CIRCULAR CLUSTERING BY MINIMUM MESSAGE LENGTH OF PROTEIN DIHEDRAL ANGLES

David L. Dowe, Lloyd Allison, Trevor I. Dix, * Lawrence Hunter,
Chris S. Wallace, Timothy Edgoose

Department of Computer Science, Monash University,
Clayton, Victoria 3168, Australia

** National Library of Medicine, Bldg. 38A, 9th floor,*
8600 Rockville Pike, Bethesda, MD 20894, U.S.A.

*e-mail: {dld,lloyd,trevor,csw,time}@cs.monash.edu.au, * hunter@nlm.nih.gov*

Early work on proteins identified the existence of helices and extended sheets in protein secondary structures, a high-level classification which remains popular today. Using the Snob program for information-theoretic Minimum Message Length (MML) intrinsic classification, we are able to take the protein dihedral angles as determined by X-ray crystallography, and cluster sets of dihedral angles into groups. Previous work by Hunter and States had applied a similar Bayesian classification method, AutoClass, to protein data with site position represented by 3 Cartesian co-ordinates for each of the α-Carbon, β-Carbon and Nitrogen, totalling 9 co-ordinates. By using the von Mises circular distribution in the Snob program rather than the Normal distribution in the Hunter and States model, we are instead able to represent local site properties by the two dihedral angles, ϕ and ψ. Since each site can be modelled as having 2 degrees of freedom, this orientation-invariant dihedral angle representation of the data is more compact than that of nine highly-correlated Cartesian co-ordinates. Using the information-theoretic message length concepts discussed in the paper, such a more concise model is more likely to represent the underlying generating process from which the data comes. We report on the results of our classification, plotting the classes in (ϕ, ψ)-space and introducing a symmetric information-theoretic distance measure to build a minimum spanning tree between the classes. We also give a transition matrix between the classes and note the existence of three classes in the region $\phi \approx -1.09$ *rad* and $\psi \approx -0.75$ *rad* which are close on the spanning tree and have high inter-transition probabilities. These properties give rise to a tight, abundant, self-perpetuating, α-helical structure.

1 Introduction

Proteins exhibit a wide variety of structural similarities with one another. Protein substructure classifications such as secondary structure are used in a variety of important applications. For example, they play an important role in fitting reasonable protein structures to electron density maps generated by X-ray crystallography (e.g. [Levi92]), and they are the target classes of many attempts to predict protein structure from sequence (e.g. [QiSe89], [SoSa94]). However, there is some controversy about whether secondary structure, as currently defined by hydrogen bonding patterns, is the best level of analysis for these tasks. Several researchers have attempted to improve on secondary structure classification for various tasks, e.g. [RoRW90], [HuSt91], [Sun93], [KaTa94], [Swin95].

Our approach to this problem is to use a quantitative version of Ockham's razor for clustering the dihedral angles of proteins taken from the Brookhaven protein structure database (PDB). We use the Minimum Message Length (MML) principle, which seeks a simple theory that fits the data well.

The Snob program for clustering and numerical taxonomy was originally developed by Wallace and Boulton [WaBo68], and was the first serious application of the MML principle for general inductive inference. The original Snob program [WaBo68, Wall86] permitted continuous data, which were modelled by Normal distributions, and discrete data, which were modelled by multi-state distributions. More recently [WaDo94b], Snob has been extended to permit data to be modelled by Poisson distributions and von Mises circular distributions. Snob is the only known program permitting cluster models of circular distributions.

Using the fact that dihedral angles along protein backbones can be modelled as coming from a circular distribution, and that angles near -179 degrees are close to and should be regarded as being close to angles around +179 degrees, it is advantageous to use the von Mises circular distribution to model our data. Additionally, our model replaces nine highly-correlated Cartesian co-ordinates with just two orientation-invariant angles, ϕ and ψ.

Our data consists of 41,731 pairs of protein (ϕ, ψ) dihedral angle pairs from the Brookhaven protein structure database (PDB). Our work follows that of Hunter and States [HuSt91, HuSt92] using AutoClass [Ch88, Ch90] with Gaussian variables.

We introduce the MML principle in Section 2. Section 3 outlines how to use MML for parameter estimation of the von Mises circular distribution. Section 4 describes the Snob clustering program and how to infer a mixture of von Mises distributions from data. The information-theoretic statistical soundness of MML is presented in Section 5. The remaining sections give the results of the secondary structure groups [KaSa83] that resulted from applying Snob to cluster our dihedral angle data.

The work reported here is preliminary in nature: we take little account of the auto-correlation of the secondary structure sequence and, less importantly, the dependence of dihedral angles upon one another.[†]

2 About Minimum Message Length

The information-theoretic Minimum Message Length (MML) principle [WaBo68(p185), BoWa70(pp63-64), WaBo75, WaFr87] of inductive inference variously states [WaDo94b] that the best conclusion to draw from data is the theory with the highest posterior probability or, equivalently, that theory which maximises the product of the prior probability of the theory with the probability of the data occurring in light of that theory. We quantify this immediately below.

Letting D be the data and H be an hypothesis (or theory) with prior probability Pr(H), we can write the posterior probability Pr(H|D) = Pr(H&D)/Pr(D) = Pr(H).Pr(D|H)/Pr(D), by repeated application of Bayes's Theorem. Since D and Pr(D) are given and we wish to infer H, we can regard the problem of maximising the posterior probability, Pr(H|D), as one of choosing H so as to maximise Pr(H).Pr(D|H) .

Also, elementary information-theoretic coding tells us that an event of probability p can be coded (e.g. by a Huffman code or an arithmetic code) by a message of length $-\log_2 p$ bits . (Negligible or no harm is done by ignoring effects of rounding up to the next positive integer.)

Since $-\log_2(Pr(H).Pr(D|H)) = -\log_2(Pr(H)) - \log_2(Pr(D|H))$, maximising the posterior probability, Pr(H|D), is equivalent to minimising $-\log_2(Pr(H)) - \log_2(Pr(D|H))$, the length of a two-part message conveying the theory and the data in light of the theory. Hence the name "minimum message length" (principle) for thus choosing a theory, H, to fit observed data, D. An alternative and equivalent way of viewing matters is that a sender wishes to transmit both H and D to a receiver who, prior to this transmission, knew only the code scheme (and hence the priors) that the sender and receiver had agreed upon. One of the earliest statements of the principle [Solo95 (Sec 5)] is due to Solomonoff [Solo64 (p20)]. The principle was re-stated and apparently first seriously applied in a series of papers by Wallace and Boulton (e.g. [WaBo68 (p185), BoWa69, BoWa70 (pp63-64), WaBo75]) dealing with model selection and parameter estimation (for Normal and multi-state variables) for problems of clustering and intrinsic classification (see [Solo95 (Sec 5)]). An important special case of the Minimum Message Length principle is an observation of Chaitin [Chai66] that data can be regarded as "random" if there is no theory, H, describing the data which results in a shorter total message length than that from the null theory.

Closely related subsequent work is the Minimum Description Length (MDL) work of Rissanen (e.g. [Riss89]). Introductory material on MML is given in [WaDo93] and a commentary on the fine technical differences between MDL and MML is given by Baxter and Oliver [BaOl95].

3 Parameter Estimation by MML

Given data $\underset{\sim}{x}$ and parameters $\underset{\sim}{\theta}$, let h($\underset{\sim}{\theta}$) be the prior probability distribution on $\underset{\sim}{\theta}$, let p($\underset{\sim}{x}|\underset{\sim}{\theta}$) be the likelihood, let $L = -\log p(\underset{\sim}{x}|\underset{\sim}{\theta})$ be the negative log-likelihood and let F be the Fisher information, the determinant of the (Fisher information) matrix of expected second partial derivatives of the negative log-likelihood. Then the MML estimate of $\underset{\sim}{\theta}$ is [WaFr87 (p245)] that value of $\underset{\sim}{\theta}$ minimising the message length, $-\log(h(\underset{\sim}{\theta})p(\underset{\sim}{x}|\underset{\sim}{\theta})/\sqrt{F(\underset{\sim}{\theta})})$ + a constant. (This is elaborated upon in [WaDo93 (pp1-3)].)

