Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry

Rachel Karchin *

Center for Biomolecular Science and Engineering

Baskin School of Engineering

University of California

Santa Cruz, CA

rachelk@soe.ucsc.edu

Melissa Cline

Affymetrix, Inc.

Emeryville, CA

melissa_cline@affymetrix.com

Yael Mandel-Gutfreund

Department of Chemistry and Biochemistry

University of California

Santa Cruz, CA

yael@biology.ucsc.edu

Kevin Karplus

Computer Engineering

Baskin School of Engineering

University of California

Santa Cruz, CA

karplus@soe.ucsc.edu

Abstract

An important problem in computational biology is predicting the structure of the large number of putative proteins discovered by genome sequencing projects. Fold-recognition methods attempt to solve the problem by relating the *target* proteins to known structures, searching for *template* proteins homologous to the target. Remote homologs which may have significant structural similarity are often not detectable by sequence similarities alone.

To address this, we incorporated predicted local structure, a generalization of secondary structure, into *two-track* profile HMMs. We did not rely on a simple helix-strand-coil definition of secondary structure,

*To whom correspondence should be addressed. Mailing address: Center for Biomolecular Science and Engineering, Jack Baskin School of Engineering, University of California, Santa Cruz, CA 95064 USA. Phone: 1-831-459-5297, Fax: 1-831-459-4829

but experimented with a variety of local structure descriptions, following a principled protocol to establish which descriptions are most useful for improving fold recognition and alignment quality.

On a test set of 1298 non-homologous proteins, HMMs incorporating a 3-letter STRIDE alphabet improved fold recognition accuracy by 15% over amino-acid-only HMMs and 23% over PSI-BLAST, measured by ROC-65 numbers.

We compared two-track HMMs to amino-acid-only HMMs on a difficult alignment test set of 200 protein pairs (structurally similar with 3-24% sequence identity). HMMs with a 6-letter STRIDE secondary track improved alignment quality by 62%, relative to DALI structural alignments, while HMMs with a STR track (an expanded DSSP alphabet that subdivides strands into six states) improved by 40% relative to CE.

Key words: protein structure prediction, two-track HMM, multi-track HMM, information theory, neural network, alignment, secondary structure

1 Introduction

Structural information about available templates has been shown to improve performance of both profile and threading fold-recognition methods [1, 2, 3, 4, 5]. In this paper, we evaluate the results of enriching HMMs, built using SAM software [6, 7], with frequency profiles derived from predicted one-dimensional structure strings, such as secondary structure. The models are designed to search a template database for structurally similar remote homologs and to align target-template pairs.

Local protein structure describes both the environment of an individual residue and its relationship to neighboring residues in three-dimensional space. A *local structure alphabet* is a discrete encoding of one or more properties of local protein structure that clusters residues with similar properties into the same state.

Most prediction of local protein structure has centered around a three-state classification of *secondary structure* that places a residue in one of three categories: *helix*, *sheet*, or *coil*. This is a broad classification, as it provides little information about the coil category that accounts for 45% of protein structure on average [8]. Although there are many methods for defining fine-grained *alphabets* of local structure [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23], there has not been much work exploring whether these alphabets can be used to improve fold recognition or alignments.

There have been previous attempts to improve fold recognition and target-template alignments by incorporating predicted secondary structure information into profile HMMs [2, 24, 25]. One group used only secondary structure information to build a template library of HMMs, covering all topology families in the CATH database [26]. Each HMM was trained on the observed secondary structures of a set of structurally similar proteins with low sequence identity. To identify the fold of a target protein, they predicted its secondary structure string (with a neural network)

and searched the library for the HMM most likely to have generated the string [2, 24]. Another group built a template library of HMMs that combined amino acid and secondary structure information. Their library covered a representative set of all proteins with known structure. For each representative protein, an HMM was trained from an HSSP alignment of homologous amino acid sequences [27]. To find the best template for a target protein, they computed the joint probability of the target's amino acid sequence and (neural-net) predicted secondary structure string, with respect to each HMM in the library [25].

Our approach differs in several important ways. We use SAM-T2K, an iterated alignment HMM method [28, 29, 30] to train an HMM from a single target (amino acid) sequence. The alignment produced by the final iteration of SAM-T2K is used both to compute the amino-acid emission probabilities of the target HMM and as the input to a secondary-structure-prediction neural network. The neural net estimates probabilities for each secondary structure state at each position in the target protein, and we incorporate these into the target HMM's match states as secondary structure emission probabilities. Instead of scoring a target sequence against a template HMM library, we score a template library of amino acid and known secondary structure sequences against a two-track target HMM. Our method has a direct probabilistic interpretation, in that it yields an alignment that best explains conservation of both amino acid sequence and secondary structure.

While other groups have relied on helix-strand-coil descriptions of secondary structure, we have explored a variety of local structure descriptions, investigating whether there is a preferred a local structure alphabet.

For our purposes, the best local structure alphabets are

- conserved in proteins having the same fold,
- predictable from amino acid sequence,
- able to improve template selection, and
- able to improve target-template alignments.

We present a protocol for evaluating local protein structure alphabets and apply the protocol to nine alphabets of protein backbone geometry. Incorporating predicted local structure information substantially increases fold-recognition and alignment accuracy of our HMM-based methods. Among the local structure alphabets tested, we find that a novel alphabet based on detailed secondary structure states, including classifications of beta strand orientation, is most useful for improving alignments. Fold recognition improvement is relatively insensitive to choice of a particular local structure alphabet, but we are currently getting best results with a simple three-state classification of secondary structure.

A webserver for two-track HMM fold recognition and target-template alignment is available at
http://www.soe.ucsc.edu/research/compbio/HMM-apps/T02-query.html

