AI Magazine Volume 25 Number 1 (2004) (© AAAI)

Representation of Protein-Sequence Information by Amino Acid Subalphabets

Claus A. F. Andersen and Søren Brunak

■ Within computational biology, algorithms are constructed with the aim of extracting knowledge from biological data, in particular, data generated by the large genome projects, where gene and protein sequences are produced in high volume. In this article, we explore new ways of representing protein-sequence information, using machine learning strategies, where the primary goal is the discovery of novel powerful representations for use in AI techniques. In the case of proteins and the 20 different amino acids they typically contain, it is also a secondary goal to discover how the current selection of amino acids—which now are common in proteins—might have emerged from simpler selections, or alphabets, in use earlier during the evolution of living organisms.

Proteins typically contain 20 different amino acids, which have been selected during evolution from a much larger pool of possibilities that exists in Nature. Protein sequences are constructed from this alphabet of 20 amino acids, and most proteins with a sequence length of 200 amino acids or more contain all 20, albeit with large differences in frequency. Some amino acids are very common, but others are rare. The human genome encodes at least 100,000 to 200,000 different protein sequences, with lengths ranging from small peptides with 5 to 10 amino acids to large proteins with several thousand amino acids.

A key problem when constructing computational methods for analysis of protein data is how to represent the sequence information (Baldi and Brunak 2001). The literature contains many different examples of how to deal with the fact that the 20 amino acids are related to one another in terms of biochemical properties—very much in analogy to natural language alphabets where two vowels might be more "similar" than any vowel-consonant pair, for example, when constructing speech-synthesis algorithms.

In this article, we do not want to cover all attempts to represent protein sequences computationally but restrict the review to recent developments in the area of amino acid subalphabets, where the idea is to discover groups of amino acids that can be lumped together, thus giving rise to alphabets with fewer than 20 symbols. These subalphabets can then be used to rewrite or reencode the original protein sequence, hopefully giving rise to better performance of an AI algorithm designed to detect a particular functional feature when receiving the simplified input. The idea is completely general, and similar approaches might be relevant in other symbol-sequence data domains, for example, in natural language processing.

It should be mentioned that alphabet expansion in some cases can also be advantageous, that is, to rewrite sequences in expanded, longer alphabets covering more than one symbol, thus encoding significant correlations between individual symbols directly into the rewritten sequence. For example, deoxyribonucleic acid (DNA) sequences contain four different nucleotides (*ACGT*), but a rewrite as dinucleotides (*AA, AC, AG*, ...), or trinucleotides (*AAA, AAC, AAG*, ...) might lead to a DNA representation where functional patterns are easier to detect by machine learning algo-

Knowledge Representation in Bioinformatics

Knowledge representations are key for the construction and performance of all AI methods, irrespective of the application domain. In bioinformatics, the knowledge representation issue is of significant importance, and there are many different types of data that should be represented in ways such that algorithms will be able to extract and integrate knowledge from the plethora of data generated by novel high-throughput techniques in biology. A large part of the data in the life science area is essentially digital because biomolecules often consist of a limited set of chemical building blocks. This is the case for deoxyribonucleic acid (DNA) and protein, where 4 nucleotides and 20 amino acids, respectively, make up the “alphabet” from which these essential molecules are constructed. Bioinformatics algorithms are typically designed to scan these symbol sequences and detect local or global functional patterns as well as structural aspects related to the function of a molecule, for example, a protein. It is clear that the chemical alphabets in biology are products of three to four billion years of evolution and that the alphabets have evolved in constrained ways where a balance has been obtained between the capacity for encoding chemical function, the complexity of sequence decoding, and energetic cost considerations.

rithms. For example, this is the case when detecting the small part of the DNA that actually encodes proteins by artificial neural networks (Hebsgaard et al. 1996). The protein-encoding part of the DNA in the human genome is a few percent of the total DNA in the chromosomes; therefore, the problem is to detect protein-encoding segments in a “sea” of noncoding DNA. This task is made easier when the sequences are also analyzed as dinucleotides (16-symbol alphabet) and trinucleotides (64-symbol alphabet) (Hebsgaard et al. 1996).

In proteins, the common 20-letter amino acid alphabet contains groups of amino acids with similar biochemical properties, which can be merged for improved computational analysis. Thus, we have subalphabets with less than 20 symbols. However, one should be careful when merging individual amino acids solely based on their biochemical properties because many functional patterns in proteins are embedded as sequence correlations. This contextual aspect is again similar to natural language, where the pronunciation of the four *A*s in the sentence *Mary had a little lamb* requires three different phonemes because the contexts of the *A*s are different (Sejnowski and Rosenberg 1987).

