PROTEINS: Structure, Function, and Bioinformatics 55:508–518 (2004)

Evaluation of Local Structure Alphabets Based on Residue Burial

Rachel Karchin,*[1] Melissa Cline,[2] and Kevin Karplus[3]
[1]*Department of Biopharmaceutical Sciences, University of California, San Francisco, California*
[2]*Affymetrix, Inc., Emeryville, California*
[3]*Center for Biomolecular Science and Engineering, Baskin School of Engineering, University of California, Santa Cruz, California*

ABSTRACT **Residue burial, which describes a protein residue's exposure to solvent and neighboring atoms, is key to protein structure prediction, modeling, and analysis. We assessed 21 alphabets representing residue burial, according to their predictability from amino acid sequence, conservation in structural alignments, and utility in one fold-recognition scenario. This follows upon our previous work in assessing nine representations of backbone geometry.[1] The alphabet found to be most effective overall has seven states and is based on a count of C_β atoms within a 14 Å-radius sphere centered at the C_β of a residue of interest. When incorporated into a hidden Markov model (HMM), this alphabet gave us a 38% performance boost in fold recognition and 23% in alignment quality. Proteins 2004;55:508–518.** © 2004 Wiley-Liss, Inc.

Key words: protein structure prediction; fold recognition; local structure alphabet; solvent accessibility; neighborhood counts; residue burial; hidden Markov model; multitrack HMM

INTRODUCTION

Residue burial has been associated with rates of evolutionary change in a residue position,[2,3] with mutations linked to changes in protein function and disease states,[3,4,5] and with enthalpy changes during protein folding.[6,7] Comparative modeling, threading, and de novo structure prediction all rely on potential functions to distinguish the native structure from other configurations. Most potential functions rely on burial either directly or indirectly. Many successful methods, such as Rosetta, use burial as an important component of their potential functions.[8] Other methods use amino acid-based potentials, which derive much of their signal from burial and hydrophobicity.[9,10] Developing a more accurate potential function, especially as related to burial, is considered a vital step toward further improvement in such methods.[11] To this aim, we have investigated many representations of residue burial to assess their value in protein structure prediction and analysis.

Burial is typically measured either in terms of decreasing *solvent-accessible surface area* or as an increasing number of intra-protein atomic contacts. In monomeric globular proteins, a surface residue has high solvent exposure and few contacts; a residue in the protein core has low (or zero) solvent exposure and many contacts.

Algorithms for computing solvent-accessible surface area have been developed by many groups.[12–18] *Contact counts* can be determined by computing actual van der Waals contacts for each residue,[19,20] but most researchers work with an approximation based on *neighborhood counts* of protein atoms within a spherical cutoff of the residue of interest (often called the contact or coordination number).[21,22,23] This is both computationally cheaper than determining van der Waals contacts and more robust to noise in protein structure data.

Our work builds on previous studies by Burkhard Rost and co-workers, which concluded that three-state relative solvent accessibility (buried, intermediate, exposed) is much less conserved than secondary structure in proteins with significant sequence identity (25–90%), that ten-state relative solvent accessibility is more predictable than three-state, and that both secondary structure and solvent accessibility are conserved in distant homologs (less than 20% sequence identity). They suggested that more conserved descriptions of a residue's relative position within a protein structure might exist but did not attempt to characterize them.[24]

We have done a detailed study to identify predictable alphabets of residue burial that are most conserved in distant homologs and to explore whether these alphabets are useful in recognizing and aligning proteins with similar folds. Our study is not focused on developing improved prediction methods for residue burial alphabets but on learning what is the most interesting description of residue burial to predict.

The residue burial alphabets are compared with the backbone-geometry alphabets explored in our earlier paper.[1] In accordance with Rost and Sander,[24] residue-burial alphabets are found to be consistently less conserved and predictable than backbone-geometry alphabets.

Grant sponsor: Department of Energy; Grant number: DE-FG0395-99ER6249.

*Correspondence to: Rachel Karchin, University of California, San Francisco, Mission Bay Genentech Hall, 600 16th Street, Suite N474-U, San Francisco, CA 94143-2240. rachelk@salilab.org

Received 1 June 2003; Accepted 7 September 2003

Published online 5 March 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20008

© 2004 WILEY-LISS, INC.

These alphabets are also found to represent substantially different information about local protein structure than backbone geometry alphabets. We have confirmed that the two-track HMM methods for protein fold-recognition and alignment described in our previous paper can be further improved by including tracks for both backbone geometry and residue burial information in a three-track HMM.

MATERIALS AND METHODS

We assessed 21 alphabets of residue burial according to their conservation and predictability. We compared the alphabets to each other to assess the extent of their overlap.[1] Finally, we applied the alphabets in a remote-homology-recognition scenario to provide perspective on their utility for template selection and target-template alignment.

Alphabet conservation measures the extent to which residue burial, as described in the alphabet, is consistent between structurally-similar proteins. For an alphabet to be useful in homology modeling or fold recognition, its level of conservation should be high. We assessed alphabet conservation by computing the average *mutual information* of letter pairs observed in the columns of FSSP structural alignments.[26,27] If letters a and b are both observed in an alignment column, mutual information describes how much the likelihood of observing b was influenced by the observation of a. The average mutual information is computed as follows:

$$M(a, b) = \sum_{a,b} p(a, b) \log \frac{p(a, b)}{p(a)p(b)}, \quad (1)$$

where $p(a)$ and $p(b)$, the probabilities of individual letters a and b, are estimated by the frequency with which they appear in the representative set of the FSSP alignments and their joint probability $p(a,b)$ is estimated by the frequency with which a and b appear together in an alignment column. Simpler measures of alignment conservation, such as average percent identity and entropy, describe the consistency of letters in an alignment column, but fail to describe the similarity between differing letters in the column. For instance, in the amino acid alphabet, the amino acids isoleucine, valine, and methionine are similar and often substitute for each other. Thus, a biologist would consider a column of these three amino acids to be more conserved than a column of three differing amino acids, such as isoleucine, glutamine, and proline. The average mutual information in the first case would be higher than the second, yet the average percent identity and entropy would both be equivalent.

Mutual information was also used to compare the alphabets to each other. In this case, for residue r, a represents r under the first alphabet, b represents r under the second, and $M(a, b)$ quantifies the similarity between the two alphabets. If $M(a, b)$ is low, the interpretation is that each alphabet represents distinct information.

