Chapter 9

SHEET CONFIGURATIONS

9.1 Introduction and Motivation

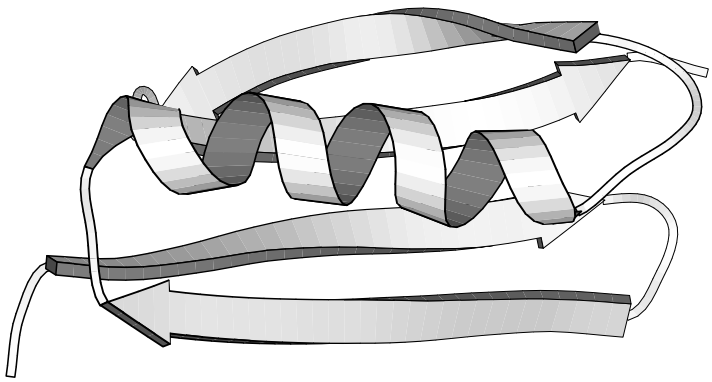
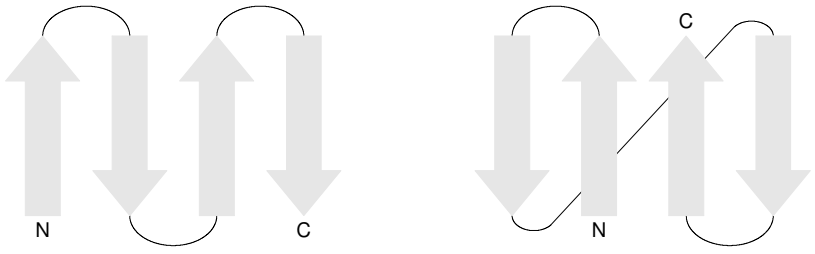


Figure 9.1: Highlighting the secondary structure elements in the three-dimensional structure of Protein L. We will use the four-stranded sheet of Protein L as example in this introduction.

When a protein is folded ab initio with ROSETTA, the score (introduced in Chapter 8) gets a big boost when two strands come together and form a sheet, or if a single strand gets attached to an already existing sheet. After generating many decoys using ROSETTA,

we found that very frequently the sheets consist of strands that are adjacent in sequence, building an ”up-down” motif as shown in Figure 9.2(a).



(a) The “up-down” motif, a common motif in ROSETTA decoys.

(b) The sheet configuration of the four strands in Protein L.

Figure 9.2: Two common configurations of four-stranded sheets.

There are plenty of proteins of known fold that actually do have such a four-stranded sheet motif. Our concern was that dis-proportionally many decoys made with ROSETTA had that up-down motif. This suspicion got confirmed when we compared local pairs of strands in sheets between ROSETTA decoys and real proteins. A local pair of strands is a pair of neighbor strands in a sheet with the two strands adjacent along the backbone of the protein. Otherwise, the pair is called non-local. The up-down motif in Figure 9.2(a) therefore has three local pairs and no non-local pair of strands, while Protein L (Figure 9.1) has a motif with two local pairs and one non-local pair of strands, see Figure 9.2(b). In Figure 9.3 we show the distributions of local versus non-local strand pairs in decoy sets we created for different proteins. This distribution seems appropriate in decoy sets for proteins that have few non-local pairs, such as the proteins with four-letter abbreviations 1pgx, 1sro, 1vif and 2ptl. The latter is the abbreviation for Protein L, and 44% of the decoys we made actually had the correct number of local and non-local pairs of strands. Matters get considerable

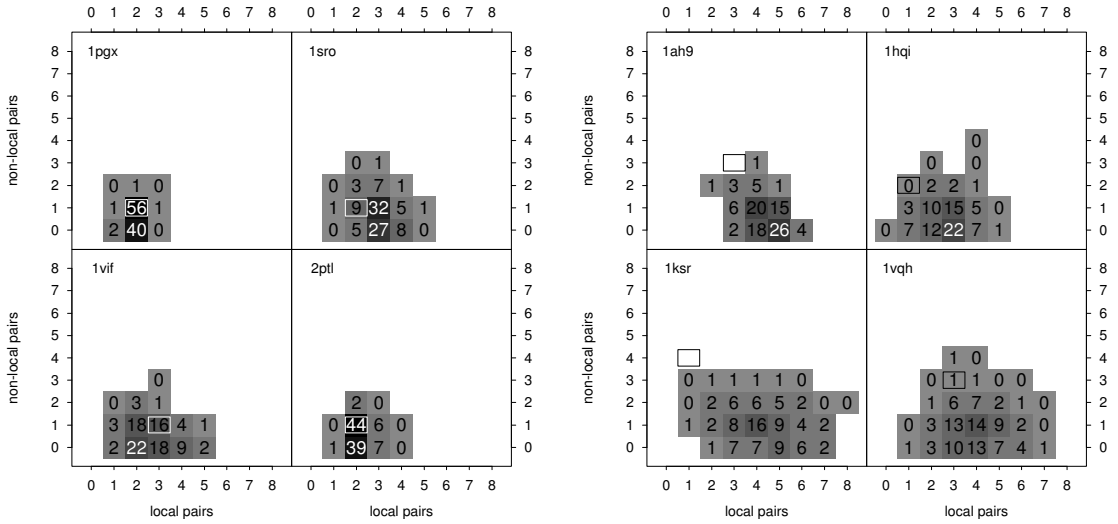


Figure 9.3: The distributions of local versus non-local strand pairs in decoy sets for various proteins. The numbers super-imposed onto the panels are rounded percentages of the frequency of decoys with the respective number of local and non-local strand pairs. A zero therefore stands for a percentage p with $0 < p < 0.5$.

worse in decoy sets for proteins with many non-local strand pairs, such as the proteins abbreviated by 1ah9, 1hqi, 1ksr and 1vqh. For 1ah9 and 1ksr we actually failed to make a single decoy (out of 10000+ decoys) with the correct number of local and non-local strand pairs!

