USING THE RADIAL DISTRIBUTIONS OF PHYSICAL FEATURES TO COMPARE AMINO ACID ENVIRONMENTS AND ALIGN AMINO ACID SEQUENCES

LIPING WEI, RUSS B. ALTMAN†, JEFFREY T. CHANG
Section on Medical Informatics
Stanford University, MSOB X215
Stanford, CA 94305-5479
{wei, altman, jchang}@smi.stanford.edu
415-725-3394
†Corresponding author.

We have performed a comprehensive analysis of the microenvironments surrounding the twenty amino acids. Our analysis includes comparison of amino acid environments with random control environments as well as with each of the other amino acid environments. We describe the amino acid environments with a set of 21 features summarizing atomic, chemical group, residue, and secondary structural features. The environments are divided into radial shells of 1 Å thickness to represent the distance of the features from the amino acid C_β atoms. We make the results of our analysis available graphically over the world wide web. To illustrate the validity and utility of our analysis, we used the amino acid comparative profiles to construct a substitution matrix, the WAC matrix, based on a simple summary of the computed environmental differences. We compared our matrix to BLOSUM62 and PAM250 in BLAST searches with query sequences selected from 39 protein families found in the PROSITE database. Although BLOSUM62 was the most sensitive matrix overall, our matrix was more sensitive for some families, and exhibited overall performance similar to PAM250. Our results suggest that the radial distribution of biochemical and biophysical features is useful for comparing amino acid environments, and that similarity matrices based on the geometric distribution of features around amino acids may produce improved search sensitivity.

Introduction

The twenty amino acids provide a useful basis set of chemical functionality for constructing protein macromolecules. Small microenvironments within proteins derive their important structural and functional characteristics from the physical interactions between the amino acids. In large part, the tolerance of protein macromolecules to mutation depends on the critical features of the lost amino acid, and the ability of the new amino acid to replace these features. Miyata et al. have shown that in proteins that have preserved their structure throughout the process of evolution, the frequency of amino acid substitution is correlated with the physico-chemical similarities between the exchanged amino acids (Miyata *et al.*, 1979). A mutation in which an amino acid is replaced by one with similar physico-chemical properties is more likely to be accepted than one in which the new environment disrupts the protein's 3D conformation. Thus, similarity matrices derived from empirical substitution frequencies often serve to confirm the physical similarity of two amino acids by showing a high frequency of substitution in protein sequences.

The tolerance of a microenvironment for one amino acid or another, however, is not directly a function of the properties of the lost amino acid (it is, after all, lost!), but is rather a direct function of the properties of the environment surrounding the lost amino acid. If a new amino acid is compatible with the environment, then it will be accepted as a replacement, and the resulting energy of interaction will be acceptable. If the new amino acid is incompatible with the environment, then the unfavorable interaction energies may significantly destabilize the protein or render it inactive. Thus, it is useful to examine the environments of amino acids in order to understand the likelihood that various substitutions will be acceptable. This can be done for single amino acid environments using molecular dynamics, in order to assess the suitability of a mutation (Grantham, 1974; Jones, 1975). It is also useful, however, to characterize the environments for multiple superimposed instances of each amino acid in order to uncover regularities and patterns that may only be discernible from averaged ensembles.

We have previously developed a general purpose system, FEATURE, for representing and characterizing protein microenvironments using a broad set of properties (Bagley & Altman, 1995). The FEATURE system takes a set of aligned sites that share some structural or functional significance, and a set of control sites without such significance (nonsites). The system creates shells of 1 Å thickness out to 10 Å and compares the distribution of values for a variety of properties (shown in Figure 1) in the shells of the sites with those of the nonsites. Significant properties within each shell are determined using the Wilcoxon Rank Sum test (Ott, 1992). The FEATURE system thus provides information about the distance of various features from a central point in the site. It is therefore somewhat different from other feature-based systems in which the relative geometry of features is not represented (Grantham, 1974; Jones, 1975). We have shown that FEATURE can also divide an environment into small three-dimensional cubes in its analysis, although at a cost of decreased statistical significance of the findings (Bagley et al, 1995). In this paper, we use the radial system to study the microenvironments surrounding ~100 instances of each of the twenty amino acids. We describe the properties of each amino acid against a background of random amino acids. We also describe all amino acids' specific properties compared to each other, thus illustrating the physico-chemical and structural differences between the environments of pairs of amino acids.

