Constructing Amino Acid Residue Substitution Classes Maximally Indicative of Local Protein Structure

Michael J. Thompson[1] **and Richard A. Goldstein**[1,2]
[1]Biophysics Research Division, [2]Department of Chemistry, University of Michigan, Ann Arbor, Michigan 48109-1055

ABSTRACT **Using an information theoretic formalism, we optimize classes of amino acid substitution to be maximally indicative of local protein structure. Our statistically-derived classes are loosely identifiable with the heuristic constructions found in previously published work. However, while these other methods provide a more rigid idealization of physicochemically constrained residue substitution, our classes provide substantially more structural information with many fewer parameters. Moreover, these substitution classes are consistent with the paradigmatic view of the sequence-to-structure relationship in globular proteins which holds that the three-dimensional architecture is predominantly determined by the arrangement of hydrophobic and polar side chains with weak constraints on the actual amino acid identities. More specific constraints are imposed on the placement of prolines, glycines, and the charged residues. These substitution classes have been used in highly accurate predictions of residue solvent accessibility. They could also be used in the identification of homologous proteins, the construction and refinement of multiple sequence alignments, and as a means of condensing and codifying the information in multiple sequence alignments for secondary structure prediction and tertiary fold recognition.**

Key words: **protein evolution, structure prediction, information theory, amino acid substitution, multiple sequence alignment**

INTRODUCTION

Although it is highly desirable to know the detailed three dimensional structure of a protein under study, such structures are few in number and laboriously determined. In contrast, amino acid sequences are often readily obtainable. When the sequences of homologous proteins are available as well, the sets of residue substitutions resulting from evolutionary differentiation can provide a rich source of information about the protein's function and structure. Structural information is contained in these families as a result of the degenerate relationship between sequence and structure. During evolution, a three-dimensional fold can persist despite significant divergence of the amino acid sequences.[1–4] In order to preserve the structural or functional integrity of the protein, important sites in the sequence must either conserve some specific characteristics requiring conservation of amino acid identity, or preserve some more general property such as polarity, charge, or size. Variable sites with little or no structural or functional constraints may evolve by random fixation of neutral or nearly neutral mutations.[5,6] As a result, alignments of homologous sequences provide more information about the architectural necessities of their commonly-adopted fold than do lone sequences, implicitly carrying information about the non-local interactions frequently invoked to explain the apparent 65% accuracy ceiling for secondary structure prediction.

Various methods have been developed for extracting information from multiple sequence alignments. The most straightforward approach has been to use the multiple members of a protein family in order to provide "signal averaging" of structure predictions over multiple homologous sequences.[7–13] This kind of simple averaging, however, does not capture the information implicit in structurally characteristic patterns of side chain variability. Another approach has been the construction of consensus or signature sequence segments characteristic of particular structures, which are often used in database searches.[12,14–20] By concentrating on the conserved residues at the expense of alignment positions which allow side chain variability, these methods lose much of the information contained in the multiple alignments, as discussed below. In addition, these particular methods create templates specific for each individual protein family, and do not give insight into the more universal patterns found in proteins.

Both the profile method of tertiary structure recognition developed by Eisenberg and coworkers[21–23]

Received July 14, 1995; revision accepted November 20, 1995.

Address reprint requests to Richard A. Goldstein, Biophysics Research Division, Department of Chemistry, University of Michigan, Ann Arbor, MI 48109-1055.

and the secondary structure and surface accessibility prediction work of Wako and Blundell[24] employed environment-dependent mutation matrices and tables of structural propensities. A recent hybrid method used similar tables.[25] Measures of side chain conservation have been used to weight secondary structure predictions[10] and to train neural networks, in combination with substitution profiles, to predict secondary structure and solvent accessibility.[26,27] While some of these more quantitative methods tend towards general applicability, they do not lend themselves to the abstraction of general principles relating protein structure and sequence. In contrast, Benner and Gerloff have developed methods for analyzing patterns of sequence variation using heuristic rules which vary as subsets of the alignments are manipulated.[28] While this work has claimed certain successes, it lacks quantitative rigor and reproducibility.[29–31] Benner et al. have also pioneered the consideration of correlated mutations as an approach to structure prediction.[32] Such efforts are still at a preliminary stage.[33–36] In spite of this work and the more common uses of multiple sequence alignments in the biochemical or crystallographic characterization of proteins, a rigorous and generally applicable classification of these structurally-constrained sets of residue substitutions is lacking.

In this article, we use information theory to construct sets of residue substitution classes which provide maximal information about local protein structure, classified into combinatorial categories of secondary structure and solvent accessibility. As this approach contains no preconceived notions concerning how the amino acid types should be clustered, beyond those proclivities present in the original mutation matrices used to construct the database of multiple sequence alignments, we have introduced no bias toward the development of any particular architecture of class membership and class linkages. To quantify the correspondence between amino acid sequence and local protein structure, we employ the concept of "mutual information." Briefly, mutual information is the amount of knowledge obtainable about one random variable by knowledge of another; in this case, how much knowledge of the local structure can be obtained by knowing the substitution class of the alignment position. This measure enables us to search the space of possible substitution classes over a database of representative protein structures and their homologs and find the set which possesses the strongest correlations with local structural environments.

The resulting substitution classes display consistency with a few well-known trends for patterns of amino acid mutations to correlate with the physicochemical properties of those amino acids. In the context of our substitution classes, however, these relationships are not nearly so highly idealized as in the hierarchical "Amino Acid Class Covering patterns" of Smith and Smith[37,38] or the Venn diagrams of Taylor.[39] This is due to the fact that our methodology contains no ad hoc elements, and allows for the unprejudiced identification of structurally informative classes of residue substitution. Our substitution classes are also conceptually similar to the clusters of "interior-indicating" and "surface-indicating" side chains used by Benner and coworkers in their prediction work.[32] In contrast to their heuristic groupings, our classes are derived by a rigorous statistical methodology. A similarly rigorous approach to generalizing the patterns of residue substitutions found in multiple sequence alignments can be found in the work of Haussler and coworkers.[40] Their work, however, is aimed at developing improved methods for multiple sequence alignment and does not investigate the relationship between allowed substitutions and local protein structure.