The two-part message describing the data thus comprises first, a theory, which is the MML parameter estimate(s), and, second, the data given this theory. While it is reasonably clear to see that a finite coding can be given when the

[†] A more concise published version of this technical report is available in [DADH96].

data is discrete or multi-state, we also acknowledge that all recorded continuous data must only be stated to finite accuracy by virtue of the fact that it was able to be (finitely) recorded. In analysing statistical data it is important to know the accuracy to which the data has been measured. Most methods implicitly assume that the data gathered has been measured to arbitrary accuracy. For an information-theoretic method such as MML, the less accurate the data the less information it conveys. In practice, we assume that, for a given continuous or circular attribute, all measurements are made to some accuracy, ε. For the Snob program (see Section 4), this accuracy is stated by the user. The accuracy should be a measure of the repeatability of a measurement.

For a physical measurement, it is presumably the accuracy of the instrument being used. This raises an important point. Using a message length criterion, if our data is very noisy, the measurement accuracy, ε, is large and the information content is therefore relatively low, then it would be wiser not to grow too many Snob classes. If the data becomes increasingly accurate, this should not decrease the number of classes which we wish to grow. We know [HuSt91] our data to have no better resolution than 2 Angstrom units in Euclidean space. We believe that this corresponds to a resolution of about 10 to 20 degrees in our dihedral angle data. The exact choice of measurement accuracy, ε, should not make a great deal of difference, as we indeed observe from experiment. We opt to assume a measurement accuracy of approximately 11.5 degrees in our data for Snob.

3.1 Continuous, Multi-state and Poisson Attributes

MML has successfully been applied to a variety of parameter estimation problems, and corrects the minor but well-known small sample bias in Maximum Likelihood estimation of the variance of a Normal distribution [WaBo68 (p190), WaDo94b (p38)].

MML also obtains sensible answers in estimating the probability parameters for a multi-state distribution [WaBo68, WaDo94b (p38)] and for the rate parameter of a Poisson distribution.

3.2 Circular Attributes and the von Mises distribution

The von Mises distribution (see, e.g. [Fish93]), $M_2(\mu, \kappa)$, with mean direction μ, and concentration parameter, κ, is a circular analogue of the Normal distribution – both being maximum entropy distributions. Letting $I_0(\kappa)$ be the relevant normalisation constant, it has probability density function (p.d.f.) $f(x|\mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa.\cos(x-\mu)}$, and corresponds to the distribution of the angle, x, of a circular pendulum in a uniform field (at angle μ) subjected to thermal fluctuations, with κ representing the ratio of field strength to temperature. For small κ, it tends to a uniform distribution and for large κ, it tends to a Normal distribution with variance $1/\kappa$. Circular data arises commonly [Fish93] in (e.g.) biology, geography, geology, geophysics, medicine, meteorology and oceanography.

MML estimation of the von Mises concentration parameter, κ, is obtained by minimising the formula earlier in this section (plus a constant) for the message length, using a uniform prior on μ in $[0, 2\pi)$ and using a Bayesian prior distribution on κ such as [WaDo93, WaDo94a] $h_3(\kappa) = \kappa/(1 + \kappa^2)^{3/2}$. Monte Carlo simulations [WaDo93 (pp12-18)] show a positive performance by the MML estimator when compared against Maximum Likelihood (ML) and other alternative methods [Scho78, Fish93]. We are currently using the $h_3(\kappa)$ prior for Snob.

As well as the fact that MML tends to provide more reliable parameter estimates for the von Mises distribution than ML, being able to associate a message length with each class enables us to use (the minimisation of) the message length as a natural metric for model selection and determining the optimal classes we think we ought to have. (See Section 4.)

In practice, the Snob program makes the reasonable assumption that the standard deviation, $\sigma \geq 0.3\varepsilon$ when dealing with data from Normal distributions. Since $M_2(\mu, \kappa) \approx N(\mu, 1/\kappa)$ for large κ, Snob also makes the similar assumption that $\kappa \leq 1/(0.3\varepsilon)^2$ when dealing with data from von Mises distributions.

The use of a circular distribution is particularly advantageous here since it can acknowledge the proximity of -179 degrees and +179 degrees in a way that a Normal distribution can not.

4 Applying MML to Intrinsic Classification – the Snob Program

Given a set of objects or "things", it is a common problem to wish to intrinsically classify (or cluster) these things into would-be natural groupings (or "classes"). Such intrinsic classification can be thought of as concept formation. The Snob program [WaBo68, Wall86, Wall90, WaDo94b] is an application of MML to this problem of numerical taxonomy.

Each continuous-valued attribute is assumed to come from a Normal distribution and each circular-valued attribute is assumed to come from a von Mises distribution. Snob uses MML for both the model selection (number of classes and assignment of things to classes) and parameter estimation (estimating means and standard deviations, etc.). Essentially, Snob tries to discern the structure in the data. Snob will prefer to hypothesise the existence of an additional class in the data precisely when the information cost of stating the parameter estimates for this additional class is more than offset by the information gain in stating the things assigned to this new class in terms of the newer, more appropriate, parameter estimates.

4.1 Stating the message

Following earlier work [WaBo68, Wall86, Wall90, WaDo94b], we suppose the data (to be intrinsically classified) is given as a matrix of D attribute values for each of N "things", with some attribute values possibly missing.

Part 1 of the message, stating the hypothesis, H, comprises several concatenated message fragments, stating in turn:

1a. The number of classes. (All numbers are considered equally likely, although this could easily be modified.)

1b. The relative abundance of each class. (Creating names or labels for each class of length $-\log_2$ of the relative abundance, via a Huffman or other code, gives us a way of referring to classes later when, e.g. we wish to say which class a particular "thing" belongs to.)

1c. For each class, the distribution parameters of each class (as discussed in Section 3). Each parameter is considered to be specified to a precision of the order of its expected estimation error or uncertainty (see, e.g. [WaDo93 (pp3-4)]). For a larger class, the parameters will be encoded to greater precision and hence by longer fragments than for a smaller class.

1d. For each "thing", the class to which it is estimated to belong. (This can be done using the Huffman code referred to in 1b above.)

Part 1 of the message states our hypothesis of how many classes there are and what the distribution parameters (μ, κ, etc.) are for each attribute for each class, and, in 1d, what class each thing is in. Part 2 of the message states the data in light of this hypothesis.

The details of the encoding and of the calculation of the length of part 1 of the message may be found elsewhere [WaBo68]. It is perhaps worth noting here that since our objective is to minimise the message length (and maximise the posterior probability), we never need construct a message – we only need be able to calculate its length.

Given that part 1d of the message told us which class each thing was estimated to belong to and that, for each class, part 1c gives us the (MML) estimates of the distribution parameters for each attribute, part 2 of the message now encodes each attribute value of each thing in turn in terms of the distribution parameters (for each attribute) for the thing's class.

It turns out that if one permits probabilistic assignment of things to classes, then [Wall86 (Sect 3), Wall90 (Sect 3), WaDo94b (Sect 3.2)] the coding is more concise and statistically more accurate. Doing this takes advantage of the fact that [WaFr87 (p241)] "if the outcomes of any random process are encoded using a code that is optimal for that process, the resulting binary string forms a completely random process". The above coding scheme codes the parameter estimates from one distribution independently of the values of the parameter estimates of another distribution. This does not assume that the dihedral angles are independently distributed (see, e.g. raw data in figure 1(a)), but it does implicitly assume that, within each Snob class, the dihedral angles are independently distributed.