In order to illustrate one possible application of our analysis, we have used a straightforward interpretation of our amino acid comparison data to generate a new similarity matrix, WAC, for sequence comparison. We compared the performance of WAC to that of BLOSUM62 (Henikoff & Henikoff, 1992) and PAM250 (Dayhoff *et al*., 1978), using the BLAST search program (Altschul *et al*., 1990). Given a query sequence and a matrix, BLAST uses the matrix to extract protein sequences, from a database, that contain high local alignment scores with the query. We also compared WAC with Rao's matrix, which is based on physico-

chemical properties of the amino acids themselves, instead of the environments around them (Rao, 1986).

Methods

Analysis of Amino Acid environments

In these experiments, we compared the average environments of each amino acid against a random set of background amino acids, as well as against the environments of the other amino acids. The original system runs in Common Lisp, but was ported to Microsoft Visual C++ v4.0 running on a Pentium Pro/200 in order to accommodate the larger data sets required for these experiments. The amino acid data were taken from a set of 20 proteins from the PDB (1ACX, 1AVR, 1F3G, 1FKF, 1GAL, 1IPD, 1PAF, 1PDA, 1PYP, 1REC, 1RHD, 1SNC, 1TEN, 2ACT, 2BP2, 2BPA, 2HIP, 3GLY, 4TMS, 9INS; Bernstein *et al.*, 1977) with resolution finer than 3Å, with no significant structural similarities (Holm & Sander, 1994), and with no homologous sequence or function. The average sequence length of the proteins in this set was 268.

Property based on:	Property Name
Atom	Atom Name
	Hydrophobicity
	Charge
	Positive Charge
	Negative Charge
	Charge with HIS
Chemical group	Hydroxyl
	Amide
	Amine
	Carbonyl
	Ring-system
	Peptide
Residue	Residue type
	Hydrophobicity Classification 1
	Hydrophobicity Classification 2
Secondary structure	Secondary Structure Classification 1
	Secondary Structure Classification 2
Other Properties	Van Der Waal volume
	B-factor
	Mobility

Figure 1 List of microenvironment properties used by FEATURE.

For each amino acid, 5 random instances were chosen from each protein. For proteins with less than 5 instances of an amino acid, all instances were used, thus

slightly less than 100 sites were sampled. Each amino acid site was centered on the C_β atom to maximize the observable effects of the side chain while still maintaining a comparable site across all 20 amino acids. The C_β atom position of Glycine was estimated based on the average position of the superimposed C_β atoms from all other amino acids. The ~100 sites of each amino acid were compared to the ~100 sites of each of the other amino acids, as well as 100 nonsites chosen randomly from the other amino acids.

From each comparison, data for 21 properties describing the biochemical and biophysical milieu were collected radially, in 10 concentric 1Å shells. Each observed feature was assigned to a volume shell depending on its distance from the site center. For example, a negative charge observed at 5.7Å would be assigned to shell 6, describing the features at 5-6Å. Since the property values do not necessarily have a normal distribution, a non-parametric test, the Wilcoxon Rank Sum test, was used to determine the statistical significance of the differences between the observed values of sites and nonsites for a property at a given radius (Ott, 1992). Properties and associated radii are reported for which the statistical significance exceeds certain threshold ($p=0.005$ in our experiments). The results are available on the WWW at *http://smi.stanford.edu/projects/helix/pubs/wacpsb/* On this page, a matrix of amino acids is presented, and selection of a cell within the matrix produces a graphical comparison of the amino acids.

Generating A Similarity Matrix

The Wilcoxon test yields a t-value that describes the difference between sites and nonsites for a property at a volume. The overall difference between two amino acids can be calculated by taking the sum of the squared t-values for each property and volume. This yields a single number that represents comparatively the differences between the features found in the amino acid pair. Some properties are slight variations of others, and to avoid double-counting, were not used in the analysis. Thus, for example, we have two different classifications of secondary structure (that divide secondary structural space slightly differently) and we only used one of them. Similarly, we used only a single hydrophobicity classification.

