# The Sequence Attribute Method for Determining Relationships Between Sequence and Protein Disorder


**Qian Xie** [1] qian@grover.chem.wsu.edu
**Gregory E. Arnold** [3 4] ge_arnold_bits@yahoo.com
**Pedro Romero** [2] promero@eecs.wsu.edu
**Zoran Obradovic** [2] zoran@eecs.wsu.edu
**Ethan Garner** [1] egarner@wsunix.wsu.edu
**A. Keith Dunker** [1] dunker@mail.wsu.edu

[1] Department of Biochemistry and Biophysics
[2] School of Electrical Engineering and Computer Science Washington State University, Pullman, WA 99164-4660, USA
[3] Biological Information Technologies, P.O. Box 1403, Richland, WA 99352, USA
[4] Present address: Amgen P Mail Stop 14-1-D, One Amgen Dr., Thousand Oaks, CA 91320, USA



**Abstract**

The conditional probability, P($s|x$), is a statement of the probability that the event, $s$, will occur given prior knowledge for the value of $x$. If $x$ is given and if $s$ is randomly distributed, then an empirical approximation of the true conditional probability can be computed by the application of Bayes' Theorem. Here $s$ represents one of two structural classes, either ordered, $s_o$, or disordered, $s_d$, and $x$ represents an attribute value calculated over a window of 21 amino acids. Plots of P($s|x$) versus $x$ provide information about the correlation between the given sequence attribute and disorder or order. These conditional probability plots allow quantitative comparisons between individual attributes for their ability to discriminate between order and disorder states. Using such quantitative comparisons, 38 different sequence attributes have been rank-ordered. Attributes based on cysteine, the aromatics, flexible tendencies, and charge were found to be the best attributes for distinguishing order and disorder among those tested so far.


## 1 Introduction

A region of protein is in the ordered state if it folds primarily into a single three dimensional structure, and it is in the disordered state if it remains mostly unfolded or partially folded into a locally flexible ensemble of structures. Despite the emerging interest in disordered states of proteins much of the structural biology community has not yet recognized the biological importance of this structural species. Recent studies demonstrate that disordered states are requisite in understanding the stabilities of proteins, their transport across membranes, their transformation into pathogenic states, and their cell signaling and regulation functions. Based on these findings, biochemical dogma will need to be revised, reflecting that amino acid sequence not only determines 3D structure [1], but also lack of structure as well [12].

Our recently reported results provide strong indication that some amino acid sequences code for disordered regions rather that structured ones, and that such disordered regions are commonly involved in function [12, 13, 14, 8]. Literature searches reveal more than 100 specific examples of proteins that are likely to be entirely disordered or contain disordered regions involved in function [16, 17, 11, 4, 12, 13, 8]. Our studies suggest that natively disordered regions have a variety of structures ranging from random-coil-like to molten-globule-like; however, further work needs to be carried out in order to investigate and characterize the structural properties of disordered states.

Towards this goal, our laboratory has developed neural networks capable of predicting, from primary sequence, disordered regions in proteins with 70% accuracy [12, 13]. A problem with the neural network predictors is that they do not provide an understanding of the factors leading to the prediction. In an effort to overcome this limitation a separate study has been initiated to explicitly investigate the relationship between disorder and primary sequence.

Described herein is application of the "the sequence attributes method" [2] to a set of sequences classified as ordered or disordered. This method, which depends on Bayesian statistics [3, 6], was developed several years ago for understanding relationships between sequence and secondary structure. This method uses a simple windowing procedure [15] to assign attribute values over stretches of 21 contiguous amino acid residues, and then partitions each window according to its attribute value. Plots of conditional probabilities versus attribute values allow quantitative comparisons between individual attributes for both their prediction potential and their ability to discriminate between order and disorder states.

# 2 Materials and Methods

## 2.1 The data set

The evaluation and testing of sequence attributes for their potential to discriminate between ordered and disordered requires prior construction of a database of sufficient size containing a balanced set of both structural classes [14]. The database was constructed from 32 proteins that were identified by literature searches, and their sequences obtained from the Swiss-protein, PIR, or PDB databases. A complete list of these proteins along with the methods used to identify their regions of disorder can be found on the Internet at, http://disorder.chem.wsu.edu/attribute.htm. A random subset of NRL_3D sequences corresponding in size and number to the identified disordered regions were used to generate a balanced database [14]. The resulting database contains 2,547 disordered and an equal number of ordered windows of size 21.