The strand-strand packing term introduced in Chapter 8 only governs how many sheets will be formed, given the number of strands. However, it does not influence how strands get arranged in sheets. Given the above, there is plenty of motivation for developing a probabilistic model of what we call sheet configurations. We would certainly like to know a priori which motifs we should expect to see in a set of decoys we fold with ROSETTA. Quite possibly, some motifs that occur frequently in the database might be rare or completely missing in the set of decoys, and at the same time there might be plenty of up-down motifs although we would not really expect to see them in the structure we are considering.

The scoring function in ROSETTA is helpful in guiding a sequence of amino acids to a reasonable, protein-like fold. But of course it does not generate the correct fold deterministically, and a variety of structures is built. Since these decoys were built using the same scoring function, they are usually fairly low-scoring, and we do not expect to see very much of a correlation between score and root mean square deviance. Hence we have to rely on some "post-filters", i. e. scoring functions somewhat independent of the ROSETTA scoring function, to select the best among the decoys. These post-filters include all-atom potentials (after adding all side chain atoms to the decoys), and clustering procedures [42]. A probabilistic model of sheet configurations certainly could be used as a post-filter. It could also be used to pre-select a subset of the decoys with a sheet configuration distribution according to the probabilistic model, basically as a "prior" distribution, and then use the above mentioned filters to search for the best decoys. Ideally, the probabilistic model of sheet configurations could be used as a feature in the scoring function, which we hope we can implement some time in the future.

