An MML Classification of Protein Structure that knows about Angles and Sequence

T. EDGOOSE, L. ALLISON, D.L. DOWE
Computer Science Department, Monash University, Clayton, VIC 3168, Australia.
eMail: {time,lloyd,dld}@cs.monash.edu.au

The MML classification program, Snob, deals with mixture modelling (or clustering) of circular data. It has recently been extended to do Markov modelling of the serial correlation between clusters such as modelling the fact that a Helix cluster favours being followed by another Helix cluster. Such a model is better known as a Hidden Markov Model. The search for the most appropriate secondary structure classification of protein data is of significant importance and was addressed by Hunter and States (1992) using the Bayesian classifier, AutoClass, on Cartesian co-ordinate data of protein residues. Dowe et al. (1996) improved upon this earlier work by using Snob to cluster dihedral angle data, with the advantage that 3x3=9 Cartesian co-ordinates can be represented by the 2 orientation-invariant angles, ϕ and ψ. The Hidden Markov Model used here is shown to be a more appropriate way again of modelling protein data and results in the selection of a simpler class model with 17 structure classes. We report on this classification, including the class transition matrix, and relate it back to the amino-acid sequence and the simple Helix, Beta, Turn classification. We find 3 types of Helix, 2 types of Beta and many types of Turn. The most numerous Turn class defines a continuous flexible structure that is negatively correlated to all the other classes.

1 Introduction

In this paper we apply a Hidden Markov Model to model the structure of a collection of known proteins. This Markov classification is able to take advantage of information implicit in the order of a sequence of observations and hence is better suited to modelling protein data than a classification model that assumes independence between observations. We use an Minimum Message Length (MML) information measure to evaluate our protein structure model which enables us to find the model best supported by the known evidence.

This work follows on from earlier unsupervised classification work modelling protein structure by Hunter and States [1] using AutoClass [2] (a Bayesian classifier) that modelled protein structure in Cartesian co-ordinates and subsequent work by Dowe et al.[3] using Snob [4,5] (an MML classifier) that modelled protein structure using dihedral angles. The dihedral angle representation requires the use of the circular von Mises distribution, but can express the same observations more compactly, which results in an improved classification.

The Markov classification model differs from the standard in that instead

Table 1: proteins in training and testing sets

1azu	1bp2	1ca2	1cc5	1ccr	1cpv	1crn	1ctx	1cy3	1cyc
1ecd	1est	1fc2	1fdh	1fdh	1fdx	1fx1	1gcn	1gcr	1gf1
1gf2	1gp1	1gp1	1hds	1hds	1hip	1lz1	1lzt	1mbd	1mbs
1mlt	1p2p	1pfc	1ppt	1rei	1rhd	1rn3	1sn3	1tgs	1tim
2app	2apr	2aza	2b5c	2cab	2ccy	2cyp	2dhb	2dhb	2gch
2gn5	2ig2	2ig2	2kai	2kai	2ldx	2lh1	2lzm	2mcp	2mcp
2pab	2rhe	2sga	2sns	2sod	2ssi	2stv	2taa	2tbv	3adk
3c2c	3cna	3fxc	3hhb	3hhb	3icb	3pcy	3pgk	3pgm	3rp2
3sgb	3tln	451c	4ape	4cts	4dfr	4fd1	4fxn	4ins	4ins
4mdh	4sbv	5cpa	5ldh	5pti	5rxn	6adh	7atc	7atc	8cat
	1abp	1acx	1hmq	1nxb	1ppd	1pyp	2act	2alp	
	2cdv	2lhb	2sbt	3gpd	3grs	6api	6api		

of assuming that the observations to be analysed are independent of one another, it assumes that they have been generated from a first order Markov process with as many states as there are classes in the model. The class of an observation generated from such a model is thus dependent on the class of the preceding observation. This permits runs of similar observations, as we expect in protein data, to be modelled more efficiently.

The model thus generated can be considered an alternate secondary structure model based only on the shape of known proteins. The results presented here are derived from a set of 100 known proteins (table 1).