5 Statistical consistency of estimates

The quotation above and the fact that general MML codes are (by definition) optimal implicitly suggest that, given sufficient data, MML will converge as closely as possible to any underlying model. Indeed, MML can be thought of as extending Chaitin's idea of randomness [Chai66] to always trying to fit given data with the shortest possible computer program (plus noise) for generating it. This general convergence result for MML has been explicitly re-stated elsewhere [Wall89, BaCo91].

The problem of model selection and parameter estimation in intrinsic classification can, at its worst, be thought of as a problem for which the number of parameters to be estimated grows with the data. It is well known (see [WaDo94b]) that Maximum Likelihood can come unstuck with such problems.

Lastly, without using the partial assignments described in Section 4.1, estimation would be guaranteed to be weakly inconsistent. This presents no problem if the underlying classes are sufficiently well-separated, but [Wall90] if the means of two Normal distributions are separated by less than 2.5 times the true standard deviation of each component, the shortest (and also maximum likelihood) explanation would be incorrectly given by a 1-component model, no matter how large the data sample. A similar comment could be made about overlapping von Mises distributions.

A discussion of the similarities between Snob applied to Normal distributions and AutoClass [Ch88] is given in [Wall90 (pp78-80)], with an old but extensive discussion of alternative algorithms for intrinsic classification having been given in [Boul75]. An alternative, popular clustering method is COBWEB [Fish87].

6 Application

Earlier applications of Snob include several to medical, biological and psychological data, with a thorough survey in [Patr91] and a more recent survey in [WaDo94b]. (A guide to using the Snob program and interpreting its output is given in [WaDo94b (pp41-42)].)

The task of applying Snob or related automated clustering methods to protein substructure data is difficult for several reasons. First, existing useful classifications (e.g. secondary structure) are of varying lengths. A beta turn may contain just 3 amino acid residues, while a long helix may have more than 20. Because all existing clustering methods require a fixed length attribute vector, previous approaches have clustered fixed length sliding windows of the data, e.g. [HuSt91].

Our overall approach to the problem of variable lengths in substructure classes is to begin by finding classes of the smallest possible unit of protein substructure (the peptide bond) and then devise a method for finding class structure in the (variable length) sequences of these unitary classes. Since analysis of variable length sequences is more tractable than structures with varying numbers of dimensions (i.e. varying numbers of constituents), we hope that this approach will ultimately address the problem. This paper reports on our results for finding structure at the unitary level. We will also describe our work characterizing the transition probabilities between the discovered unitary classes. Future work will consider a clustering method for longer protein substructures.

Hunter and States [HuSt91] attempted to address this problem by looking at short fixed length substructures, instead of looking at single peptide bonds. In order to define a uniform frame of reference for the substructures, Hunter and States used the centre of mass and the moments of inertia of the fragments. These co-ordinates cannot be defined for single amino acids, and were unstable for amino acid pairs with nearly symmetric moments of inertia. Hunter and States settled on a substructure of five amino acids for their work, although they noted the problems with that approach. Since Snob permits us to use the von Mises distribution to classify on the ϕ and ψ angles of the peptide bond, we do not need to define a specific frame of reference. We expect this will lead to better results when these unitary classes are aggregated into our target substructure classes.

A second significant problem with most clustering methods is that they tend to require that the attributes in the description vectors be independent of each other. Neither positions of atoms in Cartesian co-ordinates nor the use of ϕ and ψ angles to describe the data result in independent distributions. Violation of the independence assumption typically leads the classifiers to produce too many classes. This work also suffers somewhat from this problem, although it is an improvement over the Hunter and States work since there are only two correlated dimensions instead of nine. As described below, we have taken several steps to factor out the correlation. We are exploring the possibility of adding a correlation term to each class description (allowing the correlations within each class to vary from the correlations within other classes), which is well within the ability of the MML framework, although not yet the Snob program.

7 Methods

We have applied Snob to clustering the (ϕ, ψ) angle data from a non-redundant subset of the PDB. Our stated measurement accuracy was 0.2 radians (11.5 degrees). Varying the measurement accuracy slightly produced little or no difference in the Snob classification. This gave the following results:

Axes	Message Length (nits)	Classes
ϕ, ψ	204,270.4	27

Table 1: Von Mises distribution parameters and sizes of Snob classes.

Class	μ_ϕ (phi)	κ_ϕ (phi)	μ_ψ (psi)	κ_ψ (psi)	Size
1	-1.4846	35.17	1.4782	6.02	714
2	-2.3491	26.23	2.6347	18.61	2409
3	-1.7799	4.80	2.1944	20.01	4667
4	-1.9841	16.48	2.3979	7.94	3756
5	-1.6641	55.85	0.0703	47.21	750
6	-1.4108	56.59	-0.1708	42.71	1152
7	-1.4544	28.26	2.7944	11.41	1509
8	-1.5713	9.50	-0.4434	9.78	2264
9	-2.7258	39.09	2.8028	21.48	1050
10	-2.9697	2.76	-3.1150	10.78	671
11	-1.1634	75.08	-0.4336	37.29	2359
12	-1.1393	39.89	-0.6503	54.92	4842
13	-1.0891	164.71	-0.7455	134.32	4560
14	-0.9740	77.24	-0.8621	92.40	2074
15	-1.9805	18.09	0.2318	12.62	1146
16	-2.3384	17.80	1.3521	8.69	516
17	-1.8083	3.14	-1.0773	0.09	1095
18	-0.8727	13.41	-0.9694	17.20	729
19	1.5602	1.44	0.2542	0.24	675
20	1.6205	25.92	-0.0804	18.09	528
21	1.3191	65.40	0.2938	27.59	424
22	1.0665	85.56	0.5843	41.98	341
23	0.8884	77.93	0.8709	28.48	343
24	1.0079	34.00	-2.3054	23.40	151
25	1.5073	17.16	3.1371	6.00	251
26	-1.1953	44.97	2.5447	21.28	1969
27	-1.0072	85.85	2.3741	46.00	786

where 1 nit = $\log_2 e$ bits. Notice, properties of the logarithm function ensure that it makes no difference whether we work in bits or nits.

Snob models the classes as having attributes from independent distributions. The entries in table 1 give the size, N, of each class and the attribute parameters, μ_ϕ and κ_ϕ (for ϕ) and μ_ψ and κ_ψ (for ψ) for each class. The p.d.f. of each class is thus $M_2(\mu_\phi, \kappa_\phi).\, M_2(\mu_\psi, \kappa_\psi)$, the product of the p.d.f.'s in ϕ and ψ. The Snob model of the population is then $\sum_{\text{classes } i} N_i.\, M_2(\mu_{\phi i}, \kappa_{\phi i}).\, M_2(\mu_{\psi i}, \kappa_{\psi i})$.

Figure 1 shows (a) a histogram of the raw data and (b) the Snob model from table 1. The Snob classes are represented in figure 2 with centres at (μ_ϕ, μ_ψ) and semi-axis lengths $1/\sqrt{\kappa_\phi}$ and $1/\sqrt{\kappa_\psi}$ respectively. The semi-axis length is due to the Normal approximation for large κ (see section 3.2) and the wrap-around of some classes is due to the (ϕ, ψ) axes being toroidal.

