Definition of the geometry of Smotifs. The geometry of each Smotif is defined by four internal coordinates: a distance, D and three angles: hoist (δ), pack (θ) and meridian (ρ). D is the distance between the last and first C α of the Nt and Ct secondary structure respectively. $M1$ and $M2$ are defined by the shortest of the principal moments of inertia of the flanking secondary structures: δ , the angle between axis $M1$ and vector of D ; θ , the angle between $M1$ and $M2$; and ρ , the angle between $M2$ and the plane that contains the vector $M1$.

Figure 3. Emergence of new Smotif geometries in PDB. The plot shows the cumulative frequency distribution as a function of time for α - α (red), α - β (blue), β - α (green), and β - β (black) Smotifs.

Figure 4. Reconstruction of protein YHR087W (PDB code 1nyn) from *Saccharomyces cerevisiae* using Smotifs from unrelated structures. 1nyn was submitted to CASP5 meeting in the category of “*New Folds*”; however all the Smotifs were available from structures that were solved before.

Figure 5.

Saturation of loop conformations in databases. The plot shows the relationship between structure similarity (y-axis, measured as the RMSD of main chain atoms) as a function of sequence identity (x-axis) for loops of length 8, 9, 10, 11, 12, 13 and 14 shown in red, blue, orange, green, dark blue, brown and black, respectively. Inset shows the cumulative frequency distribution of loops that can be matched up between all known sequences (sequence space) and the available structural conformations (structure space) at a given sequence identity.

Figure 6.

Overview of ArchPRED algorithm. The prediction algorithm is composed of two major components: (i) a library of Smotifs and (ii) a selection, filtering and scoring algorithm. Smotifs that fit the geometry of the query loop are fitted in the new protein and filtered by RMSD of stem residues and steric clashes and finally scored by a composite Z-scores that combines sequence and propensity information.

Figure 7. Prediction accuracy (average RMSD of mainchain atoms in angstrom) of ArchPRED, Modloop and theoretical limit of prediction as a function of loop length. Black line represents the practical upper limit of the prediction accuracy, i.e. selecting the best available Smotif in the library. Red and blue represent ArchPRED and ModLoop prediction respectively and solid green line shows the average RMSD for random prediction, i.e. a random selection of Smotifs from the library.

ArchPRED Server
A template based loop structure prediction server

Structure Upload

Upload structure file here:

Prediction Parameters

Select loops from Search Space based on

☒ End Point Distance
☐ Geometry

Loop start position Chain ID

Loop sequence (Plain text & maximum length: 14)

Do you want to activate Clashes Filter? ☒ Yes ☐ No

of predictions to include in output Z-Score Cut-Off

Post-Prediction Parameters

Side chain reconstruction? ☒ Yes ☐ No

Fragment minimization in new protein environment? ☒ Yes ☐ No

Job Submission

Send results to (your e-mail address)

PREDICTION PARAMETERS

File: 1309470078_loop_93296.pdb
Chain: A
Start: 57
Sequence: DKLSHGA
Selection loops: 0 [0: end points distance; 1: Motif geometry]
Clash filter option: yes
Number of predictions to keep: 5
Rebuild side chains?: yes
Minimize loop?: yes

PREDICTION PROCESS

Number of selected loops 4423
Remaining loops after ranking (loops with Zscore 1442
Remaining loops after 1st. filter (RMSD stems; if N=50 -> hard top limit reached) 50
Remaining loops after 2nd. filter (clashes; if N=20 -> hard top limit reached) 18
Remaining loops after side-chains re-building 18
Remaining loops after energy minimization 18

RESULTS

CANDIDATES	Zscore	FILE
3c1y_A_16_22	3.4	1309470078_loop_93296.pdb.new_A_57_63-3c1y_A_16_22.ent.sc.mini
3eh7_A_20_26	3	1309470078_loop_93296.pdb.new_A_57_63-3eh7_A_20_26.ent.sc.mini
2ahu_A_23_29	2.9	1309470078_loop_93296.pdb.new_A_57_63-2ahu_A_23_29.ent.sc.mini
1oi8_A_287_293	2.9	1309470078_loop_93296.pdb.new_A_57_63-1oi8_A_287_293.ent.sc.mini
1hp1_A_287_293	2.9	1309470078_loop_93296.pdb.new_A_57_63-1hp1_A_287_293.ent.sc.mini

Figure 8. Snapshots of the Archpred web server. Left panel, submission page, right panel, results page.

Table 1

Accuracy of prediction and coverage for different loop lengths.

Loop length	Zscore ^a	Average RMSD ^b (Å)	Coverage ^c (%)
4	1	0.22	98
5	1	0.15	96
6	1	0.34	98
7	1	0.93	94
8	2	1.38	78
9	3	1.93	60
10	3	2.11	46
11	3	2.30	44
12	4	2.47	28
13	4	2.85	4
14	4	2.88	6

(a) Z-score cut-off to ensure higher accuracy than what is expected from ModLoop.

(b) Average RMSD of predicted loops for the given Z-score cut-off.

(c) Percentage of query loops that are modelable (i.e. a suitable Smotif can be found) at the give Z-score threshold.