For such a model to be useful it must be sufficiently complex as to avoid recklessly discarding information implicit in the observed data, but also not so complex as to be overly specific to the observed data (over fitting). A model that is too simple will describe all data badly and a model that is too complex will tend to describe the known data well, but predict unseen data badly.

In this paper we utilise the information theory based Minimum Message Length (MML) encoding proposed by Wallace[6,7] to search for a Markov classification model of protein secondary structure that avoids both of the aforementioned pitfalls. Within the MML paradigm, models are judged by their ability to reduce the length of a message transmitting all our observations to a receiver who initially shares only our prior beliefs. In this paradigm we calculate the expected length of an optimal message sending a model for the data and then actual observations given this model. The best model will minimise the length of this two-part message. The Markov classification model is significantly better able to explain our protein structure data than previous classification models that neglected the information implicit in the sequence of the observations.

In this paper we use an extension of the Snob (von Mises) classifier that implements both the Markov classification model and the von Mises distribution to better model sequences of angle data.

2 Previous Work

Protein structure is summarised by the position of the backbone atoms associated with each residue in the protein. Hunter and States [1] modelled the changes in the position of subsequent protein backbone atoms (alpha-carbon, beta-carbon and nitrogen) in Cartesian co-ordinates. For each residue, 9 values were required in order to state the relative position of each of these atoms. Unfortunately, these 9 values were highly correlated and this in turn led to proliferation in the number of classes being found along the line of correlation.

The same backbone information can be more compactly summarised by two dihedral angles, ϕ and ψ.[8] Dowe et al.[3] modelled protein structure in this way using a von Mises distribution,[9] the circular analogue of the Normal distribution. The MML coding estimates for this distribution can be found in Wallace and Dowe.[10] This more compact description of protein structure led to a simpler classification being found by Snob.

The von Mises distribution, $M(\mu, \kappa)$, has mean direction μ and concentration parameter κ. For small κ it tends to a uniform distribution and for large κ it tends to a Normal distribution with variance $1/\kappa$. This distribution is not affected by the 'wrap around' problem that the Normal distribution has when modelling angles. Snob is the only known program permitting mixture modelling of circular distributions.

Both AutoClass and Snob assume independence between the observations they model. Thus, these programs discard any sequential correlation information contained in the order of the residues. The preceding papers made some attempt to address this problem by partitioning the data into segments consecutive (ϕ, ψ) pairs however this again led to an increased number of classes being found.

We extend this earlier work by taking into account the expected autocorrelation in the secondary structure sequence and modelling it as a first order Markov process. This avoids the need to segment the data and leads to a simpler class structure being preferred. This, along with the use on the von Mises distribution to model the ϕ and ψ angle pairs, results in a more appropriate model of protein structure.

3 MML Overview

To quote C.S. Wallace, 'The best explanation of the facts is the shortest'. Taking this proposition seriously leads one into the domain of information theory and specifically Minimum Message Length (MML) encoding.

The basic idea of MML is that if we have some data (i.e. observations) and we wish to find the best model (i.e. hypothesis) given this evidence, this is equivalent to attempting to compress the data by first stating the model and then stating the data assuming the model to be true. Such an explanation can be transmitted as a message that any receiver who shares the same prior beliefs as that of the sender can decode into the original data (to a stated level of accuracy). No model that fails to compress the evidence can be considered superior to the empty model (null hypothesis).

A common problem with protein modelling is that quite complicated models are required and as the number of model parameters increases so does the risk of the model not generalising well. In the MML framework, overly complicated models are rejected by the requirement that they need to save at least as many bits in stating the data as they themselves take to be described.

Snob is a MML classifier which, given some data, attempts to compress it by constructing a classification model which would save more bits in the transmission of the data than the model itself takes to be described. Specifically, such a model could be used to transmit the 3D structure of our protein data in a more compact form. Dowe et al.[3] gives a good introduction to Snob.