The summed differences in the amino acid pairs were used to create a 24 by 24 similarity matrix, the WAC matrix. The matrix was scaled with respect to BLOSUM62 using a linear least-squares regression. The values for the B amino acid were calculated by taking the frequency-weighted average of the D and R values. The Z and X values were calculated similarly, using the E, Q values and all values, respectively. The * entries were filled with the lowest values found in the entire matrix. The frequencies of the amino acids were obtained from the SWISS-PROT Release 33.0 Release Notes.

Evaluation of the WAC Matrix

The performance of the WAC matrix was evaluated based on its sensitivity in detecting related protein sequences in the same family. 39 families were randomly picked from the PROSITE v13.1 database. Out of each family, the most distant protein sequence was chosen as the query. A BLAST search using the WAC, BLOSUM62, and PAM250 matrices was run for each query (Altschul et al., 1990). Rao's matrix, as published in (Rao, 1986), was also tested on several queries. The BLOSUM62 and PAM250 matrices were obtained from the same GCG v8.1 software package as BLAST (GCG, April 1991). The default values of E(xpect)=10 and W(ordlength)=3 were used. The Expect parameter instructs BLAST to discard sequences whose score is lower than a score that would be expected to occur 10 times by chance. The L(istsize) was increased from L=250 to L=500 so that the list of sequences found by the BLAST search was not truncated.

Results

The amino acid comparison tables are made available on the WWW. Figure 2 shows three comparison grids between two closely related amino acids, GLU and ASP, between two more distantly related amino acids, ASP and PHE, and between GLU and PHE.

The amino acid microenvironment data was used to construct the WAC amino acid similarity matrix. Since very few significant differences were found when comparing an amino acid's environment to itself (resampled), the similarity values along the diagonal are the highest in the matrix. The mean of all values is -0.945 with a standard deviation of 1.90. WAC has a correlation of 0.8136 with BLOSUM62. Figure 3 shows the position-by-position difference between WAC and BLOSUM62. WAC has a correlation of 0.7683 with Rao's matrix.

Figure 4 compares the results of the BLAST search using WAC, BLOSUM62, and PAM250. In 11 out of the 39 groups, all 3 matrices correctly identified every sequence. WAC correctly identified more sequences than BLOSUM62 in 3 groups, less in 16 groups. When compared to PAM250, WAC correctly identified more sequences in 11 groups, and less in 12. However, in the groups whose average sequence length is greater than 450 (12 groups), WAC found either the same number or more sequences than BLOSUM62. Conversely, WAC did not perform as well on shorter sequences. Rao's matrix did not perform as well as the other matrices (data not shown).

Discussion

The amino acid comparison grids show features consistent with the known differences between amino acids. In particular, the radial information is useful in distinguishing some closely related amino acids. For example, the primary difference between the environments of ASP and GLU is due to the greater length of GLU. Thus, although they share numerous features, GLU tends to have more features out to the 5-6 Å shell, while the ASP features extend to only 4-5 Å shell.

Similarly, a comparison of VAL and LYS shows that the long chain of LYS occupies different volume than VAL, and so the distribution of surrounding atoms is quite different. If we want to determine the most defining features of amino acids comparison, we could apply a stricter threshold of statistical significance than the current p=0.005.

For some amino acids, there is considerable rotational freedom in the sidechain that leads to a variety of possible positions distal to Cß. Our work has averaged over all these conformations in order to produce an "average" surrounding environment. Clearly, for particular proteins such an average would not be appropriate, and it would be preferable to model a particular subset of possible conformations. In this analysis, we have combined all orientations in order to have a general comparison of features over a wide range of proteins and microenvironments. It is remarkable that most of the key physical features are reflected in the differences in the surrounding distributions.

We focused on comparing WAC with BLOSUM62 because BLOSUM62 has been shown to have good performance (Henikoff & Henikoff, 1993; Vogt *et al.*, 1995), and it is the current default of BLAST search. In the WAC-BLOSUM62 difference matrix (Figure 3), all the diagonal entries are less than or equal to zero, while most of the off-diagonal numbers are greater than zero. This suggests that WAC is more tolerant to mutations than BLOSUM62. In database searches, WAC performed on par with PAM250 and slightly below BLOSUM62, even though WAC was not derived from amino acid substitution frequency patterns (Figure 4). The diagonal elements of our WAC matrix are all the same, because there are few significant differences between two sets of randomly chosen environments both centered around the same amino acid. We could use the comparison of individual amino acids against a random background in order to gain an estimate of the uniqueness of each amino acid environment. Such a refinement might improve the performance of our matrix with respect to BLOSUM62, which shows a clear variation in substitution likelihood along the diagonal.