For most of our applications, residue burial cannot be measured, but must be predicted from the amino acid alphabet. Thus, an alphabet's value will be influenced by its predictability. We assessed alphabet predictability by predicting burial from the amino acid sequence using feed-forward neural networks. These neural networks were built and trained as detailed in our previous work.[1] For each residue r, represented by b in the burial alphabet, the predictability of b was assessed by the average of the log likelihood ratios between the probability with which the neural network predicted b, $P_n(b)$, and the background or prior probability of b, $P_p(b)$. This log likelihood ratio, referred to as *information gain G,* is computed as follows:

$$G = \log_2 \frac{P_n(b)}{P_p(b)}. \quad (2)$$

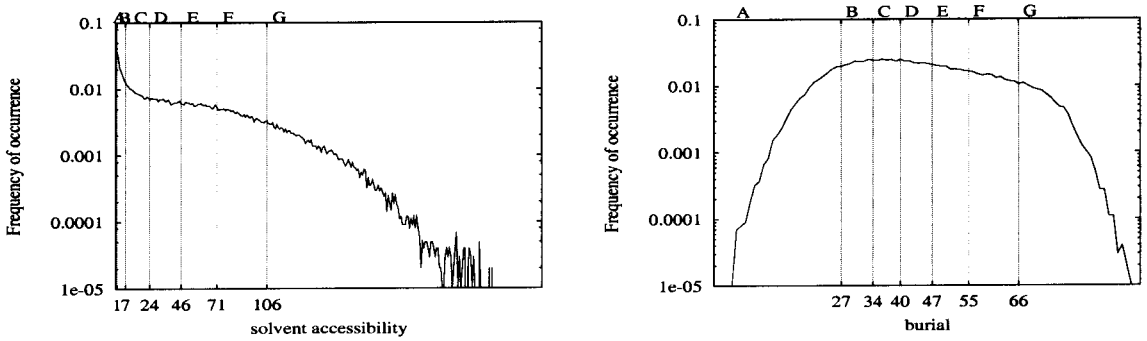
For our purposes, we have found this measure advantageous over other measures, including percent residues predicted correctly (Q_N), segment overlap measure, and correlation coefficient.[28,24,29,30,31,23] Correlation coefficient between observed and predicted alphabet strings and Q_N favor smaller alphabets, as a smaller alphabet increases the chance of making a correct prediction randomly. Thus they are not useful in comparing predictions of different-sized alphabets. The segment overlap measure,[32] developed to assess secondary structure predictions, is not applicable here because residue burial alphabets may not contain segments of repeated letters the way secondary structure strings do.

We assessed the utility of each alphabet in remote homology recognition as follows. We estimated a set of two-track HMMs for each alphabet, with the first track representing the amino acid sequence and the second track representing the burial alphabet. We compared the performance of each set of HMMs at fold recognition and alignment accuracy.[2] We assessed fold recognition performance by constructing ROC curves, a plot of true positive fraction versus true negative fraction using a sliding score threshold. The total area under the ROC curve gives the probability of a correct classification.[33] 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. For instance, ROC_{1298} describes the probability of correct classification at a threshold that permits a total of 1298 false positives in the full database query. As there are 1298 proteins in the query set, this corresponds to an average of one false positive per query.

We assessed the quality of the alignments predicted by the two-track HMMs using mean shift score (MSS). This measure is 97% correlated with *percentage of residues aligned correctly* (PRAC); unlike PRAC, it incorporates

[1]A detailed comparison of the burial alphabets' overlap with each other and with backbone-geometry alphabets is available online.[25]

[2]For more detail about these tests, see our earlier paper on backbone-geometry alphabets.[1]



(a) Residue absolute solvent accessibility.

(b) Neighborhood count of C_β with 14Å-radius sphere.

Fig. 1. Examples of seven-class, equal-sized alphabets of residue burial. Each bin in the histogram is assigned a letter so that a residue can be mapped to one of the seven states and a protein can be represented as a one-dimensional string of residue-burial states. Bin ranges are: $A = 0,0 < B < = 17, 17 < C < = 24, \cdots$ and $A < = 27, 27 < B < = 34, \cdots$.

penalties for misalignment, under- and over-alignment into a single number.[34] For two alignments, A and B (where A contains i aligned positions and B contains j aligned positions), the formula for shift-score is

$$\text{shift-score} = \frac{\sum_{i \in A} s(A(i)) + \sum_{j \in B} s(B(j))}{|A| + |B|}. \quad (3)$$

For a given pair of alignments, the alignment shift of a residue at position i ranges from $-N$ to N, where N is the length of the longer alignment. Shift is not used directly in the formula for shift-score, but is scaled to a related quantity called *s(i),* which is 1 for 0 shifts and approaches $-\varepsilon$ (ε is typically set at 0.2) for large shifts:

$$s(i) = \left\{ \begin{array}{ll} \dfrac{1+\varepsilon}{1+|\text{shift}(i)|} - \varepsilon & \text{if shift}(i) \text{ is defined} \\ 0 & \text{otherwise} \end{array} \right\}. \quad (4)$$

The scaling of shift(i) to *s(i)* avoids having a mean shift-error term dominated by large positive or negative shifts.

With ε set at 0.2, shift-score ranges from -0.2 (worst) to 1.0 (best), achieved when two alignments are identical.

Residue Burial Calculations

To compare a selection of burial alphabets, we investigated a variety of approaches for measuring burial. We calculated an estimate of residue *absolute solvent accessibility* (number of water molecules in direct contact with a residue) using the DSSP program.[35,12] *Relative solvent accessibility* was computed by dividing this area by maximum solvent-accessible area for a residue of interest, using values given by Rost.[24]

Most researchers compute neighborhood counts as the number of residues within a sphere centered at C_α or C_β, with each residue represented by its C_α or C_β.[22,23,5,36] The C_α or C_β may be referred to as the *interaction center.* We explored alternate locations for spheres using stochastic search methods, and varied sphere radii from 6 to 16 Å. Computations were done with an in-house software package called UNDERTAKER and a training set of 448 protein chains (the *monomeric-50-pc* dataset described in the "Datasets" section).

To identify the location of the most-occupied sphere (*dry spot*) for each side-chain type, we began with a randomly-chosen position in the neighborhood of each side-chain in the training set and tabulated the neighborhood counts. Positions were moved randomly and the tabulations repeated for a few thousand iterations over the training set. At the end of the search, the locations of the most-occupied spheres for each side-chain type were identified. The search was similar for the least-occupied spheres (*wet spots*), with the added constraint that spheres could not drift beyond a fixed distance from the side-chain of interest.

An alternate sphere location, known as the *generic spot,* was designed to be a more informative alternative to C_α or C_β. It was selected to be at a position on or near the side chain where the mutual information between amino acid type and sphere occupancy is maximized. This spot was obtained through a stochastic search process similar to that of the wet and dry spots, but at each candidate position, a table relating side-chain types to neighborhood counts was constructed and their mutual information was computed. After a few thousand iterations over the training set, the sphere location defining the largest mutual information was identified.

Alphabet Descriptions

We explored a set of alphabets for residue burial, testing some major alphabets from the literature as well as some of our own design. We designed 27 alphabets by constructing histograms of burial properties, using the *fssp-x* dataset (see "Datasets" section) and partitioning into seven equal-frequency bins. Each bin was assigned one alphabet letter.