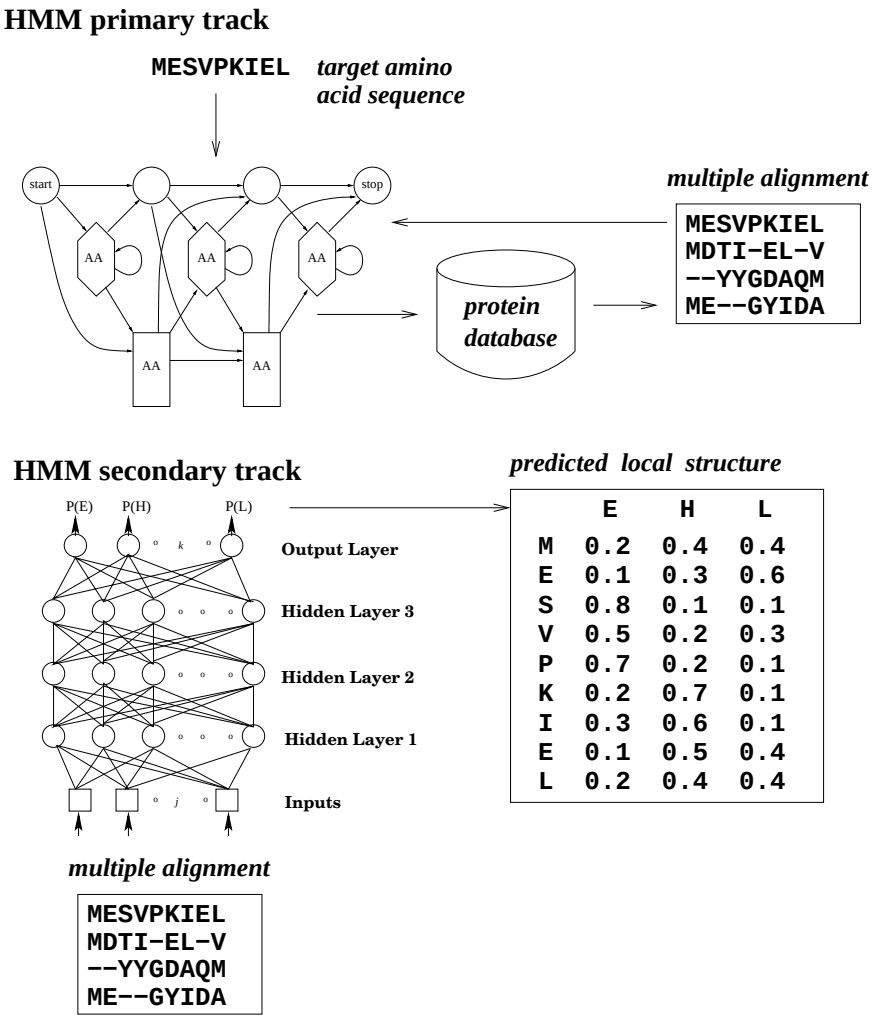


Figure 1: The 2-track HMM has a primary track of amino-acid emissions and a secondary track of local-structure-alphabet emissions. The primary track is constructed with the SAM-T2K iterated alignment algorithm. The secondary track is modeled with predicted local structure probabilities that are estimated by a neural network. Input to the neural net is the final multiple alignment generated by SAM-T2K.

2 Methods and Materials

2.1 Two-track HMMs

With two-track profile HMMs, each match node contains emission probabilities of predicted local structure information, in addition to amino-acid emission probabilities [30]. As shown in Figure 1, the primary *track* of amino-acid emission probabilities is built with the SAM-T2K iterated alignment algorithm, and the secondary track of local structure emission probabilities is estimated by a neural network. For example, the third amino acid is a serine (S), and the neural net estimates the S to be in state E with 80% probability.

To find the best match to the target in a template library, each template protein X is aligned to the target model

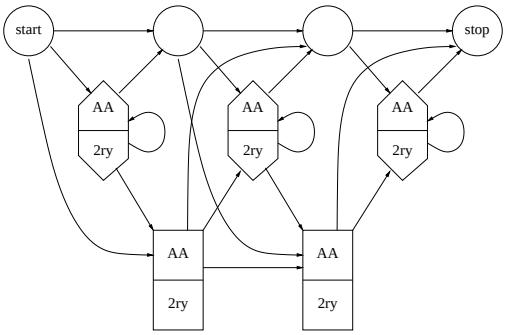


Figure 2: This picture shows how the two-track SAM-T2K target HMM is organized. The "AA" and "2ry" labels in the boxes refer to emission-probability tables for amino acids and local structure labels, respectively [30]. The match state probabilities are estimated from the training alignment and the insertion states get background probabilities. The difference from the AA-only profile HMMs used previously in SAM is that the match and the insert states describe emission probabilities for both the amino-acid and local structure alphabets. A real profile HMM has as many match states as alignment columns in the multiple alignment.

and the alignment is scored by computing the joint probability of its amino acid sequence ($A = a_1, a_2, \cdots, a_n$) and a string that encodes its local structure ($L = l_1, l_2, \cdots, l_n$), over all possible paths through the HMM:

$$\begin{aligned} P(X|M) &= P(A, L|\theta_1, \theta_2, \tau) \\ &= \sum_{s_1,\cdots,s_n} \prod_i P(a_i, l_i|s_i, \theta_1, \theta_2) P(s_i|s_{i-1}, \tau) \\ &= \sum_{s_1,\cdots,s_n} \prod_i \theta_1(a_i|s_i)\theta_2(l_i|s_i)\tau_{s_i|s_{i-1}} \end{aligned} \quad (1)$$

$s_1, \cdots, s_n$ is a path of character-emitting states through the model, θ_1 represents the amino acid emission distributions, θ_2 the local structure alphabet emission distributions, and τ the transition probabilities between character-emitting states. The alignment generated is the most probable, given both the template's amino-acid sequence and its known local structure. We deviate somewhat from a strict probabilistic interpretation of the emission probabilities, by using a weighted combination of probabilities from the two emission tables as our *match score*. That is, we use $\theta_1^{w_1}(a_i|s_i)\theta_2^{w_2}(l_i|s_i)$ instead of $\theta_1(a_i|s_i)\theta_2(l_i|s_i)$ as a strictly probabilistic interpretation would require. We have gotten the best results when most of the weight is placed on the amino-acid emission scores ($w_1 = 1, w_2 = 0.3$).

Figure 2 shows a graphical depiction of a small two-track HMM.

2.2 Alphabet Descriptions

For our evaluation experiments, we selected a sample of nine alphabets describing protein backbone geometry. These included the two most widely used secondary structure alphabets, DSSP and STRIDE, three-state reduced versions of STRIDE-EHL and DSSP-EHL; a backbone-fragment alphabet called `Protein Blocks` (PB) [23]; an alphabet (ANG)

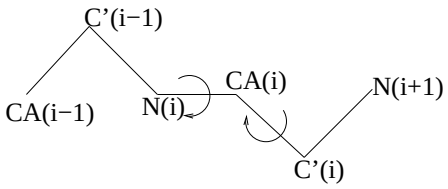


Figure 3: The pair of dihedral backbone angles ϕ and ψ are defined by the atoms C'-N-C_α-C' (ϕ) and N-C_α-C'-N (ψ).

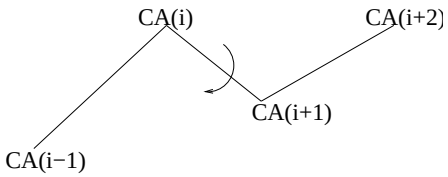


Figure 4: The backbone angle ALPHA is defined for a residue i as the virtual dihedral angle between C_α atoms of residues $i-1$, i, i+1 and i+2.

of ϕ-ψ classes developed by Bystroff [22] (ϕ and ψ are dihedral angles along the protein backbone [31], shown in Figure 3); and several novel alphabets. Two of our novel alphabets are based on: the backbone angle ALPHA, defined for a residue i as the virtual dihedral angle between C_α atoms of residues $i-1$, i, i+1 and i+2 (Figure 4); and TCO, defined for a residue i as the cosine of the dihedral angle around its carbonyl group and that of residue $i-1$ (Figure 5).

DSSP [9]. We use a seven-letter version of the secondary structure alphabet developed by Kabsch and Sander: E (beta strand), H (alpha helix), T (turn), S (bend), G (3-10 helix), B (short beta bridge) and C (random coil). We included I (pi helix) with H, since it is too rare to be predictable or to have much effect on our results. Assignments are based on patterns of hydrogen bonding.

STRIDE [16]. In this six-letter (EBGHTC) secondary structure alphabet, assignments are based on ϕ-ψ angles and H-bond energies.

DSSP-EHL, STRIDE-EHL. These are three-letter (EHL) secondary structure alphabets that reduce DSSP and STRIDE assignments to helix, strand, or coil. In this mapping, G is included in the helix class, B in the strand class, and S and T in the coil class.

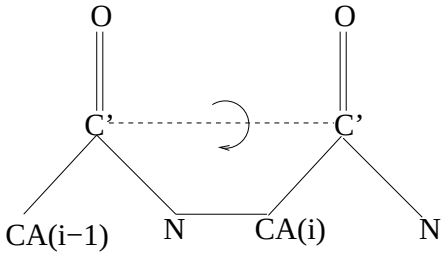


Figure 5: TCO is defined for a residue i as the cosine of the dihedral angle around its carbonyl group and that of residue $i-1$.