The WAC matrix is derived from a simple statistical difference between the biochemical, biophysical, and structural environments of amino acids. The comparative microenvironment descriptions for each amino acid are based on statistical analyses of superimposed amino acids randomly chosen from a set of nonhomologous proteins, and are presented on the web page. The use of nonhomologous proteins ensures that we get a general measure of similarity of the environments surrounding amino acids. However, we can also produce context-dependent similarity matrices by using a selected set of proteins or selected regions within proteins (Koshi & Goldstein, 1995). The broad set of properties considered give a fairly comprehensive look at the critical features of amino acids. Some of the properties are purposefully redundant because we believe that there are many alternative viewpoints from which to study a microenvironment. These redundant properties were not used in creating the summary score for the WAC matrix construction. The detailed data sets are available upon request from the authors.

The use of the PROSITE groups and BLAST to perform a comparison of the matrices is discussed in (Henikoff & Henikoff, 1993). Despite the idiosyncrasies of PROSITE, such as its very stringent requirements for family membership, it represents the most reliable collection of related sequences that form a good gold-standard for large scale similarity searches. The BLOSUM62 matrix is derived from an analysis of the substitution frequencies in the conserved blocks within the BLOCKS/PROSITE databank. Because BLAST focuses on high scoring ungapped alignments, it may be that BLOSUM62 is especially well suited for performing well in BLAST searches. Nonetheless, the wide use of BLAST and the good average performance of BLOSUM62 make this type of evaluation very attractive, and a good benchmark.

In our study, WAC generally performs well for groups that have larger average sequence lengths. One possible reason is that smaller proteins have fewer structural constrains. Miyata et al. (Miyata et al., 1979) observed that in those low-constraint regions, amino acid substitution depends less on the extent of physico-chemical properties of substituted amino acids. Thus similarity matrices based on physico-chemical properties would be expected to perform poorly. We do not yet have enough data to make any statistically significant conclusions about the range of protein sizes that are best suited for use of our matrix.

The approach we used to calculate the WAC matrix is fundamentally different from that of other similarity matrices. The difference between WAC and matrices calculated from observed and expected frequencies, such as BLOSUM62 (Henikoff & Henikoff, 1992) and PAM250 (Dayhoff et al., 1978) is obvious—WAC is computed entirely from differences in the observed average environment surrounding amino acids without respect to any multiple alignment. WAC is also different from matrices that are derived from physico-chemical parameters, such as Rao's matrix (Rao, 1986) and Miyata's matrix (Miyata et al., 1979). WAC does not simply compare the physico-chemical properties of the amino acids themselves. Instead, the numbers in WAC come from statistical difference between physico-chemical and structural features of the micro-environments of amino acids, and have a built-in representation of the radial distribution of these features. This may explain the overall better performance of WAC over Rao's matrix.

The representation and characterization system we use is general purpose. In this paper, we applied the system to study the environments of amino acids, and then used the representation to make a similarity matrix. In other work, we have used the system to study the properties of calcium binding sites, disulfide-bonding state of Cysteines, and serine proteases (Bagley & Altman, 1995; Bagley & Altman, 1996). We have also developed a scoring scheme, based on the representation system, to recognize the occurrence of protein sites in new, unannotated structures (Wei & Altman, 1996), and have achieved high accuracy in recognizing calcium binding sites, disulfide-bonding Cysteines, and ATP binding sites.