An earlier study of solvent accessibility alphabets[29] showed that uneven partitioning can produce misleadingly high prediction accuracy, while uniform partitioning simplifies the comparison of different prediction methods

TABLE I. Description of the Solvent-Accessibility Alphabets

Name	Accessibility	Letters	Selection
REL-SA-2[24]	Relative	2	RelAcc: <16%:buried RelAcc: ≥16%: exposed
REL-SA-3[24]	Relative	3	RelAcc: <9%: buried RelAcc: 9–36%: intermediate RelAcc: ≥36%: exposed
REL-SA-10[24]	Relative	10	Class = $int(\sqrt{100*RelAcc})$
REL-SA-7	Relative	7	Equal-sized classes
ABS-SA-7	Absolute	7	Equal-sized classes

TABLE II. Description of the Neighborhood-Count Alphabets. Generic, Wet, and Dry Spots are Explained in the "Residue Burial Calculations" section

Name	Centroid	Radii	Atoms counted	Classes	Selection
BURIAL-C_α-6.5-7	C_α	6.5–14 Å respectively	C_α	7	Equal-sized classes
BURIAL-C_α-8-7	C_α				
BURIAL-C_α-10-7	C_α				
BURIAL-C_α-12-7	C_α				
BURIAL-C_α-14-7	C_α				
BURIAL-C_β-6.5-7	C_β	6.5–16 Å respectively	C_β	7	Equal-sized classes
BURIAL-C_β-8-7	C_β				
BURIAL-C_β-10-7	C_β				
BURIAL-C_β-12-7	C_β				
BURIAL-C_β-14-7	C_β				
BURIAL-C_β-16-7	C_β				
BURIAL-GEN-6.5-7	Generic spot	6.5–12 Å respectively	All atoms	7	Equal-sized classes
BURIAL-GEN-8-7	Generic spot				
BURIAL-GEN-9-7	Generic spot				
BURIAL-GEN-10-7	Generic spot				
BURIAL-GEN-12-7	Generic spot				
BURIAL-WET-6.5-7	Wet spot	6.5–12 Å respectively	All atoms	7	Equal-sized classes
BURIAL-WET-8-7	Wet spot				
BURIAL-WET-10-7	Wet spot				
BURIAL-WET-12-7	Wet spot				
BURIAL-DRY-6.5-7	Dry spot	6.5–14 Å respectively	All atoms	7	Equal-sized classes
BURIAL-DRY-8-7	Dry spot				
BURIAL-DRY-10-7	Dry spot				
BURIAL-DRY-12-7	Dry spot				
BURIAL-DRY-14-7	Dry spot				

and alphabets. We set out to maximize the number of bins, as fine-grained partitioning of accessibility values has been shown to be more predictable and conserved than coarse-grained partitioning.[24] Since absolute solvent accessibility has one-seventh of the data at zero (completely buried), we chose to use seven bins. Figure 1 shows an example of two of these alphabets.

Table I presents the tested solvent accessibility alphabets. Table II presents the tested neighborhood-count alphabets. For the latter, the burial measure is based on number of atoms within a spherical cut-off of a centroid atom. We did not explore real-valued burial measures because we tested the measures using a fold-recognition framework that requires discrete burial states.

Alphabet Evaluations

The alphabets were evaluated with a protocol described in our previous paper.[1] Appendix A contains brief descriptions of five alphabets of backbone-geometry analyzed in that paper.

Datasets

FSSP-X. all X-ray structures that are representatives in FSSP.[27]

MONOMERIC-50-PC. 448 monomeric protein whole chains, taken from one of Roland Dunbrack's high-quality culled PDB lists[37] with maximum sequence identity of 50%, maximum crystallographic resolution of 2.0 Å, and maximum R-factor cut-off of 0.2.

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.[37]

PCEF.TARGET97.FSSP.PAIRS. moderate difficulty alignment test set. It contains 170 pairs of protein chains with low sequence identity (7% to 27%) but significant structural similarity (DALI Z-score ≥ 6).

TWILIGHT.PAIRS. difficult alignment test set. It contains 200 pairs of protein chains with low sequence

TABLE III. Summary of the mutual information with amino acid, conservation and predictability of 30 residue-burial alphabets. The alphabets are ordered according to their conservation in alignments of FSSP structural neighbors. For comparison purposes, the amino-acid alphabet and five backbone-geometry alphabets evaluated in our earlier paper[1] are included (see Appendix A)

Name	Alphabet size	MI w/aa	Conservation fssp-x mutual info	Predictability bits saved per position
Residue burial alphabets				
BURIAL-C_β-16-7	7	0.089	0.682	0.502
BURIAL-C_β-14-7	7	0.106	0.667	0.525
BURIAL-C_α-14-7	7	0.078	0.655	0.508
BURIAL-C_β-12-7	7	0.124	0.640	0.519
BURIAL-C_α-12-7	7	0.093	0.586	0.489
BURIAL-GEN-12-7	7	0.154	0.570	0.378
BURIAL-GEN-10-7	7	0.176	0.541	0.407
BURIAL-GEN-9-7	7	0.189	0.536	0.415
BURIAL-C_β-10-7	7	0.128	0.513	0.470
BURIAL-GEN-8-7	7	0.211	0.508	0.410
BURIAL-GEN-6.5-7	7	0.221	0.465	0.395
BURIAL-C_α-10-7	7	0.092	0.459	0.419
BURIAL-C_β-8-7	7	0.136	0.449	0.470
REL-SA-10	10	0.184	0.407	0.470
REL-SA-7	7	0.183	0.402	0.461
BURIAL-C_α-8-7	7	0.079	0.387	0.379
ABS-SA-7	7	0.250	0.382	0.447
BURIAL-DRY-14-7	7	0.030	0.363	0.301
REL-SA-3	3	0.158	0.356	0.383
BURIAL-DRY-12-7	7	0.036	0.344	0.297
REL-SA-2	2	0.133	0.322	0.300
BURIAL-DRY-10-7	7	0.044	0.291	0.256
BURIAL-C_β-6.5-7	7	0.101	0.281	0.327
BURIAL-C_α-6.5-7	7	0.051	0.272	0.359
BURIAL-WET-6.5-7	7	0.141	0.266	0.237
BURIAL-WET-10-7	7	0.122	0.260	0.214
BURIAL-WET-8-7	7	0.119	0.238	0.199
BURIAL-DRY-8-7	7	0.039	0.218	0.194
BURIAL-WET-12-7	7	0.099	0.205	0.153
BURIAL-DRY-6.5-7	7	0.039	0.172	0.209
Amino acid alphabet				
AA	20	4.184	0.197	1.771
Backbone geometry alphabets				
STR	13	0.103	1.107	1.009
STRIDE	6	0.088	0.904	0.863
STRIDE-EHL	3	0.075	0.861	0.736
ANG	10	0.228	0.678	0.736
TCO	4	0.095	0.623	0.577

identity (3% to 24%) but significant structural similarity (VAST [38] p-value ≤ 0.0001, Yale[39] RMSD ≤ 4.0 Å, and a DALI[26] Z-score ≥ 7.0).