4 The tSnob Program

The standard Snob assumes independence between all observations and so cannot directly model the extra information implicit in the order of a sequence of observations. tSnob introduces a first order Markov model into the standard Snob model. The classification of any observation can in this way be made to depend on the classification of the neighbouring observations. Thus, sequences of observations that may have been generated by a first order Markov process are well described.

It is important to note that the Markov model is part of our hypothesis and thus will be justified on information-theoretic grounds (i.e. not overly complex). The best class structure found using the Markov model may well differ markedly from the best class structure without the Markov model (in the number of classes and the specific class parameters). Thus the result from this program will, most likely, be quite different from the result obtained by first doing a standard classification and then observing the class transitions

(i.e. searching the two halves of the problem space independently).

It turns out that by the use of a simple dynamic programming method the length of an optimal message encoding a dataset given such a model can be easily calculated. The coding method and a robust search algorithm are described in Edgoose and Allison.[11]

5 Data and Results

The data that we have to work with comes from X-ray crystallography. The electron density maps generated enable the position of particular atoms within a protein to be stated to within an accuracy of about 2 Angstrom units, which translates to an error in ϕ and ψ of between 10 and 20 degrees (we use 11.5 degrees or 0.2 radians as per Dowe et al.[3]).

In this paper we use two such datasets. The first dataset contains 100 proteins (table 1) (17301 residues) from the Brookhaven database. The model and structure correlation we report relates to this dataset. The second dataset contains 229 proteins (41731 residues) and is identical to that studied previously by Dowe et al.[3]. This second dataset enables us to make a direct message length comparison with this earlier work, but unfortunately it was compiled without amino acid or secondary structure attributes.

The best Markov classification model on the primary dataset contains 17 classes with a message length of 7.13 bits/residue (the null, 1 class, model yields 9.04 bits/residue).

The class structure is shown in figure 1 with each class depicted by an ellipse with centre (μ_ϕ,μ_ψ) and dimensions $(\frac{1}{\sqrt{\kappa_\phi}}, \frac{1}{\sqrt{\kappa_\psi}})$. The actual observations are overlaid to create a Ramachandran scatter plot. This type of plot is actually a square depiction of the surface of a torus and some classes (specifically 5,9 and 16) do wrap around the edges.

The specific class parameters are listed in table 2 along with the first order Markov model describing the expected frequency of class transitions in table 3. Values in the class transition table are quoted as the log difference of the probabilities of a class occurring given the preceding class and in the global context. These values can be interpreted as the number of bits saved in stating a class by knowing the previous class. The classes have been ordered in these tables so that larger values tend to appear near the diagonal of the transition matrix. This has the effect of grouping sequentially related classes near to one another. It is interesting to note that although some of these clusters give the illusion of being correlated diagonally (for example classes 3 and 11) inspection of the transition matrix shows them to occur in significantly different contexts.

To facilitate interpretation, we also relate this classification to the known

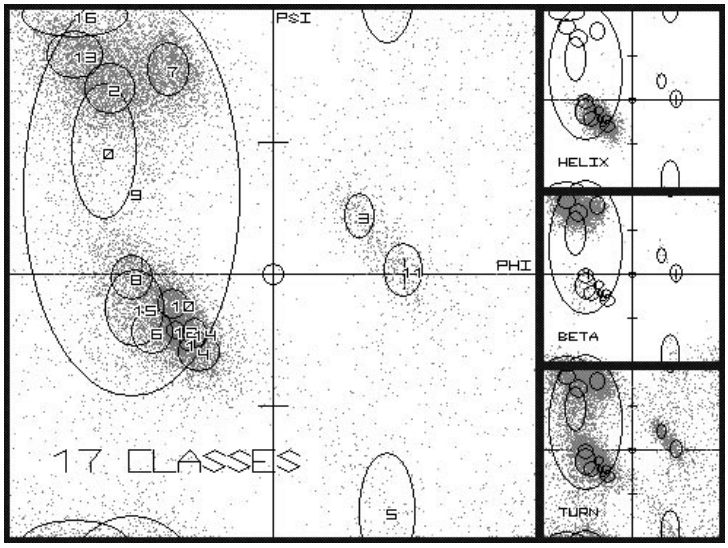

Figure 1: 17 class model

secondary structure (table 2) and amino acids sequence (table 4). This information was not, however, utilised in the classification process.