ANG	Bystroff	ϕ	ψ
H	H	-61.91	-45.20
G	G	-109.78	20.88
P	B	-70.58	147.22
E	E	-132.89	142.43
D	d	-135.03	77.26
N	b	-85.03	72.26
Y	e	-165.00	175.00
L	L	55.88	38.62
T	l	85.82	-0.03
S	x	80.00	-170.00
unused	c	cis peptide	

Table 1: Centers of the ten classes in the ANG alphabet.

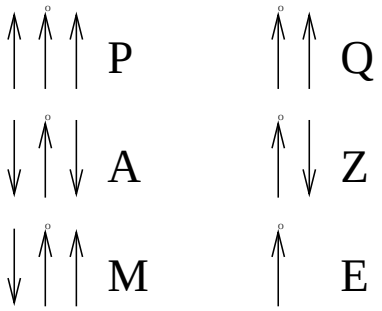


Figure 6: Six letters in the STR alphabet, which expand on the DSSP "E" or strand state. Dots indicate the strand of the residue being assigned. In a beta sheet, this strand is either surrounded by two parallel partners "P", two anti-parallel partners "A" or one anti-parallel and one parallel partner "M". Edge strands (that have only one beta strand partner) have either a parallel partner "Q" or an anti-parallel partner "Z". Finally, we retain the "E" label for strand residues to which DSSP assigns no partners (generally beta bulges).

PROTEIN BLOCKS [23]. This automatically designed alphabet looks at overlapping residue fragments of length five (chosen empirically), extracted from a non-redundant set of structures, and encodes them as sequence "windows" of ϕ-ψ pairs called *dihedral vectors*. An unsupervised *Kohnen self-organizing map* network [32, 33] was trained on the dihedral vectors with an *RMSDA* (root mean square deviations on angular values) distance metric, to produce a set of 16 clusters, each with a representative Protein Block (PB).

ANG [22]. This alphabet is based on the ϕ-ψ alphabet used in HMMSTR. Bystroff et al. partitioned the ϕ-ψ plane [31] into ten regions, using the *k-means* algorithm [34] on all trans ϕ-ψ pairs in PDB. Cluster boundaries were calculated with a *Voronoi* method. While Bystroff et al. assigned all cis residues to an eleventh cluster, we distributed the cis residues among the other 10 classes according to their ϕ-ψ values. Table 1 shows the centers of the 10 classes, and Figure 3 illustrates the ϕ-ψ angles.

STR. This novel alphabet is an enhanced version of DSSP that subdivides DSSP letter E (beta strand) into six letters (Figure 6), according to properties of a residue's relationship to its strand partners.

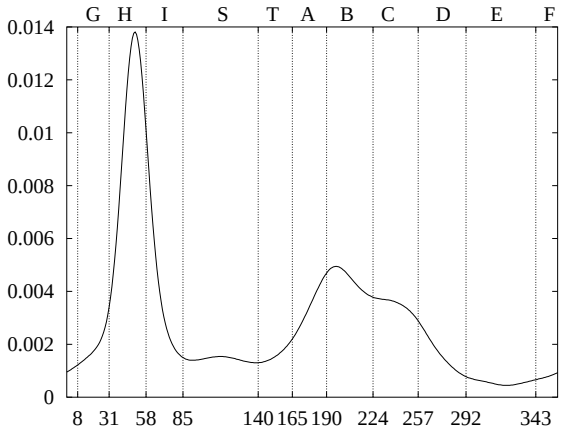


Figure 7: Smoothed histogram of ALPHA angle distribution. The 11-letter ALPHA alphabet was chosen based on break points in this curve and in the curves for $P(\text{ALPHA}|\text{amino acid})/P(\text{ALPHA})$ for each of the amino acids.

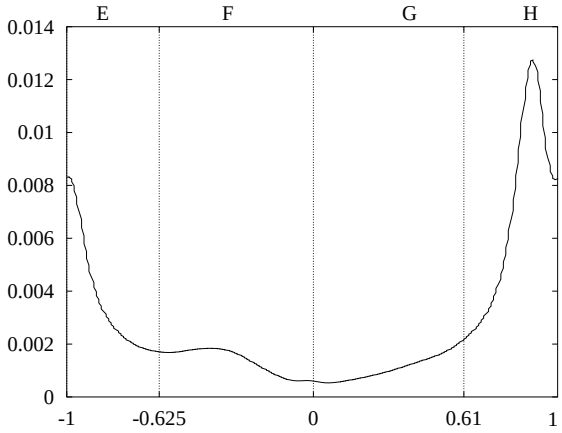


Figure 8: Smoothed histogram of the distribution of TCO cosine values in the `fssp-x` set of protein chains (all x-ray representative structures in FSSP). The 4-letter TCO alphabet was created using the k-means algorithm.

ALPHA. We created an eleven-letter alphabet by observing a smoothed histogram (shown in Figure 7) of the dihedral angle ALPHA values (see Figure 4), for all residue types, in 448 x-ray structures with resolution ≤ 2.0 Å and $\leq 50\%$ sequence identity. We manually assigned breakpoints between ALPHA classes, according to location of peaks in this histogram, and refined the breakpoints according to manual inspection of ALPHA histograms for each of the 20 amino acids.

TCO. The TCO alphabet was designed by manually dividing the distribution of TCO cosine values (see Figure 5) in the `fssp-x` dataset (all x-ray structures that were representatives in FSSP [35]) into four classes (Figure 8). The best centroids for the four classes were selected with the *k-means* algorithm.

2.3 Datasets

Two datasets of PDB chains were used in this work: all x-ray structures that are representatives in FSSP [35] (`fssp-x`), and `dunbrack-in-scop`, a high-quality set of 1298 non-homologous chains containing SCOP (version 1.55) domains. This is a modified version of the Dunbrack culled PDB set, with sequence identity cut-off of 20%, resolution cut-off of 3.0 Å and R-factor cut-off of 1.0 [36], with fragments shorter than 20 residues removed. We excluded SCOP classes e (multi-domain), i (low resolution), j (peptide), and k (designed proteins), SCOP folds a.137 (non-globular all-alpha subunits of globular proteins), d.184 (non-globular alpha-beta subunits of globular proteins), and SCOP superfamily f.2.1 (membrane all-alpha).

For each chain, we built an amino acid multiple sequence alignment with the SAM-T2K algorithm [30] and thinned the alignments to 90% sequence identity.

All of the datasets are available at http://www.soe.ucsc.edu/projects/compbio/2thmm

2.4 Estimating ALPHA and TCO

In the course of this work, we tried and tested several approaches to designing alphabets based on ALPHA and TCO, including von Mises mixture models [37, 38, 39], k-means clusters, and manual assignments. We observed very little difference between the distributions in different data sets. The ALPHA and TCO alphabets described above were the most conserved and predictable of those tested.

2.5 Alphabet Evaluation Protocol

For each alphabet, we built a structure string, representing each protein chain in our benchmark data set (`dunbrack-in-scop`), and added these to our existing library of structure strings and amino acid sequences.

2.5.1 Information Content

Alphabet compositional entropy gives an upper limit on our estimate of alphabet conservation, which is based on mutual information of letter pairs observed in equivalent positions of a structural alignment. (A simple proof of this bound is given in Durbin et al. [40].) We used the library of structure strings to estimate the compositional entropy for each alphabet A

$$-\sum_{x \in A} p(x) \log p(x) \ , \tag{2}$$

and the mutual information between all pairs of alphabets A and B (including amino acid sequence)

$$\sum_{x \in A, y \in B} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \ . \tag{3}$$

Very low mutual information between an alphabet and the amino acids indicates potential difficulty predicting the alphabet from amino acid sequences. On the other hand, very high mutual information between two alphabets means that including both in an HMM would be redundant.

