Discovering Empirically Conserved Amino Acid Substitution Groups in Databases of Protein Families

Thomas D. Wu and Douglas L. Brutlag

Department of Medicine, Section on Medical Informatics,
and Department of Biochemistry
Stanford University Medical Center
Stanford, California 94305
twu@camis.stanford.edu, brutlag@cmgm.stanford.edu

Abstract

This paper introduces a method for identifying empirically conserved amino acid substitution groups. In contrast with existing approaches that view amino acid substitution as a pairwise phenomenon, the method presented here identifies conserved groups of amino acids using a data structure called a conditional distribution matrix. The conditional distribution matrix extends the concept of a pairwise substitution matrix by changing the context of substitution from a single amino acid to a group of amino acids. The matrix tabulates information from a database of protein families that contains numerous aligned positions. Each row in the matrix contains the distribution of amino acids in those aligned positions that contain a given conditioning group of amino acids. The method converts a database of protein families into a conditional distribution matrix and then examines each possible substitution group for evidence of conservation. The algorithm is applied to the BLOCKS and HSSP databases. Twenty amino acid substitution groups are found to be conserved empirically in both databases. These groups provide insight into biochemical properties that are conserved in protein evolution.

Introduction

Protein sequences often exhibit variability in their amino acid patterns, a phenomenon that can be characterized as substitutions of amino acids for one another. Amino acid substitution has been viewed largely as a pairwise phenomenon. Typically, this phenomenon is represented by the frequency that one amino acid replaces another one; the entire set of replacement or substitution frequencies is organized into a symmetric 20 x 20 substitution matrix containing 210 distinct pairwise frequencies. Substitution matrices have been studied in depth (Altschul 1991; Gonnet, Cohen, & Benner 1992; Jones, Taylor, & Thornton 1992; Vogt, Etzold, & Argos 1995), and various matrices have been proposed, including the well-known accepted point mutation (PAM) matrix of Dayhoff and colleagues (1978). In many cases, such substitution matrices have proven quite useful for comparing, aligning, and exploring relationships between pairs of protein sequences.

However, for groupwise or consensus relationships, statistics and methods based on pairwise comparisons are often inadequate. The shift from pairwise comparisons to groupwise analyses is often challenging and non-trivial, as can be seen from the difficulties in trying to align multiple sequences (Barton 1990). Sometimes such problems can be approached using pairwise methods, but often new methods are needed. In computational biology, much attention has focused recently on groupwise or consensus analyses, such as the classification of families and superfamilies of protein sequences and structures, and the compilation of protein family databases, such as PROSITE (Bairoch 1991), BLOCKS (Henikoff & Henikoff 1991), and HSSP (Sander & Schneider 1991)

In light of these advances in computational biology, we present in this paper an empirical analysis of amino acid substitution using a group perspective. We introduce a novel method for identifying groups of amino acids that substitute for one another with high frequency. Our method identifies these substitution groups empirically from a collection of multiple sequence alignments. Although some researchers have also used multiple sequence alignments to study amino acid substitution (Henikoff & Henikoff 1992), their goal has been to derive new substitution matrices, whereas our goal is to identify substitution groups.

Various classifications of amino acids into meaningful and useful groups have been proposed in other studies, as summarized in Table 1. However, previous methods for identifying substitution groups have either analyzed pairwise data or used theoretical principles rather than empirical data. Some researchers have used pairwise substitution matrices to infer substitution groups (Dayhoff, Schwartz, & Orcutt 1978; Miyata, Miyazawa, & Yasunaga 1979). A major problem with such an approach is that substitutability is not necessarily transitive. That is, even if amino acids A and B substitute for each other in some contexts and amino acids B and C substitute for each other in other contexts, we cannot automatically conclude that amino acids A and C substitute for each other. Another problem with pairwise analyses is that they are limited in their ability to distinguish different biochemical contexts

Reference	Substitution groups
Dayhoff et al. 1978	C, FWY, HKR, DENQ, ILMV, AGPST
Miyata et al. 1979	C, FWY, HKR, DENQ, ILMV, AGPST
Jimenez-Montano & Zamora-Cortina 1981	AG, DE, KR, NQ, ST, FWY, ILMV, CFILMVWY, ADEGHKNPQRST
Taylor 1986	Approximately 70 union and intersection combinations of HKR, ILV, ACGS, HFWY, DEHKR, ACDGNPSTV, CDEHKNQRSTWY, ACFGHIKLMTVWY
Smith & Smith 1990	P, AG, DE, NQ, ST, FWY, HKR, ILV, CFILMVWY, DEHKNQRST
Mocz 1995	AEHMQY, FIKLVW, CDGNPRST
Naor et al. 1996	DN, GP, DGNP, EKQR, FILV, DEKNQR, ACFILMVWY, DEGHKNPQRST
Klingler 1996	H, K, N, AP, CF, DE, GS, KR, AGS, ILV, NQR, QTY, HMTWY, CFILMVW

Table 1 Amino acid substitution groups identified in previous studies. Substitution groups are arranged in order of increasing size. Singleton groups, consisting of a single amino acid, are listed when the study indicated that the amino acid was relatively unlikely to substitute for other amino acids.

for substitution. The concept of amino acid substitution inherently requires a context; without a context, we merely have a marginal distribution of amino acid frequencies. Therefore, substitutability essentially consists of rules of the form, "When context X is present, amino acid A substitutes with frequency *f*." With pairwise data, the context X can be specified only as a single amino acid. In contrast, in this paper, we consider more expressive and specific contexts that contain groups of amino acids rather than a single amino acid.

Other researchers have proposed substitution groups on theoretical rather than empirical grounds (Jimenez-Montano & Zamora-Cortina 1981; Kidera et al. 1985; Taylor 1986; Smith RF and Smith TF 1990; Mocz 1995). These theoretical analyses use measurements of various amino acid properties, such as volume, charge, and hydrophobicity, and then propose substitution groups that should be conserved. Unfortunately, theoretical models may not necessarily correspond to the patterns of conservation observed empirically. Moreover, amino acid properties often depend on particular biochemical environments found in protein structures, so properties of amino acid in isolation may not reflect the complexities of amino acid substitutions in particular contexts.

