PROTEINS: Structure, Function, and Bioinformatics 56:782–794 (2004)

Combining Evolutionary and Structural Information for Local Protein Structure Prediction

Jimin Pei[1] **and Nick V. Grishin**[1, 2*]
[1]*Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, Texas*
[2]*Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, Texas*

ABSTRACT **We study the effects of various factors in representing and combining evolutionary and structural information for local protein structural prediction based on fragment selection. We prepare databases of fragments from a set of nonredundant protein domains. For each fragment, evolutionary information is derived from homologous sequences and represented as estimated effective counts and frequencies of amino acids (evolutionary frequencies) at each position. Position-specific amino acid preferences called structural frequencies are derived from statistical analysis of discrete local structural environments in database structures. Our method for local structure prediction is based on ranking and selecting database fragments that are most similar to a target fragment. Using secondary structure type as a local structural property, we test our method in a number of settings. The major findings are: (1) the COMPASS-type scoring function for fragment similarity comparison gives better prediction accuracy than three other tested scoring functions for profile–profile comparison. We show that the COMPASS-type scoring function can be derived both in the probabilistic framework and in the framework of statistical potentials. (2) Using the evolutionary frequencies of database fragments gives better prediction accuracy than using structural frequencies. (3) Finer definition of local environments, such as including more side-chain solvent accessibility classes and considering the backbone conformations of neighboring residues, gives increasingly better prediction accuracy using structural frequencies. (4) Combining evolutionary and structural frequencies of database fragments, either in a linear fashion or using a pseudocount mixture formula, results in improvement of prediction accuracy. Combination at the log-odds score level is not as effective as combination at the frequency level. This suggests that there might be better ways of combining sequence and structural information than the commonly used linear combination of log-odds scores. Our method of fragment selection and frequency combination gives reasonable results of secondary structure prediction tested on 56 CASP5 targets (average SOV score 0.77), suggesting that it is a valid method for local protein structure prediction. Mixture of predicted structural frequencies and evolutionary frequencies improve the quality of local profile-to-profile alignment by COMPASS. Proteins 2004;56:782–794.**

INTRODUCTION

Protein sequences are under evolutionary selective pressure to maintain their structure and function. Although sequence contains the information about three-dimensional (3D) structure, ab initio structure prediction is a difficult task. Evolutionary information is valuable to structure prediction in that homology relationships can be inferred by sequence similarity and close homologs usually have similar tertiary structures. With sensitive similarity search tools, such as PSI-BLAST[1] and HMMer,[2] and the rapid growth of sequence and structure databases, it has become more and more convenient to make homology-based structure prediction. For distant homologs, structural conservation can be maintained at the level of general fold or architecture, but local structural details such as the conformations of turns and loops can vary considerably,[3] which remains a major challenge for comparative modeling.[4]

Prediction of local structural features can be the first step toward the prediction of 3D-structures.[5,6] The most commonly used local structural features are protein secondary structures, characterized by α-helices, β-strands, turns, and coil. There are also many local structural motifs with strong sequence signals, such as certain β-turns[7] and the α-helix capping motifs.[8] The ubiquitous presence of secondary structures and the population of similar local structural motifs in different 3D-structures reflects the common physicochemical principles governing the stability and folding of proteins and makes sequence-based local structure prediction feasible.

Secondary structure only provides a crude description of a few types of backbone conformations. Structural fragments better characterize local structural features and capture information about positional correlations within a fragment. A routine practice is to cluster fragments into a small number of structural alphabets or building blocks.[9–13] Along with the structural characterization of

*Correspondence to: Nick V. Grishin, Howard Hughes Medical Institute University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390-9050. E-mail: grishin@chop.swmed.edu

Received 25 August 2003; Accepted 20 February 2004

Published online 11 June 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20158

fragments, there have been studies on local sequence–structure relationships and efforts to predict local conformations using sequence information.[9,14–17] It is well known that evolutionary information from homologous sequences can significantly improve local structure prediction.[14,18] Analysis of sequence–structure relationship based on database structures also helps sequence-based local structure prediction.[15,17,19] The major goal of this study is to exploit information from both homologous sequences and database structures and to combine the two sources of information effectively.

While fragment clustering can provide a simplified view of local sequence and structural space, information is lost by representing objects using cluster centers. Motifs with lower frequencies of occurrence in the database can get merged into clusters with different sequence and structural properties. In this study, we build libraries of fragments of different lengths from a non-redundant set of protein domains with known structure. To fully preserve the information in them, these fragments are not clustered. Each database fragment has the local structural information as well as sequence information from a multiple sequence alignment of homologs obtained from sequence database searches. We apply a nearest-neighbor method for local structure prediction. For a target fragment with only sequence information, we select in the database those fragments that are most similar to it (nearest neighbors). The local structural properties of the target fragment are assigned using the consensus of the local structural properties of its nearest neighbors. One key component of this method is the scoring function that quantifies how the sequence information of the target fragment is compatible with the local structure of a database fragment. A better scoring function or a more effective way of representing the sequence information will result in better prediction accuracy. We explore methods to represent and utilize homology-derived sequence information and local structure-derived sequence information and try to combine these two sources in a unified framework to achieve the best performance of local structure prediction. Amino acid preferences can be derived from predicted local structures. They provide additional information to the amino acid preferences estimated from homologous sequences directly. Mixture of predicted structural frequencies and evolutionary frequencies might give better estimation of position-specific amino acid frequencies. Such frequency mixture is optimized and tested in pairwise local profile-to-profile alignments using COMPASS.[20]

METHODS

Structure Dataset and Fragment Databases

A non-redundant set of proteins with known structures was prepared (October 2002) using the program PISCES (http://www.fccc.edu/research/labs/dunbrack/pisces/)[21] satisfying the following criteria: 40 residues for minimum chain length, X-ray structures with resolution no worse than 2.5 Å, and no pairs of sequences with identity above 20%. Those structures included in the SCOP database (version 1.59) were split into domains as defined in SCOP.[22] Only domains belonging to SCOP classes 1 to 5 (alpha, beta, alpha/beta, alpha + beta, and multi-domain) were selected so that the dataset does not include membrane proteins or small proteins enriched with disulfide bonds or metal ions. Sequence redundancy of the selected SCOP domains were further checked and reduced by clustering using the program BLASTCLUST at 25% identity threshold. The final database contains 1695 SCOP domains (SCOP class alpha: 327; beta: 380; alpha/beta: 521; alpha + beta: 423; multi-domain: 44). Starting from the sequence of each domain, PSI-BLAST[1] searches were done up to five iterations over the non-redundant protein sequence database of NCBI (October 2002, 1,210,757 sequences; 386,423,161 total letters) at the e-value cutoff 0.00001. A multiple sequence alignment was derived from the final output of PSI-BLAST searches. Sequences with identity less than 25% to the query were removed from the multiple sequence alignment. Positions where the query is a gap are removed. For a query with sequence length L, we derive $L - F + 1$ overlapping fragments with length F. Each fragment has a corresponding multiple sequence alignment of length F. We consider fragment length F from 6 to 10. In total, there are 331,541 six-residue fragments, 329,843 seven-residue fragments, 328,145 eight-residue fragments, 326,447 nine-residue fragments and 324,749 ten-residue fragments in the database.