2.5.2 Predictability

We estimated how well structure strings in each alphabet were predicted by feed-forward neural networks [1]. The nets were trained with standard back-propagation, using a log loss function and have a four-layer architecture. The same inputs, architecture, and training protocol were used for all neural nets. Only the number of output units in the final layer varied according to the size of the alphabet. In general, performance of the nets has been rather insensitive to the number of layers. However, after some optimization, we selected an architecture of four layers; an input window of five alignment positions; and window widths of seven, nine, and thirteen on the three hidden layers, yielding approximately 4,000 weight terms.

We did three-fold cross-validation. The data set (`dunbrack-in-scop`) was randomly divided into three partitions, and for each alphabet, we trained a net on two-thirds of the data and tested on the remaining third. The inputs to the neural nets were the SAM-T2K multiple alignments (thinned to 90% identity). The network outputs at position i were interpreted as a probability vector $\hat{P}_i$ over the alphabet being tested, and the neural nets were trained to maximize

$$\sum_i \log \hat{P}_i(l_i) \ , \tag{4}$$

where l_i is the correct local structure code for position i. The performance of the three nets was averaged and reported according to several criteria: percent of residues correctly predicted to be in one of N states (Q_N), the fractional overlap of secondary structure segments (*SOV*) in a pairwise alignment of predicted and observed structure strings [41], and the amount of information gained (bits saved) per position in the test set:

$$G = \frac{1}{n} \sum_i^n \log \frac{\hat{P}(l_i)}{P_\emptyset(l_i)} \tag{5}$$

Information gain (Equation 5) measures how much the neural net predictions $\hat{P}$ improve on a maximum likelihood prediction based on the frequencies of each letter in the data set ($P_\emptyset$), and is reported in bits per position.

2.5.3 Conservation

We evaluate alphabet conservation between structurally similar proteins by computing the average *mutual information* of letter pairs observed in equivalent positions of FSSP structural alignments. These alignments each have a "master"

[1] Our neural networks were built with an in-house software package, PREDICT-2ND, which will be released in 2003.

belonging to FSSP's representative set (none of which share greater than 25% sequence identity) and a collection of "slaves" that are structurally aligned to the master by DALI [42, 35].

For this analysis, we built structure strings for 1609 protein chains taken from the FSSP representative set and also for their slaves. Because many sets of slaves are large and redundant, and because we are interested in conservation of properties between sequences with very low amino acid similarity, we removed chains sharing greater than 20% sequence identity from these alignments. We used only those slaves that aligned to the master with DALI Z-scores (statistical significance) ≥ 7.

Finally, we extracted all regions in the FSSP alignments where DALI has aligned a slave to the master and constructed tables of aligned residue positions. For all possible pairs of letters (X, Y) in a local structure alphabet, we counted the number of times X in the master structure was paired with Y in the slave structure in an aligned position, and computed their mutual information

$$I(X;Y) = \sum_{X,Y} P(X,Y) \log \frac{P(X,Y)}{P(X)P(Y)} \ , \tag{6}$$

where probabilities were estimated by normalized counts of letters in the dataset. Since estimates of mutual information based on small samples have been shown to be artifically inflated [43, 44], we perform a *small sample correction* on $I(X;Y)$. The small sample effect is evidenced by the fact that when aligned letter pairs are scrambled, a non-zero mutual information is still measured between randomized pairs. By repeatedly scrambling the aligned letter pairs, we can compute a distribution of "random mutual information" and correct for the over-estimate by subtracting the mean of this distribution from $I(X;Y)$ [45].

2.5.4 Fold Recognition

For our fold-recognition experiments, we used the SCOP database as a standard for identifying related proteins. We created a benchmark set (`dunbrack-in-scop`) of non-homologous, whole protein chains (detailed in Section 2.3). Our study focuses on proteins that are difficult to detect by sequence similarity, so we excluded proteins sharing greater than 20% sequence identity from the benchmark set. We define correct hits as proteins which contain the same SCOP fold (major structural similarity). Other studies have used the more stringent criterion of same SCOP superfamily (probable common evolutionary origin) [46, 47, 48, 49]. We chose a looser definition, because we are interested here in detecting structurally similar proteins, regardless of whether they are remotely homologous or non-homologous. This choice raises both the number of true positives and of false negatives [2].

We built (i) one PSI-BLAST (version 2.2.1) [50, 51] profile (using four iterations on NR [52] with threshold set at

[2] We do not see much of a difference in the relative performance of the sequence-based or local-structure methods when correct hits are defined as proteins in the same SCOP superfamily, but not in the same fold.

0.0005), (ii) one amino-acid-only HMM, and (iii) nine two-track HMMs, for each chain in the benchmark set, trying each candidate local structure alphabet as a secondary track. Each benchmark profile and HMM was used to rank all chains in the set according to E-values of the PSI-BLAST or HMM scores [53]. We chose weights of 1.0 for amino acid emissions and 0.3 for local structure emissions empirically, after experimenting with other combinations such as equal track weights ($w_1 = 0.5, w_2 = 0.5$); and all the weight on the local structure track ($w_1 = 0, w_2 = 1$). We have not yet optimized this thoroughly. Results for each alphabet could probably be improved by doing fold recognition and alignment testing with a larger number of alternate weighting schemes, and seeing which ones perform best.

In keeping with the three-fold cross-validation protocol, described in Section 2.5.2, when we built two-track HMMs for each chain in the benchmark set, we used local-structure emission probabilities predicted by a neural net trained on the two-thirds of the benchmark set which did not contain the target chain.

In this setting, we define results of a *query* as the list of E-values received by all chains in the benchmark set, with respect to a single profile or HMM. Performance of a candidate alphabet was evaluated with respect to all benchmark HMMs built with the alphabet as a secondary track (1298 queries). We report fold-recognition performance of each method in terms of ROC numbers at selected thresholds, corresponding to 0.05, 0.1, 0.5, and 1.0 false positives per query.

2.5.5 ROC numbers

The ROC curve is a plot of true positive fraction vs. true negative fraction using a sliding threshold and provides an accurate, quantitative measure of both the sensitivity and specificity of a database search.

The total area under the ROC curve gives the probability of a correct classification [54]. Because of the very large number of true negatives in a typical database search, the area is usually calculated under a truncated ROC curve, with a fixed *ROC number* (ROC_N), where N is the number of true negatives used in the calculation.

2.5.6 Alignment

We used two sets of protein pairs to evaluate alignment quality: a difficult set of 200 pairs, with high structural similarity but low sequence identity (3–24%) and a moderately difficult set of 340 pairs (homology detectable by SAM-T2K HMM or PSI-BLAST but not by BLAST).

In both test sets, two local alignments were produced for each pair, by building a SAM-T2K HMM for one pair member and aligning the other to the model using the Viterbi algorithm [55]. We tested SAM-T2K amino-acid-only HMMs; HMMs built with an FSSP structural alignment as the seed; and nine types of SAM-T2K two-track HMMs, in which one of our nine candidate local structure alphabets was used for the secondary track. For all two-track HMMs, track weights were set at 1.0 for the amino acid and 0.3 for the secondary track, as in the fold-recognition tests.

The resulting alignments were compared to structural alignments of the same pairs. To avoid possible bias, two structural alignment methods were used in the analysis: DALI [35] and CE [56]. The mean *shift score* [57, 58] between alignments produced by the HMMs and by the structural aligners was computed. Shift-score measures the disagreement of two alignments, typically a predicted candidate alignment and a reference structural alignment of the same sequence pair. It is 97% correlated with *percentage of residues aligned correctly*, but also incorporates information about alignment length, shift error, and coverage [57, 58]. The range of shift-score is from -0.2 (worst) to 1.0 (best), achieved when two alignments are identical. A detailed definition of shift-score can be found in Appendix A.