In our approach, we analyze each possible substitution group on its own merits. Hence, substitution groups may overlap or subsume one another. Our approach differs from that of some other researchers, who require that substitution groups do not overlap. For example, Mocz (1995) uses clustering techniques to identify three mutually exclusive clusters of amino acids. The mutual exclusion requirement means that each amino acid can belong only to a single substitution group. Another restriction sometimes placed is that the groups must be organized into a strict hierarchy (Smith RF & Smith TF 1990) or Venn diagram (Taylor 1986). We believe that such requirements are unnecessarily restrictive. Each amino acid has several properties and can serve different functions, depending on the biochemical context. In some contexts, the size of an amino acid may be critical; in others, its charge may be the conserved property. These different contexts will not necessarily fit into a mutually exclusive, strictly hierarchical, or set-theoretical scheme. Therefore, in our approach, we analyze each substitution group separately for empirical evidence of conservation.

In order to analyze each possible substitution group independently, we require more information than is stored in a pairwise substitution matrix. We therefore extend the idea of a substitution matrix to a larger structure called a conditional distribution matrix. This matrix provides the foundation for a consensus-based analysis of amino acid substitution. In the rest of this paper, we discuss two databases of protein families, BLOCKS and HSSP, that make a consensus-based analysis possible. We then present the concept of the conditional distribution matrix and the criteria we use to identify empirically conserved substitution groups. We then present independent analyses of substitution groups conserved empirically in the BLOCKS and HSSP databases. We find that twenty substitution groups are conserved in both databases, and we propose biochemical characteristics underlying those groups. Finally, we discuss various features of our approach and suggest how our results may be used in further work in computational biology.

Methods

Data

Our method requires a source of aligned positions; this data is readily available from databases of protein families or multiple sequence alignments. Two of largest and most widely used protein family databases are the BLOCKS and HSSP databases. Although these databases have distinct characteristics, they can still be viewed as collections of aligned positions.

The BLOCKS database (Henikoff & Henikoff 1991) contains short, highly conserved regions of protein families, represented by ungapped multiple alignments called blocks. Blocks are generated from a set of related

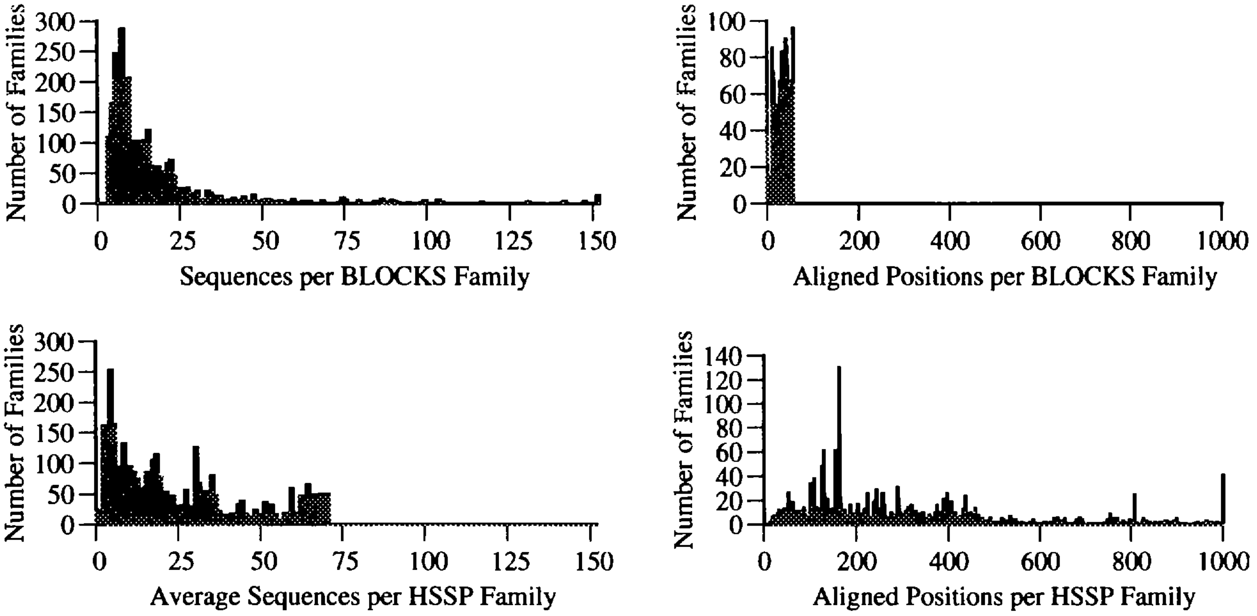

Figure 1 Comparison of BLOCKS and HSSP databases. The histograms show the number of protein families of different sizes in the two databases, where the size of a family may be measured by the number of sequences or aligned positions that it contains. The horizontal scales are selected to be the same dimension to facilitate comparison. The BLOCKS database actually contains some protein families with as many as 507 sequences, and the HSSP database contains some protein families with as many as 1983 aligned positions.

protein sequences. Conserved regions are then found within these sequences using a motif finding program (Smith HO, Annau, & Chandrasegaran 1990), and the edges of these regions are extended until a similarity score falls below some threshold. The similarity score for local alignment and extension is based on the BLOSUM 62 substitution matrix. Finally, a highly scoring set of blocks is selected from all possible conserved regions using an optimal path algorithm. In our study, we used version 8.0 of the BLOCKS database (9 August 1994), which contains 2884 blocks constructed from 770 protein groups in PROSITE version 12.0.

The HSSP (Homology-derived Secondary Structure of Proteins) database (Sander & Schneider 1991) combines structural data from the PDB (Protein Data Bank) database and sequence data from the SWISSPROT database. Each HSSP family corresponds to a PDB structure, and contains all SWISSPROT sequences that are homologous above a certain length-dependent threshold, using the Smith-Waterman alignment algorithm and a substitution matrix by McLachlan (1971). We used the version of HSSP dated 16 November 1995, which contains 3569 protein families.

The BLOCKS and HSSP databases are constructed in quite different ways for different purposes. The BLOCKS database aims to describe sequence homology, whereas the HSSP database aims to describe structural homology, inferred from sequence homology. The BLOCKS database contains families based on optimally scoring multiple sequence alignments of locally conserved regions, whereas HSSP contains families based on pairwise sequence comparisons over global regions. The BLOCKS database does not allow gaps in its alignments, whereas HSSP does. The two databases use different substitution matrices for computing alignments.

