A supersecondary structure library and search algorithm for modeling loops in protein structures

Narcis Fernandez-Fuentes, Baldomero Oliva\textsuperscript{1} and András Fiser*\textsuperscript{2}

Department of Biochemistry and Seaver Foundation Center for Bioinformatics, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, NY 10461, USA and \textsuperscript{1}Structural Bioinformatics Group (GRIB), Universitat Pompeu Fabra, C/Doctor Aiguader,80. 08003, Barcelona, Catalonia, Spain

Received January 16, 2006; Revised February 12, 2006; Accepted March 20, 2006

ABSTRACT

We present a fragment-search based method for predicting loop conformations in protein models. A hierarchical and multidimensional database has been set up that currently classifies 105,950 loop fragments and loop flanking secondary structures. Besides the length of the loops and types of bracing secondary structures the database is organized along four internal coordinates, a distance and three types of angles characterizing the geometry of stem regions. Candidate fragments are selected from this library by matching the length, the types of bracing secondary structures of the query and satisfying the geometrical restraints of the stems and subsequently inserted in the query protein framework where their fit is assessed by the root mean square deviation (r.m.s.d.) of stem regions and by the number of rigid body clashes with the environment. In the final step remaining candidate loops are ranked by a Z-score that combines information on sequence similarity and fit of predicted and observed \(\phi/\psi\) main chain dihedral angle propensities. Confidence Z-score cut-offs were determined for each loop length that identify those predicted fragments that outperform a competitive \textit{ab initio} method. A web server implements the method, regularly updates the fragment library and performs prediction. Predicted segments are returned, or optionally, these can be completed with side chain reconstruction and subsequently annealed in the environment of the query protein by conjugate gradient minimization. The prediction method was tested on artificially prepared search datasets where all trivial sequence similarities on the SCOP superfamily level were removed. Under these conditions it is possible to predict loops of length 4, 8 and 12 with coverage of 98, 78 and 28% with at least of 0.22, 1.38 and 2.47 Å of r.m.s.d. accuracy, respectively. In a head-to-head comparison on loops extracted from freshly deposited new protein folds the current method outperformed in a \(\sim\)5:1 ratio an earlier developed database search method.

INTRODUCTION

Computational analysis of protein sequences, such as the identification of conserved motifs, is often informative to learn about the possible function of a protein (1,2). However, a detailed functional characterization frequently requires the study of 3D structures and complexes of proteins (3,4). Despite recent improvements in techniques of structure determination by X-ray crystallography or NMR spectroscopy, a quick inspection of biological databases reveals a two order of magnitude difference between the number of known protein sequences [\(\sim\)3 millions; UniProt database release 5.2 (5)] and that of protein structures [\(\sim\)35 000; Protein Data Bank (PDB) database (6)]. In the absence of an experimentally described structure, computational methods, such as comparative modeling [e.g. Sali \textit{et al.} (7)], threading [e.g. Dominguez \textit{et al.} (8)] or \textit{ab initio} methods [e.g. Simons \textit{et al.} (9)] can be used to provide a useful 3D model and fill the gap between the number of sequences and structures [reviewed in (10–13)]. Comparative modeling is currently the most accurate computational approach to protein structure prediction but it is applicable only if a suitable template is found with a detectable sequence similarity over the entire length of the target protein (10,14). The applicability of comparative modeling is steadily increasing because of the worldwide efforts in Structural Genomics that aims at experimentally solving \(\sim\)5000–10,000 representative protein structures within the next few years (15). In the context of comparative modeling, the most difficult problems are the calculations of

\*To whom correspondence should be addressed. Tel: +1 718 430 3233; Fax: +1 718 430 856; Email: andras@fiserlab.org

© The Author 2006. Published by Oxford University Press. All rights reserved.

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact journals.permissions@oxfordjournals.org
an accurate alignment between the target sequence and template structure and the prediction of insertions i.e. loop structures (10,14). 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 homologs (14). Recent improvements that were observed in the performance of fold prediction and homology modeling methods throughout successive CASP experiments (16) did not extend to the performance of loop modeling techniques.

Loops often represent an important part of the protein structure. Functional differences between the members of the same protein family are usually a consequence of structural differences on the protein surface, which frequently correspond to exposed loop regions (17). Loops often determine the functional specificity of a given protein framework, contributing to active and binding sites, such as antibody complementary determining regions (18), ligand binding sites [ATP (19), calcium binding sites (20), NAD(P) (21)], DNA binding (22) or enzyme active sites [e.g. Ser-Thr kinases (23) or serine proteases (24)]. Therefore the accuracy of loop conformations often determines the usefulness of computational or experimental models.

Loop prediction can be seen as a mini protein-folding problem. The correct conformation of a given segment of a polypeptide chain has to be calculated from the sequence of the segment influenced by flanking regions that span the loop and by the structure of the rest of the protein that cradles the loop. Many loop-modeling procedures have been described in recent years. Similarly to the prediction of protein structures there are ab initio (conformational search) methods (25–27), and database search (or knowledge-based) methods (28–30). There are also procedures that combine the two (31,32). An extensive overview of published methods in loop prediction until year 2000 can be found in Fiser et al. (33).