3 Results

Table 2 presents our analysis of the compositional entropy, mutual information with amino acid, conservation, and predictability for the nine alphabets studied. The alphabets are ranked according to their conservation in alignments of non-homologous FSSP structural neighbors. According to this data, the STR alphabet best encodes conserved properties of local backbone structure, followed by PB, STRIDE, and DSSP. Because STR is an expansion of the DSSP alphabet, its higher conservation shows us that the properties of strand pairing described in STR have been preserved in remote homologs.

An alphabet's utility is limited by its predictability from amino-acid sequence, and the alphabets evaluated here are all reasonably predictable. The bits saved (information gain) measure of predictability does not depend on alphabet size and is strongly correlated with alphabet conservation (r=0.79). We have found this to be the most useful measure of predictability when comparing different alphabets. Table 3 shows the mutual information (in bits) between all pairs of the local structure alphabets. The mutual information between all of the alphabets is high, showing that there is considerable similarity in these encodings of backbone information.

Our fold recognition results are shown in Figure 9. As indicated by Figure 9(a), there is no clear separation between multi-track and single-track HMMs in the very low false positive range (0.05 false positives per query). However, if we are willing to accept between 0.1 to 1.0 false positive per query, a reasonable threshold given the low sequence similarity of the proteins in the test set, the two-track models increasingly recognize more correct folds with fewer false positives than either PSI-BLAST or SAM-T2K amino-acid HMMs (Figure 9(b)). The exception is PB, which in spite of high compositional entropy and predictability, did very poorly at fold recognition. This anomaly is an artifact of the reverse-sequence null model SAM uses to compute HMM scores. Sequence scores in SAM are normalized by taking the difference between the log-probability of a sequence and the log-probability of the sequence with a reversed HMM (equivalently, the log-probability of the reversed sequence with the HMM). The scoring system is based on the assumption that sequences and reversed sequences come from the same underlying distribution—that they are equally

Name	alphabet size	entropy	MI w/aa	conservation fssp-x mutual info	predictability bits saved per position	QN	SOV
STR	13	2.842	0.103	1.107	1.009	0.561	0.527
PB	16	3.233	0.162	0.980	1.259	0.579	0.542
STRIDE	6	2.182	0.088	0.904	0.863	0.663	0.659
DSSP	7	2.397	0.092	0.893	0.913	0.633	0.610
STRIDE-EHL	3	1.546	0.075	0.861	0.736	0.769	0.733
DSSP-EHL	3	1.545	0.079	0.831	0.717	0.763	0.732
ALPHA	11	2.965	0.087	0.688	0.711	0.469	0.375
ANG	10	2.443	0.228	0.678	0.736	0.588	0.501
TCO	4	1.810	0.095	0.623	0.577	0.649	0.547

Table 2: Summary of the information content, mutual information with amino acid, conservation and predictability of each tested alphabet. Conservation (see Section 2.5.3) was estimated by calculating the mutual information of letter pairs observed in equivalent positions of FSSP structural alignments. Prediction statistics were computed by three-fold cross-validated testing with four-layer feed-forward neural networks.

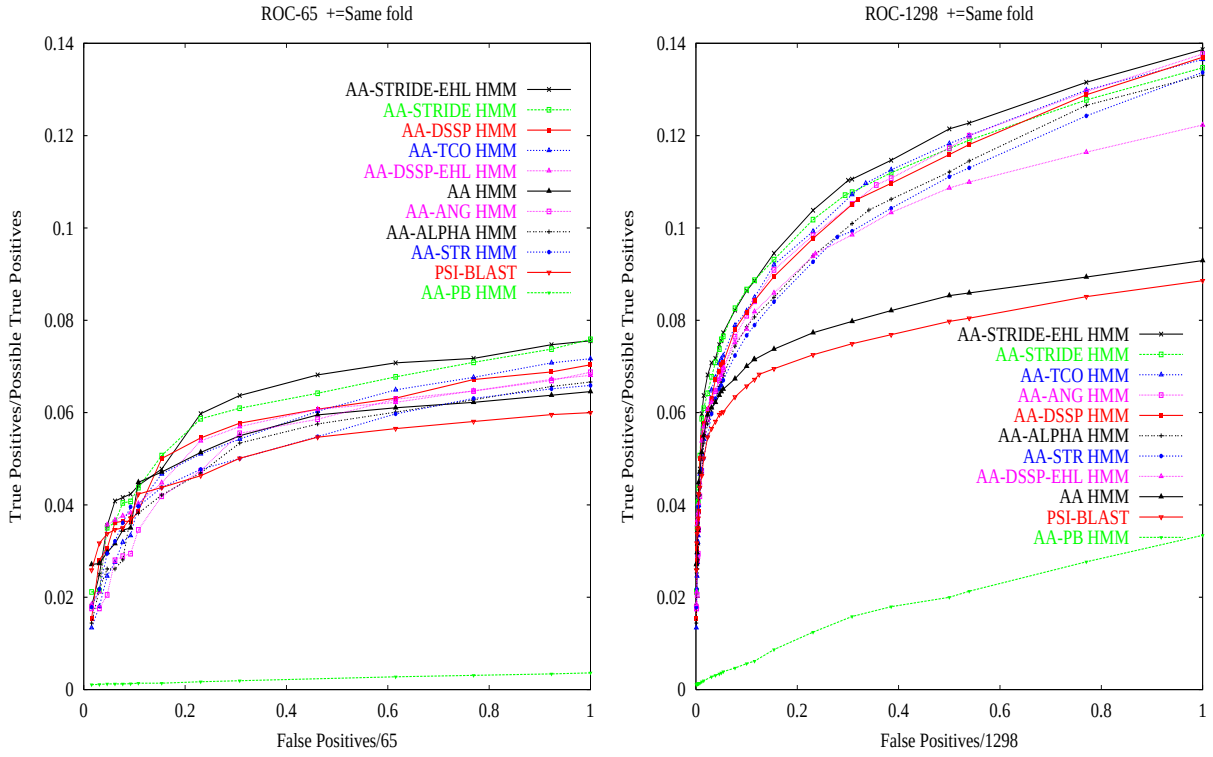
	DSSP	DSSP-EHL	STRIDE	STRIDE-EHL	PB	STR	ANG	TCO	ALPHA
DSSP	2.397								
DSSP-EHL	1.545	1.545							
STRIDE	1.557	1.271	2.182						
STRIDE-EHL	1.282	1.265	1.546	1.546					
PB	1.001	0.792	0.918	0.810	3.233				
STR	2.395	1.545	1.559	1.312	1.056	2.842			
ANG	0.889	0.743	0.790	0.749	1.124	0.962	2.443		
TCO	0.885	0.700	0.734	0.693	0.883	0.904	1.338	1.810	
ALPHA	0.849	0.673	0.727	0.653	0.944	0.878	0.955	0.761	2.965

Table 3: Matrix of mutual information (in bits) between the nine local structure alphabets. Values along the diagonal show each alphabet's compositional entropy.

protein-like. In the case of the PB alphabet, this assumption is wrong, because a distribution of PB sequences contains numerous *directed* motifs, frequently occurring combinations of letters that rarely appear in reversed form. In this situation, the reverse sequence is highly uncharacteristic and tends to have low probability. The normalization with the reversed sequence artificially inflates the scores, increasing the number of false positives. A more detailed analysis of this problem appears in our recent paper [53].

AA-STRIDE-EHL HMMs have the best accuracy by a small margin in both Figure 9(a) and Figure 9(b). AA-DSSP-EHL HMMs perform comparably to the other good two-track HMMs in the range below 0.5 false positives per query, but are less accurate in the range above 0.5 false positives per query, where protein pairs from the same SCOP class frequently score as well or better than pairs from the same fold.