These differences produce different types of multiple sequence alignments, as shown in Figure 1. A protein family can be characterized by the number of sequences and aligned positions it contains. In the HSSP database, different sequences may contribute to each aligned position, so the number of sequences should be averaged over all positions. As Figure 1 shows, the number of sequences per family in BLOCKS varies widely, ranging from 2 to 507 (mean = 15.6), whereas in HSSP, it is relatively narrow, ranging from 1 to 70 (mean = 24.9). Conversely, the number of positions per family in BLOCKS ranges only from 4 to 55 (mean = 32.8), whereas in HSSP, it ranges from 12 to 1983 (mean = 266.9). These histograms reflect the stricter requirement that BLOCKS places on each aligned position, requiring it to be conserved across all sequences. On the other hand, HSSP may include an aligned position that is conserved for some but not all homologous sequences.

Conditional Distribution Matrix

Our algorithm consists of two steps. First, we convert a database of aligned positions, such as those in BLOCKS and HSSP, into a large data structure called a **conditional distribution matrix (CDM)**. Then, we look for statistically significant groups of amino acids within this matrix. In this section, we describe the CDM.

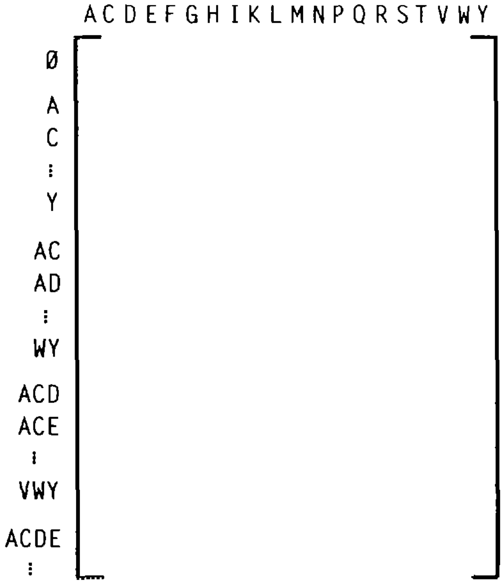

Figure 2 Conditional distribution matrix. Each row corresponds to a conditioning group, which identifies a subset of the aligned positions in the database. The entries in each row contain the distribution of amino acids in those aligned positions.

The conditional distribution matrix can be thought of as an extension of the 20 x 20 pairwise substitution matrix. The pairwise substitution matrix essentially contains different contexts for amino acid substitution, where each context is a single amino acid. For each context a, the pairwise substitution matrix contains the distribution of amino acids that substitute for a. Hence, entries in the pairwise substitution matrix are substitution frequencies $f(a'|a)$ that indicate the likelihood that amino acid a' substitutes in the context of amino acid a. If we think of the context as being on the vertical axis, then each row contains the distribution for a different context.

If we extend the idea of context from a single amino acid to a set of amino acids, we obtain the conditional distribution matrix (Figure 2). The contextual set of amino acids is called **a conditioning group**. Each conditioning group A corresponds to a subset of the aligned positions in the database. To understand the correspondence, suppose that we represent an aligned position as a set P of amino acids. The set P is the union of all amino acids in the aligned position. Then, we say that an aligned position P **satisfies** conditioning group A if $A \subseteq P$. In other words, the aligned position must contain at least one instance of every amino acid in the conditioning group. The aligned position may, of course, contain other amino acids as well, and in fact, it is the distribution of these other amino acids that we are interested in. The conditioning group A provides a context for substitution, whereas the remaining amino acids in $(P-A)$ substitute in that context.

Note that a given aligned position may satisfy several conditioning groups. For instance, consider an aligned position that contains 5 occurrences of valine, 4 occurrences of leucine, and 3 occurrences of isoleucine. Then, the aligned position satisfies the conditioning groups ILV, IL, IV, LV, I, L, V, and the null set $\varnothing$. This multiple correspondence is appropriate because it is difficult to know *a priori* the biochemical context or functionality of an aligned position and hence its substitution pattern. For instance, in this example, the underlying context might be ILV, with no substitutions Or the underlying context might be LV, with isoleucine substituting at a high frequency. Or perhaps the context might simply be V, with leucine and isoleucine both substituting at a high frequency.

Although we may not be able to infer substitution patterns from a single aligned position, sampling over several aligned positions may provide statistically meaningful answers. The conditional distribution matrix accumulates the substitution patterns for various possible contexts over numerous aligned patterns. This matrix, which is of size 2^{20} x 20, has a row for each possible conditioning group A. Each row contains the distribution of substituting amino acids in the context of a certain conditioning group. Each entry in the CDM contains a conditional count $c(a|A)$, which equals the total number of occurrences of amino acid a over all aligned positions that satisfy conditioning group A. We may think of the counting process conceptually as finding all aligned positions that satisfy conditioning group A, and then tabulating all amino acids in those positions. (In practice, however, the CDM is not constructed row by row, but by processing each aligned position in a database sequentially. For each aligned position P, we add the counts of amino acids in P to all rows A, such that $A \subseteq P$.)

Since the number and sizes of aligned positions satisfying each conditioning group varies, we normalize the conditional counts to obtain a conditional frequency:

$$f(a|A) = \frac{c(a|A)}{\sum_{a' \notin A} c(a'|A)}, \quad \text{where } a \notin A.$$

Note that the normalizing value in the denominator excludes amino acids in the conditioning group. This avoids the problem of circularity, whereby the count of an amino acid in the group is elevated simply because the group selects for it. Rather, we are interested in the distribution of amino acids outside the conditioning group.