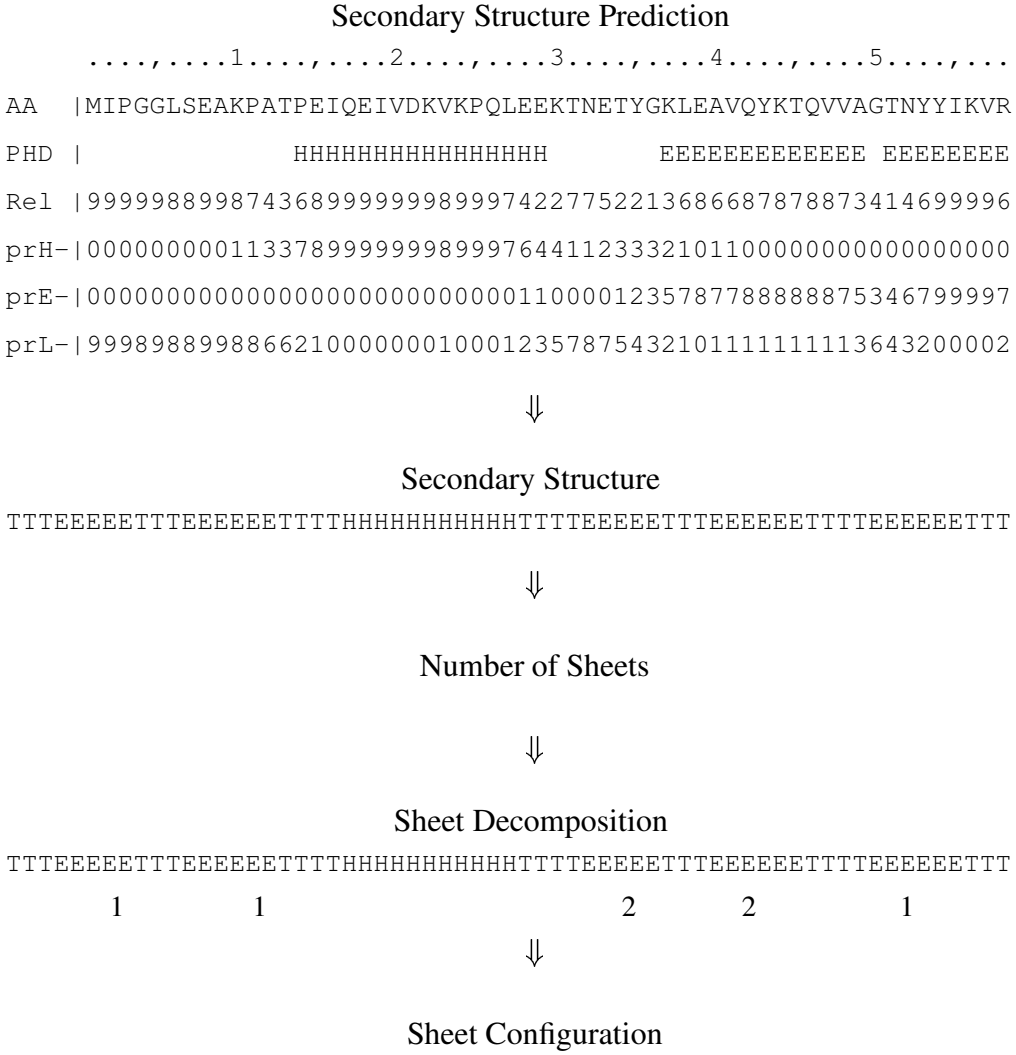
9.2 A Model for Sheet Configurations

The proteins we try to fold usually do not exceed 150 residues, and therefore the number of strands in the folds almost never reach double digits. The data we gathered from the database (see http://www.fccc.edu/research/labs/dunbrack/culledpdb.html) were from proteins of unrestricted lengths, but we only included sheets with at most ten strands in the investigation. We want to stress though that by the way we build the model, sheets of any size can be scored. Further, some proteins build barrels, i. e. secondary structure ensembles in which every strand has two neighbors. ROSETTA currently is not able to build these structures consistently, and therefore we limit ourselves to scoring decoys of proteins that do not have barrel motifs. In principle, the scoring function could easily be modified if one wanted to include the barrels.

Strictly speaking, nothing but the sequence of amino acids of the protein is known in ab initio structure prediction. To select the fragments in the move set of ROSETTA (see [44]), we make use of secondary structure predictions, a service freely available on the Web (for example under http://www.embl-heidelberg.de/predictprotein/predictprotein.html). Since the secondary structure predictions are used to generate the move set in the simulated annealing, the secondary structure elements in the decoys are usually close to those secondary structure predictions. The secondary structure predictions also include measures of certainty associated with the predictions at all positions, which can be considered fairly reliable. With a more complicated framework it might be possible to build the sheet configuration model incorporating the uncertainty about the secondary structure prediction, and build a model to quantify the probabilities of the overall configurations in the decoys (including the number of strands). However, we decided to consider the number of strands fixed (i. e. looking at one decoy at a time), and model its sheet configuration with its secondary structure given.

To help understanding the model, we give a brief general overview of the steps involved in building it. As already mentioned, we choose the fragments in ROSETTA according to the secondary structure prediction. The diagram below shows the format of such a secondary structure prediction. Given are the sequence of amino acids (AA), the prediction (here PHD) in which secondary structure element the residue is, the individual probabilities for helix (prH), strand (prE) and loop (prL), and the reliability (Rel) of that prediction. With this, we use ROSETTA to generate decoys. To score decoys, we generate a probability model for the sheet configurations, assuming the number of strands is known. Next, we need a model to decide how many sheets the strands build. Realizing that for example in proteins with 5 strands we only see a single five-stranded sheet or two sheets with two and three strands respectively, we loosely refer to this term in our model as the “Poker Hand”. Given the number of strands and sheets, the next step is to model the “sheet decomposition”, i. e. which strand belongs to which sheet. The third step in our model targets the actual

sheet configuration - knowing which strands form a sheet, in which motif do they align? The emphasis clearly is on the last step, since most small proteins only have a single sheet in which all strands get aligned.



Modeling the arrangement of strands into sheets, we assume the secondary structure to be known. Although the arrangement of strands into sheets may depend on many characteristics of the protein under consideration, we decided after an initial exploratory data analysis to only use two additional features of proteins in our model, which we can consider given together with the secondary structure.

(a) The helical status of the protein.

Using our experience from previous, unrelated work, we consider a protein to be helical if at least 20% of its residues are part of a helix.

(b) The lengths of the loops between strands.

What we call loop in this context is simply the sequence of amino acids that connects the strands under consideration. Therefore our loops can also contain residues that are part of a helix. We defined the loop between two strands as short if the number of residues was ten or less, and long otherwise. This decision is in agreement with the definition of sequence separation in Chapter 8.

We considered to use more known properties of proteins in our model, such as the length of the strands in the protein, the protein length (number of residues) itself, an indicator whether or not there is helical structure between two strands, etc. However we only included the two features described above, since they capture most of the information the other characteristics provide, and because the inclusion of more features was prohibited by the limited number of data available.

In our model, we use the following variables:

n_{S}	The number of strands in the protein.
n_{SH}	The number of sheets in the protein.
H	The helical status of the protein (either helical or non-helical).
L	The loop lengths between strands, given as indicators (either short or long).

SD The sheet decomposition, i. e. the assignment of strands into sheets.

SC The overall sheet configuration, i. e. the overall description of the arrangement of the strands.

SC^i The configuration (motif) of sheet i, i. e. the description of the arrangement of the strands in a particular sheet.

Knowing the number of strands in the protein, the lengths of the loops between the strands, and the helical status of the protein, we want to model the probability distribution of the possible sheet configurations, $P(\text{SC}|n_\text{S}, H, L)$. Using rules for conditional probabilities, we have

$$\begin{aligned} & P(\text{SC}|n_\text{S}, H, L) \\ = \; & P(\text{SD}, \text{SC}|n_\text{S}, H, L) \\ = \; & P(\text{SC}|\text{SD}, n_\text{S}, H, L) \times P(\text{SD}|n_\text{S}, H, L) \\ = \; & P(\text{SC}|\text{SD}, n_\text{S}, H, L) \times P(n_\text{SH}, \text{SD}|n_\text{S}, H, L) \\ = \; & P(\text{SC}|\text{SD}, n_\text{S}, H, L) \times P(\text{SD}|n_\text{SH}, n_\text{S}, H, L) \times P(n_\text{SH}|n_\text{S}, H, L) \end{aligned} \tag{9.1}$$