Since secondary structure is not purely a local property of the dihedral angles, we follow Hunter and States [HuSt91, HuSt92] and look at windows of consecutive sites. We ran our data through sliding windows (where subsequent windows overlap) and jumping windows (where subsequent windows are adjacent but do not overlap), both of width 2 and 3. This gave the following results (on page 8):

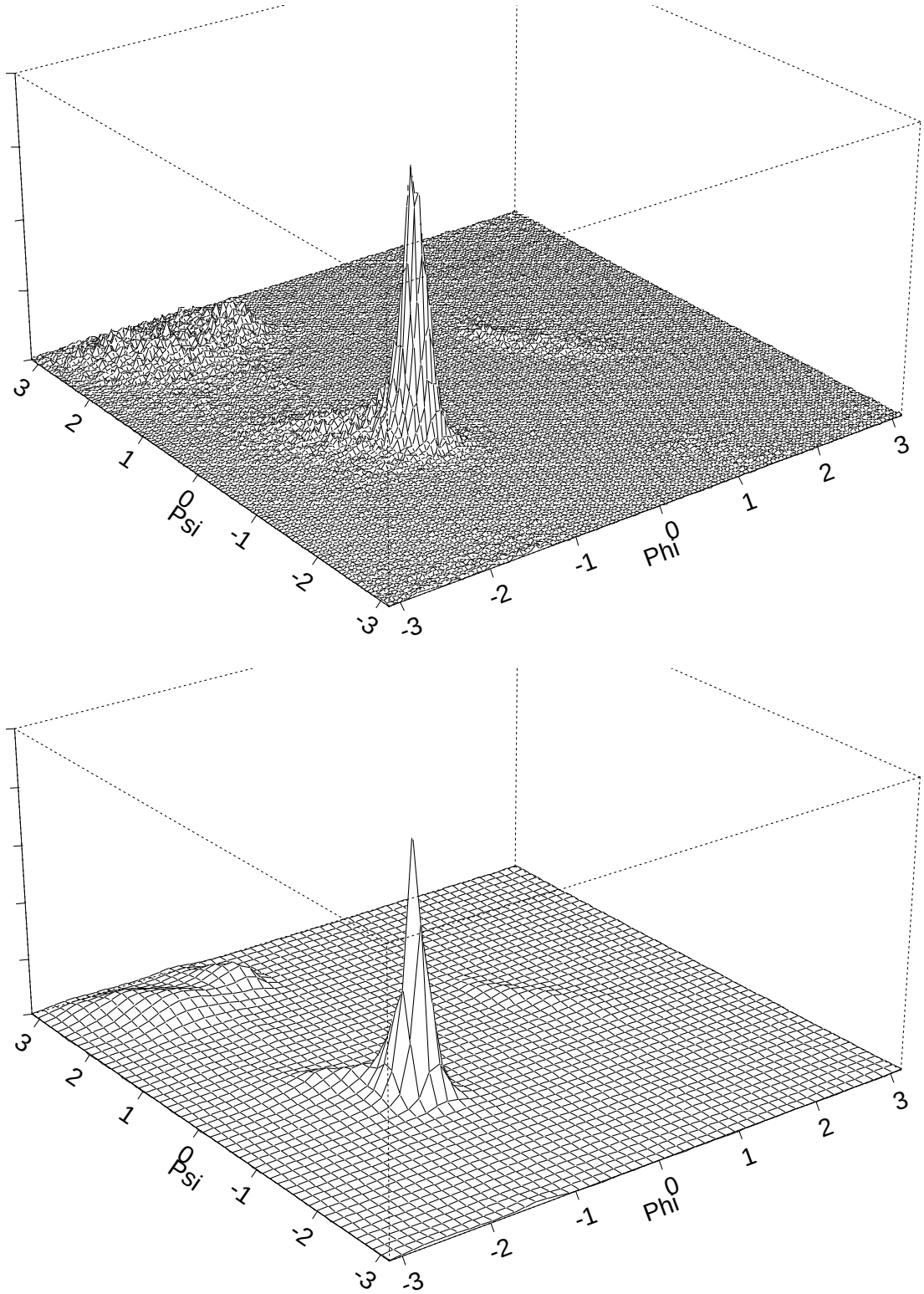

Figure 1: Histogram of (a) raw (ϕ, ψ) data and (b) Snob model.

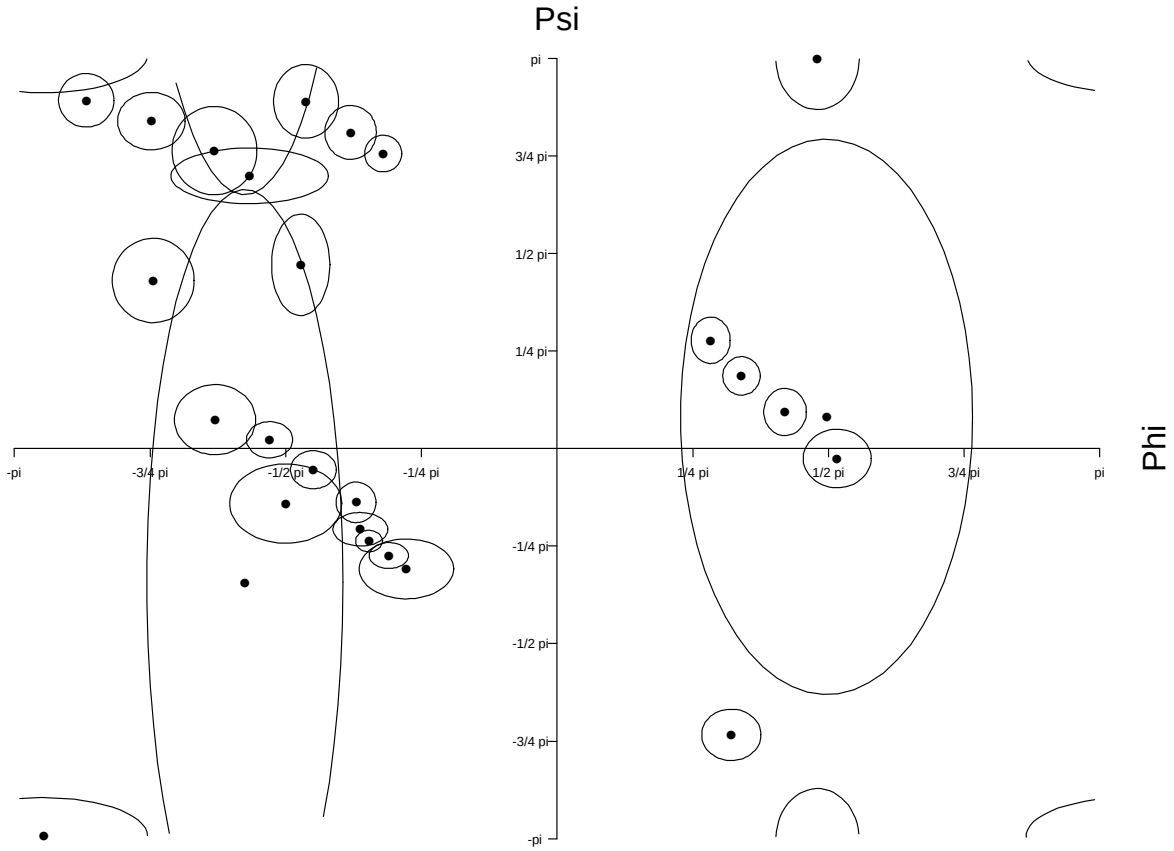

Figure 2: Snob classes in (ϕ, ψ) centred at (μ_ϕ, μ_ψ).

Axes	Message Length (nits)	Classes	Window
$(\phi, \psi)^2$	388,433.8	95	slide 2
$(\phi, \psi)^3$	580,391.3	90	slide 3
$(\phi, \psi)^2$	196,136.9	60	jump 2
$(\phi, \psi)^3$	195,844.6	83	jump 3

We note that the jumping windows of width 2 and 3 both contain essentially the same data as the original data, except the original data has 1 value of each of ϕ and ψ and the jumping windows have a half or a third as many data entries respectively.