Datasets are available at http://www.soe.ucsc.edu/research/compbio/2thmm.

RESULTS

Conservation and Predictability

For a burial measure to be useful in remote homology recognition, it should be both conserved across evolution and predictable from amino-acid sequence. If a measure is conserved, then we could expect the burial pattern of one protein to show similarity to that of its remote homologs, much as the secondary structure of two remote homologs may be similar, even when their sequences are quite divergent. Additionally, the measure must be predictable from amino-acid sequence, because, typically, novel sequences must be annotated on the basis of sequence information only.

Table III presents our analysis of the mutual information with amino acid, conservation, and predictability for 30 residue burial alphabets. Five backbone-geometry alphabets from our earlier paper[1] are included for comparison. We found alphabets based on neighborhood counts to be more conserved and predictable than those based on solvent accessibility, and neighborhood count alphabets

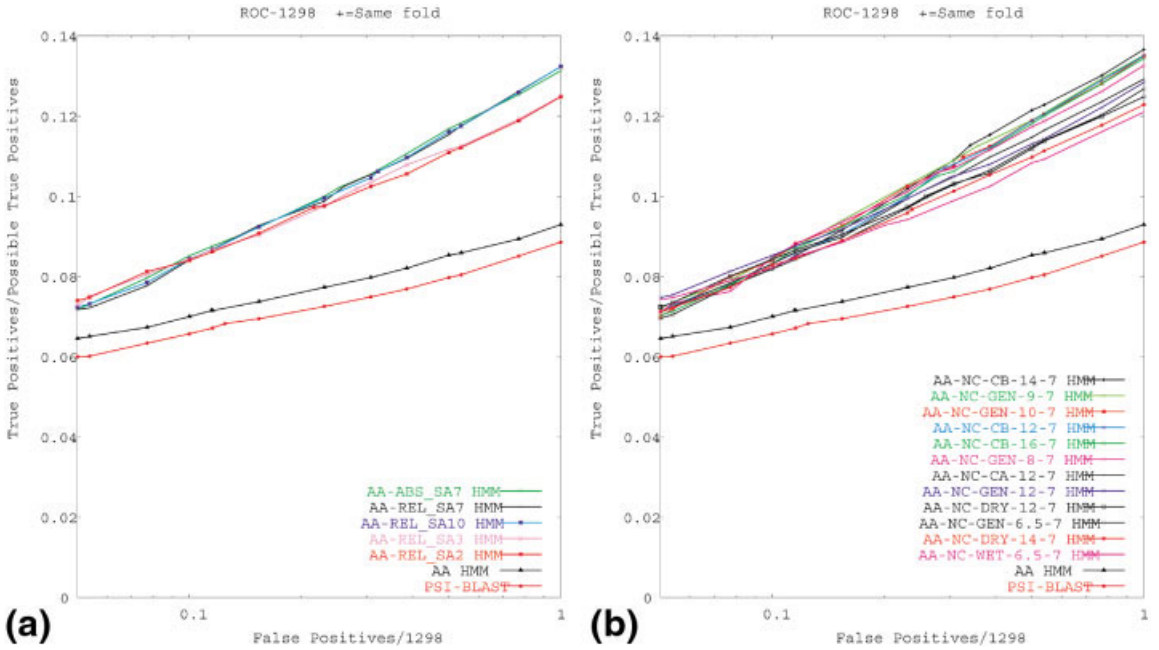


Fig. 2. (a) ROC_{1298}. Number of negatives recognized up to 1298 (1.0 false positives per query). (b) ROC_{1298}. Number of negatives recognized up to 1298 (1.0 false positives per query). Results of fold recognition tests on the benchmark dataset (*dunbrack-in-scop*), shown for two-track HMMs with an amino acid primary track and a solvent-accessibility alphabet (left) or a neighborhood-count alphabet (right) on the secondary track. The methods are ranked in the legends according to ROC_{1298} numbers. Performance of AA-only HMMs and PSI-BLAST are shown for comparison purposes.

with larger spheres (12–16 Å in radius) more conserved and predictable than those with smaller spheres. The most conserved alphabet uses a C_β interaction center with sphere radius of 16 Å, and most predictable uses C_β interaction center with sphere radius of 14 Å.

The residue-burial alphabets are less conserved and predictable than backbone-geometry alphabets. This is consistent with the "structure-more-conserved-than-sequence" paradigm. Mutations producing one tolerated side-chain substitution are more likely to alter solvent accessibility and side chain packing than the shape of the protein backbone. As multiple substitutions accumulate over time, the shape of the backbone may be affected, but this is a slower process.

The conservation and predictability measures in Table III can be compared to each other, as they share the same units (bits per position). When comparing these measures, we note that conservation tends to be higher than predictability. This is a consequence of our experimental design. Conservation of a structural property, such as burial, is measured according to structurally superimposed regions in FSSP alignments. Because these regions are structurally superimposed, one might expect them to be structurally conserved as well. In contrast, predictability should in fairness be measured over a predicted alignment; we used SAM-T2K[40] sequence alignment. Following the tradition of secondary structure prediction, we assessed predictability over all positions, not merely those in the core regions. Although sequence alignments cover a narrower evolutionary range than structural alignments, the predictability measures reflect some positions that may be structurally divergent, and thus less predictable.

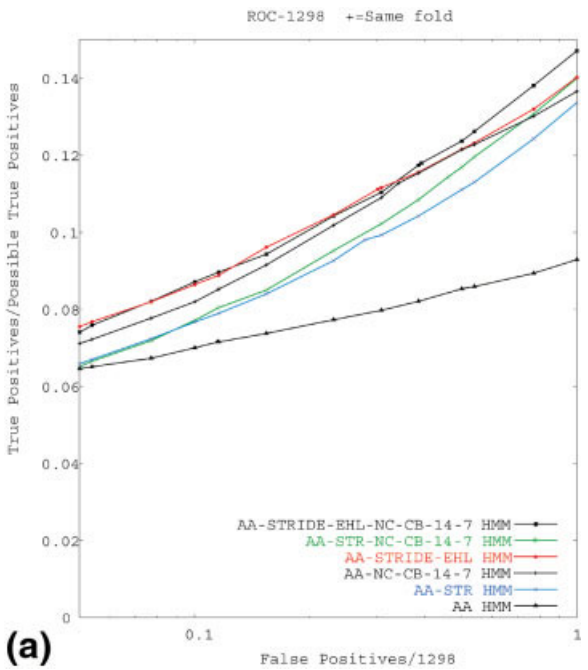


Fig. 3. (a) ROC_{1298}. Number of negatives recognized up to 1298 (1.0 false positives per query). Results of fold recognition tests on the benchmark dataset (*dunbrack-in-scop*), shown for three-track AA-STRIDE-EHL-BURIAL-C_β-14-7 and AA-STR-BURIAL-C_β-14-7 HMMs. Performance of two-track AA-STRIDE-EHL, AA-STR, AA-BURIAL-C_β-14-7 HMMs, and single-track AA-only HMMs are shown for comparison purposes. The three-track AA-STRIDE-EHL-BURIAL-C_β-14-7 HMM does better than any of the tested two-track HMMs at fold recognition. Methods are ranked in the legend according to ROC_{1298} numbers.