The first equation follows from the fact that the sheet configuration determines its sheet decomposition, and hence we have $P(\text{SD}, \text{SC}|n_\text{S}, H, L) = P(\text{SC}|n_\text{S}, H, L)$ for the specified sheet decomposition SD, and $P(\text{SD}^*, \text{SC}|n_\text{S}, H, L) = 0$ for all other sheet decompositions SD*. Next, we make the following assumption:

$$P(\text{SC}|\text{SD}, n_\text{S}, H, L) \;=\; \prod_i P(\text{SC}^i|\text{SD}, n_\text{S}, H, L) \tag{9.2}$$

This means, we assume that if we have two or more sheets in a protein, the motifs of those sheets are conditionally independent. This assumption might for example be violated in

proteins that pack sheets against each other. If two four-stranded sheets form a "sandwich", then often the two motifs are alike. However, the motifs in those sandwiches usually are motifs that are observed very frequently anyways, so that the correct topology will receive a high score, and most decoys with two different motifs in the sandwich will not. Most small proteins that we try to fold have no strands or only a single sheet, and in those cases we do not have to make use of the assumption in equation (9.2) anyways.

The model can now be written as

$$\begin{aligned} & P(\mathrm{SC}|n_\mathrm{S}, H, L) \\ = & \prod_i P(\mathrm{SC}^i|\mathrm{SD}, n_\mathrm{S}, H, L) \times P(\mathrm{SD}|n_\mathrm{SH}, n_\mathrm{S}, H, L) \times P(n_\mathrm{SH}|n_\mathrm{S}, H, L) \end{aligned} \quad (9.3)$$

We refer to $P(\mathrm{SD}|n_\mathrm{SH}, n_\mathrm{S}, H, L)$ as the "sheet decomposition" term, to $P(n_\mathrm{SH}|n_\mathrm{S}, H, L)$ as the "poker hand" term, and to $P(\mathrm{SC}^i|\mathrm{SD}, n_\mathrm{S}, H, L)$ as the "sheet configuration" term. The highest interest certainly is in the description of the model of the sheet configuration term, which is given in detail in a separate section (Section 9.2.2). Before that, we briefly describe the model for the sheet decomposition and the poker hand term in the following section. These following sections are rather technical, and tough we are making a lot of assumptions and simplifications, we can not always show the data in detail to illustrate how they support the decisions we make. However, for the convenience of the reader, we summarize the contents of the most technical parts at the end of the respective sections.

9.2.1 The Sheet Decomposition and the Poker Hand Term

The Sheet Decomposition

To model the sheet decomposition of proteins, we use the entire information in the database, although we are primarily concerned with small proteins, which most cases have no strands

at all, or only a single sheet. The distribution of the number of sheets given the number of strands is modeled in the poker hand term in the second part of this section. For the sheet decomposition, we can assume that the number of strands and sheets are known. The fact that most small proteins have either no strands or only a single sheet means that the counts used to model the decomposition term are not very high.

Investigating $P(\text{SD}|n_{\text{SH}}, n_{\text{S}}, H, L)$, we could not establish a dependency of SD on H and L, even though there are scenarios where one might expect this. We consequently simplified the decomposition term to

$$P(\text{SD}|n_{\text{SH}}, n_{\text{S}}, H, L) = P(\text{SD}|n_{\text{SH}}, n_{\text{S}}). \tag{9.4}$$

To model this term, we used the number of crossings as a surrogate. The number of crossings is defined as the number of times that, following the backbone from the N to the C-terminus of the protein, we leave a sheet and enter another sheet. Clearly, for the decomposition term we are only concerned about proteins with at least two sheets. We assume that, given the number of strands and sheets, all decompositions that yield the same number of crossings are equally likely. For example the decompositions $1-2-3-1-2-3$ and $1-2-3-2-1-3$ of 6-stranded proteins with three sheets both have 5 crossings, and are considered equally likely.

If we have n_{SH} sheets, we have at least $n_{\text{SH}}-1$ crossings. Very frequently we saw proteins in which the first sheet was completed before the second sheet got started. i. e. it is very common that folds achieve this minimum number of crossings. Some of these for example were proteins that had sandwiches, or proteins that had two separate pairs of strands. To predict the number of crossing given the number of strands and the number of sheets, we split this problem into two sub-problems. We first modeled the probability of a protein having the minimum number of crossings. For those which did not have the minimum number of crossings, we modeled the distribution of the number of crossings in excess of the minimum number.

The probability of having the minimum number of crossings: Since the outcome for this problem is binary (having the minimum number of crossings versus not having the minimum number of crossings), we used logistic regression to predict this outcome, using the number of strands and the number of sheets. We first used both the number of strands and the number of sheets as predictors, but then found that dichotomizing the number of sheets was advantageous, distinguishing proteins with two sheets from proteins with more than two sheets. The model we fit is

$$\log\left(\frac{p}{1-p}\right) = -3.372 + 0.653 \times n_{\mathrm{S}} - 1.285 \times I_{(n_{\mathrm{SH}} > 2)} \tag{9.5}$$

with I being the indicator function of the argument, here taking the value one if $n_{\mathrm{SH}} > 2$ and being zero otherwise. For illustration we show the fitted probabilities of having more than the minimum number of crossings for some combinations of sheet and strand numbers in Table 9.1.

Table 9.1: The fitted probabilities of having more than the minimum number of crossings.