## 2.2 Sequence attributes calculations

An attribute is a numerical value calculated from an amino acid sequence over a window of specified length. In this work, two types of attributes are used: compositional and numerical. A compositional attribute is the number of a specified amino acid, or the number of a set of amino acids, divided by the length of the window. A numerical attribute is value of some amino acid parameter, such as hydropathy [10] or flexibility index [18], averaged over the amino acids in a given window.

Here we have used windows of 21 amino acids. This value was selected in previous studies [12] as a reasonable compromise between the increased noisiness of smaller windows and the lower resolution of larger ones.

A detailed description of the sequence attributes method including examples of the enumeration procedure is presented elsewhere [2]. A summary of the method as applied to protein order/disorder can be found on the Internet at, http://disorder.chem.wsu.edu/attribute.htm.

# 3 Results

## 3.1 Conditional probability graph

The sequence attributes approach has been used to explore relationships between order/disorder among the 38 different compositional and numerical attributes. The dependence between structural class and a particular attribute is examined by plotting the conditional probability values, $P(s|x)$, versus the attribute value $x$. The enumeration procedure described above is computed for both ordered ($o$) and disordered ($d$) structural classes leading to two conditional probability curves, $P(s_o|x)$ and $P(s_d|x)$, per graph.

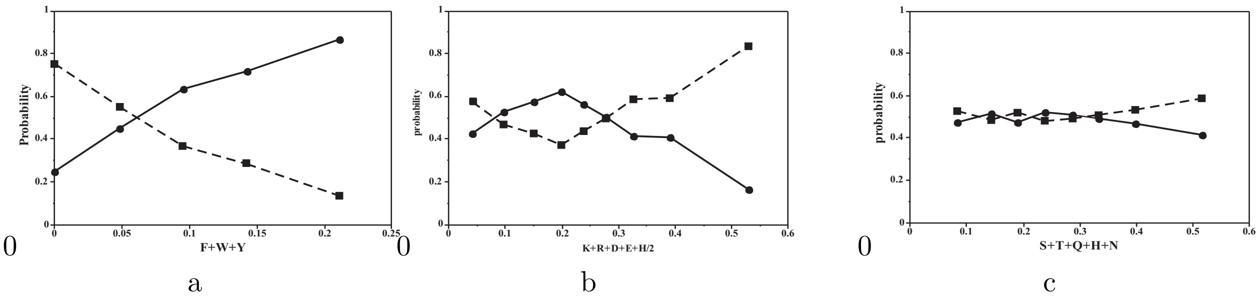


Figure 1: **Representative conditional probability plots.** Following are three example plots of P($s|x$) vs. $x$, where the curve for ordered structure, P($s_o|x$) is given by the dotted line and that for the disordered curve, P($s_d|x$) is given by the dashed line, and where the attributes are as indicated: a, aromatics (W+F+Y); b, charged (K+R+D+E+H/2); c, small and hydrophilic (S+T+Q+H+N).

Three example conditional probability graphs are shown in Fig. 1. These graphs were selected as typical examples demonstrating the capacity of various attributes to differentiate between order and disorder. The attributes in these examples are considered to have good (Fig. 1a), fair (Fig. 1b), and poor (Fig. 1c) discriminating properties. A sequence attribute, $x$, is considered useful if P($s_o|x$) and P($s_d|x$) vary reciprocally with $x$. In contrast, if P($s_o|x$) and P($s_d|x$) remain essentially constant for all values of $x$, then the attribute $x$ provides no useful information and is considered a poor attribute.

The compositional attribute W+F+Y is shown in Fig 1a. Notice there is a strong negative correlation with the W+F+Y attribute and the conditional probability for disorder, whereas the conditional probability for order exhibits the inverse relationship. This trend is concordant with the observations made by Burley and Petsko, well over a decade ago, that aromatic residues make especially strong contributions to protein stability [5]. Additionally, these 3 residues are among the most evolutionarily conserved residues found amidst homologous proteins [7]. Considering this information, the observed trend, namely that sequence windows with low aromatic content tend to be disordered while those with high aromatic content tend to be ordered, is entirely reasonable.

