

Yule Value Tables from Protein Datasets

Madhavi GANAPATHIRAJU

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213 USA
madhavi+web@cs.cmu.edu

Deborah WEISSER

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213 USA
dweisser@cs.cmu.edu

Judith KLEIN-SEETHARAMAN

Department of Pharmacology
University of Pittsburgh School of Medicine
Pittsburgh, PA 15261 USA
and
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213 USA
judithks@cs.cmu.edu

ABSTRACT

Here, we studied systematically the association between amino acids, the constituents of protein sequences in datasets of different hierarchy, i.e. genome (human), protein type (membrane proteins), protein family (specific types of membrane receptors and transporters) and transmembrane helices versus loops (either for membrane proteins in general or family-specifically). Association was estimated using Yule's Q statistics for pairs of amino acids within a window of size 4. Strong association between such nearby amino acids was observed in all the datasets studied, in contrast to the randomized datasets. Association strength increased as expected when the datasets were more specific. Strikingly, in transmembrane helices, associations were more negative than in any other dataset studied, suggesting that evolution of these helices requires suppression of occurrence of specific amino acid combinations within local range. The results have direct applicability to several areas of bioinformatics research, i.e. transmembrane helix boundary prediction, sequence alignment and understanding of design principles of membrane proteins in general. Data and access to the algorithms presented in this paper are available at <http://flan.blm.cs.cmu.edu/>

Keywords: Yule association measure, structural conservation, G protein coupled receptors, membrane proteins, interdisciplinary approach, language analogy.

INTRODUCTION

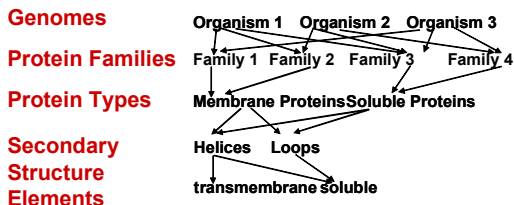
The mapping of biological sequences to form and function of proteins is conceptually similar to the mapping of words to meaning in natural language. The challenge in protein sequence analysis is to identify words that map to a specific meaning in terms of the structure of the protein – the greatest challenge being identification of word-equivalents in protein sequence language. Understanding the protein structures encoded in the human genome sequence has therefore been dubbed reading the book of life. Knowledge/rule based, machine learning (such as hidden Markov models, neural networks) and hybrid methods have been used to capture meanings from a sequence of words in natural languages and prediction of protein structure and

function alike. For example, protein secondary structure prediction and natural language processing aim at studying higher order information from composition, while tackling problems like redundancy and multiple interpretations. In recent years, jargon from natural language processing, such as text segmentation, data compressibility, Zipf's law, grammatical parsing, n-gram statistics, text categorization and classification, and linguistic complexity has become common in biological sequence processing. The exploitation of the analogy between language and the molecular biology domain has been pioneered by David Searls who developed a grammar to represent biological sequences [1]. For example, coding and non-coding regions in DNA can be distinguished as grammatically correct or incorrect and RNA secondary structure can be described by various types of grammars (reviewed in [2]). Furthermore, it was demonstrated that genome or protein sequences follow Zipf's law [3-9], but since non-deterministic sequences also follow a power law (see e.g. [10]), the extent to which amino acid sequences can be modeled stochastically in the linguistic analogy is not clear [11]. Renewed support for the hypothesis that the linguistic analogy is valid came recently from the observation that in whole-genome protein sequences the use of n-grams of length 4 can serve as genome signatures [12], suggesting some degree of word- or phrase-like character in amino acid 4-grams. Furthermore, linguistic methods have been applied specifically to the analogy between mapping words to meaning and biological sequences to structure and function of proteins. For example context-free grammars can describe models used to predict physical properties of proteins, including stable conformations and conformational changes based on computation of the partition function [13]. Furthermore, latent semantic analysis (LSA), a hybrid between signal processing and language processing techniques has been used for secondary structure prediction [14]. In natural language text processing, LSA is used to extract hidden relations between words. It is used to find words that address the same topic or are synonymous, even when such information is not explicitly

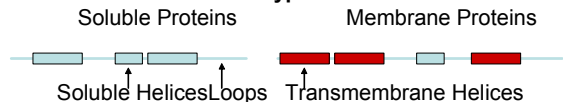
available. In the way LSA captures conceptual relations in text, based on the distribution of words across text documents, it was used to capture secondary structure propensities in protein sequences and different biological vocabularies were analyzed to characterize sequence-structure relationships in proteins. Segmentation of proteins at the domain level has also been shown to profit from language modeling techniques used in speech recognition [15].

Here, we present related work, in which we demonstrate that statistical association measures used normally for text segmentation can identify similarities between various datasets of protein sequences selected based on a hierarchy of biological meanings. In language, statistical association measures such as mutual information [16] which correlates groups of nearby words was successfully used to determine sentence boundaries [17]. All potential sentence boundaries in a sequence of words are considered and mutual information values are computed for every pair of adjacent words. Pairs with low values are candidates for sentence boundaries. The mutual information values are computed by examining overlapping fixed-sized sequences of adjacent words in a large text. In analogy to this application in language, mutual information has been used recently to identify functional building blocks in protein sequences [18]. A different association measure was used by Cai, Rosenfeld, and Wasserman [19] to categorize text. They use Yule’s measure of association (Yule’s Q statistic, or the Yule value) [20]. Given a text, which contains n different words, an $n \times n$ table of Yule values for every pair of words is computed by tabulating co-occurrences of words in the text. The distribution of Yule values in the table was shown to differ for different categories of text.