In this article, we report on the set of 28 optimal substitution classes which provides close to the maximum resolution of sequence to structure relationships without significant dataset-dependence. These classes provide much more structural information than is extractable from single sequences. They also convey more structural information with fewer parameters than the covering classes developed by Smith and Smith or the Venn diagrams constructed by Taylor. These classes can be used to address questions concerning what, if any, structure-determining properties are being conserved, and generate qualitative insight into the relationship between sequence and structure. In addition to their use in structure prediction,[41] potential applications include sequence alignment, detection of homologous proteins, and the further generation and refinement of biochemical insight.

METHODS

Data Encoding

A significant and often neglected issue is the non-uniformity of the databases of known protein sequences; a large probability of a given residue at any location may only indicate that that particular residue is characteristic of a protein subfamily over-represented in the database. Such biases can be rather extreme: mammals make up 69% of all myoglobin sequences found in the SWISS-PROT Database (release 28).[42] In order to correct for these biases, it is necessary to either cull or weight the sequences in order to get a more even distribution, a process that involves substantial approximations and assumptions.

We approach the non-uniformity of the database in a manner similar to that used in the "Amino Acid Class Covering patterns" developed by Smith and Smith[37,38] and in the "minimal sets" suggested by Taylor.[39] These methods consider only compatibility of various amino-acid residues with a given location,

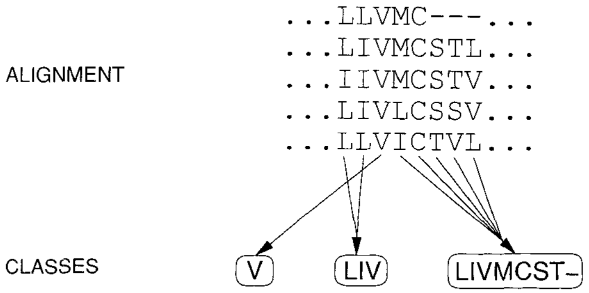

Fig. 1. Illustrative example of assigning the residue positions of a multiple sequence alignment into substitution classes.

without regard to the number of sequences that contain any particular amino acid residue. In our work, a particular position in a multiple sequence alignment is characterized by the first substitution class that is sufficiently general to account for all of the side chain types observed at that site. As shown in the simplified example of Figure 1, an alignment position with a conserved valine would be assigned to the substitution class containing only valine, while the variable positions that contained any combination of valine, leucine, or isoleucine would be assigned to the second class, and so on. As positions are assigned to the first appropriate class, the order of classes is significant. For the actual set of classes constructed by our method, the first 20 substitution classes correspond to the conserved examples of the 20 amino acid types. The remaining classes increase in generality until the last substitution class includes all possible amino acid types, so that each position in the dataset corresponds to one of the substitution classes. There could conceivably be as many as 2^{21}–2 substitution classes corresponding to all of the possible combinations of amino acid types, and the possibility of a gap. The residues observed at each position in the multiple alignments are encoded as a 21-bit binary vector, where each of the bits corresponds to the presence or absence of an amino acid type or gap at that position. Membership in a substitution class can be assigned rapidly using logical bit operations.

Theory

We perform a search for the optimal set of classes, defined as the set that maximizes the information about local structure furnished by class membership. We quantify this approach using concepts borrowed from information theory. (For a review, see reference[43]; for pioneering examples of the use of information theory in prediction efforts see references[44–51]). Consider the random variable X, which can take on the possible values $\{x_1, x_2 \ldots x_n\}$. The "Shannon entropy" of the probability distribution of X, $H(X)$, represents the uncertainty of this random variable:[52]

$$H(X) = -\sum_{i} p(x_i)\ln p(x_i). \tag{1}$$

where $p(x_i)$ is the probability of X having the value x_i. The joint entropy of two random variables X and Y, $H(X,Y)$, is calculated from their joint probability distribution:

$$H(X,Y) = -\sum_{i,j} p(x_i,y_j)\ln p(x_i,y_j). \tag{2}$$

The mutual information of these two random variables is defined as the difference between the sum of the entropies of their respective probability distributions and the entropy of their joint probability distribution.

$$M(X;Y) = H(X) + H(Y) - H(X,Y) \tag{3}$$

If knowledge of the value of one variable provides knowledge about the value of the other variable, then the information provided by knowledge of both variables will be less than the sum of the knowledge of each variable considered independently. Here, the mutual information represents how much knowledge of one variable tells us about the value of the other. Clearly, if the variables are independent, and knowledge of one provides no information about the value of the other, then $M \rightarrow 0$.

For our purposes, we write,

$$M(\{C\};\{2^{\Phi}\}) = 100 \times [H(\{C\}) + \mathrm{H}(\{2^{\Phi}\}) - \mathrm{H}(\{C\},\{2^{\Phi}\})] \tag{4}$$

where $\{2^{\Phi}\}$ indicates the set of eight local structure categories defined below, and $\{C\}$ is the set of amino acid substitution classes being considered. The values are multiplied by 100 for ease of comparison. The mutual information quantifies how much information about the local structure is contained in knowledge of the substitution class. The set of classes that maximizes the mutual information is, therefore, the best set of classes for extracting information about local protein structure.