Figs. 1b presents the order/disorder conditional probability profiles as a function of total charge. The order/disorder probability curve exhibits a reciprocal relationship over roughly half of the respective attribute domain, and no consistent mutual or reciprocal relationship over the remainder of the attribute domain. The correlation of this attribute with order and disorder is complex. Given the small separation between the two curves, this attribute was judged to be fair in its potential to differentiate between order and disorder. When single amino acid attributes are subjected to feature selection, sets of charged amino acids are among the chosen [12, 13]. The large separation between the two curves for highly charged local segments of sequence (e.g., for b, at large values of K+R+D+E+H/2) are probably the basis for this selection.

The attribute in Fig. 1c shows little correlation with order and disorder, and as such was judged to be poor attribute. There is no separation between the conditional probability curves throughout their attribute domain. The residues S+T+Q+H+N are among the least evolutionarily conserved residues [7] which may explain, in part, why these attributes exhibit virtually no relationship with either structural class.

### 3.2 Quantitative comparison of attributes

A simple quantitative method to determine the potential of an attribute to differentiate between the two structural types is by the degree of separation between the P($s_o|x$) and P($s_d|x$) curves.

The largest possible degree of separation would occur for a pair of curves that were step functions, with values of 0,1 below some threshold and 1,0, respectively, above that threshold. Such an attribute would be a perfect predictor of order and disorder. The extent to which a given attribute approaches

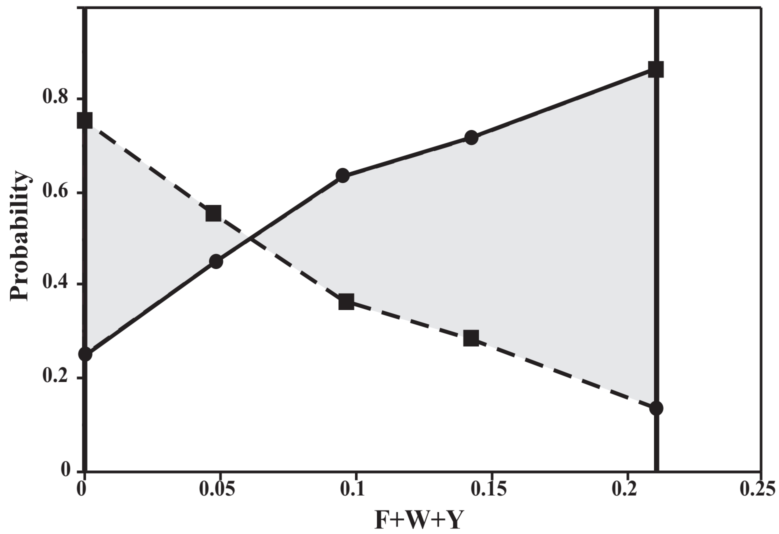


Figure 2: Evaluation of the area ratio. An example of calculating the area ratio of a P($s|x$) vs. $x$ plot is given using the F+W+Y attribute. First, vertical lines are constructed at the end points of P($s|x$) values. Next, the area of the rectangle bounded by these two lines and the upper and lower borders of the graph is calculated, call it $A_r$. Finally, the area that separates the P($s_o|x$) and P($s_d|x$) curves (shaded area) is determined, call it $A_s$. The area ratio is equal to $A_s/A_r$.

the "perfect predictor" can then be estimated by dividing the given area of separation by the total possible area as described in the figure legend; we are calling this the 'area ratio.' The area ratios were determined for the 38 attributes in this study.

These attributes and their area-ratio values are given in rank-order in Table 1. The correlation with disorder is also provided, where "+" means that the probability of disorder generally increases as the attribute value increases, where "-" means the converse, and where C, for complex, means that the two curves intersect more than one time.

The attribute set in this study included the 20 individual amino acids. This allows amino acids to be classified as "disorder-promoting" ("+" correlation), "order-promoting" ("-" correlation), or complex.

Sets of amino acids were chosen as attributes to determine the relationships between disorder/order and various classes of amino acids such as aromatic, polar, nonpolar, etc. Finally, two numerical attributes, hydropathy [10] and flexibility index [18], were included because these attributes were selected in our previous investigations as being important for predicting disorder [12].