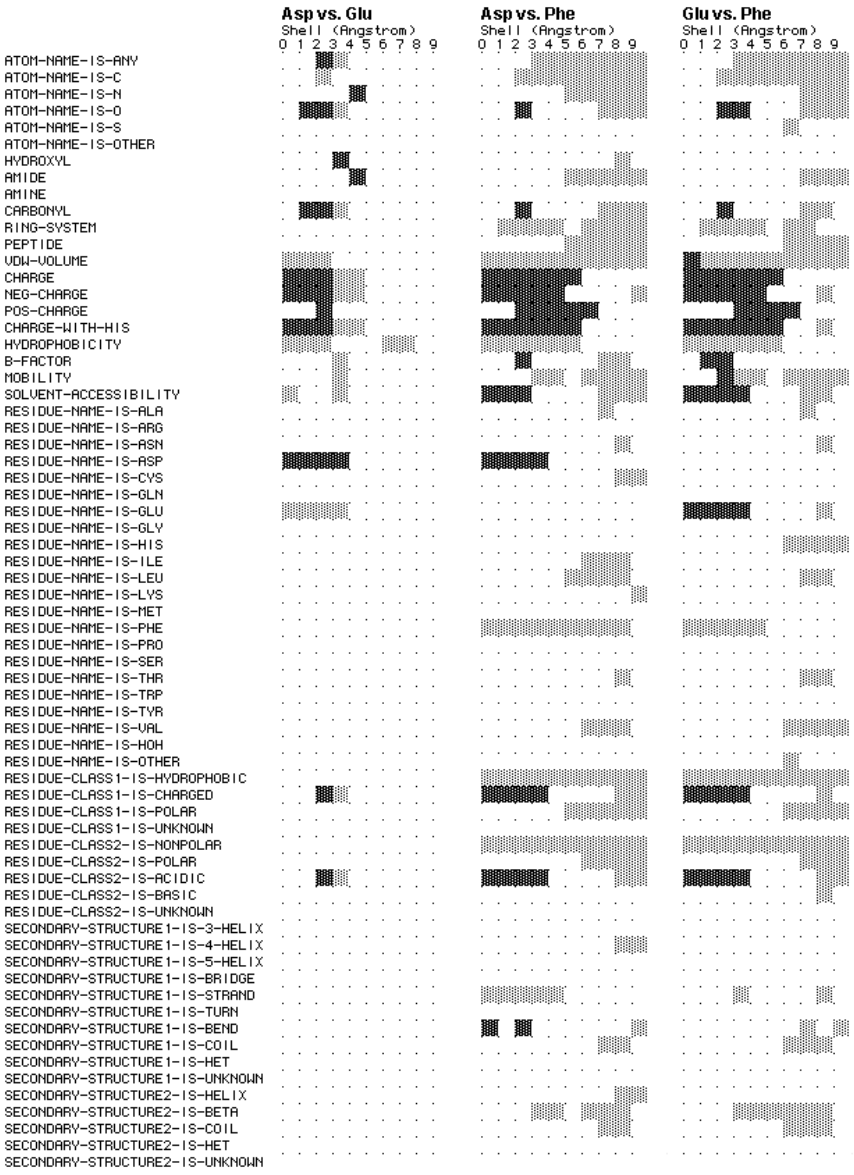

Figure 2 Sample analyses. Properties, listed along the left-hand column, were analyzed at 1Å volumes up to 10Å from the C_β carbon. The volumes for which the properties were significantly more prevalent for the first sites are marked in light grey, and dark grey for the second. Most properties are self-explanatory, and represent the presence of a particular atom, chemical group, residue type or type of secondary structure. "Asp vs. Glu" shows much fewer significantly different properties than "Asp vs. Phe" and "Glu vs. Phe," indicating that ASP and GLU are more closely related to each other than they are to PHE.