It is to be expected that the message length from the sliding window of width 2 is less than twice the message length of the jumping window of width 2, since the cost of stating the theory becomes proportionally cheaper as the number of data items increases, and a jumping window of width 2 will give Snob close to half the number of data rows that Snob would have for a sliding window of width 2. A similar comment applies to the windows of width 3.

Observing some of our original (ϕ, ψ) plots overlaid with the Snob classes inferred from this data gave us reason to believe that we would like to transform, if possible, from (ϕ, ψ) space to $(\phi + \psi, \psi)$ space, since the original (ϕ, ψ) plot (see figure 2) seemed to show a cluster of Snob classes along the lines (expressed in radians) $\phi + \psi = -\pi/2$ and $\psi = 3\pi/4$. The axis transformations considered here and below have a fairly straightforward geometrical interpretation. They also serve as something of an attempt at factor analysis.

The permissible space of angles, (ϕ, ψ), corresponds to a torus since (in radians) $-\pi$ is equivalent to $+\pi$. As such, we have some constraints if our transformation of the parameter space is to be 1-to-1 and onto. For it to be well-defined, we require that $f(\phi + 2m\pi, \psi + 2n\pi) = f(\phi, \psi)$ for integers m and n. This basically just ensures that -179 degrees and $+179$ degrees, and also 1 degree and 359 degrees, are always treated as being close to one another, and remain so under transformation. We also require the transformation, f, to be invertible. Restricting ourselves to linear

transformations, we can always change the sign of the Jacobian by reversing the direction of an axis. The invertibility condition corresponds (modulo a sign change) to the Jacobian having determinant 1. It turns out that the class of permissible linear transformations from torus to torus can be generated multiplicatively from the set of matrices $\begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$, $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, where the matrices $\begin{bmatrix} 1 & 0 \\ m & 1 \end{bmatrix}$ and $\begin{bmatrix} 1 & n \\ 0 & 1 \end{bmatrix}$ are the m^{th} and n^{th} powers respectively of the last two matrices above.

Since we are using Minimum Message Length, the message length is the criterion that we use to determine whether or not we deem a transformation of the parameter space to be a good idea. Using a "colourless", innocuous prior on some transformations, the negative logarithm of which we add to the message length, we thus look at the effect on the message length of our considered transformations. The difference in message length is the log posterior odds ratio; in other words, if theory H_1 leads to a message length 5 bits shorter than H_2 does, then H_1 is deemed to be 2^5 times more likely than H_2 a posteriori.

Following the hints in the graphical plot of the above Snob classes, we transformed the data to $(\phi + \psi, \psi)$. We earlier deemed the accuracy, ε, on ϕ and ψ to be the same. Since the Jacobian of the transformations is unity, the region of area ε^2 corresponding to measurements of ϕ and ψ should map to a region of area $1 \times \varepsilon^2 = \varepsilon^2$ in ϕ and $(\phi + \psi)$. We thus adopt ε as the measurement accuracy in both ϕ and $(\phi + \psi)$. The graphical plots of the Snob classes hint that a better choice of Snob axes than (ϕ, ψ) might be $(\phi + \psi, \psi)$. The current implementation of Snob assumes that the distributions in ϕ and ψ are independent in each class, giving each class an elliptical profile (see figure 2) whose major and minor axes align with the ϕ and ψ axes. We experimented with re-aligning the Snob axes. Transforming the data to $(\phi + \psi, \psi)$ actually increased the message length as follows:

Axes	Message Length (nits)	Classes
$\phi + \psi, \psi$	204,515.2	32

The work in the next section builds on the better results of the (ϕ, ψ) data. However, in work just completed we noticed the $(\phi + \psi, \psi)$ plot showed clusters of Snob classes roughly along the lines $\psi = (\phi + \psi) - \pi/2$ and $\psi = (\phi + \psi) + \pi/2$, lines for which we note ϕ is a constant. This led us to re-introduce the ϕ axis while retaining the $(\phi + \psi)$ axis to obtain the following results:

Axes	Message Length (nits)	Classes
$\phi, \phi + \psi$	204,249.3	23

This reduced the message length from the (ϕ, ψ) classification by approximately 21 nits and also has reduced the number of classes and needs to be investigated further.

8 Similarity measure between classes

We wish to arrive at a distance measure between classes. Just as the Dayhoff substitution measure between amino acids gives a method for determining similarity between two proteins based on their amino acid sequence, using an assignment of protein dihedral angle pairs to Snob classes, the metric below gives a method of determining similarity between two proteins based on their tertiary structure. One candidate for such a metric was the cost that Snob associates with a forced combining of the two relevant Snob classes. The shortcoming with this metric is that it is very much affected by the population sizes of the distributions. On the one hand, two very similar and largely overlapping Snob classes which had very similar parameter values could be expensive (in terms of Snob's information-theoretic bit cost) for Snob to combine if they were both very abundantly populated. On the other hand, two apparently very different classes with antipodal values of μ and large concentration parameters could be relatively inexpensive for Snob to combine if the antipodal classes had sufficiently small populations.

We conclude from the above that we would rather have a distance measure which is a function of the probability distribution parameters of the Snob classes, independent of the class sizes. For probability distributions $f(x)$ and $g(x)$, the Kullback-Leibler distance from f to g is defined by

$$dist_{K-L}(f, g) = -\int f \log g \; dx - (-\int f \log f \; dx) = \int f \log(f/g) \; dx.$$

The Kullback-Leibler distance essentially gives the difference between the cost of coding events from the probability distribution, f, using codewords of length $-\log g$ with the optimal code obtained by coding events from the probability distribution, f, using codewords of length $-\log f$. Some properties are that $dist_{K-L}(f, g) \geq 0$ and $dist_{K-L}(f, f) = 0$, although the Kullback-Leibler distance is not generally symmetric in f and g. For von Mises distributions $M_2(\mu_f, \kappa_f)$ and $M_2(\mu_g, \kappa_g)$, we get that [WaDo93]

$dist_{K-L}(f, g) = \log I_0(\kappa_g) - \log I_0(\kappa_f) + A(\kappa_f)\ (\kappa_f - \kappa_g \cos(\mu_f - \mu_g))$, where $I_p(\kappa) = \frac{1}{2\pi} \int_0^{2\pi} \cos(p\theta).\, e^{\kappa.\cos\theta}\ d\theta$, and $A(\kappa) = I_1(\kappa)/I_0(\kappa)$ is the expected value of $\cos(\theta)$ under an $M_2(0, \kappa)$ distribution.

We take a symmetrical distance measure based on the Kullback-Leibler measure which is vaguely analogous to the cost of forcing Snob to combine two classes. Let two classes have p.d.f.'s f_1 and f_2 given by $M_2(\mu_1, \kappa_1)$ and $M_2(\mu_2, \kappa_2)$ respectively.

Consider the resultant vectors of length $A(\kappa_1)$ pointing in direction μ_1 and $A(\kappa_2)$ pointing in direction μ_2 respectively. The vector sum of these gives the direction of μ_g. Halving the length of this resultant to effectively average the contribution from the two distributions, the resultant length, l, will satisfy $0 \leq l < 1$. We define κ_g to satisfy $A(\kappa_g) = l$.

Our symmetrical, "similarity", distance measure is then $d(f_1, f_2) = dist_{K-L}(1/2(f_1 + f_2), g) = 1/2\ dist_{K-L}(f_1, g) + 1/2\ dist_{K-L}(f_2, g)$.

This has several desirable properties.