A. Hierarchy of Protein Sequence Datasets



B. Schematic of Protein Type



C. G Protein Coupled Receptors Helix Arrangement

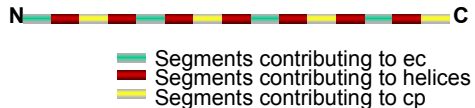


Figure 1. Schematic of the hierarchical relationship between different protein datasets studied in this paper.

The goal of the present paper is to investigate systematically the entries of Yule values in tables determined from protein sequences. Instead of words, Yule’s Q statistic in this application measures the association between amino acids, the constituents of protein sequences. What types of categories are there in protein sequences? At the highest level of generality, there are the proteins from a given organism. The presence of genome signatures described above [12] suggests that different genomes can be viewed as different languages. There are different ways in which genome sequences can be

separated into different categories. We explored (1) protein type (membrane proteins versus soluble proteins), (2) protein family (specific types of membrane receptors and transporters) and (3) secondary structure (transmembrane and soluble helices and loops), shown schematically in Figure 1. The latter categorization, secondary structure, subdivides the first two, protein type or a protein family, into further subcategories. Inspection of three-dimensional structures of proteins has revealed the presence of repeating elements of regular structure, termed secondary structure. These regular structures are stabilized by molecular interactions between atoms within the protein, the most important being the Hydrogen (H) Bond. H-bonds are non-covalent bonds formed between two electronegative atoms that share one H. There is a convention on the nomenclature designating the common patterns of H-bonds that give rise to specific secondary structure elements, the Dictionary of Secondary Structures of Proteins (DSSP) [21]. DSSP annotations mark each residue (amino acid) to be belonging to one of seven types of secondary structure: Here, we only considered two groups, helices and non-helices. These are further distinguished by the type of protein. In membrane proteins, helices can be either transmembrane (helices that span lipid bilayers in cells) or soluble (folded in the aqueous environment), while soluble proteins only contain soluble helices. Here, we focus on membrane protein families and their secondary structure elements because membrane proteins are the least well studied types of proteins due to the experimental difficulties in working with them. Yule values were computed for the human genome and compared to a specific protein family, the G protein coupled receptors (GPCR), shown schematically in Figure 1C. The helices in GPCR transverse the membrane, so the soluble loops connecting the helices can be either outside or inside the cell, referred to in the following as ec (extracellular) and cp (cytoplasmic) in the text. The GPCR family in turn was compared to all of the families of those membrane proteins with known structure [22], in particular with respect to their secondary structure elements (Figure 1C) and their relationship to soluble proteins. For each protein subset of different hierarchy, Yule’s Q statistic was computed as described below.

When applying Yule’s Q statistic to analyze the co-occurrence of nearby (within a window of size 4) amino acids in protein sequences, we observe that the distribution of Yule’s Q statistic values varies in quantifiable ways depending upon the data analyzed. This is in direct analogy to the observed differences in Yule values derived from words in texts of different categories. Therefore, the distribution of Yule Q statistic values can be used to predict sequence type and feature positions in protein sequences.

SYSTEMS AND METHODS

Yule’s Q statistic (which is referred to as the Yule value below) is defined in Eq. 1. It is a measure of association between two variables, which always takes a value between -1 and 1. A positive value implies that the variables are positively correlated. Likewise, negatively correlated variables have a negative Yule value. The Yule value was calculated by dividing the input sequence into sub-regions (windows) and calculating pairwise information for each sub-region as described [19]. The Yule value is:

$$Y = \frac{Y_{11}Y_{00} - Y_{10}Y_{01}}{Y_{11}Y_{00} + Y_{10}Y_{01}} \quad (1)$$

where,

Y_{11} = the number of times both i and j co-occur in a sub-region,

Y_{00} = the number of times neither i nor j occurs in a sub-region,

Y_{10} = the number of times i occurs and j does not occur in a sub-region,

Y_{01} = the number of times i does not occur and j occurs in a sub-region.

Note that the sum of Y_{11} , Y_{00} , Y_{10} , and Y_{01} is the number of sub-regions.

Algorithm and Implementation

Our algorithm makes two passes over amino acid sequences. The input to the first pass is not required to be the same as the input to the second pass. The first pass takes as input an amino acid sequence and outputs a 20 x 20 table of Yule values for each pair of amino acids (i,j), where, $i, j \in \{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}$. In the second pass, the input is an amino acid sequence, and the output is the Yule values for all adjacent pairs of amino acids along the input sequence. Details of the algorithm implementation are described below.

The Yule value for each pair of amino acids was computed by scanning the entire sequence by sliding windows. Let, $S = \{S_1, \dots, S_N\}$ be an amino acid sequence, and W is the window size. The first window in the sequence is the set $\{S_1, \dots, S_W\}$. The i^{th} window is $\{S_i, \dots, S_{W+i-1}\}$. The Yule value is computed by determining the following for each pair (i,j) in the sequence: Y_{11} , the number of times both i and j co-occur in a window of size W ; Y_{00} , the number of times neither i nor j occurs in a window of size W ; Y_{10} , the number of times i occurs and j does not occur in a window of size W ; Y_{01} , the number of times i does not occur and j occurs in a window of size W . Since in a sequence of length N there are $N-W+1$ windows, for each pair of amino acids (i,j), $Y_{11} + Y_{00} + Y_{10} + Y_{01} = N-W+1$.

|S|H|D|E|G|C|L|S|S|E|P|K|P|R|K|Q|S|D|S|S|T

Figure 2. Sliding window along a protein sequence.