# 4 Discussion

## 4.1 Prior Neural Network Prediction of Order and Disorder

Upon noticing that many proteins contain essential, functional regions that are disordered, it was reasoned that amino acid sequence should determine not only structure [1] but also lack of structure as well. To test this hypothesis, a small set of proteins with disordered regions was collected, and then a prediction-based approach that employed simple neural networks was used [12]. The results of those pilot studies and subsequent work provide evidence that disorder is encoded by the amino acid sequence [12, 13, 14, 8].

In the prior work supervised training of the neural networks was used, which means that inputs were

| Rank | Sequence Attribute [1] | Area Ratio | Correlation [2] |
|---|---|---|---|
| 1 | W+F+Y+C (C is low in disordered regions) | 0.517 | - |
| 2 | C (cysteine) | 0.484 | - |
| 3 | W+C+F+I+Y+V+L+H+M (least flexible, Vihinen) | 0.464 | - |
| 4 | A+T+R+G+Q+S+N+P+D+E (most flexible, Vihinen) | 0.464 | + |
| 5 | V+I+Y+F+W (best sheet formers, Chou-Fasman) | 0.439 | - |
| 6 | F+W+Y (aromatic) | 0.360 | - |
| 7 | W+F+Y+C+V+I+L+M+P (all non-polar) | 0.321 | - |
| 8 | Hydropathy | 0.302 | C |
| 9 | S+G+K+P+D+E (best sheet breakers, Chou-Fasman) | 0.295 | + |
| 10 | Flexibility (Vihinen scale) | 0.294 | + |
| 11 | W (tryptophan) | 0.279 | - |
| 12 | K+R+D+E (total charge) | 0.265 | C |
| 13 | A (alanine) | 0.265 | C |
| 14 | K+R-D-E (net charge) | 0.260 | C |
| 15 | F (phenylalanine) | 0.257 | - |
| 16 | E (glutamic acid) | 0.246 | C |
| 17 | Y (tyrosine) | 0.244 | - |
| 18 | K+R+D+E+H/2 (total charge) | 0.213 | C |
| 19 | K (lysine) | 0.210 | C |
| 20 | E+M+A+L (best helix formers, Chou-Fasman) | 0.206 | C |
| 21 | Z (glutamine or glutamate) | 0.203 | C |
| 22 | S+T+Q+H+N+D+E+R+K (all polar) | 0.190 | C |
| 23 | I (isoleucine) | 0.164 | - |
| 24 | Y+N+P+G (best helix breakers, Chou-Fasman) | 0.149 | C |
| 25 | B (asparagine or aspartate) | 0.136 | C |
| 26 | R (arginine) | 0.130 | C |
| 27 | S (serine) | 0.126 | + |
| 28 | P (proline) | 0.122 | + |
| 29 | L (leucine) | 0.108 | C |
| 30 | D (aspartic acid) | 0.091 | C |
| 31 | V (valine) | 0.083 | - |
| 32 | Q (glutamine) | 0.083 | + |
| 33 | N (asparagine) | 0.082 | - |
| 34 | T (threonine) | 0.078 | - |
| 35 | G (glycine) | 0.059 | C |
| 36 | S+T+Q+H+N (small and hydrophillic) | 0.050 | C |
| 37 | H (histidine) | 0.042 | C |
| 38 | M (methionine) | 0.041 | C |

[1] H/2 rather than H was used because the charge of H at neutral pH is typically 1/2.
[2] Correlation with disorder, where + or - indicates a single intersection of the order/disorder curves (like Fig 1a) whereas C (complex) indicates more than one intersection (like Fig. 1b, c).

Table 1: **Rank ordering of attributes according to their area ratio values.**

selected based on prior studies to find attributes or features that are predictive of order and disorder. These feature selection studies indicated some sequence attributes are important for determination of order or disorder, but these observations need to be supplemented with other information to provide understanding. For example, a feature selection process might indicate that tryptophan content is important for distinguishing order and disorder, but this observation by itself does not indicate the nature of the relationship between tryptophan content and order or disorder. Since the relationships between amino acid sequence and disorder were relatively unexplored, this problem seemed to be an excellent candidate for the application of conditional probability curves for the purpose of knowledge discovery.