The same dataset as used in Dowe et al.[3] was also modelled using the Markov classification model. This dataset contained 229 proteins (41731 residues) and the best independent classification model reported contained 27 classes with a message length of 7.06 bits/residue (the null, 1 class, model yields 8.96 bits/residue). The best Markov classification model found on this larger dataset contained 19 classes with a message length of 6.40 bits/residue. Although this 19 class model was derived from a different (and much larger) set of proteins and included two additional types of rare Turn class, it did not appear to be substantively different from the 17 class model.

The message saving from using the Markov classification on this larger dataset was about 0.66 bits/residue (or 27727 bits). To put this another way, the Markov classification model was found to be 2^{27727} times more likely as an explanation for the observed data than the standard Snob classification model. This is a strong result confirming that the Markov classification model is more appropriate for modelling this kind of data.

Table 2: class parameters and secondary structure correlation

Class	Model Parameters					Secondary-Structure		
	%	μ_ϕ	κ_ϕ	μ_ψ	κ_ψ	T 53%	B 21%	H 26%
9	12.0	-96.7	0.6	56.2	0.2	**0.93**	0.03	0.03
16	3.4	-136.3	2.6	176.9	14.4	**0.63**	0.36	0.00
15	2.9	-95.0	8.4	-23.3	5.0	**0.81**	0.17	0.02
13	9.0	-136.1	9.3	150.5	13.5	0.35	**0.65**	0.00
2	13.2	-112.3	11.4	127.4	11.4	0.26	**0.74**	0.00
11	1.9	88.5	18.3	2.5	10.5	**0.99**	0.01	0.00
7	11.8	-72.1	16.0	139.9	9.7	**0.78**	0.22	0.00
3	1.6	59.0	30.7	40.2	14.6	**0.93**	0.04	0.02
0	4.6	-115.7	6.5	84.1	1.6	**0.86**	0.12	0.01
5	1.3	77.6	9.0	-161.7	2.1	**0.87**	0.12	0.01
8	4.9	-96.9	16.6	-1.9	15.5	**0.84**	0.00	0.15
10	7.3	-68.7	35.9	-20.0	29.8	**0.60**	0.00	0.40
6	3.7	-83.0	17.6	-38.7	14.8	0.36	0.00	**0.64**
14	3.1	-53.9	47.2	-40.3	31.8	0.50	0.00	0.50
12	7.6	-67.2	104.8	-38.2	86.2	0.03	0.00	**0.97**
1	4.9	-58.9	128.4	-46.8	125.3	0.01	0.00	**0.99**
4	6.8	-51.3	17.5	-53.5	18.6	0.15	0.00	**0.85**

6 Discussion

The 17 class Markov classification model is a simpler model (fewer classes) than those found in previous work using AutoClass and Snob. The classes vary in weight from 1% (class 5) to 13% (class 2) of the data with the median at 5%, which is a relatively even spread.

From table 2 we find 3 classes (1,4 and 12) that are strongly associated with the Helix secondary structure, 2 classes (13 and 2) are most closely identified as Beta, and several classes (0,3,5,7,8,9,11 and 15) constitute different flavours of Turn. Class 16 can best be described as a transitional class between Turn and Beta and likewise the remaining classes (6,10 and 14) are transitional classes between Turn and Helix.

From the transition model (table 3) we see that, as expected, most classes are positively auto-correlated with the exception of classes 8 and 11 ($\psi \approx 0$).

The Turn class 9 (the second most abundant class, making up 12% of the residues) has an auto-correlation of 78% (expected run length of 5 residues) and very small kappa values. This class defines a very flexible continuous structure which is negatively correlated with all the other classes. The specific conformation of residues within this class is most likely completely determined by non-local structural effects. One possible explanation would be that, as the

protein folds, the Helix and extended Beta structures form first (due to the location of Proline, etc.) and then, right near the finish, parts of the protein have to flexibly twist around to accommodate these earlier Helix and Beta structures, and that this is the role of class 9.