Amino acids in proteins also do not contribute to the function or structure of proteins independently. The amino acid *alanine,* for example, is found in many different types of protein structure depending on the surrounding amino acids, and in this sense, amino acid sequences should be read in the same way that natural language sequences are read, where the short- and long-range symbol correlations are essential for the pronunciation. In proteins, relevant correlations can be long range because the sequence typically folds back on itself and is stabilized by intrachain bonds between segments far apart in the sequence.

Protein structure, which is essential for function, can be described at different levels than the complete, all-atom representation of *x, y, z* coordinates with more topology-oriented descriptions, such as protein secondary structure, where each amino acid typically is being put into one of the mutually exclusive conformational categories (from a small number of possible conformational states). The most common classification is that of α-helix, β-sheet, and coil, which is the structural level we use here in the search for novel amino acid subalphabets.

Merged amino acid subalphabets are of significant interest both in the context of evolution and in protein-structure prediction. Recently, several subalphabets have been suggested (Wang and Wang 1999); however, they all have a detrimental effect on the ability to predict structural features. The strategy for their selection has been to reduce the loss rather than to optimize the gain in terms of predictive performance.

In this article, we introduce a new computational approach for evaluating subalphabets by searching directly for sequence reencodings that improve protein secondary-structure prediction. In contrast to reduced alphabets that only support the conformation of a single-protein domain (Riddle et al. 1997), the clusters of merged amino acids found here are likely to represent substitutions that might preserve structure for proteins in general.

Using this approach, we have discovered protein alphabets composed of 13 to 19 groups that indeed increase the predictability of secondary structure from sequence.

In the paper by Wang and Wang (1999), the search for, and ranking of, subalphabets was based on the 190 amino acid substitution scores in the Miyazawa and Jernigan (1996) matrix (MJ matrix). The difference in the current approach is that subalphabets are evaluated on the basis of actual, contextual sequence data and not just on the basis of pairwise amino acid interaction scores, which cannot take commonly occurring sequence correlations into account. Such correlations are of obvious, structural importance: Parallel and antiparallel β-sheets (Wouters and Curmi 1995) and α-helixes (Peterson et al. 1999; Wan and Milner-White 1999; Wintjens, Wodak, and Rooman 1998) are supported by hydrogen bonds and require local sequence periodicities of two and four, respectively. Other correlations at larger sequence separations are also highly significant for attaining a specific structure.

In the computational approach presented here, we use a large set of 650 high-quality, nonsequence similar chains comprising 130,356 amino acids selected from the PROTEIN DATA BANK, containing experimentally solved protein structures. A given subalphabet is evaluated by measuring how well the structure can be predicted from the recoded version of the original sequence. This evaluation is performed by predicting the secondary structure from the subalphabet representation of the sequence, where the prediction is performed by a conventional neural network prediction setup (Bohr et al. 1988; Jones 1999; Qian and Sejnowski 1988; Rost and Sander 1993). A good amino acid grouping will therefore relate sequence to secondary structure, based on the assumption that the sequence uniquely determines the structure. This grouping scheme uses the protein data set as the evaluation measure, as opposed to the Wang and Wang scheme.

Using this approach, we generate good subalphabets through an iterative, one-path reduction, where we examine the prediction quality by successively merging two groups (which initially consist of individual amino acids) but keeping the best-scoring groupings joined. Thus, we start from 20 groups and evaluate the 190 possible subalphabets with 19 groups. By selecting the best-fit subalphabet, we repeat the procedure, this time going from 19 to 18 groups (171 comparisons), eventually performing 1329 evaluations to achieve a final subalphabet of only 2 groups.

Structure-Preserving Reduced Alphabets

The number of possible subalphabets, N_n, grows quite dramatically with the alphabet size, n:

$$N_n = \sum_{l=1}^{n} f_n(l)$$

$$f_n(l) = \begin{cases} 1 & \text{if } l = 1 \\ \dfrac{l^n}{l!} - \sum_{m=1}^{l} \dfrac{f(l-m)}{m!} & \text{otherwise} \end{cases}$$

where $f_n(l)$ is the number of subalphabets with l groups. For $n = 20$, this gives $N_{20} = 5 \bullet 10^{13}$ subalphabets. The number of subalphabets composed of, say, 8 groups, $f_{20}(8) = 15 \bullet 10^{12}$, is also very large, and a sampling strategy is therefore needed, as indicated earlier. In table 1, we have listed the alphabet size n versus the number of possible subalphabets N_n.