We use a Metropolis algorithm[53] with simulated annealing to search for optimal sets of substitution classes. The searches begin with a set of substitution classes with randomly constructed memberships. The mutual information is calculated for that set. One of the bits of the binary vectors representing one of the classes is altered at random. Another possible move is the random selection and switching of two of the substitution classes to rearrange the order of the classes. The mutual information of this new trial configuration is recalculated. If the new trial configuration has a mutual information value higher than the current configuration, the change is accepted and the trial configuration becomes the new current configuration. If the trial configuration's mutual information is less than that of the current configuration, the change is accepted with probability

TABLE I. List of 111 Protein Chains by PDB Identifier Code*

1aak	1ab2	1abmA	1ak3A	1apmE	1atr	1avhA	1baa
1babB	1bet	1bfg	1bmdA	1caj	1cauA	1cauB	1ccr
1cgt	1cobA	1cpcA	1cpcB	1ctm	1dlhB	1eco	1fbaA
1fha	1fkb	1fxiA	1gatA	1gd1O	1gp1A	1hbq	1hdxA
1hgeA	1hgeB	1hleA	1hsbA	1hunA	1huw	1ipd	1le4
1lenA	1lgaA	1mctA	1mfbH	1minA	1minB	1nbtA	1ndc
1nipA	1ofv	1pfkA	1pho	1php	1pk4	1pkp	1poa
1ppn	1rla2	1rec	1rfbA	1rnd	1s01	1sacA	1scmA
1scmB	1scmC	1sltA	1smrA	1spa	1tbpA	1tfd	1tgxA
1tie	1tndA	1treA	1tys	1wsyB	1xyaA	1zaaC	2acq
2azaA	2btfA	2cas	2cpl	2ctc	2gda	2hmzA	2ihl
2lh2	2mge	2mipA	2plv1	2plv3	2reb	2sn3	2snv
2tgi	3cla	3pgm	3rubS	4blmA	4gcr	4gpb	4htcI
5fbpB	5nn9	5p21	8catA	8tlnE	9ldtA	9rnt	

*The fifth letters in the codes of multichain proteins designate the chains used.

$e^{(M_{trial}-M_{current})/T}$ where T is an empirically defined decreasing function of the number of trials.

Databases

The dataset of protein chains was selected from the October 1994 PDBselect list of representative structures sharing less than 25% sequence identity between any pair, as compiled by Hobohm and Sander.[54] Alignments of homologs are extracted from "homology derived structures of proteins" (HSSP) files.[55] Application of a minimum length requirement of 60 residues for example proteins, of a minimum number of 10 homologs per sequence position and of a 40% lower bound on sequence identity between the example chain and its homologs produces the set of 111 protein chains (25,511 residues) listed in Table I. This dataset is well-distributed over the ranges of % β-strand content and % helical content observed in the protein structure databases. We do not classify the example proteins into discrete categories such as the typical "mostly α," "mostly β," "mixed α/β," and "irregular" due to the inconsistency among the various sets of classification criteria that have been proposed in the literature.[56]

Structure Definitions

In the HSSP files, there are three types of gaps: end gaps, internal gaps, and insertions. End gaps in the homolog sequences are ignored due to the ambiguity regarding whether these regions outside the boundaries of the alignment bear any structural similarity to the example protein. Insertions in the homologous sequences are also discounted because they correspond to gaps in the example protein where no secondary structure information is available. Thus, the only gaps taken into account are those within the homolog sequences.

The local structure of a residue location is assigned as one of eight categorical combinations of the four standard secondary structure states: helix, strand, turn, and coil; and two solvent accessibility states: buried and exposed. Secondary structure and solvent accessibility information is taken from the "Dictionary of Protein Secondary Structure" (DSSP) files of Kabsch and Sander[57] which were constructed based on the coordinates supplied in the Protein Data Bank (PDB) files for the proteins.[58,59] Solvent accessibility values reported for single chains of multimeric proteins were calculated based on the multi-chain complexes. The solvent accessibilities are normalized by the values obtained for glycine-X-glycine tri-peptides by Shrake and Rupley[60] in order to generate fractional exposures. The choice of threshold for distinguishing buried and exposed states and the classification of non-standard secondary structures into the common four are explained in the results section.

RESULTS

Our first use of mutual information was in determining the optimal definition of the local structure categories. The DSSP files contain eight possible structural assignments, adding the β-bridge, π-helix, 3_{10}-helix, and bend to the canonical α-helix, β-strand, turn and coil. It is not obvious how to partition these additional assignments among the four more standard structures. For example, the GOR method[61] assigns bend and β-bridge locations to coil and π-helix and 3_{10}-helix locations to helix, while Rost and Sander group 3_{10}-helix locations with helix, and combine β-bridge, turn, bend, and π-helix locations into a "loop" category.[26] Likewise, there is a problem in choosing a surface accessibility threshold, as noted by other authors.[27]

Our solution to these problems was to find the solvent accessibility threshold combined with the mapping of the four non-standard secondary structure types into the standard four which provided the maximum mutual information for a set of 40 classes (one variable class and one conserved class for each amino acid type). We chose to work with these classes for two reasons. One, these classes are "non-adjustable" and, therefore, we avoid the computational enormity of simultaneously or iteratively searching for a set of optimal substitution classes for every possible structural categorization. Two, these classes convey information about the residue identity and the basic conserved or variable nature of a sequence position. For each sequence position in the 111 protein dataset, the residue identity was taken from the example sequence while the conserved or variable nature of the position was determined by the alignment of the homologs. We exhaustively considered every possible secondary structure mapping for solvent accessibility thresholds ranging from 15 to 25%. The optimal structure categorization assigned β-bridge locations to β-strand, 3_{10}-helix and π-helix locations to helix, and bend locations to turn, for a solvent accessibility cutoff of 22%. This classified 47% of the 25,511 residues in our 111 protein chain dataset as exposed, 34% as helical, 23% as strand, 22% as turn, and 21% as coil. This local structure categorization was robust to the choice of dataset.