The 3 Helix classes (1,4 and 12) can be further partitioned by noting that transitions between class 4 and the other two are extremely rare. Also, a run of class 4 Helix is most likely preceded by a residue in class 6. Both Helix groups are most likely broken by a residue in class 10 in which Pro(P) is relatively more abundant. As expected, there is little chance of a Helix class being followed by a Beta class. The classes 6, 8 and 10 form a loose group due to the fact that residues in classes 6 and 8 are most likely preceded by a residue in class 10.

From table 4 we can see that several amino acids will almost never take on certain conformations. For example Ala(A), Glu(E), Ile(I), Pro(P), Ser(S), Thr(T), Val(V) and Trp(W) are almost certain to never occur in class 11. Class 3 contains a higher proportion than normal of Asp(D), Gly(G), His(H), and Asn(N). We also note that that Ile(I) and Val(V) imply a much higher than normal chance of class 2, which is a Beta structure. As expected, Pro(P) only occurs in classes that overlap the region where $-80^o < \phi < -40^o$.[8] It is also of interest that Gly(G), Pro(P), Ser(S) and Thr(T) are consistently less likely to occur in any of the three main Helix classes.

7 Structure Prediction - some preliminary results

We note in passing that the Markov classification model can also be used to predict secondary structure. Instead of modelling (ϕ, ψ) angle pairs, we model amino-acid and secondary structure (H,B or T) pairs directly (i.e., 2 multi-state attributes per residue). The Markov classification model was used in this manner on the smaller dataset of 100 proteins (table 1). The best model found had 6 classes and yielded a message length of 4.73 bits/residue (the null, 1 class, model yields 5.64 bits/residue).

We use this model to predict secondary structure by determining the probabilistic assignment of residues to classes based on the amino-acid sequence and the first order Markov model only. The calculation of this optimal assignment, which is influenced by the data on both sides of any particular residue, is detailed in Edgoose and Allison.[11] This optimal distribution over classes for each residue can then be used to calculate a distribution over secondary structures (H,B and T). This is our probabilistic secondary structure prediction.

Information measures provide a safe way to compare the prediction performance of different structure models. Such a measure specifies the number of bits required to send the true secondary structure given the probabilistic pre-

dictions. Overly timid predictions and overly confident predictions are both penalised. The best predictors will give the best compression.

The program was used in this way to predict the secondary structure of 15 unseen proteins (table 1) (3199 residues). The average cost for stating the true secondary structure of a particular residue in an unseen protein given the entire amino acid sequence was 1.23 bits, which is down from the 1.47 bits required without the model (a prediction accuracy of 62.2% up from 52.6%). The prediction accuracy achieved by this 6 class model is near to the bottom end of the basic nearest neighbour pattern matching algorithms which perform in the range from 63% to 68%.[12] This is reasonably good result considering the model has only 156 free parameters (i.e. each of the 6 classes requires 19 for the amino acid distribution, 2 for the secondary structures distribution and 5 for the class transition distribution) and no database lookup is required.

One interesting aspect of this prediction model is that there is no need to state any specific window size for the number of amino acids around a particular residue that can influence the prediction of its secondary structure. Although all the amino acids exert some influence, this influence becomes vanishingly small the more distant the amino acid.

8 Conclusion

Early work in the classification of protein structure using Cartesian co-ordinates by Hunter and States [1] suffered from significant inter-attribute correlation which led to a proliferation in the number of classes that were found. Subsequent work based on a von Mises modelling of the dihedral angle representation by Dowe et al.[3] improved this early work. However, both works discarded the information implicit in the sequence of the observations.

A Markov classification model has been applied successfully to the dihedral angle representation of this data and has been shown to achieve significantly higher compression and is a more likely explanation of the data observed.

The best Markov classification model found contained 17 classes with three classes of Helix (two of which are closely related), two classes of Beta, and many classes of Turn. One Turn class found was of particular note as it occurs in segments with an expected length of 5 residues and exhibits extreme flexibility. This 17 class classification is considerably simpler than the models previously found [3] and hence more amenable to further analysis and model building.