A few alphabets are less conserved than predictable. Most of these alphabets involve either solvent accessibility, or neighborhood counts with small spheres. These measures are more dependent on fine details of local interactions between side chains. Thus, these measures are easier to predict on the SAM-T2K sequence alignments, with their narrower evolutionary range than the FSSP structural alignments. The larger-radius neighborhood counts are less sensitive to small differences in the structural alignments (so can be better conserved) and depend on more of the protein (so may be harder to predict).

Alphabet Utility

We assessed the utility of these alphabets by incorporating them into two-track HMMs, with one track describing the amino acid sequence and one track describing residue burial. We tested these HMMs in all-against-all fold recognition tests on the *dunbrack-in-scop* benchmark set, a set of 1298 whole protein-chains with good-quality X-ray crystal structures and no two chains sharing more than 20% sequence identity. We assessed the results at the ROC_{1298} level of significance, which corresponds to an average of one false positive per protein chain. These results are reported in Figure 2 and Table IV. Results on amino-acid-only HMMs and PSI-BLAST are reported for comparison.

All alphabets tested provided a substantial improvement over amino-acid-only methods. The best results were obtained with a sphere of 14 Å in radius, centered at C_β, with seven burial states; this alphabet improved on the performance of an AA-only HMM by 38.4%. In general, the more fine-grained alphabets (7- and 10-state) performed better than 2- or 3-state alphabets, and spheres centered at C_β or the generic spot performed better than those centered at C_α, the dry spot, or the wet spot. C_β is known to be very informative about sidechain orientation;[36] these results suggest the same to be true of the generic spot. The relative solvent accessibility normalization had little effect, as the absolute and relative solvent accessibility alphabets were close in performance.

Table V shows the results of our alignment quality evaluations. We used the shift-score metric[34] to compare structural alignments produced by DALI and CE to predicted alignments produced by two-track HMMs with a burial alphabet on the secondary track. To assess the contribution of burial information to alignment quality, we compared the same structural alignments to predicted alignments produced by AA-only HMMs. The assessment was done on two benchmark sets of protein pairs ("difficult" and "moderate") of low sequence identity but significant structural similarity, described in the "Datasets" section.

The best alignments were produced using large-sphere neighborhood-count alphabets with C_α and C_β sphere centers and the ten-state relative solvent accessibility alphabet. The largest reported improvements in shift-score translate to an average increase of seven or eight correctly aligned positions, compared to AA-only HMMs, not as substantial as improvements contributed by backbone geometry alphabets. (The two-track AA-STR HMMs have an average increase of 16 correctly aligned positions compared to AA-only.)

TABLE IV. ROC Numbers From Fold Recognition Tests of HMMs Incorporating Side-chain Property Information, on a Difficult Set of 1298 Whole Chains[†]

Method	ROC_{1298}	ΔROC_{1298} (%)
Residue-burial alphabets		
AA-BURIAL-C_β-14-7 HMM	**0.1139**	38.4
AA-BURIAL-GEN-9-7 HMM	0.1133	37.7
AA-BURIAL-GEN-10-7 HMM	0.1127	36.9
AA-BURIAL-C_β-12-7 HMM	0.1124	36.6
AA-BURIAL-C_β-16-7 HMM	0.1120	36.1
AA-BURIAL-GEN-8-7 HMM	0.1114	35.5
AA-ABS-SA-7 HMM	0.1106	34.4
AA-REL-SA-10 HMM	0.1105	34.3
AA-REL-SA-7 HMM	0.1105	32.6
AA-BURIAL-C_α-12-7 HMM	0.1091	31.8
AA-BURIAL-GEN-12-7 HMM	0.1085	31.8
AA-BURIAL-DRY-12-7 HMM	0.1067	29.7
AA-BURIAL-GEN-6.5-7 HMM	0.1067	29.7
AA-REL-SA-3 HMM	0.1066	29.5
AA-REL-SA-2 HMM	0.1062	29.0
AA-BURIAL-C_α-14-7 HMM	0.1052	27.8
AA-BURIAL-DRY-14-7 HMM	0.1049	27.5
AA-BURIAL-WET-6.5-7 HMM	0.1035	25.8
Amino acid only		
AA HMM	0.0823	0
PSI-BLAST	0.0776	−5.7
Backbone-geometry alphabets		
AA-STRIDE-EHL HMM	0.1155	40.3
AA-TCO HMM	0.1126	36.8
AA-STRIDE HMM	0.1125	36.7
AA-ANG HMM	0.1119	36.0
AA-STR HMM	0.1065	29.4

[†]The numbers are an approximation to area under the ROC curve for thresholds that admit 1298 false positives, or an average of one false positive per chain. For comparison, results are shown for a single-track AA-only HMM, PSI-BLAST and for two-track HMMs incorporating backbone geometry informations.

This result is consistent with the lower predictability and conservation of the residue-burial alphabets, but the burial alphabets do better than expected at fold recognition; they perform similarly to the backbone-geometry alphabets. We believe this may be because residue-burial alphabets inadvertently encode chain length, a feature that has been shown to be useful in threading fold-recognition methods,[41] as similarities in length between a pair of proteins can indicate that they both belong to the same fold.

As with the backbone-geometry alphabets, alignment quality is consistently better when the HMM alignments are compared to DALI (rather than CE) reference alignments. There is one likely explanation. The two structural aligners use different objective functions; DALI attempts to minimize RMSD at the expense of inserting gaps, while CE creates longer continuous segments at the expense of slightly higher RMSD.[42] The HMMs are likely to model gaps in a similar fashion to DALI, as they use transition regular-

TABLE V. 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[†]