Alphabet size n	Possible subalphabets N_n
2	2
3	5
4	15
5	52
6	203
7	877
8	4 140
9	21 147
10	115 975
11	678 570
12	4 213 597
13	27 644 437
14	190 899 322
15	1 382 958 545
16	10 480 142 147
17	82 864 869 804
18	682 076 806 159
19	5 832 742 205 057
20	51 724 158 235 372

Table 1. The Number of Possible Subalphabets for a Given Alphabet Size.

The prediction performance is estimated from a 10-fold cross-validation. The neural network was trained using standard backpropagation (Hertz, Krogh, and Palmer 1991) with one hidden layer (45 neurons), learning rate ϵ = 0.001, entropic error function, tanh(x) as transfer function, and balanced training. To keep the computation time down, the setup has been kept simple; that is, evolutionary information, structure-to-structure networks, and ensemble predictions were not incorporated as done for today's best-performing secondary-structure prediction schemes (Jones 1999; Rost,

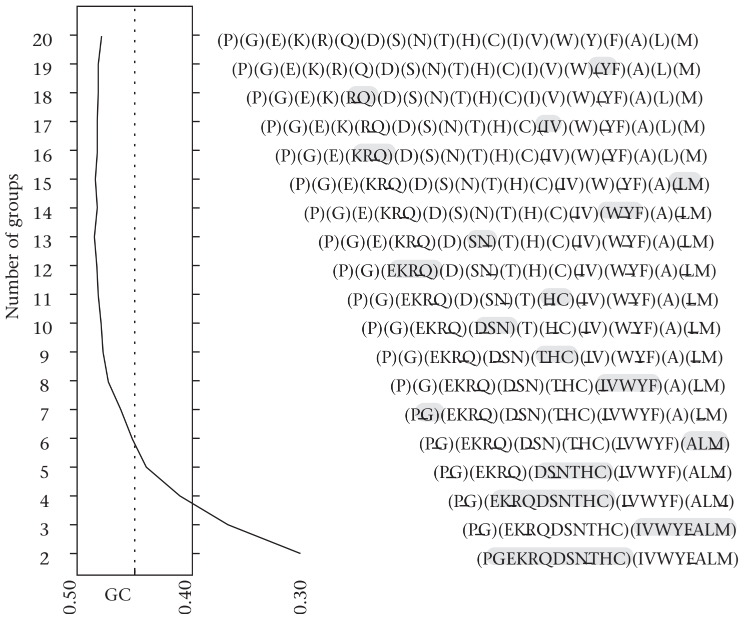

Figure 1. Reduction Tree.

By starting from the 20-letter amino acid alphabet (20 groups), the stepwise reduction by joining the highest-scoring pair and reestimating the new set of pairs is shown going down to 2 groups.

Sander, and Schneider 1994). The prediction performance is evaluated using the generalized correlation coefficient (Baldi et al. 2000) jointly for helix, sheet, and coil

$$GC = \sqrt{\frac{\sum_{ij}\left(z_{ij} - e_{ij}\right)^2}{N\left(K-1\right)}}$$

where N is the total number of observations, K is the number of classes, z_{ij} is the number of observations assigned to be in class i and predicted to be in class j, and

$$e_{ij} = \sum_i z_{ij} \cdot \frac{\sum_i z_{ij}}{N}$$

is the expected number of observations assigned as i and predicted as j. The GC is identical to the Matthews (1975) correlation coefficient for two categories, but in contrast to Matthews, it is additive when more categories are used.

Using the secondary-structure predictability as the merging principle, we found the reduction tree shown in figure 1. It shows all the optimal subalphabets, starting from the 20 amino acids going down to 2 groups. Surprisingly, we were able to find several subalphabets that increased the predictability of the secondary structure from the recoded sequence. In fact, all the optimal clusters in the range from 13 to 19 groups found using this greedy procedure increase the predictability. The path of mergers gives us a ranking of contextual, structural similarity between the amino acids based on the groups they form. In particular, we find that the aromatic group (Phe-Tyr-Trp), the aliphatic sheet amino acids (Val-Ile), the basic amino acids (Lys-Arg), and Met-Leu are merged very early in the procedure.

Interestingly, below 13 clusters, the cost of merging two groups increases quite abruptly at certain subalphabet sizes. When going below eight groups, and again when going below five groups, a fivefold and twofold increase in cost is implied, respectively. These two jumps indicate thresholds where a subalphabet no longer can represent and encode structurally important aspects in a large, diverse set of proteins. The first jump is found when merging Pro and Gly, two structurally unique amino acids, Pro for its rigidity, Gly for its flexibility. The second jump occurs when the group of predominantly small polar amino acids (DSNTHC) is joined with the group containing large polar amino acids (EKRQ). Group sizes of eight and five were also found by Wang and Wang (1999) to exhibit special characteristics, in agreement with our finding that eight groups are just about adequate to represent the most important structural properties.