Consider the example sequence shown in Figure 2, where the window size W is 4, the value used for all calculations described in this paper. In this sequence, for the pair (S,D), the components of the Yule value are as follows: $Y_{11} = 5$, $Y_{00} = 5$, $Y_{10} = 6$, $Y_{01} = 2$. For the pair (S, D), as per Eq. 1, the Yule value is 0.35.

Data Sources

A list of membrane proteins with known 3-dimensional structure was obtained from Stephen White's homepage [23]. The corresponding sequences were extracted from the protein data bank [24]. Sequences were used to retrieve family members of a given membrane protein from the Pfam database [25]. The GPCR input data was taken from the SwissProt [26] and GPCR database [27, 28]. Segments based on secondary structure (helices, transmembrane helices, loops) were extracted from the SwissProt entries.

Sample Sizes and Randomization: The output of the first pass of a Yule calculation is a 20 x 20 table of Yule values for all amino acid pairs. We studied the effects of dataset size on the distribution of values in the Yule table. We varied the sample sizes from approximately 6000 to 11,400,000 amino acids from the entire set of human protein sequences. As an

additional control, we also randomized the protein sequence datasets for all samples.

RESULTS AND DISCUSSION

The first pass of the calculation yields a 20x20 matrix of Yule values that indicate the preference of pairs of amino acids in the neighborhood of each other for the given input sequence data.

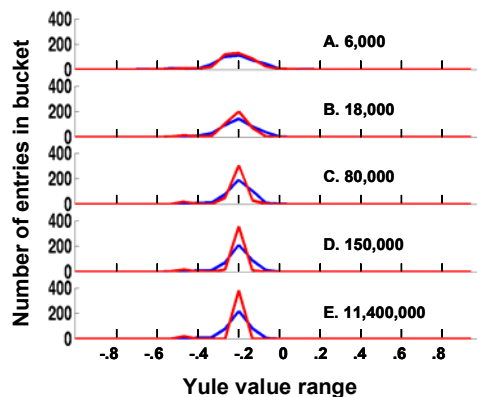


Figure 3. Histograms of Yule table values for the human genome. Different dataset sizes were studied to estimate the effect of dataset size on the Yule value distribution—human (blue), randomized human (red).

In the second pass, a specific input sequence is annotated with precomputed Yule values, corresponding to the amino acid combination occurring in the input sequence. This allows evaluating how similar or distinct a given sequence is in comparison to a well defined class of sequences, such as helices, loops, or in comparison to the characteristics of a family of proteins or a proteome of a specific organism.

Yule tables derived from an entire genome: the human genome

First, we determined the Yule table for an entire genome, i.e. the human genome. The result is shown in Figure 3E. Since the data size of a genome is much larger than any of the other datasets studied, we investigated the effect of size on the Yule table. This is plotted in Figure 3A-D, which shows the effect of dataset size on the contents of the 20x20 Yule table, which are described in the form of 30-bucket histograms. The algorithm was run on subsets of the human genome sequence with dataset sizes varying from 6000 amino acids to 11,400,000 amino acids (the entire sequence), shown in blue in Figure 3. For comparison, results for the same datasets randomized are shown in red. As expected, the entries in the Yule table tend to cluster at the mean as the dataset size increases. Also as expected, the number of entries near or at the mean is significantly higher for the randomized human sequence data than for the human sequence data. Furthermore, in the human sequence the distribution of the Yule table appears to stabilize when the dataset size reaches 150,000 amino acids for both randomized and non-randomized data.

The above qualitative observations were then studied quantitatively. The number of entries in the Yule table of the largest histogram bucket was determined as a function of sequence size for a large number of samples of a given size as follows. For a datapoint at (x_k, y_k) in a sequence of size N : x_k is the size of the subsequence. The sequence was partitioned into approximately N/x_k subsequences of size x_k . h_i is the histogram

of the i^{th} subsequence of size x_k . s is the number of buckets in the histogram [note: In our experiments, $s=30$]. h_{ki1}, \dots, h_{kis} are the values in the i^{th} histogram (i.e. h_{kij} contains the number of Yule table entries in bucket j for the histogram of the Yule table of subsequence i). $y_{ki} = \max\{h_{ki1}, \dots, h_{kis}\}$. $y_k = (y_{k1} + \dots + y_{kN/x_k}) / (N/x_k)$, i.e. y_k is the average size of largest sized histogram bucket for a subsequence of size x_k .

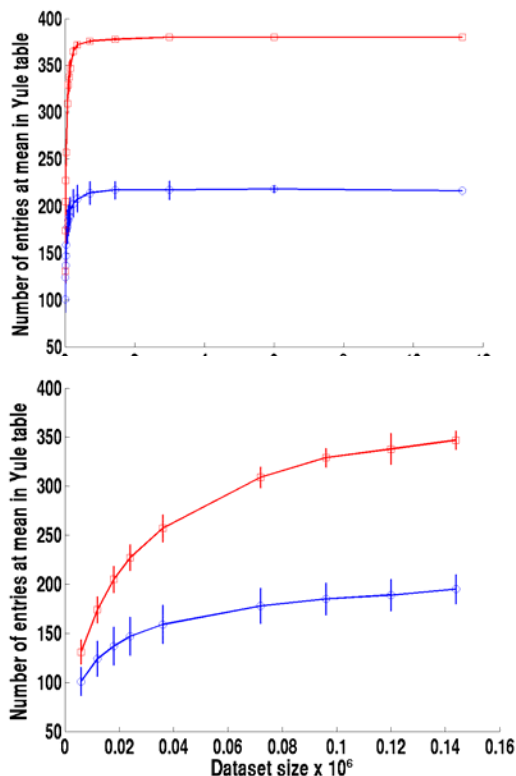


Figure 4. A. Average Yule table maximum histogram values for different dataset sizes. B. Blowup of initial region of Figure A. Vertical bars denote standard deviation. Blue lines indicate human protein sequence data, red lines indicate randomized human protein sequence data.