In order to evaluate these conditional frequencies, we require an expected value for comparison. The expected frequency comes from the marginal distribution of amino acids, that is, the distribution across all aligned positions in the database. In fact, if the null set is a considered a conditioning group, the marginal counts $c(a)$ will be stored in the CDM. The expected conditional frequency derives from the marginal counts as follows:

$$\mu(a|A) = \frac{c(a)}{\sum_{a' \notin A} c(a')}, \quad \text{where } a \notin A.$$

In addition to first-order statistical characteristics, such as the observed and expected frequencies, we may also compute second-order statistical characteristics, such as the standard error of the proportion μ:

$$\hat{\sigma}(a|A) = \sqrt{\frac{\mu(a|A)[1-\mu(a|A)]}{\sum_{a' \in A} c(a'|A)}}$$

Given these first- and second-order quantities, we can evaluate an observed conditional frequency f using the relative deviate or Z-score. The Z-score indicates the number of standard errors $\hat{\sigma}(a|A)$ that an observed frequency $f(a|A)$ differs from the expected frequency $\mu(a|A)$:

$$Z(a|A) = \frac{f(a|A) - \mu(a|A)}{\hat{\sigma}(a|A)}$$

The Z-score indicates whether an amino acid is over- or under-represented in the context of a given conditioning group. If the Z-score is positive, the amino acid is over-represented in that context; and if negative, it is under-represented. We may imagine that each conditioning group induces a frequency distribution on the other amino acids. Amino acids with positive Z-scores are positively induced, whereas those with negative Z-scores are negatively induced. For the purpose of definition, we use three standard errors as a threshold: A Z-score greater than 3 reflects positive induction; a Z-score less than –3, negative induction; and between 3 and –3, a neutral effect.

Criteria for Empirical Conservation

The conditional distribution matrix can be analyzed for evidence that a given substitution group is conserved. In particular, the Z-scores provide a basis for identifying substitution groups that are empirically conserved. Intuitively, we consider a substitution group A to be conserved empirically if amino acids within the group substitute for one another significantly more frequently than amino acids outside the group substitute for amino acids in A.

More formally, we consider a group conserved empirically if it is both compact and isolated. Compactness means that all amino acids in the group substitute for one another frequently, and isolation means that amino acids outside the group do not substitute for those within the group as frequently. We measure substitutability within the group as the Z-score of each amino acid conditioned on other members of the group, or $Z(a|A-\{a\})$. The overall compactness, or **compactness score,** is the minimum of these scores:

$$C(A) = \min_{a \in A} Z(a|A-\{a\})$$

We measure substitutability of amino acids outside the group for those within the group as the Z-score of each amino acid conditioned on the group, or $Z(a|A)$. Because a high score indicates that an amino acid outside the group should belong to the group, we define the **interference score** to be the maximum of these scores:

$$I(A) = \max_{a \notin A} Z(a|A)$$

Finally, we quantify the conservation of a substitution group by the difference between its compactness and interference scores. We call this the **separation score**:

$$S(A) = C(A) - I(A)$$

When a substitution group has a statistically significant separation score, we say that the substitution group is **conserved empirically**. We set the threshold for significance at three standard errors, which is equivalent to a significance level of 0.01. We examined each possible substitution group for a separation score greater than three standard errors.

Because we test a large number of substitution groups independently, one may ask whether the number of tests itself will yield a large number of significant results. Surprisingly, the answer is no. Let the size N be defined as the cardinality of substitution group A, that is, the number of amino acids in A. Consider all possible substitution groups of size N; there are "20 choose N" or $20!/[N!\,(20-N)!]$ such groups to be tested. If Z-scores are distributed randomly, then a group has a positive separation score whenever the Z-scores of the N amino acids in the group are all greater than the Z-scores of the $(20-N)$ outside the group. Hence, the probability of achieving a positive separation score for a group of size N is the permutation of N multiplied by the permutation of $(20-N)$, divided by all permutations of the 20 amino acids. This is simply the reciprocal of the number of groups of size N. Therefore, among all substitution groups of size N, we expect to see one group with a positive separation score by random chance. Hence, we need not make a provision, such as a Bonferroni correction, for the large number of tests.

Examples

To gain a better understanding of our method, we look at some examples. Consider the substitution group I LV; our analysis of this group using the BLOCKS database is shown in Table 2(a). The Z-scores for isoleucine, valine, and leucine are all show high rates of substitution for one another, significantly higher than any other amino acid. The closest amino acid that interferes with this group is methionine. Although methionine is positively induced by I LV, there is a clear separation of 102.4 standard errors between the substitution frequencies of I LV and M, which is highly significant. Hence, I LV is conserved empirically in the BLOCKS database. Note that I LV positively induces the hydrophobic amino acids M, F, T, A, and Y; negatively induces Q, C, S, P, W, N, R, E, D, and G; and has a neutral effect on H. The group also positively induces K slightly, which makes biochemical sense because the long side chain of lysine gives it a partially hydrophobic character.

In contrast with I LV, most substitution groups were not conserved empirically. Table 2(b) shows the analysis for the group GIM, which scored the lowest among all substitution groups of size 3 in the BLOCKS database. GIM scores poorly because it is not compact. Glycine

(a) Substitution group I L V (5328 positions) Separation score: 102.4

I	V	L	M	F	T	A	Y	K	H	Q	C	S	P	W	N	R	E	D	G
219.7	189.8	188.7	86.3	61.9	40.0	19.6	7.4	6.7	–0.9	–6.1	–7.0	–9.6	–15.4	–16.3	–16.6	–21.3	–21.9	–29.2	–51.0

(b) Substitution group G I M (606 positions) Separation score: –119.4

M	I	G	L	V	F	A	K	H	S	Q	C	W	Y	T	N	E	R	P	D
7.7	7.2	–61.9	57.5	32.9	19.9	7.2	0.9	0.3	–5.7	–6.5	–6.9	–11.7	–11.8	–13.3	–13.3	–13.7	–20.6	–25.5	–26.2

(c) Substitution group F I V (2080 positions) Separation score: –108.8

I	V	F	L	M	Y	T	H	A	K	W	C	S	Q	N	D	E	P	R	G
96.4	73.3	31.3	140.1	33.4	24.4	10.8	0.9	–2.9	–4.0	–6.2	–8.8	–14.3	–16.4	–17.3	–25.6	–29.2	–31.4	–33.2	–47.4

Table 2 Analyses of substitution groups I L V, G I M, and F I V. The amino acids in each group are separated from those outside the group by a double bar, and then sorted by Z-score. For each substitution group *A* and amino acid *a*, the values to the left of the double bar are $Z(a|A-\{a\})$, and the values to the right of the double bar are $Z(a|A)$. A single vertical bar separates amino acids that are positively, neutrally, and negatively induced by the substitution group.

substitutes only rarely in those positions with both isoleucine and methionine, as shown by its under-representation of 61.9 standard errors. Hence, G I M is not conserved empirically in the BLOCKS database, for reasons of non-compactness.