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
	-5	1	1	2	0	3	3	1	2	3	3	2	1	2	1	1	1	2	2	2	C
		0	0	1	-1	0	0	-1	-2	0	0	0	0	1	1	1	1	0	1	2	S
C	4		-1	1	0	2	1	0	0	2	1	1	1	1	1	1	0	0	1	1	T
S	0	4		-3	1	2	2	0	0	1	2	2	1	3	3	3	2	4	2	4	P
T	0	1	4		0	-1	2	0	-2	1	2	0	0	2	1	1	1	2	2	3	A
P	-1	0	0	4		-2	0	-1	0	2	1	1	1	3	2	3	1	1	1	0	G
A	0	0	0	0	4		-2	-1	0	1	-1	0	0	2	2	2	2	2	2	3	N
G	0	0	0	0	-1	4		-2	-2	-1	0	1	0	2	-1	1	0	-1	0	1	D
N	0	1	1	0	0	0	4		-1	-2	-1	-1	-2	1	-1	0	-2	-1	-2	1	E
D	-2	-1	-1	-1	-2	-2	0	4		-1	0	-1	-1	1	3	2	1	3	2	2	Q
E	-2	-2	-1	-1	-3	-2	0	0	4		-4	0	1	2	2	2	2	1	-2	2	H
Q	0	0	1	0	0	0	1	-1	0	4		-1	0	1	1	1	1	1	1	3	R
H	0	-1	-1	0	0	-1	0	-1	-1	0	4		-1	1	1	0	-1	1	0	1	K
R	-1	-1	0	0	-1	-1	0	-1	-1	0	0	4		-1	1	0	0	2	1	3	M
K	-2	0	0	0	-1	-1	0	-1	-1	0	0	2	4		0	-1	-1	0	1	3	I
M	1	0	0	1	1	0	0	-1	-1	1	0	0	0	4		0	0	1	1	3	L
I	0	-1	0	0	0	-2	-1	-4	-4	0	-1	-2	-2	2	4		0	1	1	3	V
L	0	-1	0	0	0	-1	-1	-3	-3	0	-1	-1	-2	2	1	4		-2	0	-1	F
V	0	-1	0	0	1	-2	-1	-3	-4	-1	-1	-2	-3	1	2	1	4		-3	-1	Y
F	0	-2	-2	0	0	-2	-1	-4	-4	0	0	-2	-2	2	0	1	0	4		-7	W
Y	0	-1	-1	-1	0	-2	0	-3	-4	1	0	-1	-2	0	0	0	0	1	4		
W	0	-1	-1	0	0	-2	-1	-3	-2	0	0	0	-2	2	0	1	0	2	1	4	
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Figure 3 WAC similarity matrix (Lower) and difference matrix (Upper) obtained by subtracting the BLOSUM62 matrix from the WAC matrix position by position.

The use of microenvironmental information is not limited simply to the comparison of average amino acid environments, but could be applied to specific environments within specific structural families. Bowie and Eisenberg have shown that environmental attributes can be used for a dynamic programming-based scheme for fold recognition (Bowie *et al.*, 1991). Our environmental attributes, including radial information, may also be suitable for this task. In addition, our environmental attributes could be used to create scoring functions for evaluating the detailed fit of a sequence to a structural fold in the manner of current threading techniques (Bryant & Lawrence, 1993; Jones *et al.*, 1992; Sippl & Weitckus, 1992).

Conclusions

We have performed a comprehensive analysis of the environments surrounding amino acids. Our analysis differs from previous analyses in that we maintain some information about the geometry and relative distances between features in a radial shell around C_β of the amino acids. We have shown that our analysis

Group	Description	Avg Length	Query	Sensitivity WAC	BLOSUM62	PAM250
PS00216	Sugar Transport Protein	516	Q06221	89	77	80
PS50003	PH Domain	901	P35401	11	6	1
PS00221	MIP Family	295	P42067	100	97	81
PS01080	Apoptosis regulator	220	Q07440	71	71	71
PS01047	Heavy Metal Associated Domain	657	P46839	57	57	61
PS00850	Glycine Radical	475	P13316	75	75	75
PS00687	Aldehyde Dehydrogenase	527	P30841	100	100	100
PS00667	NADH Dehydrogenase	318	Q00242	100	100	100
PS00499	C2 domain	631	P27715	90	90	56
PS00437	Catalase 1	515	P11934	100	100	100
PS00316	Thaumatin	214	P13867	100	100	100
PS00249	PDGF	212	P01128	100	100	100
PS00242	Integrin Alpha	1086	P34446	100	100	100
PS00239	Receptor Tyrosine Kinase	1051	P42159	100	100	100
PS00162	Eukaryotic CO2 Anhydrase	280	P28651	100	100	100
PS00120	Lipase Serine Active Site	435	P25234	7	7	7
PS00113	Adenylate Kinase	210	P43412	100	100	100
PS00110	Pyruvate Kinase	523	P22200	100	100	100
PS00049	Ribosomal Protein L14	125	P46767	100	100	90
PS00039	DEAD ATP helicase	514	P38719	100	100	100
PS00021	Kringle Domain	718	P00748	100	100	100
PS00013	Prokaryotic Lipoprotein attachment	286	P29722	2	2	1
PS01033	Globin	147	Q03331	1	2	5
PS00190	Cytochrome c	111	Q02469	0	1	0
PS00259	Gastrin	84	P09040	14	17	6
PS00267	Tachykinin	43	P22691	7	11	9
PS00201	Flavodoxin	157	P41050	83	87	74
PS00192	Cytochrome b	309	P15585	82	86	81
PS00233	Insect Cuticle	141	P26967	41	47	47
PS00237	G-Protein Receptor	402	P16849	34	44	32
PS00287	Cysteine protease inhibitor	226	P35479	33	45	41
PS00636	DNAJ	412	P36540	3	24	5
PS00018	EF-hand Calcium Binding	236	P39047	55	77	5
PS00418	Potex Carlavirus Coat	251	P22172	77	100	96
PS01124	Thermonuclease	272	P26950	47	73	65
PS00262	Insulin	101	P15131	60	94	92
PS00272	Snake Toxin	65	P28375	35	82	86
PS00197	Ferredoxin	214	P07771	36	85	85
PS00027	Homeobox	337	P31367	30	87	88