		Number of sheets			
		2	3	4	$\cdots$
Number of strands	4	0.32			
	5	0.47			
	6	0.63	0.32		
	7	0.77	0.48		
	8	0.86	0.64	0.64	$\cdots$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

The distribution of the number of crossings in excess of the minimum number: If the number of crossings exceeds the minimum number $n_{\mathrm{SH}} - 1$, we need a model that assigns

a probability to the number of crossings by how much it exceeds $n_{\text{SH}} - 1$. Exceeding the minimum, we have at least n_{SH} crossings. Let $E_{\max}$ be the maximum of crossings by which we can exceed $n_{\text{SH}} - 1$. We define

$$Y := \text{number of crossings } - n_{\text{SH}}, \tag{9.6}$$

and hence $Y \in \{0, \ldots, E_{\max} - 1\}$. The maximum $E_{\max}$ by which the number of crossings can exceed $n_{\text{SH}} - 1$ also depends on the number of strands in addition to the number of sheets, which makes it somewhat tricky to use a binomial model for Y. However, we found that we can simplify matters and approximate the distribution of Y using a Poisson model. Since the Poisson distribution allows counts from 0 to ∞, we used

$$P(Y = k) \propto \exp(\lambda)\frac{\lambda^k}{k!} \quad \text{for} \quad k = 0, \ldots, E_{\max} - 1. \tag{9.7}$$

The actual probability $P(Y = k)$ can be derived by dividing the right hand side of equation (9.7) by its normalizing constant, i. e.

$$P(Y = k) = \frac{\exp(-\lambda)\frac{\lambda^k}{k!}}{1 - \sum_{j=E_{\max}}^{\infty} \exp(-\lambda)\frac{\lambda^j}{j!}}, \quad k = 0, \ldots, E_{\max} - 1. \tag{9.8}$$

The term $\sum_{j=E_{\max}}^{\infty} \exp(-\lambda)\frac{\lambda^j}{j!}$ of the normalizing constant is very close to zero in most cases. We estimated the parameter λ in the probability term by

$$\log(\lambda) = -1.185 + 0.195 \times n_{\text{S}} - 0.463 \times I_{(n_{\text{SH}} > 2)}, \tag{9.9}$$

which is equivalent to

$$\lambda = 0.306 \times 1.215^{n_{\text{S}}} \times 0.629^{I_{(n_{\text{SH}} > 2)}}. \tag{9.10}$$

We summarize the model for the number of crossings. Let

X be the number of crossings in excess of $n_{\text{SH}} - 1$,

Y be the number of crossings in excess of n_{SH} (i. e. $X = Y - 1$),

Z be an indicator if the number of crossings exceeds $n_{\text{SH}} - 1$,

and denote

logistic	to be the term in equation (9.5), and
poisson	to be the term in equation (9.8).

We then have

$$P(X = 0) = P(Z = 0) = \frac{\exp(\text{logistic})}{1 + \exp(\text{logistic})} \tag{9.11}$$

and for $j \in \{1, \ldots, E_{\max}\}$ we get

$$\begin{aligned} P(X = j) &= P(X = j \wedge Z = 1) \\ &= P(X = j | Z = 1) \times P(Z = 1) \\ &= P(Y = j - 1 | Z = 1) \times P(Z = 1) \\ &= \text{poisson} \times \frac{1}{1 + \exp(\text{logistic})} \end{aligned} \tag{9.12}$$

We have a probability model $P(\text{crossings(SD)}|n_{\text{SH}}, n_{\text{S}})$ for the number of crossings, given the number of strands and sheets. Since we assumed that, given the number of strands and sheets, all decompositions that yield the same number of crossings are equally likely (their number being $\#[\text{crossings(SD)}|n_{\text{SH}}, n_{\text{S}}]$ say), we have

$$P(\text{SD}|n_{\text{SH}}, n_{\text{S}}) = \frac{P(\text{crossings(SD)})}{\#[\text{crossings(SD}, n_{\text{SH}}, n_{\text{S}})]} \tag{9.13}$$

if $n_{\text{SH}} \geq 2$, and 1 otherwise.

The Poker Hand

As in the case of the sheet decomposition model, we could not establish an additional dependency of n_{SH} given n_{S} on H and L, and simplified the poker hand term to

$$P(n_{\text{SH}}|n_{\text{S}}, H, L) = P(n_{\text{SH}}|n_{\text{S}}) \tag{9.14}$$

In the paper [45] included in Chapter 8 we already introduced a poker hand term. However, this term was used in the scoring function of the ab initio protein folding, and therefore had to allow for single strands. In real proteins single strands do not appear, and our new poker hand term needs to reflect this.