Testing Datasets

We select the first 100 domains from the 1695 domains used for fragment database as the testing dataset (the numbers of SCOP classes are alpha: 21; beta: 21; alpha/beta: 34; alpha + beta: 22; multi-domain: 2). They are listed in Table I. Only the multiple sequence alignments are used for prediction for these domains. To predict the local structure of each domain, we exclude those domains in the database that are of the same SCOP fold to it and use only the rest of the database for its prediction. In this way, we minimize the chance that the selected top-scoring fragments for prediction are homologous to the target fragment.

We also test our method on 58 target proteins in CASP5. The results are compared to methods that have shown good secondary structure prediction performance in CASP5: PSIPRED,[23] SAM-T99,[24] and SSpro2.[25] Results of these three methods are from the CASP5 website at http://predictioncenter.llnl.gov/casp5/Casp5.html. The most difficult testing cases are four domains of the CASP5 targets classified as "HARD" since more than 75% of the predictions for them are below 0.68. They are T0134_1, T0146_3, T162_2 and T0174_2.

Profile Representation of Evolutionary Sequence Information: Effective Counts and Evolutionary Frequencies of Amino Acids

For a fragment, we derive sequence profiles from the multiple sequence alignment associated with it. The numerical representation of the sequence profile includes the effective counts of amino acids and estimated frequencies

TABLE I. List of the Test Domain Protein Databank (PDB) Ids, Chain Ids and Domain Ranges

119l	1a79A_83_179	1aihA	1aq0A
12asA	1a79A_9_82	1ail	1aqcA
16pk	1a7j	1air	1aquA
16vpA	1a8l_120_226	1ajsA	1aqzA
1a12A	1a8l_1_119	1ak0	1arb
1a1x	1a8o	1ak1	1ash
1a26_662_796	1a8rA	1ako	1at0
1a26_797_1012	1a8y_127_228	1al3	1aua_4_96
1a32	1a8y_229_347	1alu	1aua_97_299
1a34A	1a8y_3_126	1am2	1aueA
1a3aA	1a9xB_1502_1652	1am7A	1auoA
1a3c	1a9xB_1653_1880	1am9A	1auvA_112_213
1a3qA_227_327	1ab8A	1amf	1auvA_214_417
1a3qA_37_226	1aba	1amm_86_174	1avaC
1a41	1ad2	1amuA	1avqA
1a48	1ae9A	1amx	1ax4A
1a4yA	1af7_11_91	1anf	1ax8
1a5t_1_207	1af7_92_284	1aoa_121_251	1axiB_131_236
1a5t_208_330	1afrA	1aoa_260_375	1axiB_32_130
1a62_1_47	1afwA_25_293	1aohA	1axn
1a62_48_125	1afwA_294_417	1aol	1ay7B
1a6q	1agjA	1aop_149_345	1ayl_1_227
1a73A	1agy	1aop_346_425	1ayl_228_540
1a76_209_316	1ah7	1aop_426_570	1ayoA
1a76_2_208	1ahsA	1aop_81_145	1ayx

of amino acids at each position,[1,20] which are described below.

Effective counts of a fragment

We use the scheme of position-specific independent counts (PSIC) to estimate the effective count of each amino acid type at a position[26,27] from the multiple sequence alignment associated with a fragment. Residue content at each position is derived from the similarity of the sequence subset that contains the given residue type at the give position.[27] After calculating the counts n_a^{PSIC} for amino acid type a in a position, the following transformation is applied[26] to derive effective count n_a:

$$n_a = -\ln \frac{20 - n_a^{PSIC}}{20} \quad (1)$$

As an effective sequence weighting scheme, PSIC provides an estimate of the number of independent observations (counts) for each amino acid type.

Evolutionary frequencies of a fragment

The multiple sequence alignment associated with a fragment is used to estimate the expected frequencies of amino acids occurring at each position. We follow the pseudocount mixture method of PSI-BLAST.[1] At each position, the observed frequencies $\{f_a\}_1^{20}$ are calculated from the effective counts:

$$f_a = \frac{n_a}{\sum_{b=1}^{20} n_b} \quad (2)$$

The expected frequencies $\{Q_a\}_1^{20}$ are calculated as the mixture of the observed frequencies $\{f_a\}_1^{20}$ and the pseudocount frequencies $\{g_a\}_1^{20}$:

$$Q_a = \frac{\alpha f_a + \beta g_a}{\alpha + \beta} \quad (3)$$

where

$$g_a = \sum_b f_b \frac{q_{ab}}{p_b} \quad (4)$$

(q_{ab} is the probability of residue pair (a, b) corresponding to the BLOSUM62 substitution matrix, p_b is the background frequency of amino acid b.) Pseudocount frequency g_a provides an estimation of the frequency of amino acid a given the observed frequencies $\{f_a\}_1^{20}$ and the prior knowledge of amino acid substitutions implied in the BLOSUM62 matrix.[28] Parameters α and β determine the proportion of the pseudocount frequencies in the mixture. α is set to $N_c - 1$, where N_c is the mean number of different symbols (including gap character) in the columns of the alignment.[1] β is set to an empirical value of 10.[1,20] We call these frequencies evolutionary frequencies since they are derived from homologous sequences.

Profile Representation of Structure-Derived Sequence Information: Structural Frequencies of Amino Acids

The equilibrium frequency of an amino acid present at a position reflects the energetic fitness of its side-chain in the local structural environment according to the Boltz-

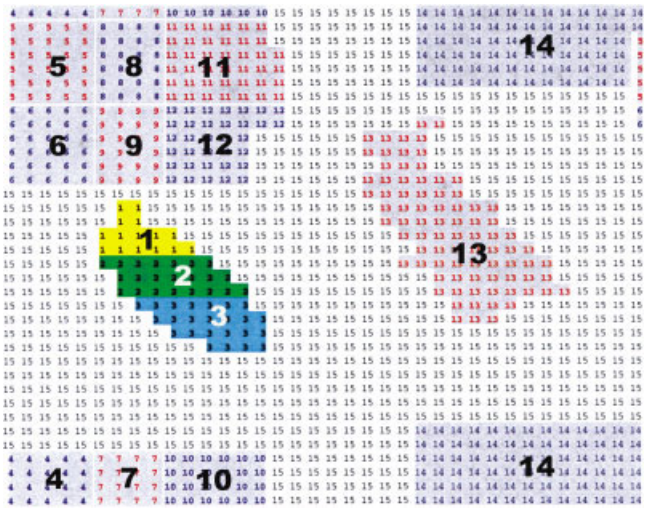

Fig. 1. Discrete representation of backbone dihedral angles by Shortle (2000). The Ramachandran plot is partitioned into 10 degree by 10 degree grids and 15 classes are defined as shown in the figure. A coarse partition consisting of 6 classes: (1,2,3), (4,5,6,7,8,9), (10,11,12), (13), (14), and (15) is also used.

mann law.[29,30] We estimate amino acid frequencies at a position based on the local structural environment of its side-chain. The scheme is similar to those used to derive statistical or knowledge-based potentials[29–32] or environment-specific substitution tables.[33,34] Two features of local structural environment are explored: backbone conformation and side-chain solvent accessibility. Due to limited sample size, we use discrete representations of backbone dihedral angles and side-chain solvent accessibilities. The backbone dihedral angles of a residue (phi and psi) have 15 classes according to the partition of the Ramachandran plot by Shortle (2002) (Fig. 1). Solvent accessibilities are calculated using the program NACCESS,[35] and the relative side-chain accessibility values are binned into a fixed number of classes (S), with each bin containing an equal number of data points. S (from 2 to 6) solvent accessibility classes combined with the 15 dihedral angles give rise to $15 \times S$ classes of local environments.