In ab initio prediction a conformational search or enumeration of conformations is conducted in a given environment, guided by a scoring or energy function (26,27). There are many such methods, exploiting different protein representations, sampling methods, energy function terms and optimization or enumeration algorithm. Recent works include ModLoop, a method that combines a pseudo-energy scoring function with molecular dynamics and simulated annealing (33); a new energy function, ‘colony energy’ (34) that combines a force-field energy and a root mean square deviation (r.m.s.d.)-dependent term to improve ranking of loop conformations; a divide-and-conquer approach to recursively decompose a target loop until the conformation of resulting conformations can be compiled analytically (35); a method that combines a fine-grained sampling of ψ/ω states and AMBER/GBSA force field for ranking (36); a low-barrier molecular dynamics simulation to improve conformational sampling and a ‘soft-core’ potential energy function to allow extensive rearrangement of loop conformations (37); a hierarchical approach, where first large number of conformations are generated that is followed by iterative cycles of clustering, side-chain optimization and energy minimization of selected conformation using all-atoms empirical potentials (38); DFIRE (39) and ROSETTA (40) are among other methods that were used to calculate loop conformations recently.

Candidate loop structures (up to 12 residues) whose conformations are similar to the native can be found if the number of loops generated is large enough (41). However, scoring functions are often not accurate enough to score the native conformation of a loop with the lowest energy (42,43). Therefore, there are two bottlenecks in conformational search approaches: (i) sampling a near native loop conformation; and (ii) constructing a scoring function that properly ranks a set of near native conformations.

Knowledge-based methods (44), also known as database search approaches, work by finding a segment that fits two stem regions of the target loop. The stems are defined as the main chain atoms that precede and follow the loop, but are not part of it. The search is performed through a database of many known protein structures, not only homologs of the modeled protein. Usually, many different alternative segments that fit the stem residues are obtained, and possibly sorted according to geometric criteria or sequence similarity between the template and target loop sequences. The selected segments are then superposed and annealed on the stem regions. Lessel and Schomburg pointed out the importance of the correct positioning of stem regions for knowledge-based loop prediction methods (45).

It has been shown by various groups that loops follow certain conformational patterns and are not random structures (46–49). Knowledge based prediction of loop structures benefit from the classification of loop conformations (32,46,50). A recent work (51) described the advantage of using HMM sequence profiles in classifying and predicting loops that are derived from ArchDB database (49). The good performance of database search methods is well established for cases when canonical loop conformations exist, as in the case of CDR loop predictions (29,52,53) but their performance is limited by the exponential increase in the number of possible conformations as a function of loop length. Although in the mid and late 90s it was argued that only segments of <7 or even only 4 residues long had most of their conceivable conformations present in structure databases (30,54), a recent update suggested sufficient coverage to model even a novel fold using fragments from the PDB, as the current database of known structures has increased enormously in the last few years (55).

Subsequently a recent work that used a compilation of fragments extracted from PDB reported good results in prediction of long loops (56). Our most current survey indicates that loops up to length 8 are essentially fully covered by known conformational segments and the structure database is rapidly saturating for longer segments as well (N. Fernandez-Fuentes and A. Fiser, manuscript submitted).

Combined methods use both database search and ab initio methods. The underlying idea is the use of database search methods to find candidate loops for a given target loop and subsequently evaluate and re-optimize it in the target protein. An example of a combined algorithm is that of Martin et al. (57), in which antibody hypervariable loops were predicted using a database search followed by ab initio reconstruction of sections of the predicted loops and side chains using the CONGEN conformational search algorithm (27). This idea has been generalized: loops were selected from a fragment databank, optimized and ranked using the CHARMM energy function (58). Deane and Blundell presented CODA (32), a combination of two algorithms: FREAD,
a knowledge-based method, and PETRA, an ab initio method (59).

Here we present the construction of a loop database (Search Space) that is an exhaustive compilation of all possible loop conformations braced in between two regular secondary structures (α-helices or β-strands) in all protein structures and a novel database search algorithm to identify loop conformations for a given sequence segment. The prediction algorithm selects a set of candidate loops from the Search Space, then subsequently filters and ranks them by various criteria. First, the Search Space is queried by the length of the loop, the type of secondary structures that span the query loop and by the geometry of the stem using various descriptions: such as a distance and various angles of the stems (48). Second, loops are filtered and discarded if the r.m.s.d. of their stem residues and the interactions between the fragment and the rest of the protein environment are unfavorable (steric clashes). Third, in the ranking step, the remaining candidate loops are sorted by a composite Z-score. The Z-score combines a sequence score, as obtained from a conformational similarity weight matrix (K3) (60), and a ϕ/ψ main chain dihedral angle propensities score (61).