Since every sheet has to have at least two strands, the maximum number of sheets is

$$n_{\mathrm{S}_{\max}} = \left\lceil \frac{n_{\mathrm{S}}}{2} \right\rceil . \tag{9.15}$$

Let X be the number of sheets in excess of the one sheet required, and define $n := n_{\mathrm{S}_{\max}} - 1$. Hence $X \in \{0, \ldots, n\}$. We modeled X as a binomial distribution, assuming

$$X \sim B(n, p(n_{\mathrm{S}})). \tag{9.16}$$

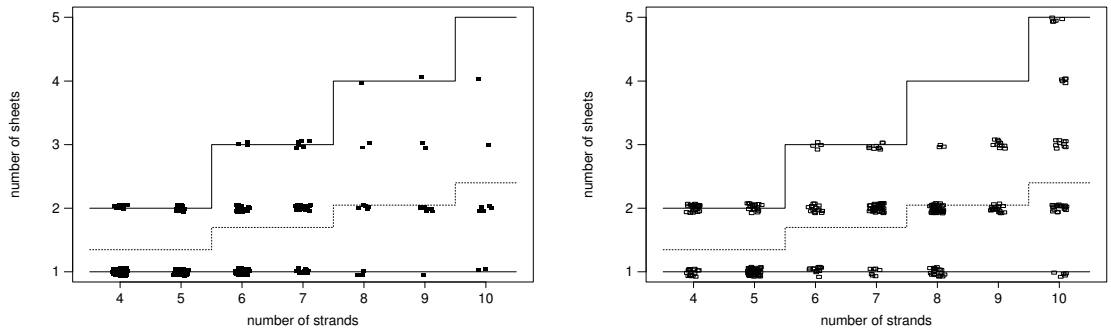
Further analyzing the data, we found that the probability in the binomial distribution does not depend on the number of strands, and estimated

$$p(n_{\mathrm{S}}) \equiv p = 0.35. \tag{9.17}$$

Figure 9.4 shows the number of strands versus the number of sheets found in small proteins with ten or fewer strands. The proteins helical and non-helical proteins are shown separately to illustrate that the distribution of the number of sheets given the number of strands is independent of the helical status, as assumed in equation (9.14).

9.2.2 The Sheet Configuration

Given the sheet decomposition, we know which strands form a sheet together. We also know the helical status of the protein and the lengths of all loops between the loops of a sheet, and want to model the distribution of the motifs the sheet can adopt. The strands in



(a) Number of strands versus number of sheets in helical proteins.

(b) Number of strands versus number of sheets in non-helical proteins.

Figure 9.4: The upper and lower limits for the number of sheets given the number of strands (solid lines), and the expected number of sheets (dotted line), using the fitted probability.

the sheet can be labeled by their number in sequence along the backbone, starting with the N-terminus of the protein.

A motif can then be described by the sequence of positions the strands take in the motif, and their directions. For a n-stranded sheet, the position information therefore is simply a permutation of the numbers 1 through n (**sequence**). Neighboring strands in the sheet are either parallel or anti-parallel, and we describe this feature (**orientation**) by a sequence of zeros and ones (up/down).

There are two axes of symmetry, as shown in Figure 9.5. The sequence for the strands in the motif of panel 1 is 2143 - the first strand is at position 2, the second strand is at position 1, etc. The orientation for the strands in the motif of panel 1 is 0110 - the first strand points up, the second strand points down, etc. Reversing the sequence in the motif (2143 becomes 3412) describes the first axis of symmetry, flipping the orientation (0110 becomes 1001) describes the second axis of symmetry. Since it does not matter from which angle we look at the protein and we can flop and spin the structure as we desire, these four motifs describe

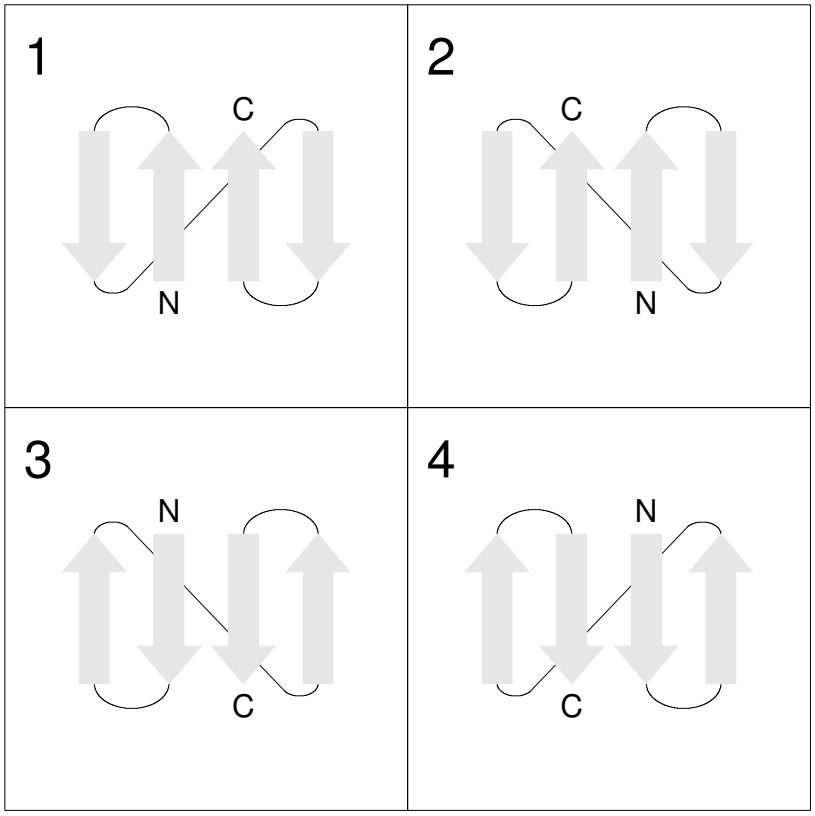


Figure 9.5: Four different motifs that represent the same sheet.

the same sheet, and we only need to consider one of the four possibilities. To uniquely characterize the sheet, we require two things:

1. The sequence starts left. For a sheet of n strands this means that the position number of the first strand is $\left\lceil \frac{n+1}{2} \right\rceil$ or less. If n is odd and the first strand has the middle position of the sheet, i. e. its position number is equal to $\left\lceil \frac{n+1}{2} \right\rceil$, the position number of the second strand in sequence has to be smaller than $\left\lceil \frac{n+1}{2} \right\rceil$.