Figure 4 BLAST Search Data. A BLAST search was performed for each matrix using a query chosen from each PROSITE group. The sensitivity of a matrix is the percentage of the group members detected.

produces detailed maps of the environmental differences between each amino acid and the other amino acids. We have further illustrated one practical use of our analysis, in the creation of a similarity matrix, the WAC matrix, based on a simple statistical summary of the differences between two sets of average amino acid environments. The performance of the WAC matrix is remarkable given the simplicity of its derivation, and suggests that refinement may lead to improved performance.

Acknowledgments

RBA is a Culpeper Medical Scholar, and this work was supported by the Culpeper Foundation and by NIH grants LM-05652 and LM-06244. The computing environment was provided by the CAMIS resource under NIH LM-05305. JTC is supported by the Howard Hughes Summer Fellowship from the Department of Biological Sciences. The BLAST searches were performed using the computing resources of CMGM. We thank Lee Kozar for his help with running BLAST search.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**(3), 403-410.

Bagley, S. C. & Altman, R. B. (1995). Characterizing the microenvironment surrounding protein sites. *Protein Sciences* **4**, 622-635.

Bagley, S. C. & Altman, R. B. (1996). Conserved Features in the active site of nonhomologous serine proteases. *Folding & Design*, in press.

Bowie, J. U., Lüthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164-170.

Bryant, S. H. & Lawrence, C. E. (1993). An empirical energy function for threading protein sequence through the folding motif. *Proteins: Structure, Function, and Genetics* **16**, 92-112.

Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure* **5, suppl. 3**, 345-352.

GCG. (April 1991). Program Manual for the GCG Package 7 edit., 575 Science Drive, Madison, WI, USA 53711.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science* **185**(154), 862-4.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 10915-10919.

Henikoff, S. & Henikoff, J. G. (1993). Performance evaluation of amino acid substitution matrices. *Proteins: Structure, Function, and Genetics* **17**, 49-61.

Holm, L. & Sander, C. (1994). The FSSP database of structurally aligned protein fold families. *Nucleic Acids Research* **22**, 3600-3609.

Jones, D. D. (1975). Amino acid properties and side-chain orientation in proteins: a cross correlation appraoch. *Journal of Theoretical Biology* **50**(1), 167-83.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature* **358**(6381), 86-9.

Koshi, J. M. & Goldstein, R. A. (1995). Context-dependent optimal substitution matrices. *Protein Eng* **8**(7), 641-5.

Miyata, T., Miyazawa, S. & Yasunaga, T. (1979). Two types of amino acid substitutions in protein evolution. *Journal of Molecular Evolution* **12**, 219-236.

Ott, L. R. (1992). *An introduction to statistical methods and data analysis*. fourth edit, Wadsworth Publishing Company, Belmont, CA 94002.

Rao, M. J. K. (1986). New scoring matrix for amino acid residue exchanges based on residue characteristic physical parameters. *International Journal of Peptide and Protein Research* **29**, 276-281.

Sippl, M. J. & Weitckus, S. (1992). Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins: Structure, Function, and Genetics* **13**, 258-271.

Vogt, G., Etzold, T. & Argos, P. (1995). An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. *Journal of Molecular Biology* **249**, 816-831.

Wei, L. & Altman, R. B. (1996). A general method for recognizing protein sites. , in preparation.