In addition, we consider the backbone conformation of the neighboring residue by considering six classes of dihedral angles of either the previous residue or the next residue,[30] resulting in $15 \times S \times 6$ local environment classes. These six classes also come from the Shortle (2002) partition of the Ramachandran plot (Fig. 1).

The probabilities of amino acids present in any of the local structural environment classes defined above are calculated from analysis of their occurrences in the 1695 SCOP domains.

$$Q_a = \frac{N_a}{\sum_{b=1}^{20} N_b} \tag{5}$$

N_a is the number of amino acid type a belonging to a certain local environmental class. We call these frequencies structural frequencies of amino acids as opposed to the evolutionary frequencies derived from a multiple sequence alignment.

The propensity of an amino acid,[30] as referenced to its background frequency, is:

$$Propensity(a) = \frac{Q_a}{p_a} \tag{6}$$

where p_a is the background frequency of amino acid a.

The COMPASS-Type Scoring Function

We need a scoring function that indicates the similarity between a target fragment and a database fragment, or from a structural perspective, the fitness of a target fragment to the local structure of a database fragment. We describe two formalisms and show that the results are equivalent. One formalism is based the log-odds scoring scheme for profile–profile comparison,[1,20] and the other formalism is based on the calculation of free energy in the framework of potentials of mean force.[29,30] We assume that positions are independent of each other in a fragment. The score for matching two fragments equals the sum of scores for matching the positions.

$$SCORE(fragment) = \sum_{p=1}^{F} SCORE(p),\ F \text{ is fragment length.}$$

The formalism based on log-odds scores

We assume that positions in a database fragment are independent amino acid generators. The probability that the effective counts ($\{n_a^1\}_1^{20}$) of amino acids at a position in a target fragment are generated by the estimated frequencies $\{Q_a^2\}_1^{20}$ at the corresponding position in a database fragment can be calculated using the multinomial distribution formula:

$$P = C(\{n_a^1\}) \cdot \prod_{a=1}^{20} (Q_a^2)^{n_a^1}. \tag{7}$$

$C(\{n_a^1\})$ is the coefficient for the multinomial distribution.

The probability of generating the effective counts by the background amino acid frequencies $\{p_a\}_1^{20}$ is:

$$P^0 = C(\{n_a^1\}) \cdot \prod_{a=1}^{20} (p_a)^{n_a^1}. \tag{8}$$

The log-odds score serves as our scoring function that characterizes the similarity between the two positions.

$$SCORE = \log(P/P^0) = \sum_{a=1}^{20} n_a^1 \log \frac{Q_a^2}{p_a} \tag{9}$$

This formula is similar to that of the scoring function used in profile–profile comparison program COMPASS,[20] which is an extension of the sequence-profile comparison program PSI-BLAST.[1]

The formalism of knowledge-based potentials

At a given position, we assume the estimated frequencies (structural or evolutionary) are equilibrium frequencies of amino acids. Under the assumptions of potentials of mean force,[29] the logarithm propensity of an amino acid approaches the free energy of moving the amino acid from an average position to a specific one.

$$E(a) = \log(Propensity(a)) = \log\left(\frac{Q_a}{p_a}\right) \tag{10}$$

Assuming the free energies are additive for independent observations of amino acids, the total free energy of transferring the observed counts of amino acids $\{n_a^1\}_1^{20}$ in a position of the target fragment, from an average position to the corresponding position in a database fragment, is:

$$E(\{n_a^1\}_1^{20}) = \sum_{a=1}^{20} n_a^1 E^2(a) = \sum_{a=1}^{20} n_a^1 \log\frac{Q_a^2}{p_a} \tag{11}$$

If $\{Q_a^2\}_1^{20}$ are the structural frequencies of the database fragment, $E(\{n_a^1\}_1^{20})$ reflects the compatibility of the target fragment amino acid counts $\{n_a^1\}_1^{20}$ to the local structural environment of the database fragment. This formula is the same as Equation (9), suggesting that the treatments from a probabilistic point of view and from the energetic point of view are equivalent.

Other Scoring Functions

We also test three other scoring functions for profile–profile comparison, as described below. Similar comparison of scoring functions has been done elsewhere.[36]

PICASSO-type scoring function

This scoring function was used in the Picasso protocol[37] for the comparison and unification of protein families and is similar to the COMPASS scoring function.[20] For two positions with observed frequencies $\{f_a^1\}_1^{20}$ and $\{f_a^2\}_1^{20}$ and effective frequencies $\{Q_a^1\}_1^{20}$ and $\{Q_a^2\}_1^{20}$, the symmetric form of the scoring function between them is

$$SCORE = \sum_{a=1}^{20} f_a^1 \log\frac{Q_a^2}{p_a} + \sum_{a=1}^{20} f_a^2 \log\frac{Q_a^1}{p_a} \tag{12}$$

Prof_sim scoring function

This scoring function was used in the prof_sim method for profile–profile comparison.[38] It involves calculation of a divergence score and a significance score, both in Jenson–Shannon entropy measure. The divergence score is

$$D = \frac{1}{2}\left[\sum_{a=1}^{20} Q_a^1 \log_2\frac{Q_a^1}{Q_a^0} + \sum_{a=1}^{20} Q_a^2 \log_2\frac{Q_a^2}{Q_a^0}\right] \tag{13}$$

where Q_a^0 is the average of Q_a^1 and Q_a^2.

The similarity score measures the Jenson–Shannon divergence between Q_a^0 and the background frequencies p_a is

$$S = \frac{1}{2}\left[\sum_{a=1}^{20} Q_a^0 \log_2\frac{Q_a^0}{R_a^0} + \sum_{a=1}^{20} p_a \log_2\frac{p_a}{R_a^0}\right] \tag{14}$$

where R_a^0 is the average of Q_a^0 and p_a.

The prof_sim score is

$$SCORE = (1 - D)(1 + S) \tag{15}$$

City-block type scoring function

This scoring function is based on the city block distance measure of two frequency vectors. This distance measure has been used for clustering and identifying local sequence motifs.[14,39]

$$SCORE = \sum_{a=1}^{20} |Q_a^1 - Q_a^2| \tag{16}$$

Fragment-Based Local Structure Prediction Using a Nearest-Neighbor Method

For the target fragment, we select those database fragments that give the top profile–profile comparison scores to make the predictions. The local structure features of these database fragments are assigned to the target fragment by a simple consensus method. Such a method can be applied to predict any local structural properties available in database fragments. In this paper, we use a three-state secondary-structure prediction as an example. Secondary-structure prediction accuracy is used to monitor the predictive power of the selected fragments. This is similar to the nearest-neighbor method for secondary-structure prediction.[40,41] The main purpose is to study the representation of sequence and structural information while secondary structure-prediction accuracy is used as an indication of the effectiveness of different representations of information and different scoring functions.