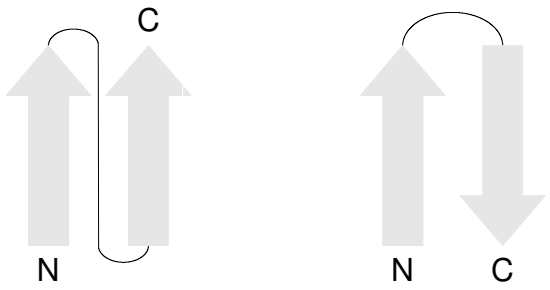
2. The first strand points up. In other words, the first number of the orientation is zero.

With those rules, we now always represent the 4-stranded sheet in Protein L by the motif in panel 1 of Figure 9.5.

There are $n!$ ways to position the strands in a sheet of size n, and 2^n possibilities for their orientations, if we ignore the axes of symmetry. Thus, taking the axes of symmetry into account, we have $\frac{1}{4} \times n! \times 2^n = n! \times 2^{n-2}$ possible n-stranded motifs. In the following three sections, we consider the probability distribution of 2-stranded, 3-stranded and 4-stranded motifs separately. Modeling those distributions without major assumptions and simplifications was feasible since there are only 2 motifs for 2-stranded sheets, 12 motifs for 3-stranded sheets, and 96 motifs for 4-stranded sheets. After these three sections we consider the probability distributions of the motifs for sheets with five or more strands.

Sheets with Two Strands

Fitting probabilities for two-stranded motifs is straightforward. There are only two ways for two strands to form a sheet: parallel (P) and anti-parallel (AP) - see Figure 9.6.



(a) A parallel pair of strands.

(b) An anti-parallel pair of strands.

Figure 9.6: The two possible configurations of two-stranded sheets.

Table 9.2 shows the counts of parallel and anti-parallel pairs of strands we found in the database, conditioning on the loop length between the two strands and the helical status

of the protein. It also displays the probabilities we use in the final model. In general, we use the term "bin" to refer to the class of structures that have a specific motif, loop length distribution, and helical status.

Table 9.2: The counts and fitted probabilities for parallel and anti-parallel pairs of strands.

	helical		non-helical	
	S	L	S	L
P	8	127	3	32
AP	609	338	278	207
P	0.01	0.27	0.01	0.13
AP	0.99	0.73	0.99	0.87

Sheets with Three Strands

There are twelve motifs for three-stranded sheets, shown and labeled in Figure 9.7. We classify the loop lengths between the three strands as short-short (L_1), short-long (L_2), long-short (L_3) and long-long (L_4). For most bins, the initially fitted probabilities were very similar, comparing helical and non-helical proteins. Using χ^2-tests for bins with sufficient counts, we determined which bins we could collapse across helical status. We removed single counts from bins and used pseudo-counts to re-fit the motif probabilities, which are shown in Table 9.3.

Since the most important features of those fitted probabilities are hard to grasp looking at the table alone, we highlighted those in Figure 9.8. Figure 9.8(a) indicates with black boxes the bins that were not collapsed across helical status. It is noteworthy that all of the bins were collapsed across helical status when both loops between the strands were short,