**MATERIALS AND METHODS**

**Construction and organization of Search Space—an exhaustive database of loop fragments**

A representative set of 6578 protein structures were selected from the February 2004 release of PDB (6). The selected proteins share >95% sequence identity and were determined by X-ray crystallography at a resolution of 2.5 Å or better. The DSSP program (62) located loop segments defined as fragments that connect two regular secondary structures. The initial dataset of loops was further filtered by various quality rules to obtain a high-quality loop library by discarding incomplete or poorly defined segments: (i) loops with missing residues and/or main chain atoms (including Cβ, except for Gly), and (ii) loops with high crystallographic B-factors were discarded. For this latter a B-factor Z-score was calculated from atomic B-factors for each residue by averaging B-factors for all atoms in the residue and comparing it with the mean and SD of B-factors of all residues in the protein. Loops containing >50% of their residues with B-factor Z-score higher than 1.0 were discarded. The final set contains 105 950 protein loops altogether with their flanking secondary structures. These loops were organized into a hierarchical and multidimensional database that we refer to as Search Space.

The Search Space of loops is representing all possible loop conformations and is organized by the definition of bracing secondary structures, loop lengths and loop geometries. Search Space is organized in a three level hierarchy: (i) at the top of the classification, loops are identified according to the type of the bracing secondary structures: αα, αβ, βα and ββ loops; (ii) at the second level, loops are grouped according to their length as defined by DSSP program (62) and (iii) at the third level, loops are grouped according to geometry of the bracing secondary structures as defined by a distance, D, between the anchor points and three angles: a hoist (δ), a packing (θ) and a meridian (ρ) (48) (Figure 2).

For distance, the interval considered for classifying all possible loops spans between 0 and 40 Å partitioned by intervals of 2 Å; for hoist and packing angles span from 0 to 180 degrees and is partitioned into 30° intervals; meridian angle spans from 0 to 360° and is partitioned into 45° intervals. This partition classifies each loop in a 4D geometrical space. The partitioning of the Search Space is optimized: very narrow, fine grain partitioning would result in numerous empty or poorly populated cells in the multi-dimensional Search Space, whereas wide bins could join highly dissimilar geometries (see below, Calibration Test-Sets and Supplementary Data).

**Selecting candidate loops**

For a given query segment, candidate loop conformations are selected from the Search Space by matching bracing secondary structures, length ±1 residue and geometrical criteria. A tolerance in loop length of ±1 residue is permitted to compensate for possible uncertainties in assigning end points to secondary structures (63,64). We refer to these selected loops as ‘candidate loops’ (Figure 1). Two loops, loop A with geometry $G_A = (D_A, \delta_A, \theta_A, \rho_A)$ and loop B with geometry $G_B = (D_B, \delta_B, \theta_B, \rho_B)$ share the same geometry if $G_{A,B} = [(|D_A - D_B|, |\delta_A - \delta_B|, |\theta_A - \theta_B|, |\rho_A - \rho_B|)]$ belongs to the 4D semi-open interval $I = [(0,0,0,0), (2,3,30,45)]$, i.e. the distance difference between the anchor points should be <2 Å, the differences between the three dihedral angles, δ, θ and ρ should be <30, 30 and 45, respectively.

The use of geometry as a descriptor for loop selection implies that the flanking secondary structures are well described (at least five residues for α-helices and two residues for β-strands). However, the current method is prepared to handle cases where secondary structures are not known and/or not well defined. In this case, Search Space is queried using only the distance of end points. The approach predicts loops between two defined regions, therefore it is not suitable for prediction of terminal fragments.

**Filtering candidate loops**

The filtering step in the algorithm discards clearly unfavorable candidates based on structural superposition of stem residues and steric violations after fitting the loop in the protein framework (Figure 1). All candidate loops are superimposed on their stem positions using the main chain atoms of two stem residues at each flanking secondary structure. Candidate segments with r.m.s.d. of stems higher than 1.0, 1.5 and 1.75 Å for loops with 4–7, 8–12, 13 and more residues, respectively, are discarded. This dynamic range of cut-off values was determined via an iterative optimization (see below, Calibration Test-Sets). The rest of the candidate loops are further filtered by exploring their conformational fit in the new protein environment in terms of number of steric violations or clashes. The conformation fit in the new environment is assessed in terms of steric clashes among main chain atoms (N, C, α and O). Two atoms are in steric clash if their distance is smaller than the 70% of sum of the respective van der Waals radii. Van der Waals radii were taken from Tsai et al. (65).
Ranking of candidate loops

The final set of candidate loops are ranked by two measures: (i) A sequence similarity score between the query and candidate loops; and (ii) $\phi/\psi$ main chain dihedral angle propensities. The sequence similarity score for a loop sequence $S_{\text{sequence}}$ is defined as the following equation:

$$S_{\text{sequence}} = \frac{1}{L} \sum_{i=1}^{L} (C_i \rightarrow Q_i),$$  \hspace{1cm} (1)$$

where $L$ is the length of aligned positions between a candidate loop and the query loop; $C_i \rightarrow Q_i$ is the value of substitution of amino acid $C$ (candidate) by the amino acid $Q$ (query) in position $i$. A number of substitution tables were tested (Calibration Test-Sets) and the conformation similarity weight matrix (K3 matrix) of Kolaskar and Kulkarni-Kale (60) was found as the best performing one.