The Markov classification model is also shown to be directly applicable to the prediction of secondary structure from amino acid sequence where a simple 6 class model (156 parameters) is found to do surprisingly well (1.23 bits/residue) without reference to a database.

Acknowledgements

Special thanks to Evan Steeg and Larry Hunter for access to the protein datasets they have compiled. The third author is supported by Australian Research Council Large Grant No. A49602504.

References

1. L. Hunter and D. J. States. Bayesian classification on protein structure. *IEEE Expert*, 7(4):67–75, 1992.
2. P. C. Cheeseman. Autoclass II conceptual clustering system. *Proceedings Machine Learning Conference*, pages 54–64, 1988.
3. D.L. Dowe, L. Allison, T.I. Dix, L. Hunter, C.S. Wallace, and T. Edgoose. Circular clustering of protein dihedral angles by Minimum Message Length. In *Proc. 1st Pacific Symp. Biocomp.*, pages 242–255, HI, U.S.A., 1996.
4. C. S. Wallace. Classification by minimum length inference. *AAAI Spring Symposium on the Theory and Application of Minimum Length Encoding, Standford*, pages 5–9, 1990.
5. C.S. Wallace and D.L. Dowe. MML mixture modelling of Multi-state, Poisson, von Mises circular and Gaussian distributions. In *Proc. 6th Int. Workshop on Artif. Intelligence and Statistics*, pages 529–536, 1997.
6. C. S. Wallace and D. M. Boulton. An information measure for classification. *Computer Journal*, 11:185–194, 1968.
7. C. S. Wallace and P. R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society (Series B)*, 49:240–252, 1987.
8. Schultz and Schirmer. *Principles of Protein Structure*. Springer-Verlag, New York, 1990.
9. N.I. Fisher. *Statistical Analysis of Circular Data*. Cambridge University Press, Cambridge, 1993.
10. C.S. Wallace and D.L. Dowe. MML estimation of the von Mises concentration parameter. Technical report TR 93/193, Dept. of Comp. Sci., Monash Univ., Clayton, Vic. 3168, Australia, 1993. prov. accepted, Aust. J. Stat.
11. T.C. Edgoose and Allison L. Mixture modelling of sequential datasets. Technical Report 285, Monash University, Computer Science Department, 1996.
12. A.A. Salamov and V. Solovyev. Prediction of protein secondary structure. *Journal of Molecular Biology*, 247:11–15, 1995.

Table 3: class transition log odds matrix ($log_2 P(class|prevClass) - log_2 P(class)$)

	Class								
Prev. Class	9 12%	16 3%	15 3%	13 9%	2 13%	11 2%	7 12%	3 2%	0 5%
9	**2.7**	**-2.8**	**-2.3**	-1.5	-1.7	**-3.2**	-1.4	**-2.4**	-1.9
16	-0.5	1.0	-0.0	**1.5**	0.0	**-3.2**	0.1	**-3.0**	0.7
15	**-2.3**	**2.6**	0.7	**2.2**	**-2.8**	-1.9	-1.6	0.2	**1.1**
13	-1.8	1.0	0.1	**2.0**	1.0	-1.9	0.6	-0.8	-1.6
2	-1.6	0.2	-0.2	-0.2	**2.1**	**-3.2**	0.3	0.5	-0.5
11	**-4.3**	**2.0**	**1.0**	-0.4	0.1	-1.9	**1.7**	**-3.0**	0.3
7	**-2.3**	0.7	**1.9**	-0.1	-0.1	**1.9**	**1.1**	0.4	0.6
3	**-3.3**	-1.3	0.8	0.9	-0.7	**2.9**	1.0	**2.4**	1.0
0	-1.0	0.3	**1.2**	0.2	**-2.0**	-1.9	0.9	0.6	0.1
5	-1.9	**1.4**	0.4	-0.4	-0.1	0.2	0.2	0.8	0.7
8	**-2.4**	-0.6	**-2.1**	-1.2	-2.0	**2.5**	**1.3**	**2.2**	**1.9**
10	**-2.5**	**-3.5**	**-2.9**	**-4.5**	**-5.0**	**-2.2**	**-2.1**	0.5	0.6
6	**-2.6**	**-2.3**	-0.3	**-4.5**	**-4.7**	-0.2	**-2.0**	**-2.0**	0.7
14	**-3.1**	**-2.3**	**-2.6**	-1.3	**-3.7**	-0.9	**-2.6**	-1.4	0.2
12	**-4.9**	**-5.1**	**-4.9**	**-5.5**	**-6.0**	**-4.2**	**-5.9**	**-4.0**	**-2.5**
1	**-3.0**	**-5.1**	**-3.9**	**-5.5**	**-6.0**	**-4.2**	**-5.9**	**-2.0**	**-5.5**
4	-0.4	**-5.1**	**-4.9**	**-5.5**	**-6.0**	-1.9	**-4.9**	**-3.0**	**-3.5**