Another substitution group, F I V, shown in Table 2(c), failed our criterion because it was not isolated. Although its compactness score is relatively high, meaning that the three amino acids each substitute for one another frequently, its interference score is even higher, because leucine substitutes in this context even more frequently. Therefore, the substitution group F I V is also not conserved empirically, for reasons of non-isolation.

Results

Our analysis of the BLOCKS database yielded 30 substitution groups that are conserved empirically, and our analysis of the HSSP database yielded 51 substitution groups. These substitution groups are listed in Tables 3 and 4, respectively. Twenty substitution groups are conserved empirically in both databases. We feel that the validation of these substitution groups by both databases provides strong evidence that they are indeed conserved in nature. We therefore consider further the biochemical characteristics of these substitution groups.

Of the 190 possible amino acid groups of size 2, nine are conserved empirically in both databases. These amino acid pairs are not evident immediately from substitution matrices. For example, the empirically conserved substitution groups have the following BLOSUM 62 scores: F Y (score of 3), I V (3), S T (1), A S (1), D E (2), K R (2), D N (1), E Q (2), and H Y (2). Conversely, the BLOSUM 62 matrix contains several positively scoring amino acid pairs that were not conserved empirically in our study: L M, I L, and W Y (all with scores of 2), and N S, N H, R Q, E K, K Q, I M, M V, L V, and F W (all with scores of 1). Hence, results from our analysis appear to go beyond substitution matrix data.

The empirically conserved substitution groups are consistent with biochemical intuition. The substitution group F Y is the most significant in both databases. Both phenylalanine and tyrosine have side chains with a single aromatic ring, and have similar volume. The group I V contains amino acids with aliphatic side chains that branch at the beta carbon. The groups D N and E Q are both acid-amide combinations with very similar side chains. In addition, the two acidic amino acids, D E, are conserved empirically. However, the two amides, N Q, do not form a substitution group empirically, even though they might seem to belong together on theoretical grounds. As we shall see later, glutamine tends to cluster more with the long-chain polar amino acids, such as lysine and arginine. The basic amino acids, K R, are conserved empirically, and both have amino groups. The group S T contains amino acids that have short hydroxy side chains. Serine is also conserved empirically with alanine (A S); both amino acids are small, each containing a single carbon atom in its side chain. Nevertheless, the smallest amino acid, glycine, does not form a group with serine or alanine, perhaps because glycine has many distinctive properties. Finally, the group H Y is conserved empirically in both databases. Both amino acids have polar ring structures, so the combination of similar volume and polarity appears to account for their conservation.

For amino acid groups of size 3, the two databases identified six empirically conserved substitution groups in common. The highest scoring amino acid triplet in both studies was I L V. All three amino acids in this group have branched aliphatic side chains. As we noted previously, isoleucine and valine form a substitution group themselves. One explanation may be that isoleucine and valine are both branched at their beta carbons, whereas leucine is branched at its gamma carbon. Apparently, branch position matters in some biochemical environments, but not in others. Another conserved amino acid triplet is F W Y. All three amino acids in this group have aromatic side chains, although tryptophan has a double ring. Since phenylalanine and tyrosine themselves form a substitution

Substitution Group	Pos	C(A)	I(A)	Sep	Pos-induced	Neutral	Neg-induced
•FY	3735	183.6	74.0	109.6	LWHIVMK		TSCRQNEPADG
•DE	5980	153.0	70.0	83.0	KQNSHTAR		PGMWLYCVFI
•KR	6453	157.3	93.0	74.3	QEHNSTD		PAMYLWFVCGI
•IV	10192	232.2	188.7	43.5	LMTFA		YCKHQWSNPERDG
•ST	7017	105.1	62.3	42.8	ANKQED	HVP	MCRYFWILG
•AS	8304	91.3	70.3	21.0	TKNQEGD	PH	VCMRWFYLI
•DN	4435	102.5	87.0	15.5	EKSQHTG	R	PAYMWFCVLI
•HY	1728	57.3	43.3	14.0	FKRQNLW	EV	SDMTCAIPG
•EQ	4856	104.3	98.9	5.4	KDHRNST	AP	MWYLFCVGI
•ILV	5328	188.7	86.3	102.4	MFTAYK	H	QCSPWNREDG
•FLY	1474	74.0	33.9	40.1	IVHMKWT	SR	QCENAPDG
•EKQ	2411	85.3	49.4	35.9	RDHNTSPA		MWLYFGCVI
•AST	3293	62.4	30.7	31.7	KNEQVH	DPM	CFRLWYIG
•KQR	2404	83.0	65.0	18.0	EHSNTDP		AYMLVFWGCI
FHY	748	40.4	28.0	12.4	LKRQWV	NSM	TECIADPG
•FWY	527	49.7	37.5	12.2	LHM	EKQIRV	DTCNASPG
•ILMV	1696	86.3	58.8	27.5	FTAYKH	Q	CSWNREPDG
FILV	1502	61.9	48.1	13.8	MYTAKH	CS	WNQDERPG
•EKQR	1277	49.4	37.4	12.0	HSDTNPA		LYFMVWGCI
DEKQ	1173	46.1	35.6	10.4	NHRSTPA		WLGYFMVCI
HKQR	746	41.4	31.2	10.2	ESNTDP	YA	LGVFMWIC
FHLY	372	28.0	24.3	3.7	KRVQNM	IWETS	CADPG
•FILMV	688	48.1	26.4	21.7	YATHK	WSCQ	NERDGP
•FILVY	509	32.0	22.0	10.0	TMHKSW	EC	AQNRDPG
EHKQR	484	31.2	25.7	5.5	SNDTPA	Y	LFMGVWIC
DEHKNQ	305	30.1	23.9	6.2	STRPA	GY	FMCVWIL
EHKQRS	351	25.7	20.5	5.2	NTDPAY		FGMVLWIC
•FILMVY	260	21.2	18.1	3.1	THKWS	EANQC	RPGD
ADEGHKNPQRST	63	19.4	14.6	4.8	Y	VMIL	FWC
ADEGHKNPQRSTY	41	6.3	2.8	3.5		LIMVF	WC

Table 3 Substitution groups conserved empirically in the BLOCKS database. Groups are arranged according to their size and sorted by separation score. Groups conserved empirically in both BLOCKS and HSSP databases are marked with a bullet. For each group, the table lists the number of positions in the database satisfying the group; the compactness, interference, and separation scores; and their effect on amino acids outside the group, listed in order of descending Z-score.