The dihedral angle propensity score measures the compatibility of observed and expected dihedral angles of each residue of the candidate loop in the corresponding position of the query. Main chain conformation definitions and propensities are defined according to the p15 propensities table of Shortle’s work (61). Similarly to the sequence score, the propensity score, $S_{\text{propensity}}$, of the query loop is obtained as it is threaded in the main chain conformation of the candidate loop (2):

$$S_{\text{propensity}} = \frac{1}{L} \sum_{i=1}^{L} \log (C_i \rightarrow Q_i),$$  \hspace{1cm} (2)$$

where $L$ is the length of aligned positions and $C_i \rightarrow Q_i$ is the propensity of the residue $C$ when it adopts the main chain conformation of the residue $Q$ in the position $i$. The two components of the scoring scheme, sequence and propensity, are combined into a composite score. First, the individual scores ($S$) were transformed into a $Z$-score using the mean ($\mu$) and

---

The flowchart illustrates the methodology for building the Search Space and performing the prediction. Figure 1: Flowchart of the methodology for building the Search Space and performing the prediction.
RESULTS AND DISCUSSION

We present a novel fragment-search based loop conformation prediction method. The approach has two parts, (i) the classification of loop fragments into an extensive library (‘Search Space’) and (ii) a three step search algorithm to Select, Filter and Rank candidate loops for a given query sequence. Five different measures are used during the prediction process. Three of the measures: motif geometry, r.m.s.d. of stems and steric clashes are used as qualitative descriptors only, to accept or to reject candidate loops through the Selection and Filtering steps. Sequence similarity and amino acid φ/ψ dihedral angle propensities were used for quantitatively rank the final set of candidate loops (Figure 1).

Search Space

The Search Space currently classifies 105,950 high quality loop structures, and it is regularly updated. Search Space is organized in a three level hierarchy: loops are identified and grouped according to (i) the type of bracing secondary structures; (ii) their length and (iii) 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) (Figure 2) (48). The third level of hierarchy is the geometrical binning of loops. It bins loops into 20 × 6 × 6 × 8 = 5760 possible cells or geometrical combinations that is obtained from the number of possible bins for the distance vector, and (δ), (θ) and (ρ) angles, respectively. Not all cells are equally populated, short loops cannot have large values of vector distance or β–β hairpin loops have a restricted geometry in terms of possible angles combinations.

![Figure 2. Definition of loop geometry.](image-url)
due to strict hydrogen bond requirements (73). For instance, in case of loops of length 4 the number of sampled cells is 614, where 225 cells have <5 loops and the most populated cell contains 681 loops. For loops of length 8 and 12 the number of sampled cells are 669 and 861, where 304 and 416 of these have <5 loops and the two most populated cells contains 110 and 93 loops, respectively. Even at longer loop lengths there are preferred geometries, in agreement with earlier observations (74).

**Selection of candidate loops from the Search Space**

Prediction of loops requires an efficient (fast and scalable) and accurate algorithm. We group our algorithm into three steps: selecting, filtering and ranking of the suitable segments from the Search Space. During selection, loops in the Search Space are queried in a stepwise manner. First, loops with similar bracing secondary structures are identified, and those having a similar length (+/− 1 residue) to the query loop are selected. The last selection step in the lookup process involves comparing one distance and three angle values, which serve as internal coordinates to describe the geometry of the stems.

Selecting loops by geometry is a quick but coarse filtering step. It is more powerful than selecting fragments from loop databases based only on end point distances (56,75) because not only a distance is considered but also the orientation of the stems as well. On the other hand it is faster than selecting fragments through superimposition and r.m.s.d. calculation of stem residues. The r.m.s.d. calculation is computationally more demanding than a simple string comparison. The initial selection of candidate loops by simple geometrical requirements quickly narrows the space to be explored by subsequent, more elaborate structural comparison. 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. This strict filtering step does not mean that good candidate loops are rejected. The second qualitative descriptor to filtering of loops explores the conformational fit of candidate loops in the new protein environment. Each candidate loop is plugged into the protein environment of the query and checked for steric clashes between the loop and its surroundings and the ones with steric clashes are removed from the candidate’s list. After these steps the average number of loop candidates decreased to 81, 35 and 5 for loops of length 4, 8 and 12, respectively.

**Filtering and ranking candidate loops**

Two qualitative descriptors are used for filtering: the fit of stem residues by superposition of main chain atoms and r.m.s.d. calculation and the evaluation of steric clashes between the loop and the rest of the protein environment. r.m.s.d. cutoffs for superposed stem residues have been applied before in loop structure prediction method either for ranking (56) or filtering (75). The r.m.s.d. fit of stem residues correlate strongly with the accuracy of prediction of short loops, but this correlation becomes less pronounced for longer loops (Supplementary Data; correlation between r.m.s.d. stem versus r.m.s.d. local of loops). The reason is that conformations that a fragment can adopt are less restricted by the stem residues in case of medium and long loops (8–14 residues) than for short loops (1–7 residues). Therefore we applied a range of r.m.s.d. stem cutoff values as a function of loop length. After this filtering step the average number of candidate loops for a given random query dropped from 181, 85 and 25 to 96, 36 and 11 for loops of length 4, 8 and 12, respectively.