Thus for a dataset of size n , the average of approximately N/n values was determined. The results are shown in Figure 4. The blue lines are the averages of the non-randomized subsets, and the red lines are the averages of the randomized sequences. The vertical lines indicate the standard deviations. Using this approach, it is possible to quantify at what data sizes the Yule table reaches stable contents. Figure 4A shows the dependence of number of entries at mean in the Yule table for the entire range of data sizes studied. Figure 4-B is an expansion of the first 150,000 entries of Figure 4-A, the approximate range in which the values stabilize.

In Figures 4-A and 4-B, notice that until the dataset size is quite large, the standard deviation remains essentially the same. This is because there are two parameters, the dataset size and the number of samples, which affect the standard deviation differently. As the number of samples increases (and thus the sample size decreases), the standard deviation tends to decrease. On the other hand, as the dataset size increases (and thus the number of samples decreases), the standard deviation tends to decrease. In other words, the sample size and the dataset size tend to have opposite effects on the standard deviation until the sample size becomes large. Since the product of the number of

samples and the dataset size is constant, it is not surprising that they cancel each other out initially.

Figure 4 also quantifies the difference between randomized and non-randomized protein sequences. The number of entries at the mean in the Yule table at small data sizes (6000 amino acids) is closer to that of the randomized data, although there is already a significant difference. Steadily, a striking discrepancy develops between real and randomized data. The low value of entries at the mean reflects a larger spread in the Yule values obtained, indicating stronger correlations. This result demonstrates that correlations significantly different from random datasets can be obtained already at relatively small dataset sizes (6000 amino acids), but larger sizes increase this difference and dataset sizes of 70,000 and higher are desirable to exploit nearly full information content from the Yule association measure.

Yule table histograms for a protein family: GPCR (type: membrane protein)

Next, we investigated Yule's association measure for the GPCR sequences and subsequences corresponding to the individual domains. Domains were defined as cp, helices and ec (Figure 1C). The sequences were retrieved from the GPCR database [28] and the feature information was extracted from the SwissProt entries for each GPCR sequence [26]. The respective segments of GPCR sequences (see Figure 1C) were concatenated. Since the termini of the sequences may have somewhat different properties from other extramembraneous regions because of the decreased restriction in conformational space, we also created ec and cp subsets lacking the N- and C-termini, respectively (referred to as ec-nt and cp-ct). The large size of the GPCR family makes this family of proteins an ideal study case for exploiting the maximum information content of Yule's association measures.

Figure 5-A shows 30-bucket histograms of Yule tables created from the different regions of GPCR (the entire GPCR, helical, cp, cp minus the C-terminus, ec, and ec minus the N-terminus). In Figure 5-B, the input sequences are randomized. As observed with the entire human protein sequences above (Figures 3 and 4), significant correlations are observed in the GPCR sequences also and randomization has the effect of eliminating correlation between pairs of amino acids.

Figure 6 shows similar information as Figure 5, but highlighting the specific entries in the Yule table. In Figure 6A, the Yule values are sorted according to the pair of amino acids (sorted alphabetically). As one can see, correlations and anti-correlations are evenly spread throughout the different combinations, and all datasets are clearly distinct from the randomized dataset, shown for the example of GPCR in brown. The randomized data cluster around -0.2, the expected Yule value, with little variation, which is very distinct from all other datasets. The human genome shows the least deviation (albeit very distinct) from random. The helices dataset shows the strongest deviation from random, in particular with respect to anti-correlations. This observation is highlighted in Figure 6B. The blue line, corresponding to GPCR helices, has a much higher slope towards negative values than the other datasets. This indicates, that evolution of sequences in GPCR helices strongly disfavors certain combinations of amino acids.

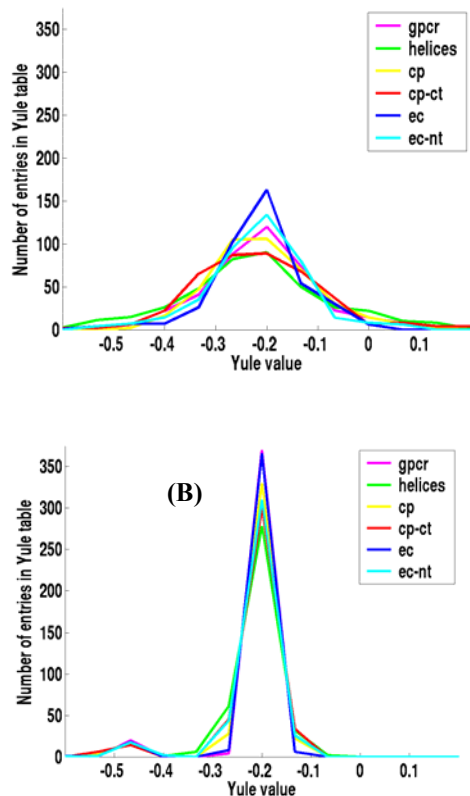


Figure 5. A. Yule table histograms for different domain segments of GPCR. B. Yule table histograms for the same segments as in A., but with randomized input data.