	Difficult set				Moderate set			
	DALI		CE		DALI		CE	
Reference alignment	MSS	Δ %	MSS	Δ %	MSS	Δ %	MSS	Δ %
Residue-burial alphabets								
Two-track BURIAL-C_β-14-7	**0.270**	22.7	0.265	21.0	**0.415**	13.7	**0.378**	16.3
Two-track BURIAL-C_α-12-7	0.269	22.3	**0.266**	21.5	0.411	12.6	0.375	15.4
Two-track BURIAL-C_α-14-7	0.266	20.9	0.261	19.2	0.407	11.5	0.372	14.5
Two-track REL-SA-10	0.265	20.5	0.258	17.8	0.402	10.1	0.358	10.2
Two-track BURIAL-C_β-16-7	0.263	19.6	0.258	17.8	0.410	12.3	0.375	15.4
Two-track BURIAL-C_β-12-7	0.263	19.6	0.262	19.6	0.411	12.6	0.375	15.4
Two-track ABS-SA-7	0.262	19.1	0.256	16.9	0.401	9.9	0.355	9.2
Two-track BURIAL-GEN-10-7	0.261	18.6	0.257	17.4	0.409	12.1	0.370	13.9
Two-track REL-SA-3	0.260	18.2	0.257	17.4	0.394	8.0	0.347	6.8
Two-track REL-SA-7	0.260	18.2	0.254	16.0	0.400	9.6	0.355	9.2
Two-track BURIAL-GEN-9-7	0.258	17.3	0.254	16.0	0.406	11.2	0.366	12.6
Two-track BURIAL-GEN-8-7	0.256	16.4	0.252	15.1	0.404	10.7	0.363	11.7
Two-track BURIAL-GEN-12-7	0.253	15.0	0.252	15.1	0.404	10.7	0.366	12.6
Two-track BURIAL-GEN-6.5-7	0.253	15.0	0.248	13.2	0.402	10.1	0.362	11.4
Two-track REL-SA-2	0.251	14.1	0.248	13.2	0.390	6.9	0.343	5.5
Two-track BURIAL-DRY-12-7	0.249	13.2	0.243	11.0	0.382	4.7	0.338	4.0
Two-track BURIAL-WET-6.5-7	0.246	11.8	0.241	10.1	0.385	5.5	0.347	6.8
Two-track BURIAL-DRY-14-7	0.246	11.8	0.241	10.1	0.384	5.2	0.347	6.8
Amino-acid only								
One-track AA	0.220	0.0	0.219	0.0	0.365	0.0	0.325	0.0
Backbone-geometry alphabets								
Two-track t2k STR	0.320	45.5	0.307	40.2	0.466	27.7	0.418	28.6
Two-track t2k STRIDE	0.357	62.3	0.292	33.3	0.452	23.8	0.400	23.1
Two-track t2k STRIDE-EHL	0.298	35.5	0.290	32.4	0.438	20.0	0.396	21.9
Two-track ANG	0.286	30.0	0.276	26.0	0.422	15.6	0.407	25.2
Two-track TCO	0.284	29.1	0.276	26.0	0.421	15.3	0.374	15.1

[†]Mean shift-scores (MSS)[34] and the percent increase in shift-score compared to AA-only HMMs (Δ %) are shown for alignments produced by two-track SAM-T2K HMMs, using 18 selected alphabets of side-chain surface accessibility and burial. For comparison, results for single-track SAM-T2K amino acid HMMs, and two-track backbone-geometry HMMs are shown.

izers (an apriori estimation of gap penalties) estimated from DALI structural alignments.

We observed BURIAL-C_β-14-7 to be the most useful representation of residue burial for both template selection and target-template alignment. To our knowledge, this has not been previously reported. It is consistent, however, with the findings of other groups. Nishikawa and Ooi showed that a 14 Å sphere radius with C_α interaction center (*Ooi14 number*) yields the highest correlation between amino acid sequence and neighborhood counts and is closely related to a residue's distance from the center of mass.[43] They believed that it reflects some important feature of globular protein tertiary structure. Williams et al. found Ooi14 to be the top-ranked property for distinguishing between ordered and disordered regions in proteins.[44] These groups do not appear to have considered the C_β interaction center, but C_β has recently been identified as the best interaction center for accessible surface potentials.[36]

Combining Burial and Backbone-Geometry Alphabets

The mutual information between the residue-burial alphabets and the amino acids ranges from 0.03 to 0.25 bits (see Table III), a similar range to that between backbone-geometry alphabets and the amino acids (0.01 to 0.228 bits).[1] As expected, mutual information with amino acids decreases as the radius of the neighborhood-count spheres increase. The alphabet that has highest mutual information with the amino acids is ABS-SA-7—absolute (but not relative) solvent accessibility is highly dependent on side-chain type, because there is a wide variation of maximum surface areas. (Values range from 84 Å^2 for glycine to 248 Å^2 for arginine.)

As shown in Table VI, the mutual information between residue-burial alphabets and backbone-geometry alphabets is small, indicating that the two classes of alphabets provide different information about the local environment of a residue. One side-chain-property alphabet (BURIAL-C_α-6.5-7) stands out as having consistently high mutual information with the backbone-geometry alphabets. This is not surprising; it has been been reported elsewhere[23] that a large neighborhood count within a small radius centered at the C_α atom frequently indicates that a residue is in a helix. The STR backbone-geometry alphabet that encodes patterns of beta-strand orientation has relatively high mutual information with the residue-burial alphabets, consistent with the theory that parallel strands are

TABLE VI. Mutual information between five backbone and 19 side-chain alphabets. Low mutual information between a pair of alphabets indicates that they contain independent information

	STRIDE	STRIDE-EHL	STR	ANG	TCO
BURIAL-C_β-16-7	0.0720	0.0677	0.1150	0.0328	0.0287
BURIAL-C_β-14-7	0.0790	0.0745	0.1240	0.0371	0.0324
BURIAL-C_α-14-7	0.0895	0.0844	0.1381	0.0403	0.0364
BURIAL-C_β-12-7	0.0915	0.0860	0.1374	0.0420	0.0365
BURIAL-C_α-12-7	0.0927	0.0878	0.1405	0.0429	0.0382
BURIAL-GEN-12-7	0.0394	0.0377	0.0605	0.0312	0.0207
BURIAL-GEN-10-7	0.0404	0.0381	0.0622	0.0334	0.0212
BURIAL-GEN-9-7	0.0423	0.0400	0.0647	0.0335	0.0214
BURIAL-GEN-8-7	0.0420	0.0390	0.0655	0.0369	0.0231
BURIAL-GEN-6.5-7	0.0438	0.0409	0.0662	0.0383	0.0238
REL-SA-10	0.0881	0.0823	0.1378	0.0529	0.0418
REL-SA-7	0.0792	0.0743	0.1281	0.0506	0.0405
ABS-SA-7	0.0624	0.0587	0.1036	0.0453	0.0319
BURIAL-DRY-14-7	0.0458	0.0427	0.0726	0.0185	0.0207
REL-SA-3	0.0568	0.0538	0.0957	0.0369	0.0301
BURIAL-DRY-12-7	0.0513	0.0477	0.0790	0.0215	0.0242
REL-sa-2	0.0426	0.0404	0.0707	0.0247	0.0207
BURIAL-WET-6.5-7	0.0240	0.0232	0.0434	0.0170	0.0126
BURIAL-C_α-6.5-7	0.2446	0.2084	0.2852	0.0732	0.0645

more likely than anti-parallel strands to be buried in the protein core.[45]

Since enhancing HMMs by adding either a backbone-geometry alphabet or a side-chain alphabet on a secondary track is useful in improving both HMM fold recognition and HMM alignment quality, we have begun to explore adding more than one secondary track to our HMMs. We would like to find out whether there is a preferred combination of alphabets that works for both fold recognition and alignment.