### 4.2 Correlations between amino acid sequence and order/disorder

The feature selection protocols used for neural network development operated on an initial pool of 24 attributes, which were the compositions of the 20 amino acids, hydropathy [10], flexibility index [18], helix moment and sheet moment [9]. Here the list has been expanded to 38 attributes, which were then rank-ordered as described above. Many of these new attributes contain the same amino acids except for just one. This provides a simplified way of asking whether the amino acids cooperate in the prediction of order and disorder. For example, a higher ranking occurs if the different amino acid acts cooperatively with the others in the set.

The conditional probability curves are, in essence, a tool to be used to better understand structure/attribute correlations. Fig. 1a shows that the probability of disorder decreases almost linearly with increasing numbers of aromatic amino acids, whereas 1b shows a more complex relationship between the charged residue attribute and order/disorder.

Although many of the relationships between order/disorder and amino acid sequence could have been predicted in advance, Table 1 contains some surprises. These may lead to new insights regarding protein structure once the physical bases for the relationships are understood.

Consider the results for glycine. Given its ability to adopt many conformations, glycine might have been expected to correlate strongly with the disordered state, but it does not. It has a ranking of 35 out of the 38 attributes tested, with an area ratio value of just 0.059 and with a complex correlation. A possible explanation for these results is provided by the work on the development of the flexibility index [18]. The flexibility index of given amino acid does not depend on its intrinsic character, but rather on its typical location in proteins, with amino acids that are usually buried having a lower index values and those that are usually on the surface having higher values. In these studies, glycine was found to depend strongly on its neighbors, having low values when next to amino acids with low values and having high values when next to amino acids with high values. The likely explanation is that glycine occurs frequently not only in flexible loops but also deep inside proteins where it allows very close contacts to be made. From these considerations, the low ranking and complex correlation for glycine do make sense after all.

As a final example, compare serine and threonine. Given its small size and hydrophilicity, serine might be expected to correlate strongly with disorder. As expected, it does exhibit a positive correlation with disorder. Furthermore, its ranking, although modest at number 27 in Table 1, is the highest for a single amino acid correlating positively with disorder. The higher ranking single amino acids either correlate negatively with disorder or exhibit complex correlations. Now consider threonine. This amino acid might be expected to be quite similar to serine, but with a lower ranking due to the increased hydrophobicity that results from the added methyl group. Threonine does have a lower ranking as expected, number 34 in Table 1, but the surprise is that threonine correlates negatively with disorder. Evidently the methyl group tips the balance and causes threonine to be "order-promoting" albeit with a very small overall propensity.

## 4.3 Applications and future developments

An obvious extension of the results presented herein would be the construction of new predictors based on the best of the new attributes. Experiments in progress suggest that formal feature selection takes many of the highest ranking attributes from Table 1 and that the resulting neural network predictors using these new attributes significantly out-perform those developed originally. Indeed, new neural networks that include some of these attributes are now achieving about 85our previous neural networks [12, 13, 14], these new ones are achieving higher success while using a substantially larger dataset and with somewhat improved generalization to completely out-of-sample proteins.

Other experiments in progress have shown that there are other attributes, such as sequence complexity [19, 20, 21, 22], that may rank high on the list and so are likely to be important for determining order and disorder. To date a rational approach has been followed in selecting input features. That is, first we identified a specific collection of attributes that seemed likely in advance to be correlated with order and disorder. Next, we tested the collection to find the best among these, initially by formal feature selection protocols (which provide little basis for understanding) and now by the sequence attributes method (which can hopefully provide more understanding). This approach assumes that we know in advance which attributes distinguish between order and disorder. A useful alternative to this rational approach would be to use the sequence attributes method as the basis for identifying useful attributes by a combinatorial procedure that sifts through very large numbers of possible compositional and numerical attributes. Thus, it should be possible to carry out random generation of attributes and then test them by the area ratio method to determine which ones are important and which ones are not. We are in the process of implementing this approach. It will be interesting to determine whether this random generation approach will uncover any useful attributes, and, if so, whether such attributes provide new insights about sequence/structure relationships, or provide improved prediction accuracies.

## 5 Summary

The results in this paper provide additional evidence that order and disorder are determined by amino acid sequence. In agreement with our earlier observations, a collection of sequence characteristics –including charge, tendency for flexibility, absence of cysteine, and absence of aromatic groups– are strong indicators of the disordered state.

## Acknowledgements

Partial support by the NSF research grant NSF-CSE-IIS-9711532 to Z. Obradovic and A.K. Dunker is gratefully acknowledged. Garner thanks Dr. John Paznokas for providing support via an institutional grant for undergraduate research from the Howard Hughes Foundation.