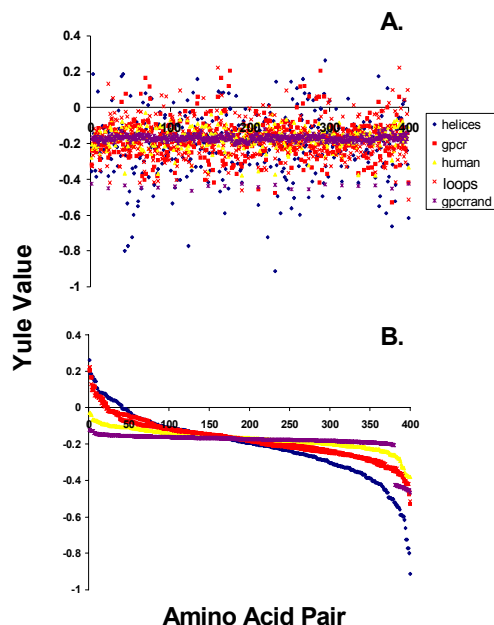


Figure 6. Entries in the Yule table shown by amino acid pair. A. The 400 amino acid pairs are sorted alphabetically. Each x-value corresponds to the same amino acid pair. B. Each Yule table is sorted from highest to lowest Yule value. Each x-value corresponds to a different amino acid pair.

Yule values obtained from different GPCR segments along specific GPCR sequences

After having explored the contents of the 20 x 20 Yule tables for different datasets, we then studied the ability of the Yule values to partition samples of entire GPCR sequences, i.e. specific full-length GPCR protein sequences, into their individual domain segments. The following set of figures shows Yule values for two representative example GPCR sequences, one sequence, gp43_human, from Class A GPCR and one sequence, casr_bovin, from Class C GPCR. These two classes of GPCR share less than 20% sequence identity in the transmembrane domains and thus, the two sequences are essentially unrelated.

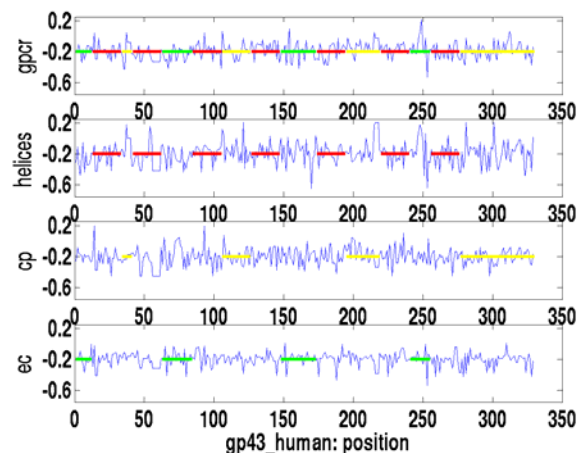


Figure 7. Yule values along gp43 human (a class A receptor) sequence using different input sources to create the Yule table. Input data was the entire GPCR dataset, helices, cp and ec domains. Horizontal lines using the same color code as in Figure 1 indicate the positions of the segments belonging to each of these domains.

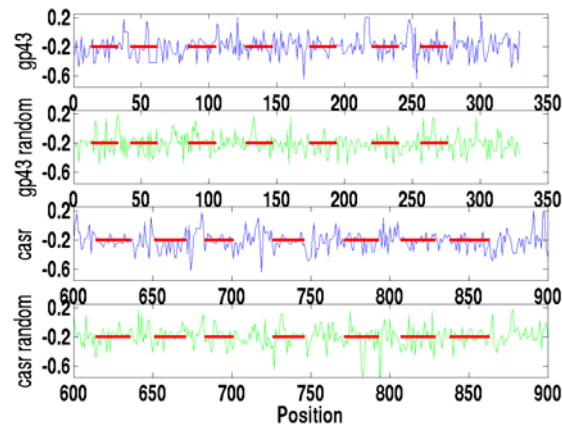


Figure 8. Comparison between Yule table generated from helices and randomized helices data along two GPCR sequences, gp43_human (a classA receptor) and casr_bovin (a class C receptor). For casr_bovin, the first 600 amino acid positions were omitted since these belong to the N-terminus.

Figure 7 shows plots of Yule values generated from the entire GPCR dataset and for helices, cp and ec subsets as described above along the gp43_human sequence in comparison to the actual positions of the segments in the sequence that were part of the respective domain dataset. cp-ct and ec-nt datasets were

also used, but differences to cp and ec datasets were small and are therefore omitted in the figure. The color code is the same as in Figure 1C. The red horizontal lines in the figure indicate the position of the helices, the green lines indicate the ec regions, and the yellow lines indicate the cp regions. We observe in Figure 7 that Yule values generated from the entire GPCR dataset tend to peak outside of the helical regions. This effect is enhanced if the helices-only dataset, but not the ec or cp datasets, is used to generate the Yule table.

In Figure 8, this observation is explored in more detail by comparing plots of Yule values using tables from helical data and randomized helical data for the two input sequences, gp43_human and casr_bovin. The first plot in each pair (shown in blue) is the Yule values along the sequence, using the Yule table created from helical GPCR regions. The second plot in each pair (shown in green) is the Yule values along the sequence using the Yule table created from randomized helical GPCR regions. In both gp43_human and casr_bovin, we note that in the non-randomized plots, the peaks almost always occur outside of helices, while in the plots using randomized helical Yule tables, the peaks occur throughout the sequence. Similar observations were made with other GPCR sequences from any of the different GPCR classes (data not shown).