Substitution Group	Pos	C(A)	I(A)	Sep	Pos-Induced	Neutral	Neg-Induced
•FY	45082	654.9	353.7	301.2	WHLMIV	NR	SQTKCAPEDG
•IV	107601	933.9	731.4	202.5	LMTFA		YWQRECKHSPNDG
•DE	92101	576.0	381.6	194.4	NQKSAPTGR		HMWYCFVIL
•ST	116853	474.4	290.1	106.3	ANKQEPDR		HMCVGYWIFL
•DN	84385	518.6	412.3	106.3	ESQKGTHR		APYMWCFIVL
•HY	26417	269.0	172.2	96.8	NFWRQKS		MDETLAPVCIG
•KR	88140	502.6	411.6	91.0	QENSTHD		PAMYGWCVFIL
•AS	121583	335.9	290.7	45.2	TPENKQDGR		CMHYVWFIL
•EQ	79815	449.6	443.4	6.1	KDNRSTAH		PMYGWCFVIL
•ILV	62336	731.4	427.2	304.2	MFATY	W	QREKHCSPNDG
•FWY	7941	353.7	130.3	223.4	LHMRISV	Q	NTCKEADPG
•EKQ	49305	433.9	267.3	166.6	RNDSTAHP		MYWGCVFIL
•AST	61564	290.1	151.6	138.5	NEQKPDR		VCMGHYWIFL
DEN	46671	381.6	272.8	98.8	SQKGATRP	H	MYWCFVIL
DNS	50077	340.4	279.8	60.6	EKQTGARPH		YMWCFIVL
•KQR	43053	372.4	315.9	56.5	ENSTHDA		MPYWGCVFLI
NST	49861	260.2	216.5	43.7	DKEQRAP	H	GYMCWVFIL
•FLY	19446	184.7	173.4	11.3	IWMHVR	NS	QTAKCEPDG
APS	30073	175.5	165.0	10.5	TDEKQNR		GHCMYWVFIL
•ILMV	19946	427.2	179.6	247.6	FATYRQ	W	KCHESNPDG
FLWY	3562	130.3	61.9	68.4	HMIRVQT	S	NKCDAEPG
•EKQR	27456	267.3	201.4	65.9	NSDTAH		PMYGWVCFLI
DENS	30264	279.8	216.9	62.9	KQTAGRP		HYMWCFVIL
FHWY	2066	107.9	56.0	51.9	LNRMQS	I	CETKVDAGP
DENQ	25113	278.9	244.5	34.4	KSRTAHGP		MYWCFVIL
APST	17550	129.1	109.7	19.4	EQNKDR		HGMYVCWFIL
AITV	18933	88.6	81.6	7.0	LSMEQKR		YNFHCPWDG
•FILMV	7079	179.6	124.0	55.6	YTRQWAH	S	KNECPDG
•FILVY	6479	180.3	134.5	45.8	MWTHARQS		CNEKPDG
FILMY	3316	143.1	114.4	28.7	VWRQTH		ASNCKEPDG
ILMTV	7248	77.2	56.2	21.0	QRFAKYS	E	WNHCPDG
FHLWY	998	56.0	35.5	20.5	RMIQNV	D	STECGAKP
DEKNQ	17742	237.1	222.4	14.7	SRATHP		GMYWCVFIL
AILMV	7901	86.4	74.4	12.0	TFRQYE	K	SCPWHNDG
DEKNQS	13239	222.1	123.8	98.3	RTAPGH		YMWVCFLI
•FILMVY	2366	114.5	39.5	75.0	WTQRHAS		NCKEPDG
EKNQRS	11898	167.3	157.8	9.5	DTAHG		PYMWVFCLI
FILVWY	1286	68.4	59.2	9.2	MQRHT	AS	NECKPDG
DEKNQRS	8083	123.8	98.2	25.6	TAGHP		YMWVFLCI
DEKNQST	8238	121.2	113.4	7.8	ARPGH		YMVWCLFI
ADEKNQST	5747	112.3	98.9	13.4	RPGH		VYMWFICL
ADEKNQRS	5413	96.2	90.8	5.4	TPHGY		MVWFLCI
FHILMNWY	102	13.2	8.5	4.7	VQG	DCPRTA	EKS
ADEKNQRST	3601	90.8	46.8	44.0	PHGY	MV	FWILC
DFGHILMWY	74	22.6	9.2	13.4	N	SVTAERPK	CQ
ADEKNPQRST	1283	46.8	18.8	28.0	HGVL		MIYFWC
DFGHILMNWY	52	9.2	2.5	6.7		TAVPESQCRK	
ADEHKNPQRST	447	18.8	8.1	10.7	VLY	GMI	FWC
ADEHIKLNPQRSTV	104	9.1	5.7	3.4	Y	FGM	WC
AEFIKLMNPQRSTVY	42	7.8	2.0	5.8		IGDWC	
ADEFGHMNPQRSTVWY	28	9.4	1.3	8.1		LCIK	

Table 4 Substitution groups conserved empirically in the HSSP database. For explanation, see caption for Table 3.

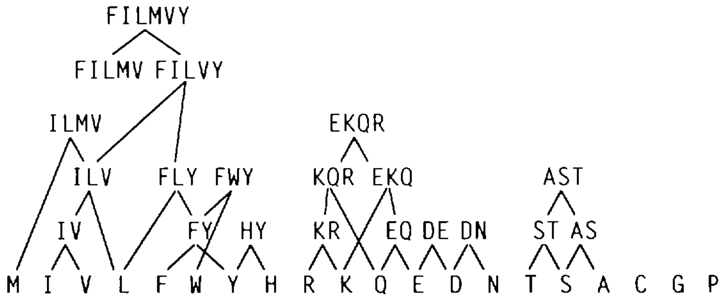

Figure 3 Classification of empirically conserved amino acid substitution groups. This classification contains all substitution groups that are conserved empirically in the BLOCKS and HSSP databases. The substitution groups are linked by subsumption relationships. Cysteine, glycine, and proline do not belong to any substitution group.

group, it appears that single-ring aromatic side chains are conserved in some contexts, but in other contexts, aromaticity itself is conserved. A closely related amino acid triplet that is also conserved empirically is FLY. The biochemical similarity for this group appears to be volume. Both phenylalanine and tyrosine have a bulky aromatic group at their gamma carbons, whereas leucine has a branched methyl group there. Perhaps in some environments, the branched methyl group provides enough volume and hydrophobicity to substitute for the aromatic ring. The amino acid triplet of AST contains amino acids that have short side chains, with either one or two carbons. The remaining amino acid triplets, EKQ and KQR, contain amino acids with relatively long polar side chains. The biochemical basis for conserving both triplets is not immediately clear. The two groups both contain lysine and glutamine, but one triplet has glutamate and the other has arginine. Arginine can donate a hydrogen bond, whereas glutamate cannot. In addition, the side chain of arginine is much larger than that of glutamate. Perhaps amino acids with long polar side chains and hydrogen bond donor capability (KQR) are conserved in different biochemical environments than amino acids with medium-length polar side chains (EKQ).