For a test sequence of length L, we derive $L - F + 1$ sequence fragments and their numerical profiles from homologous sequences as described above. Except for positions at both ends, each position in the test sequence is involved in F overlapping fragments. Positions at the ends are involved in a fewer number of overlapping fragments than F. For instance, the first position and the last position are involved in one fragment each. For each fragment, we can determine m nearest neighbors in the database of fragments based on a scoring function for profile–profile comparison. For a given position, we count the number of secondary structure types of the corresponding position in the m nearest neighbors. In the case of three-state secondary-structure types (DSSP symbols H and G for helix (H), E and B for strands (E) and others for coil (C)), we use $n_H^j(i)$, $n_E^j(i)$ and $n_C^j(i)$ to represents the counts of each type at position i for the j-th fragment that contains position i (j = 1,2,. . .F, except for positions near the ends). We observed that combination of the predictions of F consecutive fragments improves prediction accuracy. The simple way is to sum up the counts for a secondary structure type over the F fragments and compare the sums of the counts.

$$n_K(i) = \sum_{j=1}^{F} n_K^j(i), \quad K = H, E \text{ or } C \tag{17}$$

$$f_K(i) = \frac{n_K(i)}{n_H(i) + n_E(i) + n_C(i)}, \quad K = H, E \text{ or } C \tag{18}$$

$f_K(i)$ is the probability of the secondary structure type K at position i. The prediction for the secondary structure type at position i is arg $\max_K f_K(i)$. In practice, we find that the prediction accuracy is insensitive to the number of top scoring fragments selected (m) when m is between 20 and 60 (prediction accuracy variation of 0.5%). To be consistent, we use $m = 60$ for all the prediction tests. The predicted secondary structure array is compared to the real secondary structure array. We use two measures for secondary structure prediction accuracy: Q3 score and SOV (Segment Overlap) score.[42]

Combining Evolutionary and Structural Information

We describe above the derivation of position-specific amino acid frequencies for database fragments either from a multiple sequence alignment of homologs (evolutionary frequencies) or local structural environments (structural frequencies). We try to combine these two types of frequencies in the fragment selection scoring scheme. We use the 270 classes of structural frequencies considering 15 classes of backbone dihedral angles, three classes of relative side-chain solvent accessibility and six classes of backbone dihedral angles of the previous residue since such a set of structural frequencies gives the best prediction accuracy.

A simple method is to mix the frequencies through a linear combination at each position for a database fragment:

$$Q_a^M = w*Q_a^E + (1 - w)*Q_a^S \tag{19}$$

Q_a^E and Q_a^S are evolutionary frequency and structural frequency of amino acid a, respectively. w is the weight of structural frequencies and is a value between 0 and 1. The mixed frequency vector $\{Q_a^M\}_1^{20}$ can be used in the COMPASS-type scoring function [Equations (9, 11)] for fragment selection. The weight w in Equation (19) can be varied and optimized using the prediction accuracy as the target function.

Another way of mixing is similar to the way of mixing observed frequencies and pseoducount frequencies in PSI-BLAST and COMPASS [Equation (3)]. For a fragment with a small sample size of effective counts, evolutionary frequencies may not be well estimated and we want to emphasize the structural frequencies in this case.

$$Q_a^M = \frac{\alpha' Q_a^E + \beta' Q_a^S}{\alpha' + \beta'} \tag{20}$$

where Q_a^E, Q_a^S and Q_a^M are evolutionary frequency, structural frequency and mixed frequency of amino acid type a, respectively. Parameters α' and β' determine the proportion of the evolutionary frequencies in the mixture. α' is set to the total effective counts of amino acids and β' is a constant. If the sample size of the multiple alignment is large, α' is large and the evolutionary frequencies are emphasized. On the other hand, when α' is small, the structural frequencies will have a large proportion in the mixture. The mixed frequencies are used in the COMPASS-type scoring function. We test several values of β' on prediction accuracy to determine the optimal value.

The third way is to mix the COMPASS-type score generated by evolutionary frequencies (S^E) and the COMPASS score generated by structural frequencies (S^s) through a linear combination and use the mixed score (S^M) as the new scoring function for fragment selection.

$$S^M = w*S^E + (1 - w)*S^S \tag{21}$$

The weight w in Equation (21) can also be optimized using the prediction accuracy as the target function.

Derive Predicted Structural Frequencies

For a given protein sequence, a multiple sequence alignment can be obtained by PSI-BLAST searches and the fragments are derived as described previously. They are used for local structure prediction with the fragment selection method. Predicted structural frequencies in each position are derived by averaging the local structural frequencies of each predicted local structure environments in that position.

Use of Predicted Structural Frequencies in COMPASS

The predicted structural frequencies are used in local profile-to-profile alignments by COMPASS.[20] Two benchmark sets of pair-wise structural alignments are randomly selected from the FSSP database[43] with Z-score above 5: one set has 200 alignments with pair-wise sequence identity between 0 and 15%; one set has 200 alignments with sequence identity range between 15% and 30%. The first set provides the most difficult testing cases. For each sequence, PSI-BLAST is used to obtain a multiple sequence alignment for profile alignment and local structure prediction by fragment selection. Predicted structural frequencies are mixed with evolutionary frequencies using Equation (19) or Equation (20) and the parameters w and β' are optimized. When only evolutionary frequencies are used, the parameter β in Equation (3) is also optimized. The alignment produced by COMPASS is compared to the benchmark FSSP structural alignment. Three parameters are used to indicate alignment quality: the coverage of COMPASS alignment (Q_{COV}), the fraction of correctly aligned residue pairs in the region of the local alignment (Q_{LOCAL}), and the fraction of correctly aligned residue pairs compared to the full-length FSSP alignment ($Q_{DEVELOPER}$).[20]

RESULTS

Fragment Selection Using Evolutionary Sequence Profiles

One way of comparing a target fragment and a database fragment is to measure the similarity of their evolutionary sequence profiles derived from homologous proteins. Evolutionary sequence profiles are represented as the estimated

TABLE II. Comparison of Four Scoring Functions for Secondary Structure Prediction Using Eight-Residue Fragments[†]

	COMPASS	PICASSO	PROF_SIM	CITY_BLOCK
SOV score	**0.729** (0.012)	0.725 (0.012)	0.720 (0.012)	0.726 (0.012)
Q3 score	**0.752** (0.008)	0.748 (0.009)	0.739 (0.009)	0.737 (0.009)

[†]Prediction accuracy is represented as the Q3 score, standard deviation of the mean is shown in parenthesis.

effective counts and the evolutionary frequencies of amino acids (see Methods).

Scoring functions for profile–profile similarity

We test four scoring functions for profile–profile comparison. They are the COMPASS-type scoring function[20] and PICASSO,[37] which are based on log-odds scores, relative entropy measure (prof_sim)[38] and the city-block distance measure.[39] The results for secondary structure prediction accuracy for the 100 testing domains using eight-residue fragments are shown in Table II. All four methods give SOV score above 0.72 and Q3 score above 0.73. COMPASS-type scoring function gives the best average prediction accuracy, which is only slightly better than PICASSO. Q3 score using COMPASS-type scoring function is significantly better than the other three methods according to a Wilcoxon signed-rank test ($p < 0.01$). Our results are consistent with a previous comprehensive study of various profile–profile comparison methods.[36] For all the tests below we use the COMPASS-type scoring function, which is derived from a probabilistic point of view as well as from a statistical potential point of view (see Methods).

The effect of fragment length

We build fragment databases for five fragment lengths (F = 6, 7, 8, 9 and 10). Except positions at the ends, one position is involved in F consecutive fragments. The prediction for a target position combines the fragment selection results for all the fragments that contain that position (see Methods). The accuracy is determined by a window of $2*F - 1$ positions, except at the ends. At the ends, the first position uses information from one fragment and F positions; the second position uses information from two fragments and $F + 1$ positions and so on. The prediction accuracy (SOV score) is 0.722 using six-residue fragments and reaches the highest value using nine-residue fragments (0.738) (Fig. 2).