We also used mutual information to determine how distant homologs should be included in the multiple sequence alignments. Inclusion of more distant homologs increases the probability that the range of possible amino acid residues that can occupy each location is well defined. However, it also increases the probability that a rare but possible mutation may be included, and given the nature of the substitution classes, be accorded as much weight as a more likely residue. The accidental inclusion of nonhomologous sequences becomes a concern as well as the general reliability of the multiple sequence alignments. Using the same type of search procedure with non-adjustable classes as before, we found that the mutual information plateaued in the range of 40–55% sequence identity. We chose to work with the dataset produced by the 40% lower bound as it provided the largest number of example proteins.

For all searches for optimal sets of residue substitution classes, 20 non-adjustable classes were explicitly considered, representing conserved examples of each of the 20 amino acid types. While a variable set of residues may represent conservation of a generic property of side chains associated with a particular local protein conformation, conservation of a certain side chain type more likely represents a strict functional constraint which may or may not be coupled with a conserved structural feature. Also, the structural propensities of conserved positions of a particular side chain are likely to be different than those of variable positions which allow that side chain. These twenty classes allow for separation of such differences without creating any additional complexity in the optimization searches.

With a very small number of residue substitution classes, only very generic physicochemical properties and their correlations with local protein structures are resolved. As the number of classes, N, increases, more specific residue properties and their correlations with particular local protein conformations are distinguished. As can be seen in Figure 2, the mutual information begins to saturate in the range of a relatively small number of classes. Therefore, the insignificant gain in mutual information provided by larger sets of optimal substitution classes is obtained at increasingly prohibitive computational costs as the number of possible configurations roughly grows as 2^{21N}. There is also an increasing risk that the resulting classes represent memorization of dataset-specific sequence to structure correlations. This memorization reaches a maximum when each alignment position in the database which has a unique set of residues is allowed to be a substitution class itself. There were 7,705 such positions in our database of alignments. Therefore, the mutual information value calculated for these 7,705 classes—86.22—was the maximum attainable value for this dataset. Here we report on the optimal set of 28 classes. At this resolution, many of the well-known relationships between generic residue properties and local protein structure are distinguished with a consequent large initial gain in mutual information. As shown in Figure 3, the Metropolis search scheme was capable of finding the optimal set of 28 substitution classes in a consistent and rapid fashion.

To address the issue of dataset dependence, we first split the database of 111 protein chains into subsets of 55 and 56 examples. The optimal sets of 28 classes derived for each dataset-half could be mapped 1 to 1 onto the 28 classes optimized for the entire 111 protein dataset. Changes between the subset-optimized and globally optimized classes involved small changes in the residue membership of some classes and a shift in the order of classes. For the half-dataset classes, 82% of the alignment positions were assigned to the same classes as for the globally optimal set. This indicates that while the exact appearance of the optimal set of classes shows some dataset dependence, the relationships between sequence and structure represented by these sets of substitution classes are relatively robust.

We separately addressed the possibility of memorization of particular residue substitution patterns by dividing the 111 protein chain dataset into fifths. The optimal set of substitution classes was found for 4/5s of the dataset and a mutual information value was calculated over the remaining 1/5 of the dataset using these classes. This was done cyclically for all permutations of the dataset-fifths, and an average mutual information was calculated. Similarly, the set of substitution classes optimized for the entire 111 protein chain database was used to calculate a mutual information value over each 1/5 of the database. These values were also averaged. The differences between the two averages were negligible for numbers of substitution classes from 21 to 28. This

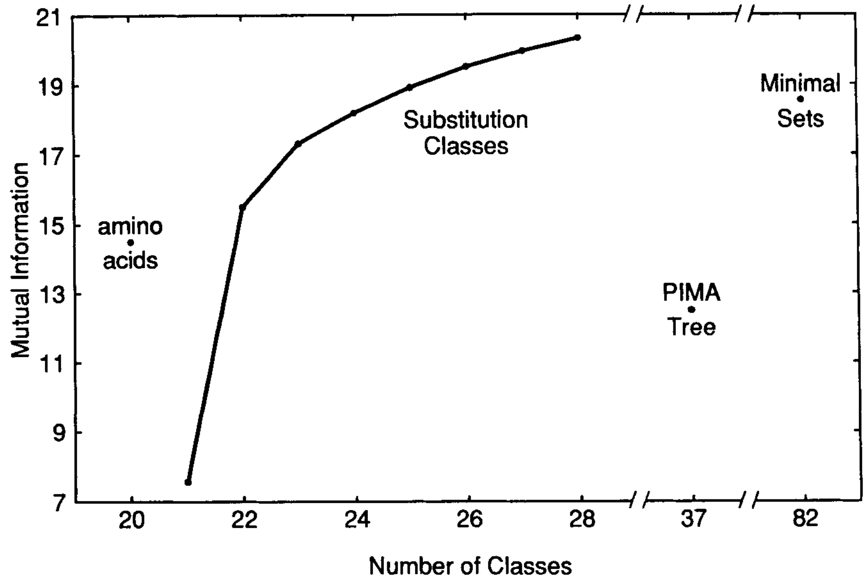

Fig. 2. Curve of mutual information for increasing numbers of substitution classes. In comparison to the value of 20.32 calculated for the optimal 28 substitution classes, we calculated values of 14.51 for the 20 single-residue amino acid classes based on single sequence data, 12.49 for the 37 class-nodes of the PIMA tree of Smith and Smith, and 18.54 for a set of 86 classes which include the "minimal sets" of Taylor. The maximum mutual information attainable for our 111 protein dataset, corresponding to complete memorization of the dataset, was 86.22.