One might ask what amino acid will be able to best represent the merged cluster of amino acids? A representative amino acid for each cluster (underlined in figure 1) was identified using the neural network trained on the complete 20-letter alphabet (the original PDB sequences). This network was used to decide which of the amino acids in a cluster would be the representative one according to how well it mapped the sequence to the structure relationship. Our five-letter alphabet

(PG)(EKRQ)(DSNTHC)(IVWYF)(ALM)

has only one representative amino acid in common with the cluster found by Riddle et al. (1997) and confirmed by Wang and Wang

(PG)(SKNRQ)(DE)(IVLMCWYF)(ATH)

but the representatives are all very similar. Ala and Met are similar hydrophobic amino acids, Gly and Pro both influence structure a lot and have their separate group, Glu and Ser are both positive polar amino acids, and Lys and Arg are similar basic amino acids. We therefore propose the PRISM alphabet as an improved set of representatives, which also reflects structural and contextual importance. This is in contrast to the GKEIA alphabet, which is solely based on the 190 amino acid interaction energies in the MJ matrix.

Some proteins have a more biased composition than others. One might ask whether the variability of amino acids used in a given protein influences the sequence-specific performance in terms of prediction accuracy? When the sequence-specific performance for the optimal subalphabet of 13 groups (measured by GC) was plotted against the amino acid distribution entropy, $\Sigma_l\, p_l \log_2(p_l)$, it was clearly demonstrated that lower-complexity sequences with bias in the composition were not easier to predict (data not shown). This observation again indicates that the reduced alphabets found here reflect a clustering that is of general nature and not a reduction that is optimized for a small subset of the sequence space.

The Impact of Sequence Correlations

A comparison of the alphabets found by Wang and Wang (using the MJ matrix) to the alphabets found here should show explicitly how protein structure and sequence context change the outcome. In figure 2a, we compare the 10 highest-scoring five-group Wang and Wang subalphabets and the Wang and Wang mismatch score M_{ab} to the structural predictability scores used in our evaluation. It is clear that we have found a five-group alphabet that better preserves structural information when compared to any of the highest-scoring Wang and Wang alphabets. Otherwise, there is a high degree of agreement in the ranking; in fact, only one subalphabet receives a significantly different ranking.

Where does the MJ matrix fail to preserve structural information? To compare amino acids in the MJ matrix, we calculated the root mean square distance between all pairs of amino acid contact energy vectors, which estimates whether two amino acids have similar interactions with all the other amino acids, instead of just strong mutual contact energies. We refer to this transformed matrix as the MJRMS (MJ root mean square) matrix.

Using a numeric evaluation of all subalphabets of 19 groups, we have produced an analogous structure-preserving amino acid–grouping matrix (SPAA matrix).

The scatter plot in figure 3 shows that the two matrixes are quite different. For example, the two extreme pairwise values in the MJRMS

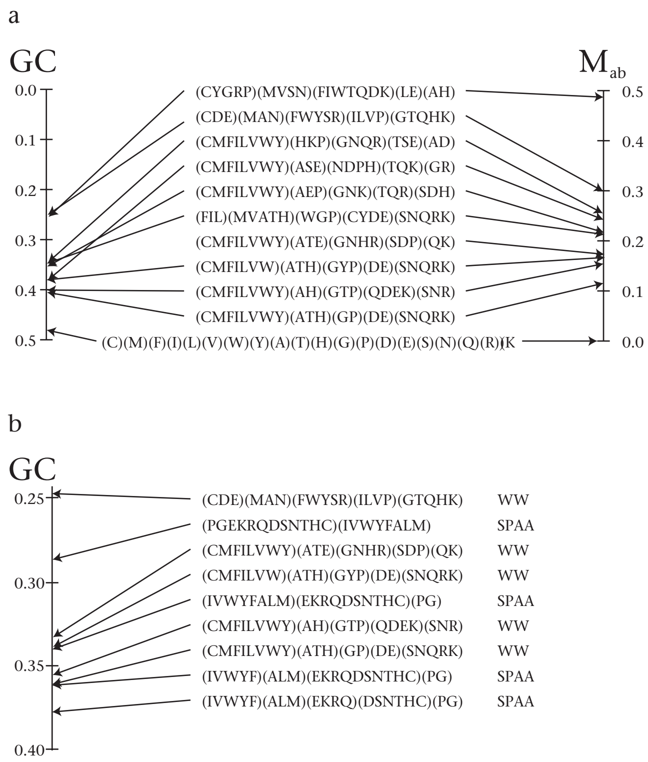

Figure 2. Comparison of Different Subalphabets.