The effect of using multiple sequences

The quality of the sequence profile is directly related to the diversity and quality of the multiple sequence alignment. We assume that homologous proteins have similar local structures as the template query sequence. However, there can be much change in the local structures in remote homologs, especially in the turn and loop regions. The alignment quality for remote homologs can also be poor. For this reason, we discard those sequences that show less than 25% overall sequence identity to the query sequence from the PSI-BLAST alignment. This procedure gives about 0.5% increase in the prediction accuracy (results not shown), suggesting that very remote homologs hinder the prediction of local structures in our methods. Using a multiple sequence alignment instead of the single sequence increases the sample size of the effective counts and also results in more accurate estimation of the amino acid frequencies at a position. If multiple sequence information is not used for the target fragments, the prediction accuracy (SOV score) is only 0.628 with fragment length 8, which is much worse than when using the multiple sequence alignment (0.729). If multiple sequence information is not used for the database fragments, the accuracy is also drastically decreased to 0.670. Using single sequences for both database fragments and target fragments, the prediction accuracy is only 0.597.

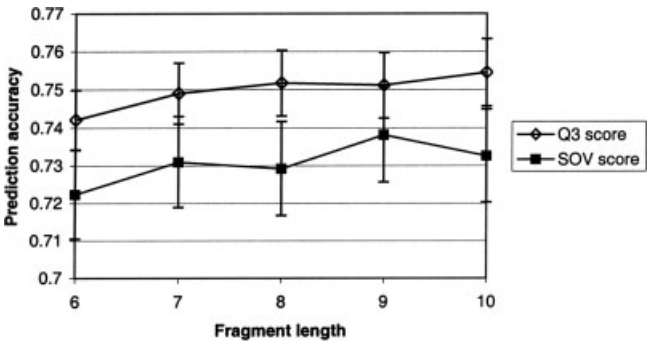

Fig. 2. The dependence of prediction accuracy on fragment length.

The effect of database size

For each fragment length (F = 6, 7, 8, 9, and 10), a fragment database is constructed from 1695 SCOP domains and contains more than 300,000 fragments. The quality of the sequence profile is related to the number of the effective counts in the multiple sequence alignment. About half of the database fragments have the average effective counts over the positions less than 14.33 for eight-residue fragments. The prediction accuracy (SOV score) is 0.709 using only fragments with average effective counts less than 14.33 in the database, which is 2% less than when using the whole database. The prediction accuracy is 0.730 when using only fragments with average effective counts greater than 14.33, which is as good as using the whole database (0.729). This suggests that the database can be reduced to half of the original size without deterioration of prediction power by keeping only those fragments with large effective counts of amino acids.

Fragment Selection Using Structural Frequencies

Another way of ranking and selecting top-scoring fragments is to measure how the sequence information of a

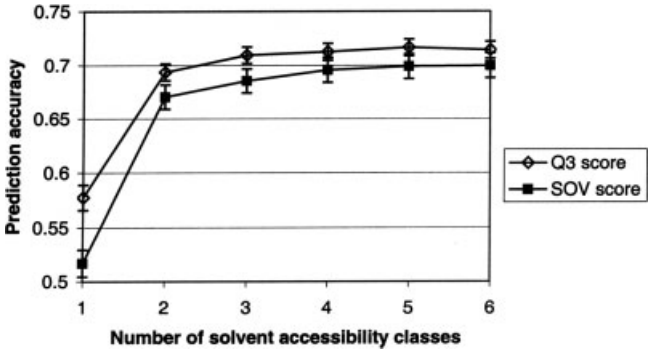

Fig. 3. The dependence of prediction accuracy on the number of side-chain solvent accessibility classes.

target fragment is compatible with the local structural environment of a database fragment. The fitness of a residue type in a position depends on the interaction of its side-chain with the surrounding residues (backbone and side-chains) as well as the solvent. It is well known that there is significant dependency of amino acid propensities on the backbone conformations.[19,30,44,45] Solvation also plays a key role in the stability of folded conformations. Structure-derived amino acid propensities[19] or substitution tables[33,34] have been used extensively in sequence–structure compatibility measurements for secondary-structure prediction and fold recognition.

In this study, we define local structural environmental classes using discrete representations of backbone torsion angles[30] and relative side-chain solvent accessible surface area.[35] We calculate amino acid frequencies and propensities for each environmental classes based on a statistical analysis of the database structures. These frequencies are used for fragment selection based on the COMPASS-type scoring function [Equations (9) and (11)]. From the viewpoint of structure-derived statistical potentials,[29] the COMPASS-type scoring function measures the free energy of transferring the observed counts of amino acids in a target fragment to the local structural environment of a database fragments (see Methods).

The effect of side-chain solvent accessibility

Considering 15 classes of backbone conformation (phi and psi dihedral angles) and S classes of relative side-chain solvent accessibility, we have $15*S$ classes of local environment. If solvent accessibility is not considered ($S = 1$), the prediction accuracy is only 0.517 with 15 classes of backbone conformation. As the number of solvent accessibility classes (S) increases, the prediction accuracy gets higher, as shown in Figure 3. Two solvent accessibility classes already give rise to about 12% increase in prediction accuracy. Prediction accuracy reaches a plateau after three classes of solvent accessibility. Further increase of the number of classes does not lead to significant increase in prediction accuracy, probably because the description of the side-chain solvent accessibility is crude.

The effect of neighboring residue conformations

The backbone dihedral angles of a position capture the interaction of the side-chain with the two nearest backbone peptide groups. To take into account the interaction of the side-chain with the next nearest peptide groups, we consider six classes of backbone conformations of the previous position ($i - 1$) or the next position ($i + 1$), together with the 15 classes of backbone conformations of the target position (i) and three classes of side-chain solvent accessibilities. This gives rise to $15*3*6 = 270$ classes of local structural environments. The results are shown in Table III. Considering six classes of backbone conformations of the previous residue ($i - 1$) gives significant increase (~3%) in prediction accuracy, while the effect of adding six classes of backbone conformations of the next residue ($i + 1$) is less significant (~0.5% accuracy increase).

Combining Evolutionary Frequencies and Structural Frequencies

Evolutionary frequencies and structural frequencies come from two independent sources. Evolutionary frequencies, derived from homologous sequences, reflect the evolutionary constraints on mutations in natural selection. Structural frequencies are derived from a statistical analysis of amino acid occurrences in different local environmental classes from a non-redundant database of structures. They reflect the effect of backbone conformation and side-chain solvent accessibility on amino acid preferences. We try to improve prediction accuracy by integrating evolutionary information and structural information. A simple way is to use a mixed frequency vector in COMPASS-scoring that is a linear combination of the evolutionary frequencies and structural frequencies [Equation (19)]. In Figure 4, we observe that such a mixing scheme does give increase in prediction accuracy. The optimal weight for evolutionary frequencies is about 0.4 as measured by SOV score and 0.6 as measured by Q3 score. At weight 0.4, the prediction accuracy is 0.746 using eight-residue fragment selection, which is significantly better than using evolutionary frequencies only (0.729) or structural frequencies only (0.716).