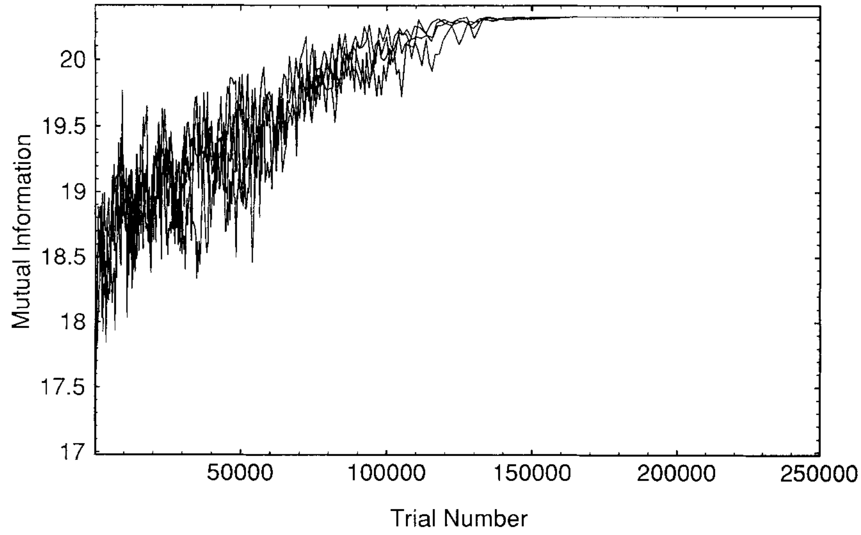

Fig. 3. Five sample searches for 28 optimal residue substitution classes, each seeded with a different random number.

demonstrated that there was effectively no memorization occurring at the resolution of sequence to structure correlations represented by the 28 substitution classes described in this article. In the space of possible sets of substitution classes, there are very many near-optimal sets forming a large, but broad and flat plateau in the mutual information landscape. The small dataset-dependence of the substitution classes discussed above corresponds to small peaks atop that plateau. Whatever dataset dependence exists does not substantially affect the information available when the classes are applied to other datasets.

As marked in Figure 2, the mutual information value for the 20 amino acid residues taken from the sequences of the 111 example proteins only—neglecting information from homologous sequences—was 14.51. The value attained for the optimal set of 28 classes was 20.32—a clear gain in information regarding local protein structure. Another compar-

TABLE II. Structural likelihood Ratios and Residue Memberships for the Optimal Set of 28 Residue Substitution Classes*

Class	Buried				Exposed			
	α	β	*t*	*c*	α	β	*t*	*c*
L	74	40	22	32	−133	−116	−159	−108
I	49	79	−6	22	−91	−180	−166	−122
V	34	97	−23	−10	−184	−11	−140	−139
F	44	62	53	37	−183	−90	−147	−90
M	45	55	39	−10	−106	−94	−102	−20
C	−24	52	40	79	−97	−24	−101	−29
A	69	23	32	0	−34	−108	−109	−114
W	44	72	78	5	−323	−63	−158	−114
Y	21	65	−27	34	−77	29	−106	−95
T	−21	35	52	59	−117	50	−86	−18
S	−9	33	13	68	−73	−41	−21	−28
H	30	22	40	55	−61	−4	−81	−87
Q	15	7	−11	21	−18	30	−67	11
N	0	3	25	63	−73	−61	−10	6
E	17	−25	−10	−27	44	21	−24	−42
D	−22	−2	18	68	−40	−64	−4	15
K	−28	−23	−70	6	50	62	−29	−11
R	21	−16	−17	2	3	63	−71	−4
P	−76	−39	77	89	−128	−41	−2	45
G	−52	1	114	51	−203	−105	37	−37
—IVFMCA —	63	78	−16	5	−160	−75	−181	−127
LIVFMCAWYTSH G—	40	40	24	20	−89	−38	−59	−49
LIVFMCAWYTSHQNE R —	0	−28	0	−20	21	33	−11	9
LIVFMCAWYTSH D RPG—	−30	−47	19	45	−49	−15	27	43
LIVFMC WYTSHQNE KR —	−58	−77	−55	−81	56	90	0	34
LIVFMC WYTSHQN DKRPG—	−131	−126	−20	−35	−10	16	83	69
LIVFMCAWYTSHQNEDKR	−90	−139	−131	−132	108	29	22	4
LIVFMCAWYTSHQNEDKRPG—	−94	−182	−72	−67	54	−14	76	49

*— indicates gap, α indicates helix, β indicates strand, *t* indicates turn, and *c* indicates coil.

ative value was obtained by encoding the PIMA tree of Smith and Smith[38] as a set of 37 classes and calculating the mutual information with the eight local structure categories over the dataset of 111 protein chains. This resulted in a value of 12.49 which is less than that achieved by our set of 22 substitution classes or even that obtained from single sequence information. While the PIMA structure may be useful in the alignment applications suggested by its authors, it is relatively inadequate for structure-based applications. Similarly, we coded the 65 "minimal sets" of amino acids proposed by Taylor to characterize alignment positions and " . . . capture virtually all the useful information that can be extracted from a number of aligned sequences"[39] with 20 single-residue classes for the twenty conserved amino acids and one class with full amino acid and gap membership to account for all those alignment positions which failed to fit one of these 85 classes. The mutual information calculated for this combined set of 86 classes was 18.54—less than that calculated for our set of 25 substitution classes and providing only slightly more information than the 20 amino acid classes which do not include any information from homologous sequences. In comparison to other methods, our classification scheme provides more information with fewer parameters.

We can represent the propensity for various residues and substitution classes to exist in any structural context through the use of log-likelihood ratios. $L(C,2^{\Phi})$, the log-likelihood ratio for class C to be in context 2^{Φ}, is defined by

$$L(C,2^{\Phi}) = 100 \times \ln\left(\frac{p(C,2^{\Phi})}{p(C) \times p(2^{\Phi})}\right). \qquad (5)$$

This ratio represents how much more likely it is for a given position described by a substitution class C to be in local structure 2^{Φ}, compared with what would be expected at random. A listing of these log likelihood ratios and the actual amino acid memberships of the optimal 28 substitution classes can be found in Table II.