	Class								
Prev. Class	5 1%	8 5%	10 7%	6 4%	14 3%	12 8%	1 5%	4 7%	
9	-0.8	**-4.0**	**-3.4**	-1.4	**-3.9**	**-6.3**	**-2.6**	-0.4	
16	-0.3	0.2	0.2	**-2.6**	0.1	**-2.4**	-0.9	**-3.8**	
15	**2.2**	-1.0	**-2.7**	**-3.2**	**-3.3**	**-5.3**	**-4.6**	-1.9	
13	0.9	**-3.3**	**-3.4**	**-4.2**	-1.5	**-6.3**	**-5.6**	**-4.5**	
2	0.4	**-2.4**	**-2.4**	**-5.2**	**-2.6**	**-6.3**	**-62.1**	**-5.1**	
11	-0.8	0.8	**-2.3**	-1.8	-0.9	**-2.7**	**-4.0**	**-5.1**	
7	**-2.1**	-1.6	-0.1	**-5.2**	**1.5**	**-3.4**	**-4.6**	**-2.6**	
3	-0.7	-0.1	-1.1	**-3.2**	-0.1	**-5.3**	**-4.6**	**-4.5**	
0	-0.1	-0.9	**1.3**	**-3.6**	**2.0**	**-2.1**	**-3.0**	-1.9	
5	**1.6**	**1.6**	**1.1**	-1.3	**-2.9**	**-2.9**	**-2.4**	**-3.8**	
8	**1.3**	-0.3	**-3.6**	**1.2**	0.6	**-4.7**	**-2.6**	**-3.1**	
10	-0.2	**2.9**	**1.1**	**3.0**	**-3.9**	**-2.3**	**-5.6**	**-4.1**	
6	**1.4**	0.6	**1.5**	**1.8**	**1.8**	0.2	-0.3	1.0	
14	-0.1	**1.1**	**2.2**	**-2.1**	**1.9**	**1.8**	**-2.0**	**-4.1**	
12	**-3.7**	-0.6	0.4	-0.8	**-3.9**	**3.1**	**1.8**	**-3.5**	
1	**-3.7**	**-3.3**	0.6	-1.4	**-2.1**	**1.4**	**3.7**	**-2.0**	
4	**-3.7**	**-2.1**	0.1	-1.3	-1.3	**-2.7**	**-2.1**	**3.5**	

Table 4: amino acid log odds given class ($log_2 P(AA|class) - log_2 P(AA)$)