The two databases identified only two empirically conserved substitution groups of size 4 in common. One group, ILMV, is a well-recognized group of small hydrophobic amino acids. The other group, EKQR, contains amino acids with long polar side chains. This group subsumes the triplets EKQ and KQR discussed previously. These amino acids have been observed to participate in salt bridges on the surfaces of proteins, which help stabilize protein structure (Goldman 1995).

For substitution groups of size 5, the two databases identified two empirically conserved groups in common. One group, FILMV, contains what are referred to as the major hydrophobic amino acids. The other group, FILVY, demonstrates that tyrosine sometimes acts as a hydrophobic amino acid.

Finally, both databases identified the six-member amino acid group FILMVY as being conserved empirically. This group is a combination of the two substitution groups of size 5 and contains amino acids with hydrophobic characteristics.

For the larger substitution groups, the two databases correlated less well than for smaller substitution groups. Conservation of large substitution groups is difficult to identify because few aligned positions in the BLOCKS and HSSP databases satisfy large conditioning groups. In fact, some of these findings are based on fewer than 100 aligned positions and may not be reliable. Moreover, aligned positions that do satisfy a large conditioning group must contain many sequences and many different amino acids, meaning that the position may not be conserved well. In addition, the large number of sequences means that a few protein families could bias the results. As databases grow larger, we might expect to obtain more accurate results for large groups.

Nevertheless, the two databases showed a near-match with the groups ADEGHKNPQRST in BLOCKS and ADEHKNPQRST in HSSP. The two groups differ on whether glycine belongs, with BLOCKS including glycine and HSSP excluding it. Furthermore, in these larger substitution groups, revealing insights can be obtained from examining the amino acids that they induce negatively. For instance, both databases identify substitution groups that negatively induce the hydrophobic amino acids. These substitution groups differ slightly between the databases, perhaps reflecting the heterogeneity of hydrophilic environments. In addition, both databases identify substitution groups that negatively induce tryptophan, cysteine, and sometimes phenylalanine. These large substitution groups might therefore be defined in a negative sense, by specifying the absence of certain amino acids.

Our set of empirically conserved substitution groups is generally different from those found in the literature, although some similarities can be found. Of the 20 groups found in both BLOCKS and HSSP, eight have been proposed explicitly in previous analyses: DE, KR, ST, DN, ILV, FWY, ILMV, and EKQR. In addition, previous analyses have proposed the groups FILV and ADEGHKNPQRST, which are conserved empirically only in BLOCKS, and the group DENQ, which is conserved empirically only in HSSP.

The twenty substitution groups conserved empirically in both databases can be organized into a classification hierarchy, as shown in Figure 3. In this hierarchy, the amino acids are divided into three major classes. One class, MIVLFWYH, contains hydrophobic amino acids; another class, RKQEDN, contains charged or polar amino acids; and the third class, AST, contains small amino acids. In addition, three amino acids—cysteine, glycine, and proline—do not belong to any substitution group. These amino acids have unique properties that cannot be easily fulfilled by other amino acids. Cysteine can form disulfide bridges. Glycine is the smallest amino acid, having only a hydrogen atom for its side chain. And proline has a distinctive cyclical side chain that causes it to form bends in helices and strands.

Discussion

Because our approach to amino acid substitution is empirical, it enjoys the same advantages and suffers the same limitations as all empirical studies. One feature of our approach, which could be viewed as either an advantage or limitation, is that our analysis is general. Although our analysis conditions on specific groups that represent specific biochemical contexts, the substitution groups in our study are nevertheless conserved empirically across an entire database. In contrast, many models for describing conservation, such as motifs (Bairoch 1991), profiles (Gribskov, McLachlan, & Eisenberg 1987), and hidden Markov models (Krogh et al. 1994), characterize specific protein families. In those models, each protein family has its own pattern of conservation. The issue is whether generalized biochemical contexts exist and whether selecting aligned positions across an entire database adequately specifies a single biochemical context.

Our opinion is that general patterns of conservation do exist, and that understanding them is critical to understanding specific protein families. We believe that nature is likely to use the same patterns over and over. A strategy of finding general patterns of conservation may be especially fruitful because most protein families are relatively small and the biochemical context of each position is not known. By drawing upon a large amount of data, a general approach is more likely to minimize statistical noise and extract meaningful signals.

Nevertheless, we acknowledge that specific patterns of substitution may occur only in specific protein families. Unfortunately, selecting those protein families may be problematic. An intermediate approach based on secondary structure might prove fruitful, since alpha-helices and beta-strands likely exist in different biochemical environments. Data sets of alpha-helices and beta-strands may generate different sets of substitution groups, which might otherwise be obscured in the entire protein family database. Another strategy would be to analyze large families of proteins with similar function, such as the globins or kinases. In future work, we plan to apply our method to such specialized data sets.

Empirical studies, such as this one, must also consider the issue of sampling bias. Protein families often contain sequences closely related protein sequences that are over-represented, and distantly related sequences that are under-represented. Several methods have been proposed to weight the sequences to remove this bias (Altschul, Carroll, & Lipman 1989; Sibbald & Argos 1990; Vingron & Sibbald 1993). We believe that such weighting methods might help to strengthen our results, and we intend to study the effect of sequence weighting on our analysis of amino acid conservation.

Another characteristic of our work is that our analysis has been biased towards regions that are highly conserved. In particular, the BLOCKS database contains regions with relatively low rates of mutability. Our analysis therefore might miss acceptable substitutions that occur in poorly conserved, highly variable regions. However, our goal has not been to find acceptable substitution groups that are weakly conserved, but rather to identify groups that are strongly conserved and have empirical evidence to support them. We feel that it is these groups that give the best insight into the biochemical principles that are important in protein structure and function.