The second qualitative descriptor to filtering of loops explores the conformational fit of candidate loops in the new protein environment. Each candidate loop is plugged into the protein environment of the query and checked for steric clashes between the loop and its surroundings and the ones with steric clashes are removed from the candidate’s list. After these steps the average number of loop candidates decreased to 81, 35 and 5 for loops of length 4, 8 and 12, respectively.

**Ranking candidate loops by sequence and main chain dihedral angle propensity comparisons**

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 criteria.

Sequence Z-score gauges the similarity between the sequences 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 (60) 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 a specific environment. The environment is defined by the 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) (61). 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 (61). The composite Z-score is defined as the sum of the two types of Z-scores.

There is a (negative) correlation between the composite Z-score and the r.m.s.d. local for all the three calibration test sets (Figure 3A–C). The distribution of sequence Z-score versus propensity Z-score (Figure 3D–F) for all candidate loops in the calibration test shows that in most of the cases the sign and magnitude of Z-score is related. For instance, if a candidate loop has a high sequence Z-score, the propensity Z-score is also high and vice versa. Also, candidates with good r.m.s.d. local have both positive and large Z-score (Figure 3D–F).

**Performance of loop prediction**

Benchmarking loop prediction approaches using database methods is not straightforward. Some sort of artificially filtered input fragment dataset 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. We compare our results with (i) using variety of pre-filtered Search Spaces (ii) the performance of a competitive and freely available *ab initio* prediction method, ModLoop (76); (iii) the theoretical minimum r.m.s.d. which depends on the database applied (Search Space) thus informing on the practical limits of the method; (iv) with the expected r.m.s.d. of a prediction made by random selection of loops segments (Figure 4); and (v) by directly comparing with an earlier developed, publicly available fragment search based method (32).

The minimum value of r.m.s.d. that can be obtained with loops available in the Search Space (i.e. the loop with the smallest r.m.s.d.) 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 (Figure 4). This indicates that there are candidate loops at all loop length that outperform the accuracy of the *ab initio* approach. This supports the conclusion of Du et al. (55) who found that even for long loops (up to 15 residues) there is a 90% probability that a non-homologous structure within 2 Å r.m.s.d. exists. Therefore the bottleneck in fragment search 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 outperforms the current prediction method at 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 (Figure 4). However, averages of both methods fall within the boundaries of 1 Å SD. The accuracy obtained with the current method is clearly much higher than the accuracy obtained with a random prediction (Figure 4). Differences between the current method and ModLoop are smaller in case the comparison is based on r.m.s.d. global measures (Figure 4). Global r.m.s.d. measures the accuracy of the orientation of the loop altogether with its local conformation. Better global r.m.s.d. values as compared with local r.m.s.d. imply that candidate loops are selected with proper orientation. This is probably due to the filter that is applied on steric clashes.

Figure 3. (A) Correlation of composite Z-score and accuracy of prediction (average r.m.s.d. local) for calibration test-sets of (A) 4, (B) 8 and (C) 12 residues long loops, respectively. SDs are shown around the averages. (B) Distribution of Z-scores (propensity Z-score versus sequence Z-score) for all candidates (x) and for the candidates with the best r.m.s.d. local among top 10 hits (closed circle) for (D) 4, (E) 8 and (F) for 12 residues long loops, respectively.
Coverage versus accuracy: identifying confidence

**Z-score thresholds**

We explored the performance of the method as a function of Z-score cut-offs. The r.m.s.d. values decrease as the composite Z-scores and the accuracy of predictions increase. Meantime the corresponding coverage of the prediction decreases (Figure 5).

It is important to assign confidence values to a prediction. Table 1 lists r.m.s.d. local and coverage results versus Z-score cutoffs. Z-score cutoffs were defined in such a way that fragments selected with more significant Z-scores will have equal or better accuracy than the average accuracy of fragments obtained by ModLoop. For instance, for loops between lengths of 4–7 residues a Z-score of 1.0 gives an equal or better

Figure 4. Average (A) r.m.s.d. local and (B) r.m.s.d. global as a function of loop length. (closed diamond) indicates the practical limit of the prediction; (closed circle) shows the average r.m.s.d. of ModLoop predictions; (closed triangle) shows the average r.m.s.d. of candidates with the highest composite Z-score; and (closed square) shows the average r.m.s.d. for random prediction.

Figure 5. Average r.m.s.d. local (A) and coverage (B) as a function of Z-score threshold for all loop lengths.
performance than ModLoop with a corresponding coverage of 90% (Figure 5). For loops between 8 and 11 residues a Z-score larger than 2–3 is required and the average coverage is around 50–60%. For longer loops, beyond 12 residues long the coverage rapidly drops (Table 1).