A. A comparison of the Wang and Wang mismatch measure M_{ab} for protein subalphabet evaluation and our structural predictability measure (generalized correlation coefficient GC [figure 1]) used in this work. B. The predictive test performance (generalized correlation coefficient GC) for various alphabets on the independent Pfam-A data of protein families. The figure shows that the SPAA alphabets preserve the sequence-structure relationship better than any of the Wang and Wang alphabets.

Data-Set Preparation

The 650 protein chains extracted from PDB were selected to have low sequence similarity (25 percent or less); a minimum length of 30 amino acids; no transmembrane segments; and, for X-ray data, a resolution better than 2.5 Å. The Pfam-A data set consisting of protein families was selected, running each family's profile hidden Markov model (Bateman et al. 1999) against each protein in the 650-protein data set and removing families with alignment *E*-values below 100. Furthermore, only families containing X-ray structures with a resolution better than 2.5Å were accepted. All members in a family were assigned the consensus secondary-structure field from the Pfam alignment. This field contains the DSSP consensus assignment of the PDB entries present in the family. The 90 selected families were given equal weighting in the final result.

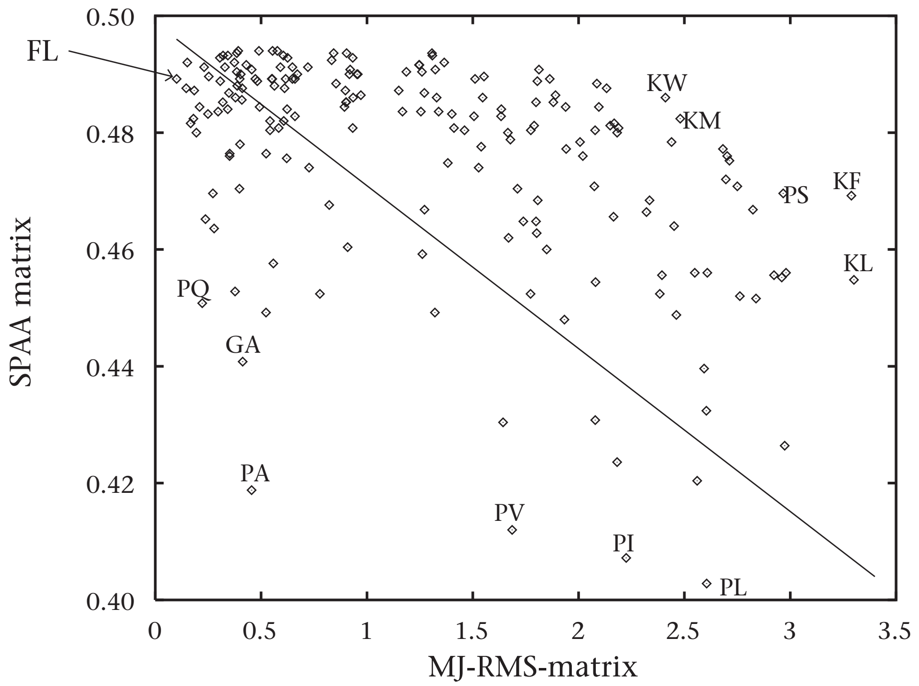

Figure 3. Comparison of Substitution Matrixes.

The scatter plot compares all values in the MJRMS matrix (Wang and Wang 1999) with the structural amino acid–grouping matrix presented in this article. The amino acid pairs most far away from the diagonal highlight differences between contact versus structural comparison of the amino acids. Pro and Gly are structurally unique amino acids, which makes them stand out, but Lys has a unique contact energy motif.

matrix (Phe/Leu and Lys/Leu) have similar scores in the SPAA matrix. The differences can also be illustrated in the form of unrooted trees (figure 4), where a comparison to the equivalent BLOSUM62 (Henikoff and Henikoff 1992) grouping is also shown. For the MJRMS matrix, Pro, Gly, and Ala fall into the hydrophilic group even though they are nonpolar, which has also been observed by Chan (1999). A more peculiar feature of the MJ matrix is that Pro and Gly, structurally two very special amino acids, are not identified as such. Pro and Gly have individual, distant branches in the BLOSUM62 matrix, and by the SPAA matrix, they are even placed in a separate cluster.