Another way to mix the two types of frequencies [Equation (20)] is similar to the scheme mixing observed frequencies and pseudocount frequencies in PSI-BLAST [Equation (3)]. The weight of the evolutionary frequencies is proportional to the total amino acid effective counts, since the quality of evolutionary frequencies is closely related to the total effective counts of the multiple sequence alignment. The weight of the structural frequencies is set to a fixed value of β'. The best prediction accuracy (SOV: 0.748, Q3: 0.765) is achieved where β' is about 10 (Fig. 5).

On the other hand, a linear combination at the COMPASS log-odds score level [Equation (21)] improves SOV score less significantly at the optimal weight 0.4 (0.737) and does not give any improvement in Q3 score (Fig. 6).

A Test on CASP5 Targets

Our previous focus is to study the effect of various factors on the fragment selection. We do not intend to design our method specifically for secondary structure

TABLE III. The Effect of Neighboring Residues on Prediction Accuracy Using Eight-Residue Fragments

	$15(i)*3(S)$ residue i	$15(i)*3(S)*6(i-1)$ residues i and $i-1$	$15(i)*3(S)*6(i+1)$ residues i and $i+1$
SOV score	0.686 (0.011)	**0.716** (0.011)	0.691 (0.012)
Q3 score	0.709 (0.008)	**0.732** (0.008)	0.710 (0.008)

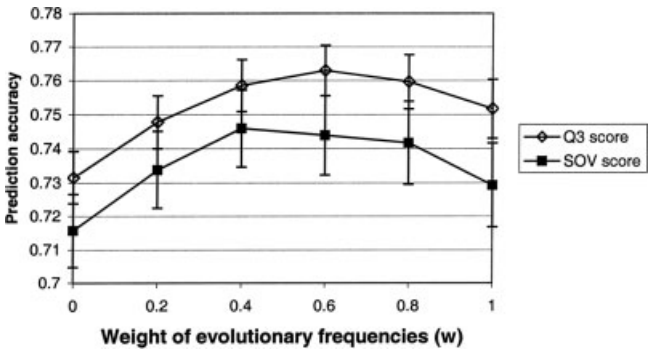

Fig. 4. The dependence of prediction accuracy on the weight of evolutionary frequencies in Equation (19).

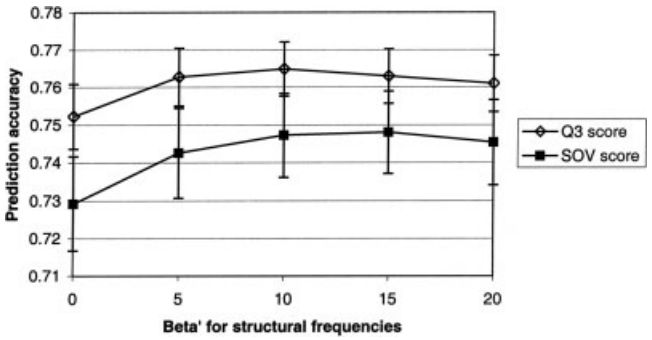

Fig. 5. The dependence of prediction accuracy on the constant β′ in Equation (20).

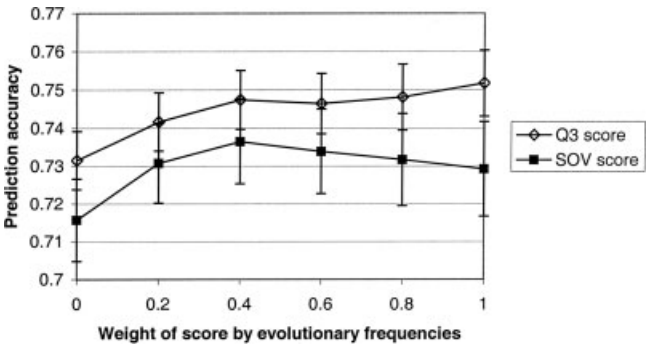

Fig. 6. The dependence of prediction accuracy on the weight of COMPASS-type score calculated using evolutionary frequencies in Equation (21).

prediction. As mentioned above, our method can be used to predict any local structural properties residing in structural fragments. Secondary structure prediction accuracy is rather used to monitor the prediction power when varying factors such as scoring functions and determine the optimal weight for frequency combination. However, it is valuable to compare our method to other commonly used methods to see if it can give reasonable results. We test our method on 56 target proteins in CASP5. The results of our method and three other methods (PSIPRED, SAM-T99, and SSPRO2) are shown in Table IV. In our method, combining evolutionary frequencies and structural frequencies again gives improvement in SOV scores over using evolutionary frequencies only. The best average SOV score is 0.77 for our method, only slightly worse than the three methods based on neural network training, which are more complex and based on optimization of many parameters that are hard to interpret. Tested on four difficult domains as classified by the CASP5 assessor, our method also gives results comparable to the other three methods. Although our method is simpler than others, it still gives reasonable secondary structure prediction results tested on CASP5 targets.

Predicted Structural Frequencies in Local Profile-to-Profile Alignment by COMPASS

Table V shows that for the difficult testing cases (sequence identity between 0 and 15%), the best average coverage of COMPASS alignments by mixing the predicted structural frequencies and evolutionary frequencies by Equation (20) (optimization of β′) is 0.440, which gives rise to almost 50% increase over the best average coverage (0.30) using only evolutionary frequencies [Equation (3), optimization of β]. The alignment accuracies (Q_{LOCAL} and $Q_{DEVELOPER}$) also get higher. This effect is not as prominent in the easier testing case (sequence identity between 15% and 30%). This suggests that predicted structural information is most useful when aligning sequences that are very divergent. Mixture by Equation (19) is also tested and the results are similar to that of Equation (20) (data not shown).

DISCUSSION

The rapidly growing sequence and structure databases provide us valuable evolutionary and structural information about proteins. While it is important to study protein structure and function from the basic physicochemical principles, knowledge-based approaches have gained popularity and made contributions in various fields of computational biology such as structure prediction,[46] protein folding,[47] comparative modeling,[48] and sequence design.[49] Ab initio prediction of three-dimensional structure still remains a very complex problem despite recent advances.[50,51] Prediction of local structure features is much easier and can provide valuable structural information for new protein families. In this study we apply a knowledge-based

TABLE IV. Performance on CASP5 Targets

	Prof	Mix_freq 0.6[a]	Mix_beta′ 10[b]	PSIPRED	SAM-T99	Sspro2
SOV score (all targets)	0.763 (0.012)	0.771 (0.013)	0.769 (0.012)	0.795 (0.011)	0.782 (0.011)	0.778 (0.011)
Q3 score (all targets)	0.769 (0.010)	0.772 (0.010)	0.773 (0.010)	0.802 (0.009)	0.789 (0.008)	0.797 (0.009)
SOV score (4 hard domains)	0.493	0.527	0.538	0.509	0.530	0.490
Q3 score (4 hard domains)	0.576	0.575	0.571	0.600	0.609	0.600

[a]Equation (19), $w = 0.6$.
[b]Equation (21), $\beta' = 10$.

TABLE V. Pairwise Profile-to-Profile Alignment Coverage and Accuracy by COMPASS

	Q_{COV}	Q_{LOCAL}	$Q_{DEVELOPER}$
Identity range: 0 ~ 15%			
Evolutionary frequencies only[a]	0.30	0.333	0.172
Mix with predicted structural frequencies[b]	0.440	0.346	0.215
Identity range: 15 ~ 30%			
Evolutionary frequencies only[a]	0.630	0.59	0.455
Mix with predicted structural frequencies[b]	0.695	0.59	0.475

[a]The parameter β is optimized in Equation (3) to achieve the best coverage or accuracy.
[b]The parameter β′ is optimized in Equation (20) to achieve the best coverage or accuracy.
The values are averaged over the 200 alignments in each set.

approach to the prediction of local protein structures. We explore information of positional amino acid preferences from both homologous sequences and database structures and attempt to combine these two information sources in a meaningful and effective way.