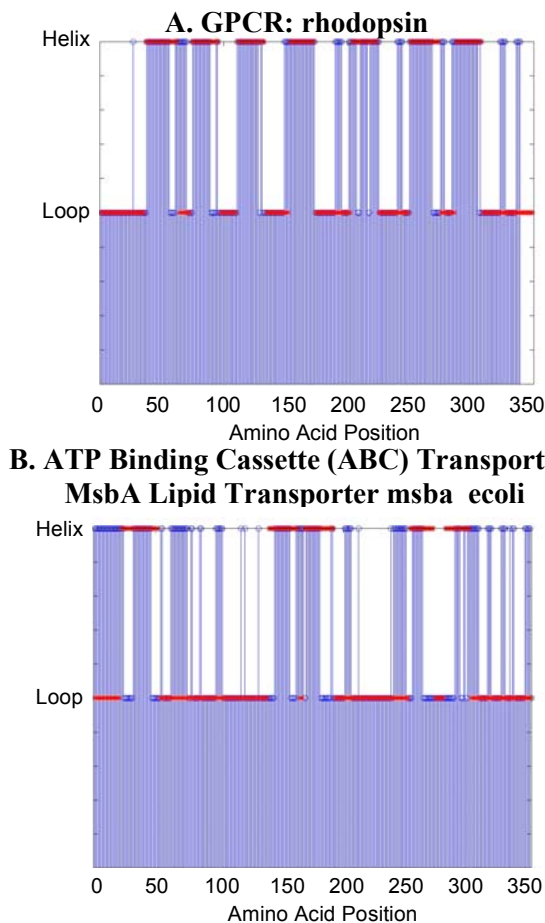


Figure 9. Segmentation of members of two protein families. A. G-protein coupled receptors, example rhodopsin (pdb file 1F88) using Yule's Q statistics. B. Transporter, MsbA Lipid Transporter (pdb file 1JSQ).

Prediction of segment boundaries in various membrane protein families

Next, we designed an algorithm for the prediction of helices and loops in membrane proteins based on the Yule values. The Yule values were computed for amino acid pairs in neighboring positions, when these pairs occur in helices, and when they occur in loops as described above for the GPCR family. An initial classification is made for each residue position in the protein depending on whether the Yule value is high for helices or loops for the amino acid pair appearing at that position. The classification is then smoothed by sliding a window of size 5, and using a majority scoring within each window. In Figure 9, predictions using this method are shown in blue, and the actual labelling of helices and loops is shown in red. A value of 1 indicates loop and a value of 2 indicates helix in both predicted and actual labels. Figure 9A shows the result for the GPCR rhodopsin (sequence ID opsd_human). As one can see, there is strong correlation between the positions of helices and those predicted by the algorithm.

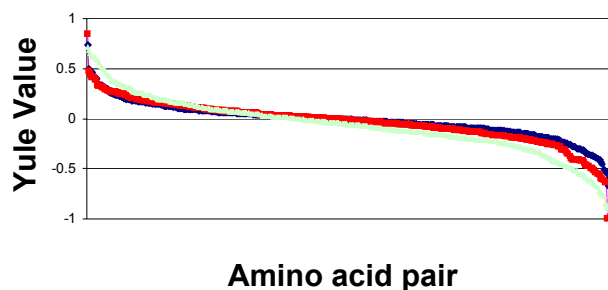


Figure 10. Comparison of transmembrane (blue), soluble (red) and GPCR (green) helices.

Comparison of position-specific Yule table entries for helices of different type: transmembrane, soluble and GPCR helices

Figure 10 shows the Yule values sorted for each of three datasets, helices from soluble proteins (red), transmembrane helices from membrane proteins (blue) and transmembrane helices specifically from the GPCR family (green). However, instead of computing Yule values over the window of size 4, we now investigated pairs with position-specific information, i.e. AB, A*B, A**B, A***B, where A and B are the amino acids to be correlated and * is a wild-card character. Only the curve for AB is shown in Figure 10, but the other three curves reproduced the trends seen: the family specific dataset shows stronger correlations than either helices from membrane or from soluble proteins. In particular, the negative correlations are much stronger in the GPCR helices dataset. This confirms the previous observations with the Yule values computed over a window of size 4.

The extent of correlations can be seen in more detail in Figure 11(A-B), in which the numbers of entries with Yule values above 0.25 (positive correlations) and below -0.25 (negative correlations) are reported. In all three datasets, the negative correlations (upper panel of the graph) outweigh the positive correlations (lower panel of the graph). The strongest negative correlations are invariably found in the soluble and transmembrane helices datasets, regardless of the number of wild-cards separating the two amino acids. Figure 11(C-D) shows the degree of overlap of amino acid pairs in the most

SUMMARY AND CONCLUSIONS

highly correlated pairs shown in Figure 10. The number of pairs found in each comparison is strongly dependent on the relative position of the two amino acids with respect to each other and whether under-represented (Figure 11A) or over-represented (Figure 11B) pairs were studied. While overall the overlap between transmembrane and GPCR helices is greatest, there is a very significant overlap between GPCR helices and soluble helices also. This striking observation can be explained by the fact that transmembrane helices have amphipathic character, i.e. the residues facing the inside of a helical bundle are often facing aqueous cavities or polar residues from the other helices. The conservation of polar residues in multiple sequence alignments of membrane proteins has been used for example in the case of the GPCR family to deduce relative orientations of the helices with respect to each other [29].