Completeness of fragment database

Knowledge based approaches are limited by the completeness of the database they are based on. In our benchmarking process we artificially impoverished our Search Space by discarding loops that belong to the same SCOP superfamily as the query. This simulates a situation where no similar structures to our query protein are available on the SCOP superfamily level when attempting to predict its loop conformations. We also explored the performance of the prediction algorithm using a more dynamic range of pre-filtering of the Search Space. We have studied three additional scenarios by removing loop fragments that shared >75, 50 or 25% of sequence identity with the query and re-run the prediction (Figure 6). As the sequence identity filter is less restrictive better candidates can be selected, and better accuracy is achieved. While a sequence filtering at 25% resembles the performance of our default filtering approach on the SCOP superfamily level, at 50% the performance of the current approach becomes competitive with ModLoop method (33), while at 75% it exceeds its performance. The coverage of loop fragments in PDB has been analyzed in a separate work (N. Fernandez-Fuentes and A. Fiser, manuscript submitted) and it has been found that current PDB supplies us with loop fragments up to 14 residues long that are on average 40–60% identical to any observed fragment in the sequence databases. This suggests that the 50% filtering of Search Space might actually be the one that resembles true application scenarios.

The performance of all benchmarks discussed so far were in the context of sequence signal only. Using the full power of the current approach the accuracy of prediction can be significantly increased at each loop length (Figure 7). These improvements become more significant as we apply the prediction for more filtered Search Spaces, where sequence signal has less influence.

Comparing performance to FREAD database search method

We performed a head-to-head comparison of performances between the current ArchPRED and the FREAD methods (32). To avoid a trivial exercise we used only new structural releases

Table 1. Accuracy and coverage of prediction for different loop lengths and Z-score thresholds

<table>
<thead>
<tr>
<th>Loop length</th>
<th>Confidence Z-score</th>
<th>Average r.m.s.d. local (%)</th>
<th>Coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>≥1</td>
<td>0.22</td>
<td>98</td>
</tr>
<tr>
<td>5</td>
<td>≥1</td>
<td>0.15</td>
<td>96</td>
</tr>
<tr>
<td>6</td>
<td>≥1</td>
<td>0.34</td>
<td>98</td>
</tr>
<tr>
<td>7</td>
<td>≥1</td>
<td>0.93</td>
<td>94</td>
</tr>
<tr>
<td>8</td>
<td>≥2</td>
<td>1.38</td>
<td>78</td>
</tr>
<tr>
<td>9</td>
<td>≥3</td>
<td>1.93</td>
<td>60</td>
</tr>
<tr>
<td>10</td>
<td>≥3</td>
<td>2.11</td>
<td>46</td>
</tr>
<tr>
<td>11</td>
<td>≥3</td>
<td>2.30</td>
<td>44</td>
</tr>
<tr>
<td>12</td>
<td>≥4</td>
<td>2.47</td>
<td>28</td>
</tr>
<tr>
<td>13</td>
<td>≥4</td>
<td>2.85</td>
<td>4</td>
</tr>
<tr>
<td>14</td>
<td>≥4</td>
<td>2.88</td>
<td>6</td>
</tr>
</tbody>
</table>

aZ-score thresholds were defined to guarantee that the selected segments are at least as accurate on average as the corresponding prediction of ModLoop.  
bAverage local r.m.s.d. for a given Z-score threshold.  
cAverage local r.m.s.d. for a given coverage.

Figure 6. Average r.m.s.d. local as a function of loop lengths under different condition of pre-filtering Search Space: selecting from candidates that belong to a different SCOP superfamily than the query loop (closed triangle); using ModLoop (closed circle); selecting among candidates that have >75% (closed diamond), 50% (open square) and 25% (closed square) sequence identity with query loop.

Figure 7. Average r.m.s.d. local versus loop length under different condition of selection using information only sequence identity or the full algorithm for prediction.
from PDB (6), which could not yet enter the classification schemes of either methods and we tracked these new PDB structures for two weeks. Among the new structures we identified new folds by removing all proteins with sequence (>40% sequence identity) and structural similarity [DALI (77) Z-score >3] to any known PDB structures. From the remaining 6 novel fold structures we located 35 loop regions and submitted the sequences of these fragments to our method and to the FREAD server. The predicted loops were superposed with the experimental solution and r.m.s.d. values obtained (Table 2). The current method, ArchPRED not only provides a higher coverage (it predicted all segments, while FREAD did not return answer for four cases) but also on average it returned more accurate predictions in 23 out of 28 cases, while in three cases they returned identical solutions (Table 2).

Examples of predicted loops

We present three examples as illustrations, to predict a short, a medium and a long loop. For a short loop we predicted a loop with a length of four residues (extracted from structure 1g29 chain 1, between residues 37 and 40). The loop spans two β-strands forming a β-β hairpin motif. The top three fragment candidates and the experimental solution structure are shown in Figure 8A. All of the candidates fit with a similar r.m.s.d. of stems and without clashes in the new protein framework. If ranking was based only on sequence signals, the candidate loop in green color would be the top choice. However, red candidate has the highest composite Z-score (Z = 2.85 versus 1.45 and 1.01) and is the most accurate fragment (r.m.s.d.local = 0.2 Å versus 0.4 and 0.6 Å).