Incorporating sequence correlations into a protein subalphabet evaluation is impossible when the approach is based on a simple matrix. In the structural amino acid grouping presented here, position-specific interactions greatly influence the predictability; for example, instead of separating amino acids that are acidic and basic, the separation of polar amino acids is based on their size (large/long or small/short), as discussed earlier. This clustering makes a lot of sense when compared to work showing which amino acids are involved in specific types of interactions with side-chain hydrogen bonds stabilizing secondary structure (α-helixes and β-sheets) (Bordo and Argos 1994). The amino acid counts for each type of interaction group the amino acids in similar ways, as found in our work, that is, mixing acidic and basic amino acids and separating large and small amino acids.

We have furthermore tested the reduced alphabets on a set of 90 unrelated Pfam protein families (Bateman et al. 1999) (figure 2b). The aim was to validate and select structure-preserving reduced alphabets that are applicable to proteins in general instead of those just being optimal for individual families. The test families used were all unrelated to the data set of 650 proteins used for neural network training, and they had, in addition, a manually aligned core (seed) alignment (Pfam-A) and a

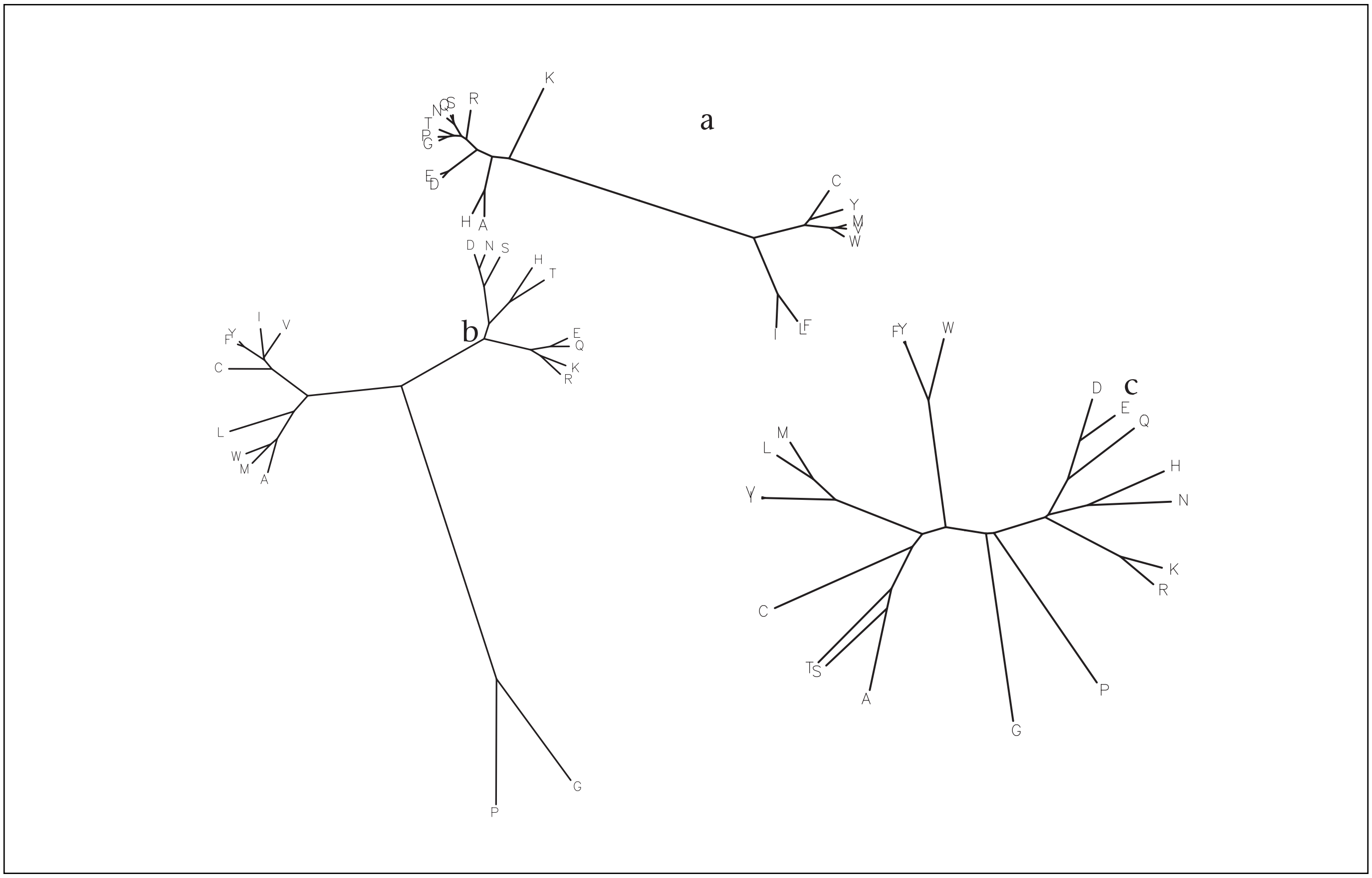

Figure 4. Comparison of the MJRMS (top), SPAA (bottom left), and BLOSUM62 (bottom right) Matrixes by Unrooted Trees.