One approach for combining alphabets is to select the most informative combination of "orthogonal" alphabets. We can estimate this by adding the average predictable information per position in alphabet A to that in B and subtracting the average mutual information between observed occurrences of letters in A and B (see Table VI) from the sum. An alternative is to select alphabets that have been shown to work best independently in our two-track HMM experiments.

Since the best-performing residue-burial alphabet BURIAL-C_β-14-7 has very low mutual information (0.07 bits) with the three-state secondary-structure alphabet STRIDE-EHL, the alphabet of backbone geometry most useful for fold-recognition,[1] we chose this combination for our first three-track HMM experiment. The average predictable information per position in this alphabet combination is approximately $0.736 + 0.525 - 0.07 = 1.12$ bits. We also built a three-track HMM that combined BURIAL-C_β-14-7 with STR, reported in our earlier paper as the backbone-geometry alphabet most useful for alignment (combined predictable information is approximately $1.009 + 0.525 - 0.124 = 1.41$ bits per position). For both fold-recognition and alignment tests, we empirically chose track weights[1] of $w_1 = 1.0$ for amino-acid emissions, $w_2 = 0.3$ for backbone-geometry emissions, and $w_3 = 0.2$ for residue-burial emissions, after trying other combinations such as ($w_1 = 1.0$, $w_2 = 0.15$, $w_3 = 0.15$) and ($w_1 = 1.0$, $w_2 = 0.3$, $w_3 = 0.3$).

The three-track AA-STRIDE-EHL-BURIAL-C_β-14-7 HMM does better than any of the tested two-track HMMs at fold recognition, with ROC_{1298} of 0.1187, a 44% increase over AA-only HMM performance. The three-track AA-STR-BURIAL-C_β-14-7 HMM has ROC_{1298} of 0.1110, a 35% increase over AA-only, but lower than the ROC_{1298} for a two-track AA-BURIAL-C_β-14-7 HMM (see Table IV). Figure 3 shows the performance of AA-STRIDE-EHL-BURIAL-C_β-14-7 and AA-STR-BURIAL-C_β-14-7 HMMs, compared with two-track AA-STRIDE-EHL, AA-STR, AA-BURIAL-C_β-14-7 HMMs, and single-track AA-only.

For alignment quality, both the AA-STR-BURIAL-C_β-14-7 and AA-STRIDE-EHL-BURIAL-C_β-14-7 three-track HMMs do better than any of the two-track HMMs (excepting two-track AA-STR for moderate set and CE reference), but AA-STR-BURIAL-C_β-14-7 is consistently the best combination (see Table VII). The shift scores of AA-STR-BURIAL-C_β-14-7 HMMs represent an average gain of 21 correctly aligned positions on the difficult test set and 18 correctly aligned positions on the moderate test set, with respect to AA-only HMMs.

Although we have been able to identify alphabet combinations that work well on either remote homolog identification or target-template alignment, we have not found a preferred combination that works for both problems. Finding optimal combinations of alphabets and track weights for three-track HMMs remains an open problem, and we plan to do extensive testing to discover how to build the most effective three-track HMMs.

CONCLUSION

We evaluated many ways of describing residue burial in a local structure alphabet, according to conservation in structural alignments, predictability from amino acid sequence, and utility in fold recognition and sequence align-

TABLE VII. Alignment Quality Improvement From AA-STR-BURIAL-C_β-14-7 and AA-STRIDE-EHL-BURIAL-C_β-14-7 Three-track HMMs[†]

	Difficult set mean shift-score		Moderate set mean shift-score	
Reference alignment	DALI	CE	DALI	CE
Three-track AA-STR-BURIAL-C_β-14-7	**0.340**	**0.322**	**0.480**	**0.432**
Three-track AA-STRIDE-EHL-BURIAL-C_β-14-7	0.323	0.308	0.457	0.417
Two-track AA-STR	0.320	0.307	0.466	0.418
Two-track AA-STRIDE-EHL	0.298	0.290	0.438	0.396
Two-track AA-BURIAL-C_β-14-7	0.270	0.265	0.415	0.378
	Δ Mean shift-score (%)		Δ Mean shift-score (%)	
Three-track AA-STR-BURIAL-C_β-14-7	**54.6**	**47.0**	**31.5**	**33.0**
Three-track AA-STRIDE-EHL-BURIAL-C_β-14-7	46.8	40.6	25.2	28.3
Two-track AA-STR	45.5	40.2	27.7	28.6
Two-track AA-STRIDE-EHL	35.5	32.4	20.0	21.9
Two-track AA-BURIAL-C_β-14-7	22.7	21.0	13.7	16.3

[†]Shift scores and percent increase in shift score from AA-only HMMs are shown. Performance of two-track AA-STRIDE-EHL, AA-STR, AA-BURIAL-C_β-14-7 HMMs are shown for comparison purposes. The best shift-scores (for AA-STR-BURIAL-C_β-14-7) represent an average gain of 21 correctly aligned positions on the difficult test set and 18 correctly aligned positions on the moderate test set, with respect to AA-only HMMs.

ment. The alphabets that were the most predictable and conserved also yielded the best performance in fold recognition and sequence alignment, validating that these are good indicators of alphabet performance.

When fold recognition methods using both amino acids and burial information were compared to AA-only methods, all burial alphabets yielded a substantial improvement. The best results were achieved by a neighborhood count alphabet with seven states, a large-radius sphere (14 Å) and a C_β interaction center; this improved fold-recognition accuracy by 38% and alignment accuracy by 23%. In general, the more fine-grained alphabets (7- and 10-state) performed better than 2- or 3-state alphabets, a finding supported by prior work.[24]

By comparing backbone-geometry alphabets and residue-burial alphabets with a mutual information measure, we have been able to quantify the amount of distinct information they contain. The quantification is useful in selecting pairs of non-redundant alphabets to be combined in a fold-recognition or homology-modeling method. Within the context of an HMM-based method, we have been able to identify combinations of alphabets that work well for either fold recognition or alignment, but have not yet found a combination that is best for both.

We plan to improve on our crude estimates of how much predictable information is available in pairs of local structure alphabets and to better identify good combinations of alphabets. We are interested in incorporating these alphabets into three-track HMMs and also in exploring ways of combining scores produced by "juries" of several two-track HMMs, with different alphabets on their secondary tracks.