Table 4 shows the fold recognition ROC_{65}, ROC_{130}, ROC_{649} and ROC_{1298} numbers (see Section 2.5.5) for two-track HMMs, single-track HMMs and PSI-BLAST. These were computed by estimating the area under the corresponding ROC curves with a trapezoidal method. The thresholds correspond to 0.05 FP/Q (false positives per query) at ROC_{65},



(a) ROC_{65}. Number of negatives recognized fixed at 65 (0.05 false positives per query).

(b) ROC_{1298}. Number of negatives recognized fixed at 1298 (1.0 false positives per query).

Figure 9: Results of three-fold cross-validated fold recognition tests on our benchmark dataset (`dunbrack-in-scop`) shown for two-track HMMs with an amino acid primary track and STRIDE-EHL, STRIDE, DSSP, TCO, DSSP-EHL, ANG, ALPHA, STR, and PB (Protein Blocks) on the secondary track, amino-acid only HMMs, and PSI-BLAST run with four iterations. The methods are ranked in the legends according to ROC_N score. As shown in Figure 9(a), with a strict threshold of 0.05 false positives per query, AA-STRIDE-EHL HMMs have the best accuracy by a small margin. With looser thresholds of 0.1 to 1.0 false positives per query, shown in Figure 9(b), the accuracy of all the local structure HMMs is significantly better than that of the AA-only methods. AA-DSSP-EHL HMMs perform comparably to the other good two-track HMMs in the range below 0.5 false positives per query, but are less accurate in the range above 0.5 false positives per query. The poor performance of AA-PB HMMs is an artifact of the reverse-sequence null model SAM uses to compute HMM scores [53].

0.1 FP/Q at ROC_{130}, 0.5 FP/Q at ROC_{649}, and 1.0 FP/Q at ROC_{1298}. The addition of backbone geometry information in the SAM-T2K HMMs clearly improves fold recognition performance, but choice of alphabet for the secondary track does not make much of a difference. Although the two-track STRIDE-EHL HMM has the best ROC numbers, the advantage is very small. This is consistent with the similarity of backbone geometry alphabets. However, these alphabets are not equally useful for alignment quality. The alphabets associated with the best alignments have the highest compositional entropy and conservation (see Table 2 and Table 5).

Table 5 shows results of our alignment quality tests on a difficult set of 200 protein pairs, with high structural similarity but low sequence identity (3–24%) and a moderately difficult set of 340 protein pairs (homology detectable

Method	ROC_{65}	ROC_{130}	ROC_{649}	ROC_{1298}
AA-STRIDE-EHL HMM	**0.0632**	**0.0724**	**0.1010**	**0.1155**
AA-STRIDE HMM	0.0615	0.0716	0.0985	0.1125
AA-DSSP HMM	0.0579	0.0673	0.0952	0.1113
AA-TCO HMM	0.0571	0.0673	0.0970	0.1126
AA-DSSP-EHL HMM	0.0567	0.0654	0.0906	0.1032
AA-ANG HMM	0.0547	0.0650	0.0955	0.1119
AA-ALPHA HMM	0.0537	0.0635	0.0914	0.1078
AA-STR HMM	0.0533	0.0625	0.0901	0.1065
AA HMM	0.0552	0.0612	0.0755	0.0823
PSI-BLAST	0.0514	0.0572	0.0708	0.0776
AA-PB HMM	0.0024	0.0035	0.0122	0.0196

Table 4: ROC numbers (see Section 2.5.5) from three-fold cross-validated fold recognition tests on a difficult set of 1298 whole chains. The numbers are an approximation to area under the ROC curve for thresholds of 65, 130, 649 and 1298 negatives recognized. The thresholds correspond to 0.05 FP/Q (false positives per query) at ROC_{65}, 0.1 FP/Q at ROC_{130}, 0.5 FP/Q at ROC_{649}, 1.0 FP/Q at ROC_{1298}.

	Difficult set mean shift-score		Moderate set mean shift-score	
Reference alignment	DALI	CE	DALI	CE
DALI	1.000	0.607	1.000	0.616
CE	0.607	1.000	0.616	1.000
two-track t2k STR	0.320	**0.307**	**0.466**	**0.418**
two-track t2k PB	0.309	0.303	0.435	0.395
two-track t2k DSSP	0.306	0.295	0.454	0.402
two-track t2k STRIDE	**0.357**	0.292	0.452	0.400
two-track t2k STRIDE-EHL	0.298	0.290	0.438	0.396
two-track t2k DSSP-EHL	0.297	0.287	0.435	0.391
two-track ALPHA	0.288	0.279	0.429	0.387
two-track ANG	0.286	0.276	0.422	0.407
two-track TCO	0.284	0.276	0.421	0.374
one-track AA	0.220	0.219	0.365	0.325
one-track AA FSSP seed	0.219	0.192	0.415	0.330

Table 5: Evaluation of alignment quality for a difficult set of 200 protein pairs with high structural similarity but low sequence identity (3-24%) and a moderately difficult set of 340 protein pairs. See Appendix A for a definition of shift-score. To give an idea of the magnitude of the differences, the two structural aligners are compared using mean shift score. Mean shift-scores [58, 57] are shown for alignments done with single-track SAM-T2K amino acid HMMs, single-track amino acid HMMs trained on FSSP structural alignments, and two-track SAM-T2K HMMs with the following track combinations: amino acid-ANG, amino acid-TCO, amino acid-ALPHA, amino acid-PB, amino acid-STRIDE, amino acid-DSSP, amino acid-STRIDE-EHL, amino acid-DSSP-EHL, and amino acid-STR.

by SAM-T2K HMM or PSI-BLAST but not by BLAST). To avoid possible bias due to arbitrary choices made by a structural aligner, we tested the two-track HMM alignments with respect to structural alignments produced by both DALI and CE.

Previous to this work, our best quality alignments were produced by amino-acid HMMs trained on an initial FSSP

structural alignment[3]. As shown by the reported mean shift-scores, the two-track SAM-T2K HMMs produce better quality alignments than SAM-T2K amino-acid-only HMMs and the FSSP-seeded amino-acid-only HMMs. These results are reasonably consistent, regardless of which structural aligner is used as reference.

On the moderately difficult set, the two-track STR-HMMs produce the best quality alignments, as measured by mean shift-score, with respect to both DALI and CE structural alignments of the same pairs. On the difficult set, the STR HMMs produce the best alignments when compared to CE, but STRIDE-HMMs do better when compared to DALI. Overall, the highly informative, 13-letter STR alphabet, which encodes detailed predictions about beta-sheet structure, is the best choice for HMM alignment. However, when there is very low sequence identity between a pair of proteins, the 6-letter STRIDE alphabet may be a better choice.

Although the test alignments were produced with a local alignment method, they often contained more aligned residues than the reference structural alignments, indicating that their coverage of the structural alignments is not limited to highly conserved positions. As expected, the mean shift-scores are higher for the moderately difficult set than for the difficult set.

4 Conclusion

We have shown that adding predicted local structure information to profile HMMs can improve detection and alignment of structurally similar proteins, even when there is very little sequence relationship. A simplified helix-strand-coil representation of secondary structure works well for fold recognition, but not so well for alignments, where greater benefit is found by using a more detailed alphabet of local protein structure. The high degree of conservation and predictability in the STR alphabet suggests that there is useful information in the patterns of strand orientation found in beta sheets [60]. The success of this alphabet at aligning structurally similar proteins with low sequence identity shows that these patterns are both predictable from amino acid sequence and conserved in remote homologs.