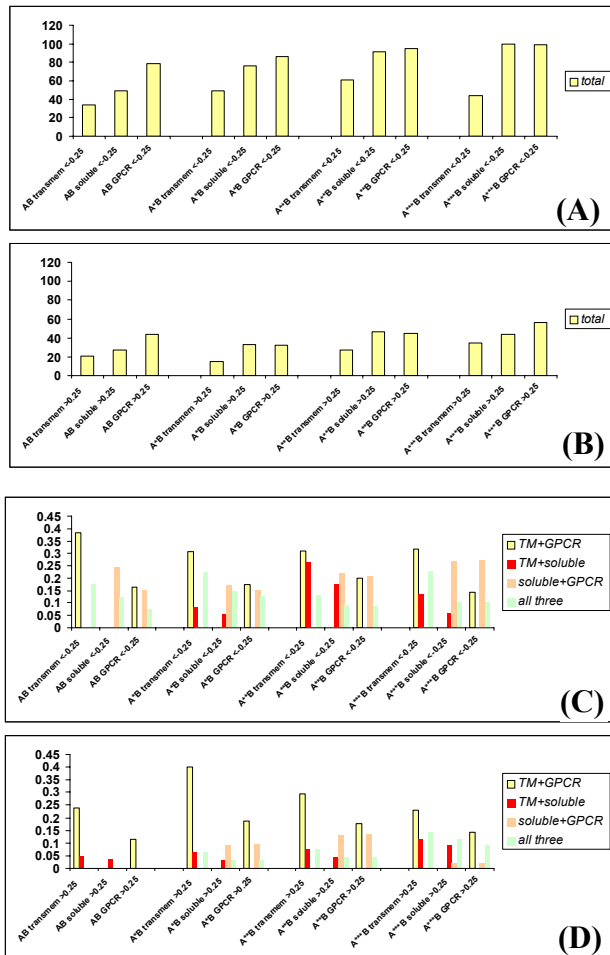


Figure 11. (A-B) Most over and underrepresented pairs. The axis shows the percent of entries in upper and lower part of the graphs shown in Figure 10. Number of entries in the upper and lower part of the graph shown in Figure 9. The entries with values below -0.25 and above 0.25 are plotted in the (A) and (B) respectively. **(C-D). Overlap between the helices datasets.** For each highly over- or under-represented pair from Figure 10, the degree of overlap between transmembrane and GPCR helices (yellow), between transmembrane and soluble helices (red) between soluble and GPCR helices (orange) and between all three datasets (green) are indicated.

The derivation of features from primary protein sequences that are indicative of structural conservation where there is little sequence conservation is an important and open problem in structural and functional proteomics. Even where comparatively abundant structural information is available, such as for the common folds attained by soluble proteins, there is little consensus as to what features of the sequence determine the particular fold [30-33]. Here we attempt to derive similarities between protein sequences without alignment of the related proteins. The approach is based on a statistical association measure, Yule's measure of association. We studied the Yule tables derived from different categories of protein sequences. Categories include genome, membrane proteins, secondary structure elements and protein families, and various combinations of the Venn diagrams describing the relationship between these categories (Figure 1). We focus particularly on membrane proteins because they are notoriously difficult to study experimentally because of their hydrophobic nature. Since little structural information is available for individual members of membrane protein families, computational approaches are particularly needed to derive hypotheses about structure-function relationships in these proteins from their sequences. The most urgent question generally in the field of study of a particular protein family, such as the GPCR family, regards the degree of generality of structure and mechanisms of action of individual members. Membrane proteins are usually pharmacologically important because drugs often interact with proteins accessible from the extracellular surface, and identifying common principles of structure and action of members in a given family has important implications for rational drug design directed toward these pharmacologically relevant families of proteins.

Yule's measure of association between nearby amino acids in protein sequences varies in quantifiable ways depending upon the category studied. This result indicates that local contacts can be derived from sequence information alone and can be used as features of protein sequences. These features, extracted from primary amino acid sequence data, capture a more structural level of conservation than is evident in the primary amino acid sequence directly. Furthermore, this approach can now be used to determine feature boundaries in given sequences such as GPCR which share little or no significant sequence homology. This will be useful in aiding experimental studies aimed at testing the generality hypothesis of GPCR structure and function by generating chimeric GPCR proteins with ec, helices and/or cp domains swapped between members of the GPCR family that show little sequence similarity.

Since the pairwise amino acid sequence features, extracted here from primary amino acid sequence data, capture a more structural level of conservation than primary amino acid sequence directly, the results are expected to have impact in a number of bioinformatics research areas. These include secondary (soluble and transmembrane) structure prediction, sequence alignment of remote homologues and possibly tertiary contact prediction. The approach presented here is complementary to existing secondary structure prediction methods for soluble and transmembrane proteins, because it can capture differences and general similarities between properties in specific areas of a protein sequence (for a review of secondary structure prediction methods see [34] and for transmembrane prediction [35, 36]). In particular, it will be interesting to compare for existing membrane protein structures, where correlations that are compatible with most

transmembrane protein structures are with those where the similarity to soluble helices is more prominent. Finally, the identification of family-specific preferences using the Yule tables will likely identify sites important for the specific function of the protein family under study.

ACKNOWLEDGEMENTS

We would like to thank the National Science Foundation for financial support from ITR grant 0225656.