1. If $\mu_1 = \mu_2 = \mu$ and $\kappa_1 = \kappa_2 = \kappa$, then $\mu_g = \mu$ and $\kappa_g = \kappa$ and $d(f_1, f_2) = 0$.
2. If $\mu_1 - \mu_2 = \pi\ mod\ 2\pi$ and $\kappa_1 = \kappa_2$, i.e. if the distributions are antipodal mirror reflections, then $\kappa_g = 0$ and $d(f_1, f_2) = dist_{K-L}(f_1, g) = dist_{K-L}(f_2, g)$.
3. It is additive across multiplicatively independent distributions. In other words, thinking of Snob classes of window width 1 as being of the form $f(\phi, \psi) = f_\phi(\phi) f_\psi(\psi)$, we have $d(f_1, f_2) = d(f_{\phi,1}, f_{\phi,2}) + d(f_{\psi,1}, f_{\psi,2})$.

We have extended the Snob program [WaDo94b] to carry out these calculations. In order to better visualize the unitary classification, we have used these calculations to build a minimum spanning tree from our Snob classes as follows: Iteratively work out which two of the current classes are closest together under this distance measure. Then reduce the number of classes by one, replacing these two classes with distributions f_1 and f_2 by a class having distribution g as above. The node referring to g then becomes the parent node of the nodes referring to f_1 and f_2. This is iterated until we are left with the root node. Backtracking gives the minimum spanning tree. Notice that our original flat classification is based on the MML principle. The hierarchical classification of the von Mises classes here is based on an information-theoretic distance measure, but the spanning tree does not have a messsage length directly associated with it.

Figure 3 shows the minimum spanning tree for the 27 classes in table 1. This can be viewed as a form of hierarchical cluster, with the caveat that the formula used in section 7 for calculating g from f_1 and f_2 would then be assuming that all classes are arbitrarily large. The tree has five sub-trees at intermediate nodes 33, 37, 39, 41 and 52 which partition all but four of the classes into groups such that the classes in each group seem fairly close using our symmetric distance measure.

Table 2 gives the transition matrix whose entry in row i, column j is $\log(Pr(Class_j \mid predecessor\ in\ Class_i) / Pr(Class_j))$. The entries for classes 12, 13 and 14 show that class 13 prefers most to be preceded by class 13 and has a positive inclination to be preceded by classes 12 and 14. Class 13 has a disinclination to be preceded by every other class. Classes 12, 13 and 14 have close proximity on the minimum spanning tree. Class 13 has large concentration parameters, κ_ϕ and κ_ψ (see table 1), corresponding to standard deviations in ϕ and ψ of 4.5 and 4.9 degrees respectively. Also, approximately 11% of the population is in this class. Classes 12, 13 and 14 show a preference to be preceded by one another and together contain some 27.5% of the population. This suggests that class 13 is an abundant, tight (having large values of κ) and self-perpetuating (highly auto-correlated) structure, with μ values also corresponding to those of α-helices. A similar comment applies to the group of classes 12, 13 and 14 together. Notice these classes are below the intermediate node 33 in the minimum spanning tree.

We note in passing from the transition matrix that class 5 prefers not to succeed all but five of the classes. It prefers not to succeed itself and of the five classes it prefers to succeed, only one of these (class 24) prefers to succeed it. This suggests a possible transition conformation.

9 Conclusions

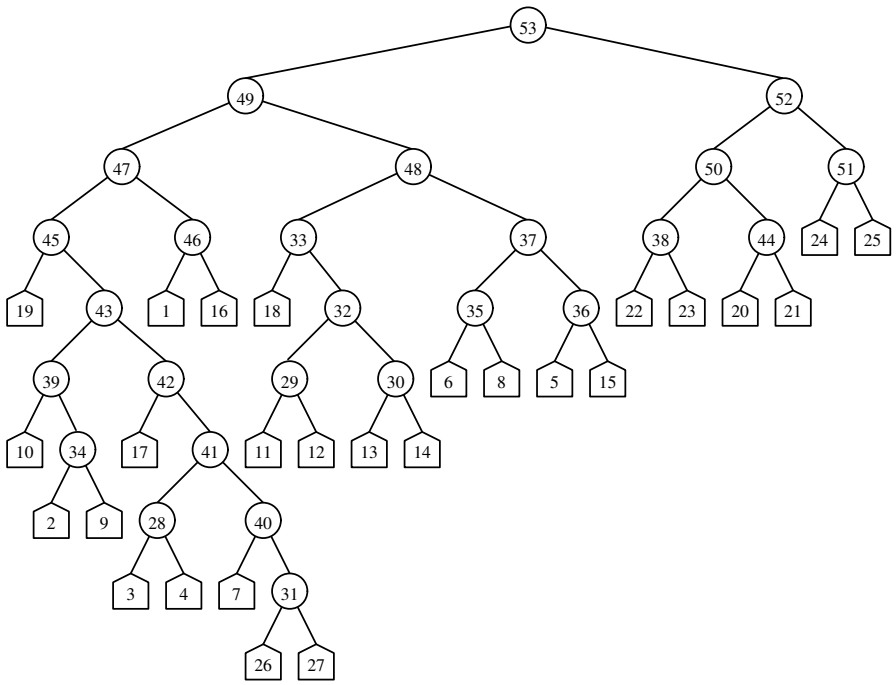

Figure 3: Minimum spanning tree formed by iteratively merging closest classes.

Hunter and States [HuSt91, HuSt92] had earlier clustered protein secondary structure conformations using nine highly correlated Cartesian co-ordinates per secondary structure site. Using recently developed code for clustering of circular data, we have been able to carry out similar work to Hunter and States, but instead using only two orientation-invariant angles as our data per site. This work using the Snob program is founded on the Minimum Message Length principle, whose firm statistical basis we have alluded to. Presuming a measurement accuracy of 11.5 degrees, Snob arrived at 27 classes in (ϕ, ψ).

We formed a minimum spanning tree of the classes and also a transition matrix between the classes. The highlight was one very tight class containing approximately 11% of the population. Its members have a strong disinclination to be preceded in a sequence by members of any class other than itself or two of its nearest neighbours on the minimum spanning tree. We also found that this set of three classes contained approximately 27.5% of the population and tended to prefer being preceded by its own members, suggesting an abundant, tight, self-perpetuating, α-helical structure.

The theory behind MML single linear factor analysis [WaFr92] and multiple linear factor analysis [Wa95] has been completed. The program currently implicitly assumes that variables are uncorrelated and does not yet use the MML factor analysis[WaFr92, Wa95]. Where there is correlation, linear factor analysis (which permits axis rotation) should enable the data to be better compressed.

As a Ramachandran plot of the raw (ϕ, ψ) data demonstrates, the distributions in ϕ and ψ are not independent of one another. We have investigated axis transformations as a way of dealing with this. One alternative, currently in the early stages, is to extend the theory of MML factor analysis (on Normal distributions) [WaFr92, Wa95] to single factor analysis for von Mises distributions. Another alternative is to permit the individual Snob classes to have their own, separate, axis transformations. We also wish to extend the MML decision graph [OlDW92] work on inferring a probabilistic theory of {Extended, Helix and Other} secondary structure from the primary amino acid sequence to inferring a theory using this larger number of secondary structure classes here. Given the clear serial correlation discussed and observed between some of the classes, it would be desirable to extend Snob to permit a more explicit model of this. We note in passing that Snob has also been used to cluster protein spectral data into classes [ZaCD].