Class	Amino-Acid									
	A	C	D	E	F	G	H	I	K	L
	8%	2%	6%	5%	4%	9%	2%	5%	6%	8%
9	-0.2	-0.2	0.4	-0.1	-0.4	0.7	0.3	-0.7	0.2	-0.4
16	-0.5	0.3	-0.4	-0.8	0.0	**1.6**	-0.2	-1.4	-0.5	-1.3
15	-0.6	-0.9	0.5	-0.4	-0.4	-1.0	-0.2	0.2	-0.0	-0.1
13	0.0	0.6	-1.2	-0.3	0.7	-0.8	0.1	0.5	-0.3	-0.4
2	-0.8	0.4	-0.8	-0.5	0.3	**-2.4**	-0.4	**1.1**	-0.4	0.4
11	**-3.9**	**-2.3**	**-2.3**	**-3.2**	**-2.8**	**3.3**	-0.5	**-4.1**	-1.3	**-2.9**
7	-0.0	0.2	0.1	-0.4	-0.3	-1.1	-0.3	-0.5	-0.4	0.0
3	-0.9	**-2.0**	0.6	-1.5	-1.2	**1.8**	**1.3**	**-2.2**	0.2	**-2.3**
0	-0.8	-0.1	**1.0**	-0.2	0.6	-1.0	0.8	-0.2	-0.2	-0.3
5	-1.9	**-2.4**	-0.6	**-2.4**	**-2.6**	**3.1**	-1.3	**-2.6**	-1.1	-1.9
8	-0.2	0.2	0.8	0.2	-0.0	-1.0	0.5	-0.7	0.1	-0.1
10	0.4	-0.2	0.1	0.4	-0.8	-1.1	0.1	-0.9	0.3	0.0
6	0.3	0.6	-0.0	0.5	0.1	-1.2	0.3	-0.2	0.4	0.4
14	0.6	-0.7	0.0	0.5	-0.4	-1.2	-1.2	-0.4	-0.1	-0.1
12	0.7	-0.0	-0.1	0.7	0.2	-1.1	-0.2	-0.1	0.3	0.4
1	0.6	-0.3	-0.2	0.6	0.4	-1.2	-0.3	0.3	0.4	0.4
4	0.5	-0.8	-0.1	0.4	-0.1	-0.8	0.1	0.1	0.4	0.4
Class	Amino-Acid									
	M	N	P	Q	R	S	T	V	W	Y
	2%	5%	5%	3%	3%	8%	7%	7%	2%	4%
9	-0.6	0.2	0.2	-0.1	-0.1	0.2	-0.1	-0.5	-0.9	-0.3
16	-1.0	0.1	-0.8	-0.4	-0.1	0.6	0.4	-1.1	0.1	-0.1
15	-0.7	0.0	0.3	-0.2	0.1	0.7	0.7	0.1	0.0	-0.6
13	0.4	-1.0	**-4.0**	0.2	0.3	0.5	0.4	0.3	0.6	0.7
2	0.1	-0.5	**-2.1**	-0.1	0.1	-0.5	0.5	**1.1**	0.4	0.7
11	-1.7	-0.1	**-5.0**	**-2.6**	**-2.0**	**-3.3**	**-3.6**	**-5.2**	**-3.3**	-1.1
7	-0.1	0.0	**1.8**	-0.0	-0.3	0.1	-0.2	-0.3	-0.4	-0.2
3	0.0	**2.0**	**-4.8**	0.1	0.1	-1.0	**-2.6**	**-2.6**	**-3.1**	-0.1
0	0.4	**1.2**	-1.8	-0.5	0.0	-0.1	-0.0	-0.3	0.3	0.2
5	-0.9	-0.6	**-3.4**	-1.9	**-3.5**	-1.1	-1.2	**-3.0**	**-2.7**	**-2.3**
8	-0.2	0.5	-1.4	0.4	0.1	0.4	0.4	-1.2	0.3	0.3
10	-0.1	-0.3	1.0	0.1	0.1	0.4	-0.1	-0.6	-0.1	-0.6
6	0.3	-0.3	**-2.9**	0.5	0.2	-0.3	-0.1	0.3	-0.0	-0.2
14	-0.5	-0.9	**1.6**	-0.3	-0.1	0.3	-0.4	-0.3	-0.0	-0.6
12	0.5	-0.2	-1.7	0.5	-0.0	-0.8	-0.4	0.1	0.0	-0.7
1	0.5	-0.6	-1.6	-0.1	0.3	-1.0	-0.4	0.1	0.2	0.1
4	0.5	-0.0	-0.8	0.2	0.1	-0.5	-0.6	-0.0	0.7	-0.2