REFERENCES

- [1] D. Freedman, **AI helps researchers find meaning in molecules**, *Science*, vol. 261, pp. 844-5, 1993.
- [2] D. B. Searls, **The language of genes**, *Nature*, vol. 420, pp. 211-7, 2002.
- [3] O. Weiss, M. A. Jimenez-Montano, and H. Herzel, **Information content of protein sequences**, *J Theor Biol*, vol. 206, pp. 379-86., 2000.
- [4] W. Li, **Statistical properties of open reading frames in complete genome sequences**, *Comput Chem*, vol. 23, pp. 283-301., 1999.
- [5] A. Czirok, R. N. Mantegna, S. Havlin, and H. E. Stanley, **Correlations in binary sequences and a generalized Zipf analysis**, *Physical Review. E. Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, vol. 52, pp. 446-452., 1995.
- [6] N. E. Israeloff, M. Kagalenko, and K. Chan, **Can Zipf distinguish language from noise in noncoding DNA?**, *Physical Review Letters*, vol. 76, pp. 1976., 1996.
- [7] A. K. Konopka and C. Martindale, **Noncoding DNA, Zipf's law, and language**, *Science*, vol. 268, pp. 789., 1995.
- [8] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C. K. Peng, M. Simons, and H. E. Stanley, **Linguistic features of noncoding DNA sequences**, *Phys Rev Lett*, vol. 73, pp. 3169-72, 1994.
- [9] A. A. Tsonis, J. B. Elsner, and P. A. Tsonis, **Is DNA a language?**, *J Theor Biol*, vol. 184, pp. 25-9., 1997.
- [10] C. A. Chatzidimitriou-Dreismann, R. M. Streffer, and D. Larhammar, **Lack of biological significance in the 'linguistic features' of noncoding DNA--a quantitative analysis**, *Nucleic Acids Res*, vol. 24, pp. 1676-81., 1996.
- [11] D. B. Searls, **Linguistic approaches to biological sequences**, *Comput Appl Biosci*, vol. 13, pp. 333-344, 1997.
- [12] M. Ganapathiraju, D. Weisser, R. Rosenfeld, J. Carbonell, R. Reddy and J. Klein-Seetharaman, **Comparative n-gram analysis of whole-genome protein sequences**, presented at Human Language Technologies Conference, San Diego, 2002.
- [13] D. Chiang and A. K. Joshi, **Formal grammars for estimating partition functions of double-stranded chain molecules**, presented at Proc. of the Human Language Technologies Conference, San Diego, CA, 2002.
- [14] M. Ganapathiraju, J. Klein-Seetharaman, N. Balakrishnan, and R. Reddy, **Characterization of Protein Secondary Structure using Latent Semantic Analysis**, *IEEE Signal Processing Magazine*, vol. 21, pp. 78-87, 2004.
- [15] L. Coin, A. Bateman, and R. Durbin, **Enhanced protein domain discovery by using language modeling techniques from speech recognition**, *Proc. Natl. Acad. Sci. USA*, vol. 100, pp. 4516-4520, 2003.
- [16] R. Fano, **Transmission of information**. New York, NY: The MIT Press, 1961.
- [17] D. M. Magerman and M. P. Marcus, **Parsing a Natural Language Using Mutual Information Statistics.**, presented at Proceedings, Eighth National Conference on Artificial Intelligence, 1990.
- [18] D. Weisser and J. Klein-Seetharaman, **Identification of Fundamental Building Blocks in Protein Sequences Using Statistical Association Measures**, Proc. ACM Symposium on Applied Computing, Nicosia, Cyprus, 2004.
- [19] C. Cai, R. Rosenfeld, and L. Wasserman, **Exponential Language Models, Logistic Regression, and Semantic Coherence.**, presented at Proc. NIST/DARPA Speech Transcription Workshop, 2000.
- [20] Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland, **Discrete Multivariate Analysis: Theory and Practice**. Cambridge, MA: The MIT Press, 1975.
- [21] W. Kabsch and C. Sander, **Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features**, *Biopolymers*, vol. 22, pp. 2577-2637, 1983.
- [22] <http://compbio.ornl.gov/Grail-1.3/>.
- [23] http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html.
- [24] <http://www.rcsb.org/pdb/>.
- [25] A. Bateman, E. Birney, L. Cerutti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, and E. L. L. Sonnhammer, **The Pfam protein families database**, *Nucleic Acids Res*, vol. 30, pp. 276-280, 2002.
- [26] <http://us.expasy.org/sprot/>.
- [27] F. Horn, R. Bywater, G. Krause, W. Kuipers, L. Oliveira, A. C. Paiva, C. Sander, and G. Vriend, **The interaction of class B G protein-coupled receptors with their hormones**, *Receptors Channels*, vol. 5, pp. 305-14, 1998.
- [28] <http://www.gpcr.org>.
- [29] J. M. Baldwin, G. F. X. Schertler, and V. M. Unger, **An Alpha-Carbon Template for the Transmembrane Helices in the Rhodopsin Family of G-Protein-Coupled Receptors**, *J. Mol. Biol.*, vol. 272, pp. 144-164, 1997.
- [30] P. Bork, L. Holm, and C. Sander, **The immunoglobulin fold. Structural classification, sequence patterns and common core**, *J Mol Biol*, vol. 242, pp. 309-20, 1994.
- [31] D. M. Halaby, A. Poupon, and J. Mornon, **The immunoglobulin fold family: sequence analysis and 3D structure comparisons**, *Protein Eng*, vol. 12, pp. 563-71, 1999.
- [32] R. B. Russell and G. J. Barton, **Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility**, *J Mol Biol*, vol. 244, pp. 332-50, 1994.
- [33] L. A. Mirny and E. I. Shakhnovich, **Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function**, *J Mol Biol*, vol. 291, pp. 177-96, 1999.
- [34] B. Rost, **Protein secondary structure prediction continues to rise**, *J. Struct. Biol.*, vol. 134, pp. 204-218, 2001.
- [35] C. P. Chen, A. Kernysky, and B. Rost, **Transmembrane helix predictions revisited**, *Protein Sci*, vol. 11, pp. 2774-91, 2002.
- [36] C. P. Chen and B. Rost, **State-of-the-art in membrane protein prediction**, *Applied Bioinformatics*, vol. 1, pp. 21-35, 2002.