Table 2: Log of number of observed transitions from class in row i to class column j divided by expected number of transitions if classes were uncorrelated.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
1	0.6	0.4	-0.2	-0.1	-0.1	-0.2	0.6	-0.1	0.8	0.4	0.3	-0.3	-1.5	-0.3	0.5	0.8	1.0	0.6	1.2	-0.9	-1.1	-0.5	-0.8	-0.1	0.3	-0.1	0.2
2	-0.0	1.0	0.7	0.7	-1.1	-0.8	0.4	-1.4	0.6	0.7	-1.1	-1.4	-2.4	-1.9	-0.8	-0.5	-0.6	-0.8	-0.7	-2.1	-1.0	-0.7	-0.2	0.5	0.2	0.4	0.1
3	0.4	0.3	0.8	0.8	-0.4	-0.8	0.3	-0.2	-0.2	0.2	-0.5	-1.1	-2.3	-1.7	-0.1	0.0	0.2	-0.2	-0.1	-0.1	-0.7	-0.3	0.3	0.4	-0.2	0.3	0.3
4	0.2	0.8	0.7	0.7	-1.3	-1.0	0.4	-0.9	0.5	0.3	-0.8	-1.2	-2.2	-1.4	-0.8	-0.1	-0.1	-1.1	-0.4	-2.2	-1.1	-0.3	-0.2	0.2	0.1	0.4	0.5
5	-0.0	-0.3	-0.1	-0.3	-0.6	-0.8	0.6	0.0	-0.0	0.3	-1.2	-0.9	-2.3	-0.4	0.2	0.5	0.0	-0.1	0.6	1.5	1.9	1.9	1.6	0.2	0.8	0.6	0.6
6	0.5	-0.6	-0.2	-0.5	0.9	0.3	-0.0	0.9	-0.4	0.3	-0.5	-0.8	-1.9	-0.8	1.0	0.9	0.2	0.0	-0.1	1.3	1.6	1.4	1.3	0.6	0.8	-0.0	0.2
7	0.2	0.3	0.1	-0.1	-0.5	-0.1	0.58	-0.3	0.3	0.2	0.2	-0.3	-0.8	-0.3	-0.1	-0.2	0.1	0.0	-0.3	-1.0	-0.5	0.3	-0.2	-0.5	0.2	0.7	0.9
8	0.4	0.3	-0.3	-0.1	-0.4	-0.4	-0.2	0.8	0.9	0.6	-0.6	-0.3	-1.4	-0.1	0.6	0.9	0.6	0.5	0.6	0.6	0.4	-0.0	0.2	0.8	0.8	-0.2	0.1
9	-0.2	0.8	0.4	0.7	-0.6	-0.8	0.3	-0.7	1.1	0.9	-0.7	-1.3	-2.0	-1.9	0.1	0.3	0.0	-1.7	-0.3	-1.4	-1.3	-0.8	0.3	0.7	0.9	0.3	0.1
10	0.4	0.9	0.2	0.5	-2.1	-0.0	0.3	-0.7	0.8	0.7	-0.5	-0.7	-2.3	-1.3	0.5	-0.1	0.8	-0.4	0.7	-0.7	-2.5	-1.1	-1.1	0.1	0.9	0.4	-0.2
11	-0.5	-1.6	-1.3	-1.3	1.8	1.4	-1.4	1.0	-0.8	-0.7	0.8	0.3	-0.8	-1.0	1.1	-0.1	-0.9	-0.1	-0.4	-0.5	0.3	-0.3	-0.0	-0.2	-0.1	-1.4	-1.5
12	-1.0	-1.6	-1.9	-1.9	0.6	0.7	-2.3	0.5	-0.7	-1.4	0.6	0.8	0.7	0.4	-0.1	0.0	-1.6	-0.0	-1.3	-1.7	-1.3	-2.0	-1.1	-0.8	-0.4	-2.2	-3.1
13	-2.0	-3.2	-3.1	-3.6	-1.0	-0.4	-3.4	-1.0	-2.4	-2.8	0.1	0.7	1.3	0.9	-1.6	-1.2	-2.1	-0.4	-2.5	-3.1	-2.5	-3.5	-2.7	-2.8	-2.8	-4.0	-3.3
14	-2.0	-3.0	-2.8	-2.6	-1.3	-0.0	-2.7	-1.1	-2.3	-2.3	0.5	0.7	0.9	1.4	-1.0	-1.1	-1.1	1.1	-0.7	-2.0	-2.2	-1.9	-2.4	-1.7	-1.7	-2.8	-3.3
15	0.1	0.1	-0.0	0.1	-0.5	-0.5	0.5	-0.1	0.5	0.6	-0.5	-0.6	-1.5	-0.7	0.1	0.4	0.5	-0.4	0.7	0.5	0.9	1.5	1.5	0.3	1.1	0.5	0.4
16	0.9	-0.8	-0.2	-0.4	-1.0	0.1	0.6	-0.6	-1.0	0.5	0.8	0.1	-1.9	-0.7	-0.8	-0.2	1.4	-0.1	0.7	-0.2	-0.3	-0.0	0.5	1.5	0.5	0.8	0.9
17	0.5	-0.5	0.0	-0.2	-0.3	-1.0	-0.2	0.4	0.3	1.0	-0.6	-0.7	-1.6	-0.4	0.6	0.6	1.8	0.8	2.5	-0.3	-0.7	-2.1	0.2	0.9	0.7	-0.3	0.1
18	0.5	-1.8	-1.4	-1.9	0.2	-0.0	-1.9	0.4	-1.1	-0.9	0.4	0.5	-0.3	1.0	0.5	0.5	0.9	2.4	0.7	-0.6	-0.4	-2.1	-0.5	0.6	-0.9	-2.0	-3.0
19	0.6	-0.2	-0.2	-0.4	-1.5	-0.5	0.5	0.4	0.5	0.7	-0.5	-1.1	-2.1	-0.4	0.4	1.0	1.7	0.4	2.3	0.3	-1.3	-1.1	0.9	0.8	1.1	0.4	-0.1
20	0.3	-0.2	0.3	0.3	-0.3	-0.4	1.5	0.2	-0.3	-1.1	-1.3	-1.1	-2.1	-1.2	-0.1	0.1	0.4	-0.2	-0.0	-0.5	-0.9	-1.2	-1.1	0.4	-0.5	1.3	0.4
21	0.3	0.4	0.3	0.8	-0.5	-1.0	0.9	0.1	0.2	-1.4	-1.0	-0.8	-2.1	-2.2	0.6	-0.5	0.1	-0.4	-2.4	-0.3	-0.1	0.3	-0.1	-0.2	-0.3	0.6	0.5
22	0.7	0.1	0.1	0.6	-0.1	0.1	0.4	-0.2	-0.6	0.1	-0.8	-0.9	-2.0	-2.5	0.6	0.3	0.6	-2.2	-1.1	1.4	1.6	1.8	0.4	-1.1	-0.6	0.3	0.2
23	-0.1	0.1	-0.2	0.1	-0.2	-0.4	0.7	-0.3	0.1	0.2	-0.1	-1.2	-3.0	-1.0	0.2	0.2	0.1	-2.1	0.4	2.0	2.1	1.5	0.4	-1.1	-0.5	0.6	0.5
24	1.3	-1.2	-1.9	-0.9	2.0	1.5	-2.5	-0.3	-0.3	0.1	1.2	-0.4	-2.0	-1.8	1.4	0.4	-1.4	0.2	0.8	-0.5	-1.4	-1.1	-1.1	-0.4	1.0	-1.9	-0.4
25	1.1	0.2	-0.0	0.1	-1.0	0.1	0.4	-0.0	0.6	0.6	0.5	-0.6	-1.3	-0.3	-0.5	-2.1	1.0	-0.4	0.7	0.1	-0.8	-0.1	-0.0	1.3	-0.4	0.7	-0.3
26	0.2	0.3	0.2	0.2	-0.5	-0.5	0.6	0.0	0.3	0.2	-0.1	-0.5	-1.4	-0.7	-0.1	-0.1	-0.2	-0.5	-1.1	0.4	-0.1	0.3	-0.3	-0.7	-0.2	0.8	0.9
27	0.5	-0.4	-0.1	-0.1	-0.4	-0.1	0.1	0.2	0.1	0.3	0.1	-0.5	-2.4	-0.8	-0.0	-0.4	0.3	-0.4	0.2	2.4	1.8	1.2	0.0	0.0	-0.3	0.1	0.7

Availability of the Snob program

The current version of the Snob program (written in Fortran 77) is freely available for not-for-profit, academic research, and not for re-distribution, from <URL:ftp://ftp.cs.monash.edu.au/pub/snob/> or directly from C.S. Wallace. Usage restrictions are given in [WaDo94b].