High mutual information with amino acid sequence does not guarantee that a local structure alphabet is highly predictable. Of the alphabets tested, ANG had the highest mutual information with the amino acids, but was less predictable (in bits saved per position) than several other alphabets (see Table 2). Most of the mutual information between ANG and amino acid comes from two amino acids: proline and glycine. Because the proline side-chain is bonded to a backbone nitrogen atom, it does not have the flexibility to assume many of the ANG states, and appears predominantly in P and H (see Table 1). At the other extreme, glycine is so flexible that it can assume the ANG states S, T, and L that are rare or impossible for the other amino acids. Our neural nets make their predictions based on alignment columns rather than single residues and can miss the special character of the proline and glycine positions,

[3] Although these HMMs yield strong alignments, prior work by our group and others [59] indicates that these HMMs do not perform well on fold recognition. This could be due to the small size and high sequence diversity in structural alignments. Yet while the alignments are more varied, they emphasize the key structurally-conserved positions, which in turn reduces the chance of misalignment.

if these residues are not conserved.

We suspect that highly informative, detailed alphabets like STR and PB have greater potential for fold recognition than simple alphabets like STRIDE-EHL and TCO, and that best results will be achieved by combining several alphabets that encode different kinds of information about local protein structure. Our inability to achieve these results, in the present work, is due to the limitations of the null models we currently use to compute HMM scores and our methods of estimating the statistical significance of these scores, which work best for amino acid distributions and simple three-state secondary structure distributions.

In the future, we will focus on discovering the most predictable and conserved alphabets of side chain properties, on developing HMM scoring methods that work with detailed local structure alphabets, on combining backbone and sidechain properties into multi-track HMMs, and on finding better ways to calibrate our HMMs, to use local structure most effectively.

Acknowledgements

This work was supported in part by DOE grant DE-FG0395-99ER62849 and a National Physical Sciences Consortium graduate fellowship. We are grateful to David Haussler and Anders Krogh for starting the hidden Markov model work at UCSC, to Lydia Gregoret for her work on beta-sheet patterns, and to Richard Hughey and Mark Diekhans for contributing to the software used in our experiments.

References

[1] B. Rost, R. Schneider, and C. Sander. Protein fold recognition by prediction-based threading. *Journal of Molecular Biology*, 270:471–480, 1997.

[2] V. Di Francesco, V. Geetha, J. Garnier, and P.J. Munson. Fold recognition using predicted secondary structure sequences and hidden Markov models of protein folds. *Proteins: Structure, Function and Genetics, Suppl.*, 1:123–128, 1997.

[3] D.W. Rice and D. Eisenberg. A 3d-1d substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *Journal of Molecular Biology*, 267(4):1026–1038, 1997.

[4] X. De La Cruz and J.M. Thornton. Factors limiting the performance of prediction-based fold recognition methods. *Protein Science*, 8:750–759, 1999.

[5] L.A. Kelley, R.M. MacCallum, and M.J.E. Sternberg. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *Journal of Molecular Biology*, 299:501–522, 2000.

[6] R. Hughey and A. Krogh. SAM: Sequence alignment and modeling software system. Technical Report UCSC-CRL-95-7, University of California, Santa Cruz, Computer Engineering, UC Santa Cruz, CA 95064, 1995.

[7] Richard Hughey, Kevin Karplus, and Anders Krogh. SAM: Sequence alignment and modeling software system, version 3. Technical Report UCSC-CRL-99-11, University of California, Santa Cruz, Computer Engineering, UC Santa Cruz, CA 95064, October 1999. Available from `http://www.soe.ucsc.edu/research/compbio/sam.html`.

[8] J.S. Fetrow, M.J. Palumbo, and G. Berg. Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme. *Proteins: Structure, Function, and Genetics*, 27:249–271, 1997.

[9] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, Dec 1983.

[10] R. Unger, D. Harel, Wherland S., and J. Sussman. A 3d building blocks approach to analyzing and predicting structure of proteins. *Proteins*, 5:355–373, 1989.

[11] R. Unger and J.L Sussman. The importance of short structural motifs in protein structure analysis. *J Comput Aided Mol Des*, 7(4):457–472, 1993.

[12] M.J. Rooman, J. Rodriguez, and S.J. Wodak. Automatic definition of recurrent local structure motifs in proteins. *Journal of Molecular Biology*, 213:327–336, 1990.

[13] M.J. Rooman, J. Rodriguez, and S.J. Wodak. Relations between protein sequence and structure and their significance. *Journal of Molecular Biology*, 213:337–350, 1990.

[14] X. Zhang, J.S. Fetrow, W.A. Rennie, D.L. Waltz, and G. Berg. Automatic derivation of substructures yields novel structural building blocks in globular proteins. *Proceedings, 1st International Conference on Intelligent Systems for Molecular Biology*, pages 438–446, 1993.

[15] H.S. Kang, N.A. Kurochkina, and B. Lee. Estimation and use of protein backbone angle probabilities. *Journal of Molecular Biology*, 229:448–460, 1993.

[16] Dimitrij Frishman and Patrick Argos. Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Genetics*, 23:566–579, 1995.

[17] M.B. Swindells, M.W. MacArthur, and J.M. Thornton. Intrinsic phi,psi propensities of amino acids, derived from the coil regions of known structures. *Nat. Struct. Biol.*, 2(7):596–603, Jul 1995.

[18] M.J. Thompson and R.A. Goldstein. Predicting protein secondary structure with probabilistic schemata of evolutionarily derived information. *Protein Science*, 6:1963–1975, 1997.

[19] C. Bystroff and D. Baker. Prediction of local structure in proteins using a library of sequence-structure motifs. *Journal of Molecular Biology*, 281:565–577, 1998.

[20] A.C. Camproux, P. Tuffery, J.P. Chevrolat, J.F. Boisvieux, and S. Hazout. Hidden Markov model approach for identifying the modular framework of the protein backbone. *Protein Eng.*, 12(12):1063–1073, Dec 1999.

[21] S.M. King and W.C. Johnson. Assigning secondary structure from protein coordinate data. *Proteins: Structure, Function, and Genetics*, 35:313–320, 1999.

[22] C. Bystroff, V. Thorsson, and D. Baker. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *Journal of Molecular Biology*, 301(1):173–190, Aug 2000.

[23] A.G. de Brevern, C. Etchebest, and S. Hazout. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins: Structure, Function and Genetics*, 41:271–287, 2000.

[24] V. Di Franceso, P.J. Munson, and J. Garnier. FORESST: fold recognition from secondary structure predictions of proteins. *Bioinformatics*, 15(2):131–140, 1999.

[25] J. Hargbo and A. Elofsson. Hidden Markov models that use predicted secondary structures for fold recognition. *Proteins: Structure, Function, and Genetics*, 36:68–76, 1999.

[26] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. CATH— a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–108, August 1997.

[27] C. Sander and R. Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Structure, Function, and Genetics*, 9(1):56–68, 1991.

[28] Kevin Karplus, Kimmen Sjölander, Christian Barrett, Melissa Cline, David Haussler, Richard Hughey, Liisa Holm, and Chris Sander. Predicting protein structure using hidden Markov models. *Proteins: Structure, Function, and Genetics*, Suppl. 1:134–139, 1997.

[29] Kevin Karplus, Christian Barrett, Melissa Cline, Mark Diekhans, Leslie Grate, and Richard Hughey. Predicting protein structure using only sequence information. *Proteins: Structure, Function, and Genetics*, Supplement 3(1):121–125, 1999.

[30] Kevin Karplus, Rachel Karchin, Christian Barrett, Spencer Tu, Melissa Cline, Mark Diekhans, Leslie Grate, Jonathan Casper, and Richard Hughey. What is the value added by human intervention in protein structure prediction? *Proteins: Structure, Function, and Genetics*, 2001. accepted.