General Methodology of Local Structure Prediction by Fragment Selection

Our method is based on fragment ranking and selection. The idea of using database nearest neighbors has already been applied in many local structure prediction methods.[29,40,41,52,53] The rationale of this method is that there is a large portion of local structures with strong sequence signals that are reused in non-homologous structures.[14,39] For these local structures, sequence similarities often suggest structural similarities. Given a target fragment with only sequence information, we identify database fragments of the same length that are most likely to give the sequence patterns of the target fragments. The local structure features of the top scoring database fragments are used to assign the local structure features to the target fragment by a simple consensus method. In this study, we use secondary structure types as a local structural feature, while other features such as backbone dihedral angles, solvent accessibilities, and the classes of protein building blocks[11] can be predicted in the same way. Rather than just providing another method of secondary structure prediction, our major goal is to use the secondary structure prediction accuracy to monitor the effects of various factors in representing the homologous and structural information in the fragment selection process. The selected top-scoring fragments can also serve as the initial point for ab initio structure prediction methods based on fragment assembly.[54]

The key components of our method are the profile representation of sequence information and the scoring function for measuring fragment similarity. The sequence profile representation is based on a probabilistic model of positional amino acid usages. We use the COMPASS-type scoring function of profile–profile comparison[20] for ranking and selecting fragments. There are many other types of scoring functions for comparison of positional sequence profiles.[37–39,55–57] The four scoring functions tested in this study give similar secondary structure prediction accuracy above 0.73, suggesting they are all valid in reflecting profile similarity. The COMPASS-type scoring function gives significantly better results than the other three scoring functions in eight-residue fragment-based secondary-structure prediction as measured by Q3 score, justifying our model based on a probabilistic representation of sequence information.

Representation of Evolutionary and Structural Information

The numerical representation of sequence profiles contains two parts: the effective counts of amino acids and the estimated frequencies of amino acids. The effective counts are estimated from a multiple sequence alignment of homologous sequences, taking into account the uneven sampling of the sequence space for naturally occurring sequences.[26] The positional amino acid frequencies are estimated either from homologous sequences (evolutionary frequencies) or the local structural environment (structural frequencies). The derivation of evolutionary frequencies follows the pseudocount strategy that has been demonstrated to be effective in integrating prior knowledge of amino acid replacements.[1,20,28,58] We show that the effective number of homologous sequences contributes much to the prediction accuracy when using evolutionary frequencies. Prediction accuracy gets much worse when multiple sequence information is not used in either the target fragments or the database fragments. Exclusion of fragments with low effective counts from the database does not compromise prediction accuracy.

The estimation of positional amino acid frequencies from local structural environment is based on the derivation of statistical potentials of mean force under the Boltzmann

assumption.[29,59,60] The logarithm of the propensity of an amino acid in a local environment can then be viewed as the free energy of exchanging this amino acid from an average local environment to a specific one.[30] The COMPASS-type scoring function for ranking the database fragments can be interpreted as the total free energy of the observed amino acids in a target fragment. We show that the derivation of the COMPASS-type scoring function is consistent in the probabilistic formalism and the formalism of structure-derived statistical potentials for position-specific amino acid preferences.

The structural properties considered in constructing the local environment classes are simply the backbone conformation and side-chain solvent accessibility. More specific interactions are ignored. However, these two properties capture the most important information for specifying the fitness of a certain type of side-chain to its local environment.[44,61] Due to limited sample size, we use a discrete representation of local structure environments. We show that side-chain solvent accessibility provides much additional information to backbone conformations. Finer description of local environments by increasing the number of solvent accessibility classes or considering neighboring residue backbone conformation achieves better prediction accuracy. We find that the backbone conformation of the preceding residue has a larger positive effect than the following residue (Table III). This is consistent with the observation that the side-chains of most amino acids prefer the gauche + $\chi 1$ rotamer which points toward the preceding residue.[31,62] The best set of structural frequencies of the 270 local environment classes (15 classes of backbone conformation, three classes of side-chain accessibility, and six classes of backbone conformation of the previous residue) gives the SOV prediction accuracy of 0.716, only slightly worse than using evolutionary frequencies (0.729) using eight-residue fragments.

Combining Evolutionary Information and Structural Information

Evolutionary frequencies and structural frequencies have advantages and limits and they can potentially compliment each other to achieve better estimation of positional amino acid preferences.

Evolutionary frequencies reflect positional amino acid usages in naturally occurring sequences that are homologous to the template structure. These naturally occurring sequences all show compatibility with the structure. However, the quality of evolutionary frequencies depends greatly on the effective counts of amino acids and the quality of multiple sequence alignment. The effective estimation of evolutionary frequencies requires a relatively large sample of non-redundant homologous sequences. Functional constraints in homologous sequences can be another problem with evolutionary frequencies. It has been shown that functional residues are often not optimized for the structure.[63] Amino acid frequency biases caused by functional constraints may hinder the selection of structurally similar fragments from non-homologous structures.

Structural frequencies are estimated from structures and they do not rely on other homologous sequences. However, the disadvantage of structural frequencies is that they are estimated from a coarse description of local environments consisting of only a limited number of discrete environmental classes. In practice, structural frequencies perform worse than evolutionary frequencies in prediction accuracy. We show that mixing the evolutionary frequencies and structural frequencies, either by a simple linear combination or taking into account the effective counts of the multiple sequence alignment, results in improvement in prediction accuracy. This suggests that mixed frequencies are a better estimation of position-specific amino acid preferences. We also show that mixing the log-odds scores generated by evolutionary frequencies and structural frequencies using the COMPASS formula does not help much in prediction accuracy. However, many existing scoring functions used in structure prediction combine sequence and structural information by a linear combination of log-odds form scores.[64,65] They have been shown to give better results than using sequence information alone. Our findings suggest that there might be better ways to combine sequence and structure information than just a linear combination of log-odds form scores.

Predicted Structural Information Applied in Sequence Alignment

The rationale of using predicted structural information in alignment is that local structures can retain similarity while sequences diverge in evolution. Predicted structural frequencies contain additional information not available from sequences directly. First, the local structure prediction method is based on selection of similar fragments from a fragment database. The structural frequencies for a number of local structural environment classes are also derived from statistical analysis of database structures. Thus, predicted structural frequencies contain information from database structures. Second, a window of $2*F - 1$ positions are used to predict the local structure for each position (except at the sequence termini). This takes into account positional correlations in local structures. Thus predicted structural frequencies are influenced not only by the amino acid preferences at a given position, but also by the neighboring positions. Mixtures of predicted structural frequencies and evolutionary frequencies might provide better estimation of position-specific amino acid preferences. The tests on local profile-to-profile alignments by COMPASS demonstrate that predicted structural frequencies do help to improve local alignment quality in terms of coverage and accuracy.

ACKNOWLEDGMENT

We thank Ruslan Sadreyev for helpful discussions, and James Wrabl and Sara Cheek for critical reading of the manuscript.