The MJ matrix has been transformed according to the root mean square between the amino acid contact energy vectors to show the similarity between amino acids based on all interactions instead of just their pairwise contact energies. To produce the structural amino acid–grouping matrix, we have evaluated all subalphabets with 19 groups in the manner described earlier. This gives a GC for each subalphabet, which describes how predictable the protein secondary structure is from that subalphabet. The unrooted trees shown here were built by the neighbor-joining algorithm from the Phylip package (Felsenstein 1989). The MJRMS matrix is derived from the MJ matrix by calculating the root mean square between amino acid vectors:

$$MJ_{RMS}(i,j)=\sqrt{\frac{1}{20}\sum_{k}^{20}\left(MJ(i,k)-MJ(j,k)\right)^{2}}$$

resolution better than 2.5 Å. The result of this validation shows that the SPAA alphabets are better at preserving structural information across families than are the Wang and Wang alphabets. The four-letter SPAA alphabet is even slightly better in terms of predictive performance than the best of the Wang and Wang five-letter alphabets. Furthermore, the two-letter SPAA alphabet is seen to surpass one of the (nonoptimal) Wang and Wang five-letter alphabets, which has merged polar amino acids (SR) with nonpolar amino acids (FWY). This fact limits the ability of this particular alphabet to preserve structure considerably, again underscoring the importance of the hydrophobic aspect. In figure 2a, the GCs obtained are lower than in figure 2b, which is caused by the variation in structure within each Pfam family. When restricting the validation to those family members with known PDB structure, the GC levels are essentially the same as those reported earlier (data not shown).

Other amino acid-grouping methodologies have been tried on the MJ matrix (Cieplak et al. 2001). These clustering methods basically reflect the clusters seen in the unrooted MJRMS tree (figure 4) and therefore completely neglect the structural importance of Pro and Gly.

In conclusion, we believe that the structure-preserving amino acid groupings found here are more likely to be relevant in relation to protein folding and that they indeed offer a different perspective than groupings based on individual contact energies between amino acids. Special amino acids might very well be required in the active site to maintain functional specificity, as observed by Riddle et al. (1997).

Acknowledgments

The work was supported by a grant from the Danish National Research Foundation and a Ph.D. grant from the Technical University of Denmark.

References

Baldi, P., and Brunak, S. 2001. *Bioinformatics: The Machine Learning Approach.* 2d ed. Cambridge, Mass.: MIT Press.

Baldi, P.; Chauvin, Y.; Andersen, C. A.; Nielsen, H.; and Brunak, S. 2000. Assessing the Accuracy of Prediction Algorithms for Classification: An Overview. *Bioinformatics* 16(5): 412–424.

Bateman, A.; Birney, E.; Durbin, R.; Eddy, S. R.; Finn, R. D.; and Sonnhammer, E. L. 1999. Pfam 3.1: 1313 Multiple Alignments Match the Majority of Proteins. *Nucleic Acids Research* 27(1): 260–262.

Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; and Bourne, P. E. 2000. The Protein Data Bank. *Nucleic Acids Research* 28(1): 235–242.

Bohr, H.; Bohr, J.; Brunak, S.; Cotterill, R. M.; Lautrup, B.; Nørskov, L.; Olsen, O. H.; and Petersen, S. B. 1988. Protein Secondary Structures and Homology by Neural Networks: The Alpha-Helices in Rhodopsin. *Federation of European Biochemical Societies Letters* 241(1–2): 223–228.

Bordo, D., and Argos, P. 1994. The Role of Sidechain Hydrogen Bonds in the Formation and Stabilization of Secondary Structure in Soluble Proteins. *Journal of Molecular Biology* 243(3): 504–519.

Chan, H. S. 1999. Folding Alphabets. *Nature Structural Bio*logy 6(11): 994–996.

Cieplak, M.; Holter, N. S.; Maritan, A.; and Banavar, J. R. 2001. Amino Acid Classes and the Protein Folding Problem. *Journal of Chemical Physics* 114:1420–1423.

Felsenstein, J. 1989. Phylip—Phylogeny Inference Package (Version 3.2). *Cladistics* 5(2): 164–166.