DISCUSSION

The first 20 classes listed in Table II are single-residue classes which account for conserved instances of the 20 amino acids. The structural propensities of these 20 classes in the standard four secondary structure states are highly correlated with those calculated in the GOR method[61] (data not shown). Small statistical differences result from our use of different categorizations of local protein structure and a different dataset. Greater disparities can be understood by considering that our classes represent only conserved examples of each of the 20 amino acids while the GOR method makes no such distinction. Thus, while the structural propensities calculated by the GOR method follow general trends in protein biochemistry, the structural propensities calculated for our 20 conserved classes reflect more specific structural or functional constraints.

The broad range of acceptable side chains displayed in the eight multi-residue classes is consistent with conclusions reached in mutagenesis experiments involving multiple structural locations in several proteins.[62,63] Clearly, most of these classes do not display strict conservation of any of the physicochemical properties often used to typify the 20 amino acids, such as side-chain flexibility, bulk, or polarity. In fact, there are no one-to-one correspondences between our multi-residue substitution classes and the nodes of the PIMA tree of Smith and Smith or the "minimal sets" derived from the Venn diagram of Taylor. Examination of the eight multi-residue substitution classes reveals that they are apparently built-up as combinations of a few single amino acid types and "blocks" of multiple amino acids. These "blocks" display a degree of clustering of amino acids with similar physicochemical attributes. For instance, one block which is found in all the multi-residue classes consists of leucine, isoleucine, valine, phenylalanine, methionine, and cysteine. Aside from the presence of cysteine, this block is similar to the "aliphatic.or.large_non-polar" set ({LIVFM}) defined by Taylor.

We also find a block of aromatic and short residues, including tyrosine, tryptophan, histidine, serine, and threonine. In the set-logic nomenclature of Taylor, the simplest description would be polar_aromatic.or.S.or.T. Using a set-logic description based on other common physicochemical characters not included in Taylor's Venn diagrams, it would perhaps be more illuminating to describe this block of residues as hydrogen-bonding_non-charged_non-flexible (ignoring the variable charge of histidine). The pair of residues, glutamine and asparagine, also either appears together or not at all, and could be regarded as the hydrogen-bonding_non-charged-_flexible set—a partial complement to the preceding block. The remaining residues, which appear in various combinations with one another and the blocks of residues just described include the "default"[64] amino acid alanine, the "breakers" of repetitive secondary structure proline and glycine, and the charged residues aspartic acid, glutamic acid, lysine, and arginine.

Due to the broad membership of most of the multi-residue classes, the amino acid members of these classes do not have high probabilities of pairwise mutations amongst themselves as indicated by mutation matrices.[65] Interestingly, there is a general lack of preferential mutations between members of the smaller constitutive blocks as well. While the first block, {LIVFMC}, contains the relatively exchangeable residues {LIVM} with phenylalanine (F) near to belonging to this group, cysteine is rather distinct and has low probabilities of mutation to several of the other members of this block. No significant pattern can be found in the Dayhoff matrix *for* either of the two remaining blocks ({WYTSH} and {QN}).

Consideration of the structural propensities of the eight multi-residue classes indicates that the first two classes are strongly associated with solvent-inaccessible structures while the remaining six are associated with surface structures. This simple distinction, however, is certainly not the limit of information being conveyed by this small number of multi-residue classes. As seen in Figure 2, the point for 21 classes (20 conserved and one variable) corresponds to the point where the basic distinction is made between conserved and variable alignment positions and, hence, between buried and exposed structure locations. The mutual information value at this point is 7.57—not particularly high. With the addition of a few more multi-residue classes, however, the mutual information increases dramatically. This indicates that these classes which represent generalized patterns of amino acid substitutability are accounting for very significant correlations between amino acid substitution patterns and local protein structure. From the curve in Figure 2 we can also conclude that, on the whole, the information provided by patterns of variability as captured by these multi-residue classes is significantly greater than that provided by conserved residue locations.

CONCLUSION

It is unfortunate that biological systems resist the kind of reductionism that would yield "folding rules" specifying exact amino acid residue identities required for producing exact local protein structures. What rules might be set forth are so full of exceptions and caveats that they hardly qualify as rules. Instead we must develop descriptions based on propensities and inclinations. Such probabilistic descriptions are the natural province of information theory. Here, we have used information theory to extract information about protein structure from a dataset of multiple sequence alignments for a structurally representative set of protein chains. This information is represented by a set of amino acid residue substitution classes.

The structurally indicative residue substitution classes obtained by our optimization method do not correspond to the areas of Taylor's Venn Diagrams,[39] the nodes of the PIMA tree of Smith and

Smith,[38] or the clusters of "surface indicating" and "interior indicating" side chains used by Benner and coworkers[32] in any simple and consistent fashion. This is due to the fact that these other groupings of amino acids are heuristic idealizations of the residue substitution data. In contrast, our substitution classes are derived though a rigorous statistical procedure that introduces no bias concerning how the amino acids "should" be grouped. Therefore, while our classes display segregation of the amino acids according to some generic physicochemical characteristics, mostly in the form of the "blocks" of residues as discussed above, this segregation is not particularly strict or defining. Although there is no simple mapping of our structurally-informative classes to the sequence-derived Dirichlet mixture priors of Haussler and coworkers,[40] the results of that statistical approach also display only an emphasis of certain residues or common physicochemical properties.

Our results indicate that conserved alignment positions, on the whole, are less informative with regard to local protein structure than are variable alignment positions. This suggests that any method of sequence profiling of structural motifs should include consideration of alignment positions which display side chain variability. Since the mutual information begins to saturate with a small number of substitution classes, any potential application will benefit from this small number of parameters needed to characterize protein structure as subdivided into our eight structural categories.