Figure 8B shows an example of predicting a medium-size loop of 8 residues, between positions 107 and 114 in the 1srp structure. This example illustrates the usefulness of filtering by steric clashes. All three candidate loops, shown in red, green and blue, have approximately the same r.m.s.d. for stems residues, around 1.1 Å. The loop in green has the highest Z-score for sequence signal (3.2, 2.8 and 1.9 for green, red and blue loops, respectively). However, the green bumps against a neighboring β-strand in the new protein environment, therefore it is removed from the list of putative candidates by the prediction method. Remaining candidates, red and blue, both fit without steric clashes, but the composite Z-score for red loop is higher than the blue loop (4.7 versus 3.6) in good agreement with a superior r.m.s.d.local (1.3 Å versus 1.9 Å).

The third example is a prediction of a loop with 12 residues extracted from structure 1j85 chain A, between residues 121 and 132 (Figure 8C). Out of the three different candidate loops the one in blue is discarded because of steric conflicts with the protein framework. Both loops in red and green have a comparable sequence Z-score, (2.3 and 2.5) but the main chain dihedral angle propensity Z-score is more favorable for the

Figure 8. Cartoon representations for three examples of predicted loops. N- and C-termini of loops are marked. The experimental structure is in gray while in red, blue and yellow are the candidate loops. Red colored candidate loops are with the highest composite Z-score. Gray shaded spheres indicate regions with repulsive contacts (clashes) between the candidate loop and the protein framework. (A) Prediction of a loop of length 4, PDB 1g29, chain 1, residues 37–40. (B) Prediction of a loop of length 8, PDB 1srp, residues 107–114. (C) Prediction of a loop of length 12, PDB 1j85, chain A residues 121–132. All figures were generated using Pymol (http://pymol.sourceforge.net).
Table 2. Comparing prediction performances of ArchPRED and FREAD programs

<table>
<thead>
<tr>
<th>PDB Chain</th>
<th>Start</th>
<th>Sequence</th>
<th>ArchPRED r.m.s.d.,local(A)</th>
<th>FREAD r.m.s.d.,local(A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2d0 A</td>
<td>17</td>
<td>GHLLEDVYVYVSSD</td>
<td>4.06</td>
<td>3.27</td>
</tr>
<tr>
<td>2d0 A</td>
<td>54</td>
<td>AFPADIDFAD</td>
<td>2.23</td>
<td>2.80</td>
</tr>
<tr>
<td>2ef A</td>
<td>94</td>
<td>GLA</td>
<td>0.15</td>
<td>0.22</td>
</tr>
<tr>
<td>2ef A</td>
<td>22</td>
<td>LARTDRAPRKNID</td>
<td>2.30</td>
<td>3.29</td>
</tr>
<tr>
<td>2ef A</td>
<td>172</td>
<td>VLQFDTD</td>
<td>1.62</td>
<td>1.79</td>
</tr>
<tr>
<td>2ef A</td>
<td>268</td>
<td>GWDD</td>
<td>0.40</td>
<td>0.91</td>
</tr>
<tr>
<td>2ef A</td>
<td>60</td>
<td>QPEA AQ</td>
<td>1.41</td>
<td>1.24</td>
</tr>
<tr>
<td>2ef A</td>
<td>235</td>
<td>QLS</td>
<td>0.06</td>
<td>No hit</td>
</tr>
<tr>
<td>2ef A</td>
<td>205</td>
<td>GQL</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>2ef A</td>
<td>276</td>
<td>ASPHYL</td>
<td>0.27</td>
<td>2.02</td>
</tr>
<tr>
<td>2d13 A</td>
<td>211</td>
<td>DGLS</td>
<td>0.22</td>
<td>0.26</td>
</tr>
<tr>
<td>2d13 A</td>
<td>194</td>
<td>MPFK</td>
<td>0.36</td>
<td>1.32</td>
</tr>
<tr>
<td>2d13 A</td>
<td>11</td>
<td>YSGG</td>
<td>0.10</td>
<td>1.15</td>
</tr>
<tr>
<td>2d13 A</td>
<td>96</td>
<td>AGALAS</td>
<td>0.93</td>
<td>1.79</td>
</tr>
<tr>
<td>2d13 A</td>
<td>121</td>
<td>TP AWEDK</td>
<td>0.82</td>
<td>No hit</td>
</tr>
<tr>
<td>2d13 A</td>
<td>138</td>
<td>LGF</td>
<td>0.12</td>
<td>0.10</td>
</tr>
<tr>
<td>2d13 A</td>
<td>27</td>
<td>SGL</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>2d13 A</td>
<td>88</td>
<td>GLKVD</td>
<td>0.21</td>
<td>0.86</td>
</tr>
<tr>
<td>2d13 A</td>
<td>63</td>
<td>GIP</td>
<td>0.16</td>
<td>0.16</td>
</tr>
<tr>
<td>2d13 A</td>
<td>177</td>
<td>GIHIAGEGGE</td>
<td>2.69</td>
<td>3.84</td>
</tr>
<tr>
<td>2fgg A</td>
<td>44</td>
<td>RLDPADEPVA</td>
<td>2.30</td>
<td>2.96</td>
</tr>
<tr>
<td>2fgg A</td>
<td>20</td>
<td>HDER</td>
<td>0.94</td>
<td>1.00</td>
</tr>
<tr>
<td>2fgg A</td>
<td>48</td>
<td>DMTGPVPTQEDI</td>
<td>3.32</td>
<td>No hit</td>
</tr>
<tr>
<td>2fgg A</td>
<td>76</td>
<td>TGEI LRP</td>
<td>1.79</td>
<td>3.23</td>
</tr>
<tr>
<td>2fgg A</td>
<td>65</td>
<td>ESQIE</td>
<td>0.30</td>
<td>1.92</td>
</tr>
<tr>
<td>2fgg A</td>
<td>94</td>
<td>KRI</td>
<td>0.10</td>
<td>0.07</td>
</tr>
<tr>
<td>2fgg A</td>
<td>41</td>
<td>EGHPM</td>
<td>0.35</td>
<td>0.55</td>
</tr>
<tr>
<td>2fgg A</td>
<td>97</td>
<td>GAW</td>
<td>0.18</td>
<td>0.60</td>
</tr>
<tr>
<td>2fgg A</td>
<td>109</td>
<td>GYP APAN</td>
<td>1.33</td>
<td>2.25</td>
</tr>
<tr>
<td>2fgg A</td>
<td>142</td>
<td>SQF DLP AAL</td>
<td>1.22</td>
<td>No hit</td>
</tr>
<tr>
<td>2fgg A</td>
<td>83</td>
<td>HGM PL A</td>
<td>0.52</td>
<td>1.88</td>
</tr>
<tr>
<td>2fgg A</td>
<td>212</td>
<td>SLCAC</td>
<td>0.20</td>
<td>0.15</td>
</tr>
<tr>
<td>2fgg A</td>
<td>257</td>
<td>PSCQ</td>
<td>0.07</td>
<td>0.15</td>
</tr>
<tr>
<td>2fgg A</td>
<td>268</td>
<td>LEFDRHAD</td>
<td>1.98</td>
<td>2.48</td>
</tr>
<tr>
<td>2fgg A</td>
<td>293</td>
<td>DGY</td>
<td>0.08</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Thirty-five loops were collected from six recently deposited novel fold structures and used as test cases. The table shows the PDB code, chain identifier, loop starting position, loop sequence, backbone r.m.s.d.,local value calculated after the superposition of predicted and experimental solution structure.