REFERENCES

1. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation

of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
2. Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14: 755–763.
3. D'Alfonso G, Tramontano A, Lahm A. Structural conservation in single-domain proteins: implications for homology modeling. J Struct Biol 2001;134:246–256.
4. Jones TA, Kleywegt GJ. CASP3 comparative modeling evaluation. Proteins 1999;Suppl 3:30–46.
5. Fischer D, Eisenberg D. Protein fold recognition using sequence-derived predictions. Protein Sci 1996;5:947–955.
6. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
7. Hutchinson EG, Thornton JM. A revised set of potentials for beta-turn formation in proteins. Protein Sci 1994;3:2207–2216.
8. Aurora R, Rose GD. Helix capping. Protein Sci 1998;7:21–38.
9. Unger R, Harel D, Wherland S, Sussman JL. A 3D building blocks approach to analyzing and predicting structure of proteins. Proteins 1989;5:355–373.
10. Rooman MJ, Rodriguez J, Wodak SJ. Automatic definition of recurrent local structure motifs in proteins. J Mol Biol 1990;213: 327–336.
11. de Brevern AG, Etchebest C, Hazout S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins 2000;41:271–287.
12. Kolodny R, Koehl P, Guibas L, Levitt M. Small libraries of protein fragments model native protein structures accurately. J Mol Biol 2002;323:297–307.
13. Hunter CG, Subramaniam S. Protein fragment clustering and canonical local shapes. Proteins 2003;50:580–588.
14. Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. J Mol Biol 1998;281:565–577.
15. Bystroff C, Thorsson V, Baker D. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J Mol Biol 2000;301:173–190.
16. Yang AS, Wang LY. Local structure-based sequence profile database for local and global protein structure predictions. Bioinformatics 2002;18:1650–1657.
17. Yang AS, Wang LY. Local structure prediction with local structure-based sequence profiles. Bioinformatics 2003;19:1267–1274.
18. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
19. Chou PY, Fasman GD. Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol Relat Areas Mol Biol 1978;47:45–148.
20. Sadreyev R, Grishin N. COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. J Mol Biol 2003;326:317–336.
21. Wang G, Dunbrack RL, Jr. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
22. 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;247:536–540.
23. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
24. Karplus K, Karchin R, Barrett C, Tu S, Cline M, Diekhans M, Grate L, Casper J, Hughey R. What is the value added by human intervention in protein structure prediction? Proteins 2001;Suppl 5:86–91.
25. Pollastri G, Przybylski D, Rost B, Baldi P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins 2002;47:228–235.
26. Pei J, Grishin NV. AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 2001;17:700–712.
27. Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber B, Tumanyan VG, Kuznetsov EN. PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. Protein Eng 1999;12:387–394.
28. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
29. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 1990;213: 859–883.
30. Shortle D. Composites of local structure propensities: evidence for local encoding of long-range structure. Protein Sci 2002;11:18–26.
31. Dunbrack RL Jr, Karplus M. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. J Mol Biol 1993;230:543–574.
32. Kortemme T, Morozov AV, Baker D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. J Mol Biol 2003;326:1239–1259.
33. Rice DW, Eisenberg D. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. J Mol Biol 1997;267:1026–1038.
34. Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 2001;310: 243–257.
35. Hubbard SJ, Campbell SF, Thornton JM. Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. J Mol Biol 1991;220:507–530.
36. Mittelman D, Sadreyev R, Grishin N. Probabilistic scoring measures for profile–profile comparison yield more accurate short seed alignments. Bioinformatics 2003;19:1531–1539.
37. Heger A, Holm L. Picasso: generating a covering set of protein family profiles. Bioinformatics 2001;17:272–279.
38. Yona G, Levitt M. Within the twilight zone: a sensitive profile–profile comparison tool based on information theory. J Mol Biol 2002;315:1257–1275.
39. Han KF, Baker D. Recurring local sequence motifs in proteins. J Mol Biol 1995;251:176–187.
40. Yi TM, Lander ES. Protein secondary structure prediction using nearest-neighbor methods. J Mol Biol 1993;232:1117–1129.
41. Salamov AA, Solovyev VV. Protein secondary structure prediction using local alignments. J Mol Biol 1997;268:31–36.
42. Zemla A, Venclovas C, Fidelis K, Rost B. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins 1999;34:220–223.
43. Holm L, Sander C. Touring protein fold space with Dali/FSSP. Nucleic Acids Res 1998;26:316–319.
44. Srinivasan R, Rose GD. A physical basis for protein secondary structure. Proc Natl Acad Sci USA 1999;96:14258–14263.
45. Street AG, Mayo SL. Intrinsic beta-sheet propensities result from van der Waals interactions between side-chains and the local backbone. Proc Natl Acad Sci USA 1999;96:9074–9076.
46. Sippl MJ. Knowledge-based potentials for proteins. Curr Opin Struct Biol 1995;5:229–235.
47. Kussell E, Shimada J, Shakhnovich EI. A structure-based method for derivation of all-atom potentials for protein folding. Proc Natl Acad Sci USA;99:5343–5348.
48. Fiser A, Feig M, Brooks CL, 3rd, Sali A. Evolution and physics in comparative protein structure modeling. Acc Chem Res 2002;35: 413–421.
49. Russ WP, Ranganathan R. Knowledge-based potential functions in protein design. Curr Opin Struct Biol 2002;12:447–452.
50. Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CE, Baker D. Rosetta in CASP4: progress in ab initio protein structure prediction. Proteins 2001;Suppl 5:119–126.
51. Skolnick J, Kolinski A, Kihara D, Betancourt M, Rotkiewicz P, Boniecki M. Ab initio protein structure prediction via a combination of threading, lattice folding, clustering, and structure refinement. Proteins 2001;Suppl 5:149–156.
52. Topham CM, McLeod A, Eisenmenger F, Overington JP, Johnson MS, Blundell TL. Fragment ranking in modelling of protein structure. Conformationally constrained environmental amino acid substitution tables. J Mol Biol 1993;229:194–220.
53. van Vlijmen HW, Karplus M. PDB-based protein loop prediction: parameters for selection and methods for optimization. J Mol Biol 1997;267:975–1001.
54. Bonneau R, Baker D. Ab initio protein structure prediction: progress and prospects. Annu Rev Biophys Biomol Struct 2001;30: 173–189.
55. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.
56. Pietrokovski S. Searching databases of conserved sequence re-

gions by aligning protein multiple-alignments. Nucleic Acids Res 1996;24:3836–3845.
57. Jaroszewski L, Rychlewski L, Godzik A. Improving the quality of twilight-zone alignments. Protein Sci 2000;9:1487–1496.
58. Tatusov RL, Altschul SF, Koonin EV. Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. Proc Natl Acad Sci USA 1994;91:12091–12095.
59. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 1998;275:895–916.
60. Miyazawa S, Jernigan RL. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. Proteins 1999;34:49–68.
61. Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. Nature 1986;319:199–203.
62. Ponder JW, Richards FM. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. J Mol Biol 1987;193:775–791.
63. Shoichet BK, Baase WA, Kuroki R, Matthews BW. A relationship between protein stability and protein function. Proc Natl Acad Sci USA 1995;92:452–456.
64. Fischer D. Hybrid fold recognition: combining sequence derived properties with evolutionary information. Pac Symp Biocomput 2000:119–130.
65. Shan Y, Wang G, Zhou HX. Fold recognition and accurate query-template alignment by a combination of PSI-BLAST and threading. Proteins 2001;42:23–37.