We have demonstrated the utility of these classes in the prediction of solvent accessibility.[41] They could also be used in improved methods of multiple sequence alignment, in the sequence profiling of structural patterns, and in improved secondary structure prediction. Such classes of side chain substitution could clearly be made increasingly general through the use of larger sets of proteins. They could also be improved through a self-consistent iterative approach wherein they would be used to construct better alignments from which more optimal classes could be derived.

ACKNOWLEDGMENTS

We thank Mark Saper for helpful guidance, and Kurt Hillig for computational assistance. Peter Wolynes made the original suggestion to use mutual information to optimize structure prediction methods. We extend a general thanks to those who solve protein structures and make this information available and to those who construct and maintain databases of protein sequences and structures. Financial support was provided by the College of Literature, Science, and the Arts, the Program in Protein Structure and Design, the Horace H. Rackham School of Graduate Studies at the University of Michigan, and NIH grant R29 LM05770.

REFERENCES

1. Chothia, C. Principles that determine the structure of proteins. Annu. Rev. Biochem. 53:537–572, 1984.
2. Chothia, C., Lesk, A.M. The relation between the divergence of sequence and structure in proteins. EMBO J. 5:823–826, 1986.
3. Dickerson, R.E., Timkovich, R., Almassy, R.J. The cytochrome fold and the evolution of bacterial energy metabolism. J. Mol. Biol. 100:473–491, 1976.
4. Pastore, A., Lesk, A.M. Comparison of the structures of globins and phycocyanins: Evidence for evolutionary relationship. Proteins 8:133–155, 1990.
5. Kimura, M. Evolutionary rate at the molecular level. Nature 217:624–626, 1968.
6. King, J.L., Jukes, T.H. Non-Darwinian evolution. Science 164:788–798, 1969.
7. Maxfield, F.R., Scheraga, H.A. Improvements in the prediction of protein backbone topography by reduction of statistical errors. Biochem. 18:697, 1979.
8. Sawyer, L., Fothergill-Gilmore, L.A., Russel, G.A. The predicted secondary structure of enolase. Biochem. J. 236: 127–130, 1986.
9. Crawford, I.P., Niermann, T., Kirschner, K. Prediction of secondary structure by evolutionary comparison. Proteins 2:118–129, 1987.
10. Zvelebil, M.J., Barton, G.J., Taylor, W.R., Sternberg, M.J.E. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J. Mol. Biol. 195:957–961, 1987.
11. Perkins, S.J., Haris, P.I., Sim, R.B., Chapman, D. A study of the structure of human complement component factor H by Fourier transform infrared spectroscopy and secondary structure averaging methods. Biochemistry 27:4004–4012, 1988.
12. Levin, J.M., Pascarella, S., Argos, P., Garnier, J. Quantification of secondary structure prediction improvement using multiple alignments. Protein Eng 6:849–854, 1993.
13. Goldstein, R.A., Katzenellenbogen, J.A., Luthey-Schulten, Z.A., Seielstad, D.A., Wolynes, P.G. Three-dimensional model for the hormone binding domains of steroid receptors. Proc. Natl. Acad. Sci., U.S.A. 90:9949–9953, 1993.
14. Taylor, W., Thornton, J. Recognition of super-secondary structure in proteins. J. Mol. Biol. 173:487–514, 1984.
15. Wierenga, R.K., Terpstra, P., Hol, W.G.J. Prediction of the occurrence of the ADP-binding βαβ-fold in proteins, using an amino acid sequence fingerprint. J.Mol. Biol. 187:101–107, 1986.
16. Webster, T.A., Lathrop, R.H., Smith, T.F. Prediction of a common structural domain in aminoacyl-tRNA synthetases through use of a new pattern-directed inference system. Biochemistry 26:6950–6957, 1987.
17. Barton, G.J., Sternberg, M.J.E. Flexible protein sequence patterns: A sensitive method to detect weak structural homologies. J. Mol. Biol. 212:389–402, 1987.
18. Bairoch, A. Prosite: A dictionary of sites and patterns in proteins. Nucleic Acids Res 19:2241–2245, 1991.
19. Pickett, S.D., Saqi, M.A.S., Sternberg, M.J.E. Evaluation of the sequence template method for protein structure prediction. J. Mol. Biol. 228:170–187, 1992.
20. Bork, P. Mobile modules and motifs. Curr. Opin. Struct. Biol. 2:413–421, 1992.
21. Gribskov, M., McLachlan, A.D., Eisenberg, D. Profile analysis: Detection of distantly related proteins. Proc. Natl. Acad. Sci., U.S.A. 84:4355–4358, 1987.
22. Lüthy, R., McLachlan, A.D., Eisenberg, D. Secondary structure-based profiles: Use of structure conserving scoring tables in searching protein sequence databases for structural similarities. Proteins 10:229–239, 1991.
23. Bowie, J.U., Lüthy, R., Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–170, 1991.
24. Wako, H., Blundell, T. Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. I. Solvent accessibility classes. J. Mol. Biol. 238:682–692, 1994.
25. Salamov, A., Solovyev, V. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. J. Mol. Biol. 247:11–15, 1995.