Hebsgaard, S. M.; Korning, P. G.; Tolstrup, N.; Engelbrecht, J.; Rouzé, P.; and Brunak, S. 1996. Splice Site Prediction in *Arabidopsis thaliana* PrE–mRNA by Combining Local and Global Sequence Information. *Nucleic Acids Research* 24(17): 3439–3452.

Henikoff, S., and Henikoff, J. G. 1992. Amino Acid Substitution Matrixes from Protein Blocks. *Proceedings of the National Academy of Sciences* 89(22): 10915–10919.

Hertz, J.; Krogh, A.; and Palmer, R. G. 1991. *Introduction to the Theory of Neural Computation.* Redwood City, Calif.: Addison-Wesley.

Jones, D. T. 1999. Protein Secondary-Structure Prediction Based on Position-Specific Scoring Matrixes. *Journal of Molecular Biology* 292(2): 195–202.

Kabsch, W., and Sander, C. 1983. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* 22(12): 2577–2637.

Matthews, B. W. 1975. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochimica et Biophysica Acta* 405(2): 442–451.

Miyazawa, S., and Jernigan, R. L. 1996. Residue-Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading. *Journal of Molecular Biology* 256(3): 623–644.

Peterson, R. W.; Nicholson, E. M.; Thapar, R.; Klevit, R. E.; and Scholtz, J. M. 1999. Increased Helix and Protein Stability through the Introduction of a New Tertiary Hydrogen Bond. *Journal of Molecular Biology* 286(5): 1609–1619.

Qian, N., and Sejnowski, T. J. 1988. Predicting the Secondary Structure of Globular Proteins Using Neural Network Models. *Journal of Molecular Biology* 202(4): 865–884.

Riddle, D. S.; Santiago, J. V.; Bray-Hall, S. T.; Doshi, N.; Grantcharova, V. P.; Yi, Q.; and Baker, D. 1997. Functional Rapidly Folding Proteins from Simplified Amino Acid Sequences. *Nature Structural Biology* 4(11): 805–809.

Rost, B., and Sander, C. 1993. Prediction of Protein Secondary Structure at Better Than 70% Accuracy. *Journal of Molecular Biology* 232(2): 584–599.

Rost, B.; Sander, C.; and Schneider, R. 1994. Redefining the Goals of Protein Secondary-Structure Prediction. *Journal of Molecular Biology* 235(1): 13–26.

Sander, C., and Schneider, R. 1991. Database of Homology-Derived Protein Structures and the Structural Meaning of Sequence Alignment. *Proteins* 9(1): 56–68.

Sejnowski, T. J., and Rosenberg, C. R. 1987. Parallel Networks That Learn to Pronounce English Text. *Complex Systems* 1(1): 145–168.

Wan, W. Y., and Milner-White, E. J. 1999. A Recurring Two-Hydrogen-Bond Motif Incorporating a Serine or Threonine Residue Is Found Both at Alpha-Helical *N-termini* and in Other Situations. *Journal of Molecular Biology* 286(5): 1651–1662.

Wang, J., and Wang, W. 1999. A Computational Approach to Simplifying the Protein-Folding Alphabet. *Nature Structural Biology* 6(11): 1033–1038.

Wintjens, R.; Wodak, S. J.; and Rooman, M. 1998. Typical Interaction Patterns in fffi and fiff Turn Motifs. *Protein Engineering* 11(7): 505–522.

Wouters, M. A., and Curmi, P. M. 1995. An Analysis of Side-Chain Interaction and Pair Correlation within Anti-Parallel Beta-Sheets: The Difference between Backbone Hydrogen-Bonded and Non–Hydrogen-Bonded Residue Pairs. *Proteins* 22(2): 119-–131.

Claus A. F. Andersen has an MSc in engineering from the Technical University of Denmark (DTU) (1998) and a Ph.D. in bioinformatics, also from DTU (2001). His university research has focused on protein structure and function analysis and prediction. His industry research has dealt with large-scale EST and expression analysis, high-throughput mass spectrometry and two-dimensionnal–gel data interpretation, and disease-focused pathway analysis for regulatory and metabolic pathopathways. Andersen works for Siena Biotech SpA. His e-mail address is ca2@cbs.dtu.dk.

Søren Brunak has an MSc in physics from the University of Copenhagen (1987) and a Ph. D. in computational biology from the Technical University of Denmark. He is a professor of bioinformatics at the Center for Biological Sequence Analysis, the Technical University of Denmark, which he founded in 1993 with a grant from the Danish National Research Foundation. Brunak has more than 12 years of experience in computational biology. He is a reviewer for numerous scientific journals; has coauthored more than 100 scientific articles; and has coauthored or edited 6 books on bioinformatics, machine learning, and protein folding. His e-mail address is brunak@cbs.dtu.dk.