We anticipate that a set of empirically conserved substitution groups may find several potential applications. First, such substitution groups may provide the basis for new methods for aligning multiple sequences. Most existing methods for aligning multiple sequences rely upon pairwise substitution frequencies. However, substitution groups may provide a more appropriate model for groupwise or consensus relationships. Second, substitution groups might provide an alphabet to describe discrete protein motifs. Most discrete motifs, such as those in PROSITE, are constructed manually, although automated methods have been developed recently (Wu & Brutlag 1995). Both automated and manual methods for building discrete motifs would benefit from having a set of standardized substitution groups. Analogously, our analysis might even provide a basis for probabilistic motifs, such as hidden Markov models (Krogh et al. 1994). The conditional distribution matrix essentially contains probabilistic amino acid profiles for various biochemical contexts. Selected amino acid distributions from the matrix could serve as canonical distributions for hidden Markov models. These distributions might provide general, idealized models of amino acid substitution instead of the empirically tailored distributions based on a specific protein family that are currently used. Recent work on Dirichlet mixture priors (Brown et al. 1993) also tries to find idealized amino acid distributions. Finally, because substitution groups attempt to capture important amino acid properties, they might be helpful in predicting the secondary and tertiary structure of protein sequences. Many researchers have tried to generalize amino acid sequences in terms of their properties (Bork 1989); substitution groups may provide insight into properties that are conserved empirically.

Aside from these applications, though, we hope that our study leads to an improved understanding of amino acid substitution. Amino acid substitution is a central principle in molecular biology. Improved knowledge about amino acid substitution may ultimately lead to better understanding of protein structure and function. Patterns of amino acid substitution represent static evidence of the dynamic process of amino acid evolution and conservation. Findings such as those in this study are central to our understanding of protein structure, function, and evolution.

Acknowledgments

This work was supported in part by grants LM 05716 and LM 07033 from the National Library of Medicine. The authors thank Tod Klingler for valuable discussions.

References

Altschul, S. F. 1991. Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol* 219:555–565.

Altschul, S. F., Carroll, R. J., and Lipman, D. J. 1989. Weights for data related by a tree. *J Mol Biol* 207:647–653.

Bairoch, A. 1991. PROSITE: A dictionary of sites and patterns in proteins. *Nucl Acids Res* 19:2241–2245.

Barton, G. J. 1990. Protein multiple sequence alignment and flexible pattern matching. *Meth Enzymol* 183:403–427.

Bork, P. 1989. Recognition of functional regions in primary structures using a set of property patterns. *Febs Letters* 257:191–195.

Brown, M., Hughey, R., Krogh, A., Mian, I. S., Sjölander, K., and Haussler, D. 1993. 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, 47–55. Menlo Park, Calif.: AAAI Press.

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. 1978. A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Function*, Nat. Biomed. Research Foundation, pages 345–352.

Goldman, A. 1995. How to make my blood boil. *Structure* 3:1277–1279.

Gonnet, G. H., Cohen, M. A., and Benner, S. A. 1992. Exhaustive matching of the entire protein sequence database. *Science* 256:1443–1445.

Gribskov, M., McLachlan, A. D., and Eisenberg, D. 1987. Profile analysis: Detection of distantly related proteins. *Proc Natl Acad Sci USA* 84:4355–4358.

Henikoff, S. and Henikoff, J. G. 1991. Automated assembly of protein blocks for database searching. *Nucl Acids Res* 19:6565–6572.

Henikoff, S. and Henikoff, J. G. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 89:10915–10919.

Jimenez-Montano, M. A., and Zamora-Cortina, L. 1981. Evolutionary model for the generation of amino acid sequences and its application to the study of mammal alpha-hemoglobin chains. In Proceedings of the Seventh International Biophysics Congress, Mexico City.

Jones, D. T., Taylor, W. R., and Thornton, J. M. 1992. The rapid generation of mutation data matrices from protein sequences. *Comp Appl Biosci* 8:275–282.

Kidera, A., Yonishi, Y., Masahito, O., Ooi, T., and Scheraga, H. A. 1985. Statistical analysis of the physical properties of the twenty naturally occurring amino acids. *J Prot Chem* 4:23–55.

Krogh, A., Brown, M., Mian, I. S., Sjölander, K., and Haussler, D. 1994. Hidden Markov models in computational biology: Applications to protein modeling. *J Mol Biol* 235:1501–1531.

Klingler, T. M. 1996. Structural inference from correlations in biological sequences. Ph.D. diss., Program in Medical Informatics, Stanford Univ.

McLachlan, A. D. 1971. Tests for comparing related amino acid sequences. *J Mol Biol* 61:409–424.

Miyata, T., Miyazawa, S., and Yasunaga, T. 1979. Two types of amino acid substitution in protein evolution. *J Mol Evol* 12:219–236.

Mocz, G. 1995. Fuzzy cluster analysis of simple physicochemical properties of amino acids for recognizing secondary structure in proteins. *Protein Sci* 4:1178–1187.

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

Sibbald, P. R., and Argos, P. 1990. Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *J Mol Biol* 216:813–818.

Smith, H. O., Annau, T. M., and Chandrasegaran, S. 1990. Finding sequence motifs in groups of functionally related proteins. *Proc Natl Acad Sci USA* 87:826–830.

Smith, R. F. and Smith, T. F. Automatic generation of primary sequence patterns from sets of related protein sequences. *Proc Natl Acad Sci USA* 87:118–122, 1990.

Taylor, W. R. The classification of amino acid conservation. *J Theor Biol* 119:205–218, 1986.

Vingron, M., and Sibbald, P. R. 1993. Weighting in sequence space: a comparison of methods in terms of generalized sequences. *Proc Natl Acad Sci USA* 90:8777–8781.

Vogt, G., Etzold, T., and Argos, P. An assessment of amino acid exchange matrices in aligning protein sequences: The twilight zone revisited. *J Mol Biol* 249:816–831, 1995.

Wu, T. D., and Brutlag, D. L. 1995. Identification of protein motifs using conserved amino acid properties and partitioning techniques. In Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology, 402–410. Menlo Park, Calif.: AAAI Press.