[31] G.N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7:95–99, 1963.

[32] T. Kohonen. Self-organizing formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982.

[33] T. Kohonen. *Self-organization and associative memory*. Springer-Verlag, third edition, 1989.

[34] J. MacQueen. Some methods for classification and analysis of multivariate observations. In L.M. Le Cam and J. Neyman, editors, *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. University of California Press, Berkeley, 1967.

[35] Liisa Holm and Chris Sander. Mapping the protein universe. *Science*, 273(5275):595–603, Aug 2 1996.

[36] R. Dunbrack. Culling the PDB by resolution and sequence identity, 2001. `http://www.fccc.edu/research/labs/dunbrack/culledpdb.html`.

[37] D.L. Dowe, L. Allison, T.I. Dix, L. Hunter, C.S. Wallace, and T. Edgoose. Circular clustering of protein dihedral angles by minimum message length. In *Pac. Symp. Biocomput.*, pages 242–255, 1996.

[38] C. Wallace and D. Dowe. Intrinsic classification by MML - the snob program. In *Proc. 7th Australian Joint Conference on Artificial Intelligence*, pages 37–44, Armidale NSW, Australia, November 1994. LINE, World Scientific.

[39] J.S. Fetrow and G. Berg. Using information theory to discover side chain rotamer classes: analysis of the effects of local backbone structure. In *Pac. Symp. Biocomput.*, pages 278–289, jan 1999.

[40] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[41] B. Rost, C. Sander, and R. Schneider. Redefining the goals of protein secondary structure prediction. *Journal of Molecular Biology*, 235:13–26, 1994.

[42] Liisa Holm and Chris Sander. Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology*, 233(1):123–138, 5 Sept 1993.

[43] D. H. Wolpert and D. R. Wolf. Estimating functions of probability distributions from a finite set of samples. *Physical Review E*, 52(6):6841–54, December 1995.

[44] D. H. Wolpert and D. R. Wolf. Estimating functions of probability distributions from a finite set of samples; part 2: Bayes estimators for mutual information, chi-squared, covariance, and other statistics. Technical Report LA-UR-93-833,TR-93-07-047, Los Alamos National Lab, Santa Fe Institute, 1993.

[45] Melissa S. Cline, Kevin Karplus, Richard H. Lathrop, Temple F. Smith, Robert G. Rogers Jr., and David Haussler. Information-theoretic dissection of pairwise contact potentials. *Proteins: Structure, Function, and Genetics*, 49(1):7–14, 1 October 2002.

[46] J. Park, K. Karplus, C. Barrett, R. Hughey, D. Haussler, T. Hubbard, and C. Chothia. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *Journal of Molecular Biology*, 284(4):1201–1210, 1998.

[47] T. Jaakkola, M. Diekhans, and D. Haussler. A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 7(1/2):95–114, February-April 2000. Available from `http://www.soe.ucsc.edu/research/compbio/research.html`.

[48] J. Gough, K. Karplus, R. Hughey, and C. Chothia. Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. *Journal of Molecular Biology*, 2001. in press.

[49] A. Raval, Z. Ghahramani, and D.L. Wild. A Bayesian network model for protein fold and remote homologue recognition. *Bioinformatics*, 18:788–801, 2002.

[50] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25:3899–3402, 1997.

[51] Alejandro A. Schäffer, L. Aravind, Thomas L. Madden, Sergei Shavirin, John L. Spouge, Yuri I. Wolf, Eugene Koonin, and Stephen F. Altschul. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research*, 29(14):2994–3005, 2001.

[52] NRP (Non-Redundant Protein) Database. Distributed on the Internet via anonymous FTP from ftp.ncifcrf.gov, under the auspices of the National Cancer Institute's Frederick Biomedical Supercomputing Center., 1998.

[53] K. Karplus, R. Karchin, and R. Hughey. Calibrating E-values for hidden Markov models with reverse-sequence null models, 2002.

[54] M. Gribskov and N.L. Robinson. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Computers Chem.*, 20(1):25–33, 1996.

[55] A. J. Viterbi. Error bounds for convolution codes and an asymptotically optimal decoding algorithm. *IEEE Trans. Informat. Theory*, 13:260–269, 1967.

[56] I. N. Shindyalov and P. E. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*, 11(9):739–47, 1998.

[57] Melissa Cline and Kevin Karplus. On alignment shift and its measures. Technical Report UCSC-CRL-97-27, University of California, Santa Cruz, Jack Baskin School of Engineering, UC Santa Cruz, CA 95064, February 1998.

[58] Melissa Cline, Richard Hughey, and Kevin Karplus. Predicting reliable regions in protein sequence alignments. *Bioinformatics*, 18(2):306–314, 2002.

[59] S. Griffiths-Jones and A. Bateman. The use of structure information to increase alignment accuracy does not aid homlogue detection with profile HMMs. *Bioinformatics*, 18(9):1243–1249, 2002.

[60] Y. Mandel-Gutfreund and L. Gregoret. On the significance of alternating patterns of polar and nonpolar residues in beta strands. *Journal of Molecular Biology*, 323(3):453–461, 2002.

[61] Melissa Cline. *Protein sequence alignment reliablity: prediction and measurement*. PhD thesis, University of California, Computer Science, UC Santa Cruz, CA 95064, 2000.

Appendix A Shift-Score

The shift-score of two alignments X and Y is computed as follows [61]:

$$\text{shift\_score} = \frac{\sum_{i=1}^{|X|} cs(X_i)}{|X| + |Y|} \tag{7}$$

where

$$\begin{aligned} \epsilon &= \text{small-valued algorithmic parameter,} \\ &\quad \text{typically set to } 0.2 \\ |X| &= \text{Number of aligned residue pairs in} \\ &\quad \text{alignment } X \end{aligned}$$

Basic depiction of alignment shift

Reference	
template	ABCD--EFG
target	L-MNOPQR-

Candidate	
template	-AB-CDEFG
target	LMNOP--QR

Target Residue	Template residue aligned to in Reference alignment	Template residue aligned to in Candidate alignment	Shift
M	C	A	-2
N	D	B	-2
Q	E	F	+1
R	F	G	+1

Figure 10: Illustration of the shift of a single residue. Shift is measured for target residues aligned in both alignments, and refers to the number of template residues between its position in the two alignments [61].

$$
\begin{aligned}
X_i &= \text{Aligned residue pair } i \text{ in alignment } X \\
s(r_i) &= \text{Subscore for residue } r_i \\
&= \left\{ \begin{array}{ll} \frac{1+\epsilon}{1+|\text{shift}(r_i)|} - \epsilon & \text{if shift}(r_i) \text{ is defined} \\ 0 & \text{otherwise} \end{array} \right\} \\
X_i(A) &= \text{Sequence } A \text{ residue aligned in column } X_i \\
cs(X_i) &= \text{Column score for column } i \text{ in alignment } X \\
&= \left\{ \begin{array}{l} s(X_i(A)) + s(X_i(B)) \\ \text{if column } X_i \text{ aligns } X_i(A) \text{ and } X_i(B) \\ 0 \text{ otherwise} \end{array} \right\}
\end{aligned}
$$

Consider Figure 10 and the pair of residues *C* and *M* aligned in the reference alignment. In the candidate alignment, target residue *M* is aligned to template residue *A* rather than *C*. The *shift* of *M* is defined as -2, the number of positions between *A* and *C* in the template sequence. Note that the shift can be positive or negative, depending on the direction of the shift. A positive shift moves residue *M* closer to the C-terminus of the sequences, to the right in Figure 10. If a residue is not aligned in either alignment, its shift is undefined. Hence, in Figure 10, no shift is listed for target residues *L, O,* or *P* [61].