red (2.2 versus 1.4), resulting in an overall higher composite Z-score, in agreement with the overall accuracy, or r.m.s.d.,local (1.6 versus 2.7 Å).

Loop prediction web server

The Search Space and the prediction method described here are implemented in a web server. The user provides the query structure in PDB format that contains the missing loop(s) and defines its sequential location. The interface of the web server provides all the controls of the method: searching loops by end-point distance only, or by geometry; if by geometry than by two types of bracing regular secondary structural elements; if these elements are beta strands than further distinguished by hairpin or link types. Once the prediction is completed results are sent by email in form of a link pointing to temporary web pages. Optionally the best loop fragment located is built in the query structure and a conjugate gradient minimization is applied to smoothly anneal the stems in the protein framework. The server is accessible at http://www.fiserlab.org/servers/archpred.

CONCLUSION

The number of experimentally solved structures has grown dramatically in the last few years. More importantly, due to the ongoing Structural Genomics efforts an increasing number of new folds or remotely related proteins are being solved, amalgamating the library of conformational segments (15). Studies in 1994 and 1997 concluded that database search methods were limited to predict loops up to four residues long (seven considering three stem residues) (30,54). However, our analysis (N. Fernandez-Fuentes and A. Fiser, submitted) in agreement with other recent reports (55) suggests that there is a sufficient sampling of short segments in the PDB to efficiently use database search methods to predict loops currently up to 9–12 residues. If a good fragment is found in the database it could be used straightforward or as a starting conformation for subsequent optimization. In both approaches, the presence of a suitable segment permits one to avoid computationally more demanding and riskier ab initio approaches. To assess the usefulness of a given predicted segment it is necessary to define confidence values for fragment based approaches. We tackled the problem of defining confidence values in the current method by calibrating Z-score cutoff values that ensure a superior solution to a competitive ab initio approach.

The accuracy and coverage of the current method implies that database search approaches rapidly gain importance in loop prediction and the bottleneck in these approaches does not appear to be the sampling (database completeness of segments), but the search algorithm and scoring function to locate these segments. With the advance of structural genomics efforts (15) we expect that this trend will be further accentuated in the coming years.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors acknowledge all Fiser lab members for their insightful comments on the work, especially Dr D. Rykunov. N.F.F. was partially supported by a Boehringer fellowship. Financial support provided by NIH GM62519-04 and the Seaver Foundation. Funding to pay the Open Access publication charges for this article was provided by NIH GM62519-04 and MEC B102005-00533.

Conflict of interest statement. None declared.

REFERENCES