26. Rost, B., Sander, C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 19:55–72, 1994.
27. Rost, B., Sander, C. Conservation and prediction of solvent accessibility in protein families. Proteins 20:216–226, 1994.
28. Benner, S.A., Gerloff, D. Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: The catalytic domain of protein kinases. Adv. Enzyme Regul. 31:121–181, 1991.
29. Rost, B., Sander, C. Jury returns on structure prediction. Nature 360:540, 1992.
30. Barton, G.J., Russel, R.B. Protein structure prediction. Nature 361:505–506, 1993.
31. Robson, B., Garnier, J. Protein structure prediction. Nature 361:506, 1993.
32. Benner, S.A., Badcoe, I., Cohen, M.A., Gerloff, D.L. Bona Fide prediction of aspects of protein conformation. J. Mol. Biol. 235:926–958, 1994.
33. Göbel, U., Sander, C., Schneider, R., Valencia, A. Correlated mutations and residue contacts in proteins. Proteins 18:309–317, 1994.
34. Neher, E. How frequent are correlated changes in families of protein sequences. Proc. Natl. Acad. Sci., U.S.A. 91:98–102, 1994.
35. Shindyalov, I., Kolchanov, N., Sander, C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations. Protein Eng. 7:349–358, 1994.
36. Taylor, W.R., Hatrick, K. Compensating changes in protein multiple sequence alignments. Protein Eng. 7:341–348, 1994.
37. Smith, R.F., Smith, T.F. Automatic generation of primary sequence patterns from sets of related protein sequences. Proc. Natl. Acad. Sci., U.S.A. 87:118–122, 1990.
38. Smith, R.F., Smith, T.F. Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling. Protein Eng. 5:35–41, 1992.
39. Taylor, W. The classification of amino acid conservation. J. Theor. Biol. 119:205–218, 1986.
40. Brown, M., Hughey, R., Krogh, A., Mian, I., Sjölander, K., Haussler, D. Using Dirichlet mixture priors to derive hidden Markov models for protein families. In: "Proceedings of the First International Conference on Intelligent Systems for Molecular Biology." Hunter, L., Searles, D., Shavlik, J. (ed.). Menlo Park, CA: AAAI/MIT Press 1993, 47–55.
41. Thompson, M.J., Goldstein, R.A. Predicting solvent accessibility: Higher accuracy using Bayesian statistics and optimized residue substitution classes. Proteins 25:38–47, 1996.
42. Bairoch, A., Boeckmann, B. The SWISS-PROT protein sequence data bank: Current status. Nucleic Acids Res. 22: 3578–3580, 1994.
43. Cover, T.M., Thomas, J.A. "Elements of Information Theory." New York: John Wiley & Sons, Inc., 1991.
44. Garnier, J., Osguthorpe, D., Robson, B. Analysis of accuracy and implications of simple methods for predicting secondary structure of globular proteins. J. Mol. Biol. 120:97–120, 1978.
45. Schneider, T.D., Stormo, G.D., Gold, L. Information content of binding sites on nucleotide sequences. J. Mol. Biol. 188:415–431, 1986.
46. Gibrat, J.F., Garnier, J., Robson, B. Further developments of protein secondary structure prediction using information theory. J. Mol. Biol. 198:425–443, 1987.
47. Bowie, J.U., Clarke, N.D., Pabo, C.O., Sauer, R.T. Identification of protein folds: Matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. Proteins 7:257–264, 1990.
48. Stephens, R.M., Schneider, T.D. Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. J. Mol. Biol. 228:1124–1136, 1992.
49. Farber, R., Lapedes, A., Sirotkin, K. Determination of eukaryotic protein coding regions using neural networks and information theory. J. Mol. Biol. 226:471–479, 1992.
50. Papp, P.P., Chattoraj, D.K., Schneider, T.D. Information analysis of sequences that bind the replication initiator repA. J. Mol. Biol. 233:219–230, 1993.
51. Williamson, R.M. Information theory analysis of the relationship between primary sequence structure and ligand recognition among a class of facilitated transporters. J. Theor. Biol. 174:179–188, 1995.
52. Shannon, C.E., Weaver, W. "The Mathematical Theory of Communication." Urbana, IL: University of Illinois Press, 1949.
53. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E. Equation of state calculations for fast computing machines. J. Chem. Phys. 21:1087, 1953.
54. Hobohm, U., Sander, C. Enlarged representative set of protein structures. Protein Sci. 3:522–524, 1994.
55. Sander, C., Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9:56–68, 1991.
56. Eisenhaber, F., Persson, B., Argos, P. Protein structure prediction: Recognition of primary, secondary, and tertiary structural features from amino acid sequence. Crit. Rev. Biochem. Mol. Biol. 30:1–94, 1995.
57. Kabsch, W., Sander, C. Dictionary of protein secondary structures. Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
58. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. Protein data bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
59. Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F., Weng, J. Protein data bank. In: "Crystallographic Databases—Information Content, Software Systems, Scientific Applications." Allen, F.H., Bergerhoff, G., Sievers, R. (eds.). Bonn: Data Commission of the International Union of Crystallography, 1987:107–132.
60. Shrake, A., Rupley, J.A. Environment and exposure to solvent of protein atoms: Lysozyme and insulin. J. Mol. Biol. 79:351–371, 1973.
61. Garnier, J., Robson, B. The GOR method for predicting secondary structure in proteins. In: "Prediction of Protein Structure and the Principles of Protein Conformation." Fasman, G.D. (ed.). New York: Plenum Press, 1989:417–465.
62. Reidhaar-Olson, J.F., Sauer, R.T. Combinatorial cassette mutagenesis as a probe of the informational content of protein sequences. Science 241:53–57, 1988.
63. Kamtekar, S., Schiffer, J.M., Xiong, H., Babik, J.M., Hecht, M.H. Protein design by binary patterning of polar and nonpolar amino acids. Science 262:1680–1685, 1993.
64. Richardson, J.S., Richardson, D.C. Principles and patterns of protein conformation. In: "Prediction of Protein Structure and the Principles of Protein Conformation." Fasman, G.D. (ed.). New York: Plenum Press, 1989:1–98.
65. Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C. A model of evolutionary change in proteins. In: "Atlas of Protein Sequence and Structure." Vol. 5, suppl. 3. Dayhoff, M.O. (ed.). Washington, DC: National Biomedical Research Foundation, 1979:353–358.