## References

[1] Anfinsen, C.B., Principles that govern the folding of protein chains, *Science*, 181(96):223–230, 1973.

[2] Arnold, G.E., Dunker, A.K., Johns, S.J., Douthart, R.J., Use of conditional probabilities for determining relationships between amino acid sequence and protein secondary structure, *Proteins: Structure, Function and Genetics*, 12(4):382–399, 1992.

[3] Barlow, R.E., Proshan, F., *Statistical Theory of Reliability and Hypothesis Testing,* McArdle Press, Inc., Silver Springs, MD, 1981.

[4] Boelens, R., Vis H., Vorgias C.E., Wilson K.S., Kaptein R., Structure and Dynamics of the DNA Binding Protein HU from Bacillus Stearothermophilus by NMR Spectroscopy, *Biopolymers*, 40(5):553–559, 1997.

[5] Burley, S.K., Petsko, G.A., Aromatic-aromatic interaction: a mechanism of protein structure stabilization, *Science*, 229(4708):23–28, 1985.

[6] Charniak, E., McDermott, D., *Introduction to Artificial Intelligence*, Addison-Wesley Publishing Co., Reading, Massachusetts, 1987.

[7] Doolittle, R.F., *Of URFS and ORFS*, University Science Books, Mill Valley, CA, 1986.

[8] Dunker, A.K., Garner, E., Guilliot, S., Romero, P., Albrecht, K., Hart, J., Obradovic, Z., Kissinger, C., et al., Protein Disorder and the Evolution of Molecular Recognition: Theory, Predictions and Observations, *Pacific Symposium on Biocomputing*, 3:471–782, 1998.

[9] Eisenberg, D., Weiss, R.M., Terwilliger, T.C., The helical hydrophobic moment: a measure of the amphiphilicity of a helix, *Nature*, 299(5881):371–374, 1982.

[10] Kyte, J., Doolittle, R.F., A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol.*, 157(1):105–132, 1982.

[11] Lewis, M., Chang, G., Horton, N.C., Kercher, M.A., Pace, H.C., Schumacher, M.A., Brennan, R.G., Lu, P., Crystal structure of the lactose operon repressor and its complexes with DNA and inducer [see comments], *Science*, 271(5253):1247–1254, 1996.

[12] Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., Dunker, A.K., Identifying Disordered Regions in Proteins from Amino Acid Sequences, *Proc. I.E.E.E. International Conference on Neural Networks*, 1:90–95, 1997.

[13] Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., Guilliot, S., Garner, E., Dunker, A.K., Thousands of Proteins Likely to have Long Disordered Regions, *Pacific Symposium on Biocomputing*, 4–9 January 1998.

[14] Romero, P.Z., Obradovic, C., Dunker, A.K., Intelligent data analysis for protein disorder prediction, *Artificial Intelligence Review*, :in press, 1998.

[15] Rose, G.D., Prediction of chain turns in globular proteins on a hydrophobic basis, *Nature*, 272:586–590, 1978.

[16] Schulz, G.E., Nucleotide Binding Proteins, *Molecular Mechanism of Biological Recognition*, Elsevier/North-Holland Biomedical Press:79–94, 1979.

[17] Spolar, R.S., Record II, M.T., Coupling of Local Folding to Site-Specific Binding of Proteins to DNA, *Science*, 263:777–784, 1994.

[18] Vihinen, M., Torkkila, E., Riikonen, P., Accuracy of Protein Flexibility Predictions, *Proteins: Structure, Function, and Genetics*, 19:141–149, 1994.

[19] Wootton, J.C., Statistic of Local Complexity in Amino Acid Sequences and Sequence Databases, *Comput. Chem.*, 17(2):149–163, 1993.

[20] Wootton, J.C., Sequences with 'Unusual' Amino Acid Compositions, *Current Opinion is Structural Biology*, 4:413–421, 1994.

[21] Wootton, J.C., Non-globular domains in protein sequences: automated segmentation using complexity measures, *Comput. Chem.*, 18(3):269–285, 1994.

[22] Wootton, J.C., Federhen, S., Analysis of Compositionally Biased Regions in Sequence Databases, *Methods in Enzymology*, 266:554–571, 1996.