ACKNOWLEDGMENTS

This work was supported in part by 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 Richard Hughey and Mark Diekhans for contributing to the software used in our experiments, to Yael Mandel-Gutfreund for valuable conversations, and to Spencer Tu for assembling one of our alignment quality test sets.

REFERENCES

1. Karchin R, Cline M, Mandel-Gutfreund Y, Karplus K. Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. Proteins 2003;51:504–514.
2. Dean AM, Golding GB. Enzyme evolution explained (sort of). In Pacific Symposium on Biocomputing, 2000, p 6–17.
3. Terp BN, Cooper DN, Christensen IT, Jorgensen FS, Bross P, Gregersen N, Krawczak M. Assessing the relative importance of the biophysical properties of amino acid substitutions associated with human genetic disease. Human Mutation 2002;20:98–109.
4. Sunyaev S, Ramensky V, Koch I, Lathe III W, Kondrashov AS, Bork P. Prediction of deleterious human alleles. Human Molecular Genetics 2001;10:591–597.
5. Saunders CT, Baker D. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. J Mol Biol 2002;322:891–901.
6. Chothia C. Principles that determine the structure of proteins. Annu Rev Biochem 1984;55:537–572.
7. Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. Nature 1986;319:199–203.
8. Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. Protein 1999; Suppl 3:171–176.
9. Betancourt MR, Thirumalai D. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein Sci 1999;8:361–319.
10. Cline MS, Karplus K, Lathrop RH, Smith TF, Rogers Jr. RG, Haussler D. Information-theoretic dissection of pairwise contact potentials. Proteins 2002;49:7–14.
11. Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001;294:93–96.
12. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971;55:151–176.
13. Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms. J Mol Biol 1973;79:351–371.

14. Richards FM. Areas, volumes, packing and protein structure. Ann Rev Biophys Bioeng 1977;6:151–176.
15. Wodak SJ, Janin J. Analytical approximation to the accessible surface area of proteins. Proc Nat Acad Sci USA 1980;77:1736–1740.
16. Richmond DJ. Solvent accessible surface area and excluded volume in proteins. J Mol Biol 1984;177:63–89.
17. Wang Y, Levinthal C. A vectorized algorithm for calculating the accessible surface area of macromolecules. J Comput Chem 1991; 12:868–871.
18. Le Grand SM, Merz JKM. Rapid approximation to molecular surface area via the use of boolean logic and lookup tables. J Comput Chem 1993;14:349–352.
19. Barrett C. Investigation of non-pairwise protein structure score functions using sets of decoy structures, PhD thesis, University of California, Santa Cruz, 2001.
20. Berrera M, Molinari H, Fogolari F. Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. BMC Bioinformatics 2003;4:8.
21. Flöckner H, Braxenthaler M, Lackner P, Jaitz M, Ortner M, Sippl MJ. Progress in fold recognition. Proteins 1995;3:376–386.
22. Fariselli P, Casadio R. Prediction of the number of residue contacts in proteins. In Altman R, Bailey TL, Bourne P, Gribskov M, Lengauer T, Shindyalov IN, Ten Eych LF, Weissig H, editors. Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, p 146–151. San Diego, CA: AAAI Press, 2000.
23. Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of coordination number and relative solvent accessibility in proteins. Proteins 2002;47:142–153.
24. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216–226.
25. Karchin R. Evaluating local structure alphabets for protein structure prediction. PhD thesis, University of California, 2003. available at http://www.soe.ucsc.edu/rachelk/pubs/rk-thesis.pdf.
26. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
27. Holm L, Sander C. Mapping the protein universe. Science 1996;273: 595–603.
28. Rost B. PHD: predicting one-dimensional protein structure by profile based neural networks. Meth Enzymol 1996;266:525–539.
29. Thompson MJ, Goldstein RA. Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. Protein 1996;25:38–47.
30. Richardson CJ, Barlow DJ. The bottom line for prediction of residue solvent accessibility. Protein Eng 1999;12:1051–1054.
31. Naderi-Manesh H, Sadeghi M, Sharhriar A, Moosavi Movahedi AA. Prediction of protein surface accessibility with information theory. Proteins 2001;42:452–459.
32. Rost B, Sander C, Schneider R. Redefining the goals of protein secondary structure prediction. J Mol Biol 1994;235:13–26.
33. Gribskov M, Robinson NL. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. Compu Chem 1996;20:25–33.
34. Cline M, Hughey R, Karplus K. Predicting reliable regions in protein sequence alignments. Bioinformatics 2002;18:306–314.
35. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
36. Melo F, Sanchez R, Sali A. Statistical potentials for fold assessment. Protein Sci 2002;11:430–448.
37. Dunbrack R. Culling the PDB by resolution and sequence identity, 2001. http://www.fccc.edu/research/ labs/dunbrack/culledpdb.html. Accessed 21 August 2001.
38. Gilbrat J, Madej T, Bryant S. Surprising similarities in structure comparison. Curr Opin Struct Biol 1996;6:377–85.
39. Gerstein M, Levitt M. Comprehensive assessment of automatic structural alignment against a manual standard, the SCOP classification of proteins. Protein Sci 1998;7:445–456.
40. 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;45:86–91.
41. McGuffin LJ, Jones DT. Improvement of the genTHREADER method for genomic fold recognition. Bioinformatics 2003;19:874–881.
42. Shindyalov IN, Bourne PE. An alternative view of protein fold space. Proteins 2000;38:247–260.
43. Nishikawa K, Ooi T. Radial locations of amino acid residues in a globular protein: correlation with the sequence. J Biochem 1986; 100:1043–1047.
44. Williams RM, Obradovi Z, Mathura V, Braun W, Garner EC, Young J, Takayama S, Brown CJ, Dunker AK. The protein non-folding problem: amino acid determinants of intrinsic order and disorder. In Pac Symp Biocomput 2001, p 89–100.
45. Mandel-Gutfreund Y, Gregoret L. On the significance of alternating patterns of polar and nonpolar residues in beta strands. J Mol Biol 2002;323:453–461.
46. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins 1995;23:566–579.
47. 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.
48. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J Mol Biol 1963;7: 95–99.

APPENDIX A

Backbone Geometry Alphabets

Five of the backbone geometry alphabets were evaluated in our earlier paper:[1]

STRIDE.[46] Six-letter secondary structure alphabet: E (beta strand), B (short beta bridge), G (3–10 helix), H (alpha helix), T (turn), and L (loop).

STRIDE-EHL. Three-letter (EHL) secondary structure alphabet that reduces STRIDE assignments to helix, strand, or loop. In this mapping, G is included in the helix class, B in the strand class, and S and T in the loop class.

ANG.[47] Based on the ϕ-ψ alphabet used in HMMSTR. The ϕ-ψ plane[48] is partitioned into ten regions.

STR. An enhanced version of DSSP that subdivides DSSP letter E (beta strand) into six letters, according to properties of a residue's relationship to its strand partners.

TCO. A four-letter alphabet of cosine-of-carbonyls values.