Acknowledgements

This work was supported by the Australian Research Council (Grant A49330656 and Small Grant 9103169), and Monash University Faculty of Computing and Information Technology funding (to assist L. Hunter).

References

A.R. Barron and T.M. Cover, Minimum Complexity Density Estimation, *IEEE Transactions on Information Theory*, 37, 1034-1054, 1991.

R.A. Baxter and J.J. Oliver, MDL and MML: Similarities and Differences, Technical Report 95/207, Dept of Computer Science, Monash University, Australia, 1995.

D.M. Boulton, The Information Criterion for Intrinsic Classification, PhD thesis, Dept. of Computer Science, Monash University, Australia, 1975.

D.M. Boulton and C.S. Wallace, The Information Content of a Multistate Distribution, *J. Theoret. Biol.*, 23, 269-278, 1969.

D.M. Boulton and C.S. Wallace, A Program for Numerical Classification, *The Computer Journal*, 13(1), 63-69, 1970.

G.J. Chaitin, On the length of programs for computing finite sequences, *J. Assoc. Comp. Mach.*, 13(4), 547-549, 1966.

P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Taylor, D. Freeman, AutoClass: A Bayesian Classification System, *Proc. 5th International Conference on Machine Learning*, Ann Arbor, MI, U.S.A., Morgan Kaufmann, 1988.

P. Cheeseman, J. Stutz, R. Hanson, W. Taylor, AutoClass III, Research Institute for Advanced Computer Science, NASA Ames Research Centre, USA, 1990.

D.L. Dowe, L. Allison, T.I. Dix, C.S. Wallace, T. Edgoose, Circular Clustering of Protein Dihedral Angles by Minimum Message Length, to appear in *Proc. of First Pacific Symposium on Biocomputing*, HI, USA, 1996.

D.L. Dowe, J.J. Oliver, T.I. Dix, L. Allison & C.S. Wallace, A Decision Graph Explanation of Protein Secondary Structure Prediction, *Proc. of 26th Hawaii International Conf. on System Sciences* 1, 669-678, Maui, HI, USA, Jan 1993.

D. Fisher, Knowledge Acquisition Via Incremental Conceptual Clustering, *Machine Learning*, 2, 139-172, 1987.

N.I. Fisher, *Statistical Analysis of Circular Data*, Cambridge Univ. Press, Cambridge, 1993.

L. Hunter, D.J. States, Applying Bayesian Classification to Protein Structure, *Proc. 7th IEEE Conference on Artificial Intelligence Applications*, 10-16, Feb. 1991.

L. Hunter, D.J. States, Bayesian Classification on Protein Structure, *IEEE Expert*, 7(4), 67-75, 1992.

W. Kabsch and C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, 22(12), 2577-2637, 1983.

M. Kamimura and Y. Takahashi, $\phi - \psi$ conformational pattern clustering of protein amino acid residues using the potential function method, *CABIOS*, 10(2), 163-169, 1994.

M. Levitt, Accurate modeling of protein conformation by automatic segment matching, *J Mol Biol*, 226(2), 507-533, 1992.

J. Oliver, D.L. Dowe and C.S. Wallace, Inferring Decision Graphs Using the Minimum Message Length Principle, in *Proc. of Austn. Artificial Intelligence Conference '92*, Hobart, Tasmania, 361-367, 1992.

J.D. Patrick, Snob: A program for discriminating between classes, Technical Report 91/151, Dept of Computer Science, Monash University, Australia, 1991.

N. Qian and T. Sejnowski, Predicting the Secondary Structure of Globular Proteins using Neural Network Models, *J Mol Bio*, 202(4), 865-884, 1988.

J. Rissanen, *Stochastic complexity in statistical inquiry*, World Scientific Series in Computer Science, 15, 1989.

M.J. Rooman, J. Rodriguez and S.J. Wodak, Automatic definition of recurrent local structure motifs in proteins, *J. Mol.*

Biol., 213,327-336, 1990.

G. Schou, Estimation of the concentration parameter in von Mises-Fisher distributions, *Biometrika*, 65(1), 369-77, 1978.

R.J. Solomonoff, A formal theory of inductive inference (I and II), *Information and Control*, 7, 1-22 and 224-254, 1964.

R.J. Solomonoff, The Discovery of Algorithmic Probability: A Guide for the Programming of True Creativity, in *Computational Learning Theory: EuroCOLT'95*, ed. P. Vitanyi, Springer-Verlag, 1-22, March 1995.

V.V. Solovyev and A.A. Salamov, Predicting alpha-helix and beta-strand segments of globular proteins, *Comput Appl Biosci*, 10(6), 661-669, 1994.

S. Sun, Reduced representation model of protein structure prediction: statistical potential and genetic algorithms, *Protein Sci*, 2(5), 762-785, 1993.

M.B. Swindells, A procedure for detecting structural domains in proteins, *Protein Sci*, 4(1), 103-112, 1995.

C.S. Wallace, An Improved Program for Classification, *9th Australian Computer Science Conference (ACSC-9)*, 8(1), 357-366, 1986.

C.S. Wallace, False Oracles and SMML Estimators, Technical Report 89/128, Department of Computer Science, Monash University, Australia, June 1989.

C.S. Wallace, Classification by Minimum-Message-Length Inference, in *Advances in Computing and Information - ICCI'90*, S.G. Akl et al (eds.) *LNCS* 468, Springer-Verlag, 72-81, 1990.

C.S. Wallace, Multiple Factor Analysis by MML Estimation, Technical Report 95/218, Dept of Computer Science, Monash University, Australia, 1995.

C.S. Wallace and D.M. Boulton, An Information Measure for Classification, *Computer Journal*, 11(2), 185-194, 1968.

C.S. Wallace and D.M. Boulton, An Invariant Bayes Method for Point Estimation, *Classification Society Bulletin*, 3(3), 11-34, 1975.

C.S. Wallace and D.L. Dowe, MML estimation of the von Mises concentration parameter, Technical Report 93/193, Dept of Computer Science, Monash University, Australia, 1993; submitted to Aust. J. Statistics.

C.S. Wallace and D.L. Dowe, Estimation of the von Mises concentration parameter using Minimum Message Length, *Proc. 12th Australian Statistical Society Conference*, Monash University, Australia, 1994.

C.S Wallace and D.L. Dowe, Intrinsic classification by MML - the Snob program, *Proc. 7th Australian Joint Conference on Artificial Intelligence*, Armidale, NSW, Australia, World Scientific, 37-44, Nov 1994.

C.S. Wallace and P.R. Freeman, Estimation and Inference by Compact Coding, *Journal Royal Statistical Society, Series B, Methodology*, 49(3), 240-265, 1987.

C.S. Wallace and P.R. Freeman, Single Factor Analysis by MML Estimation, *J.R. Statist. Soc. B*, 54(1), 195-209, 1992.

J.D. Zakis, I. Cosic and D.L. Dowe, Classification of protein spectra derived for the Resonant Resonant Recognition model using the Minimum Message Length principle, *17th Australian Computer Science Conference (ACSC-17)*, Christchurch, NZ, 209-216, Jan 1994.