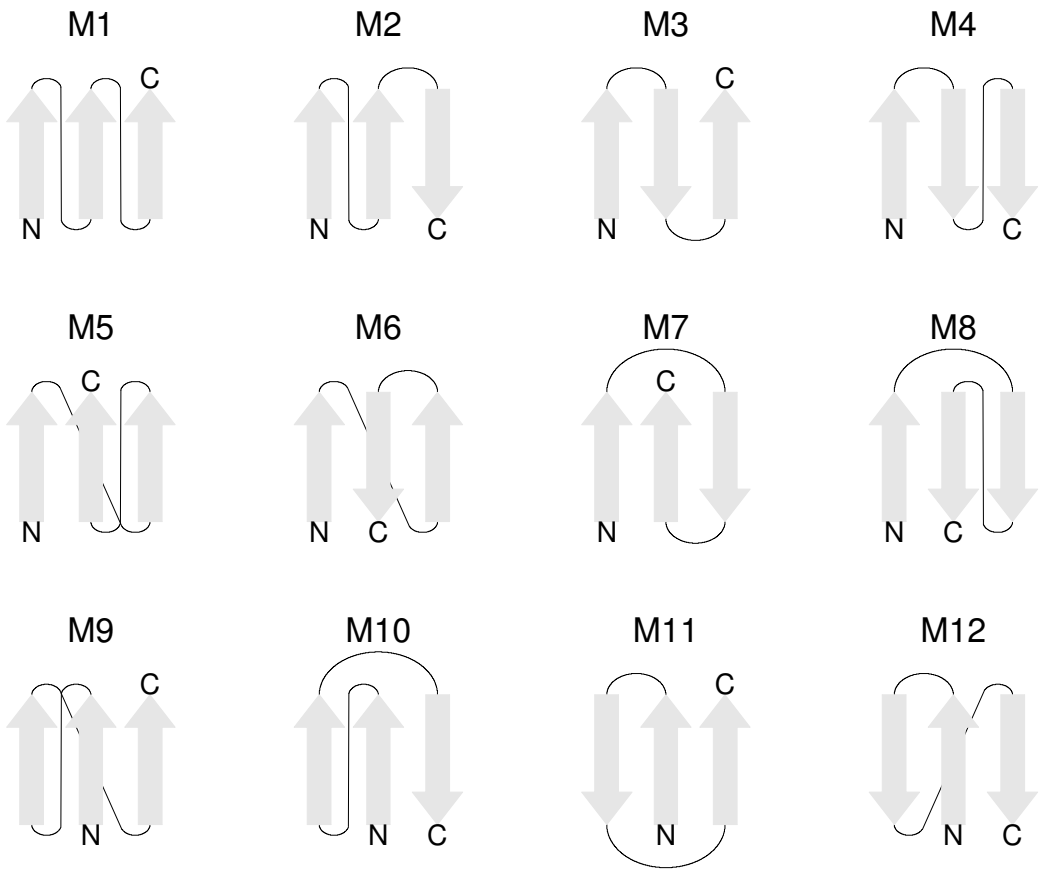
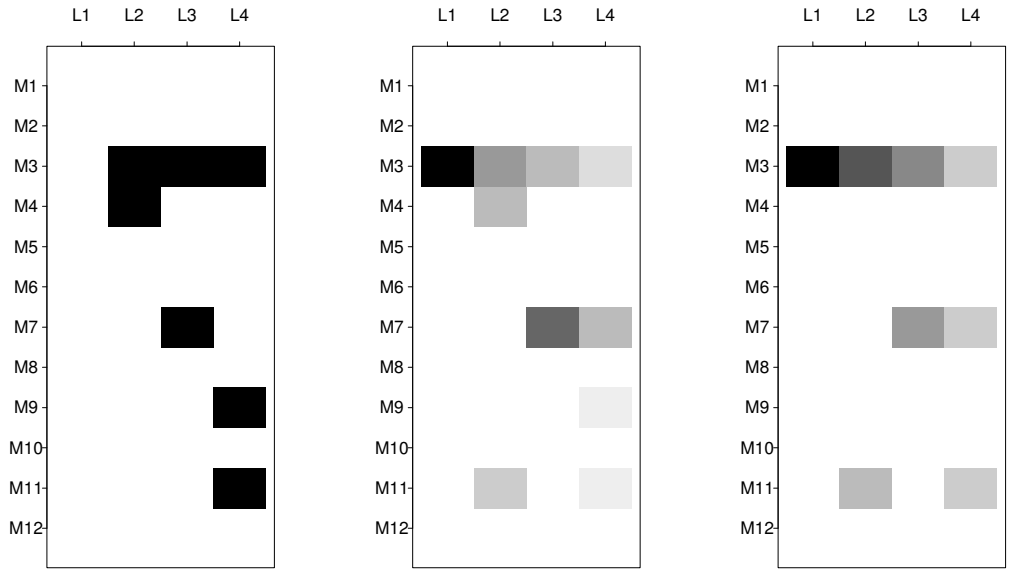


Figure 9.7: The twelve possible configurations of three stranded sheets.

and almost all motifs were up-down-up (M_3) in the case of both loops being short. Only seven bins in total were not collapsed, three of those for motif M_3, the "up-down-up" motif. Motif probabilities bigger than 10% within each length bin are highlighted for helical and non-helical proteins (Figure 9.8(b) and 9.8(c) respectively).

Table 9.3: The fitted probabilities for three-stranded motifs.

	helical				non-helical			
	L_1	L_2	L_3	L_4	L_1	L_2	L_3	L_4
M_1	0.0043	0.0056	0.0051	0.0491	0.0043	0.0058	0.0051	0.0416
M_2	0.0043	0.0056	0.0829	0.0803	0.0043	0.0058	0.0830	0.0681
M_3	0.8970	0.4014	0.2761	0.1621	0.8970	0.6107	0.4220	0.2517
M_4	0.0043	0.2622	0.0051	0.0285	0.0043	0.0423	0.0051	0.0242
M_5	0.0043	0.0056	0.0051	0.0190	0.0043	0.0058	0.0051	0.0161
M_6	0.0364	0.0115	0.5472	0.2822	0.0364	0.0118	0.4011	0.2394
M_7	0.0043	0.0056	0.0481	0.0315	0.0043	0.0058	0.0481	0.0267
M_8	0.0043	0.0056	0.0051	0.0142	0.0043	0.0058	0.0051	0.0121
M_9	0.0043	0.0056	0.0051	0.1144	0.0043	0.0058	0.0051	0.0121
M_{10}	0.0043	0.0056	0.0051	0.0348	0.0043	0.0058	0.0051	0.0295
M_{11}	0.0043	0.0270	0.0051	0.0315	0.0043	0.0279	0.0051	0.0267
M_{12}	0.0279	0.2587	0.0103	0.1525	0.0279	0.2668	0.0103	0.2517



(a) Bins that were not collapsed across helical status.

(b) Motif probabilities bigger than 10% for helical proteins.

(c) Motif probabilities bigger than 10% for non-helical proteins.

Figure 9.8: A visual display of the most important features of the fitted probabilities of three-stranded motifs.

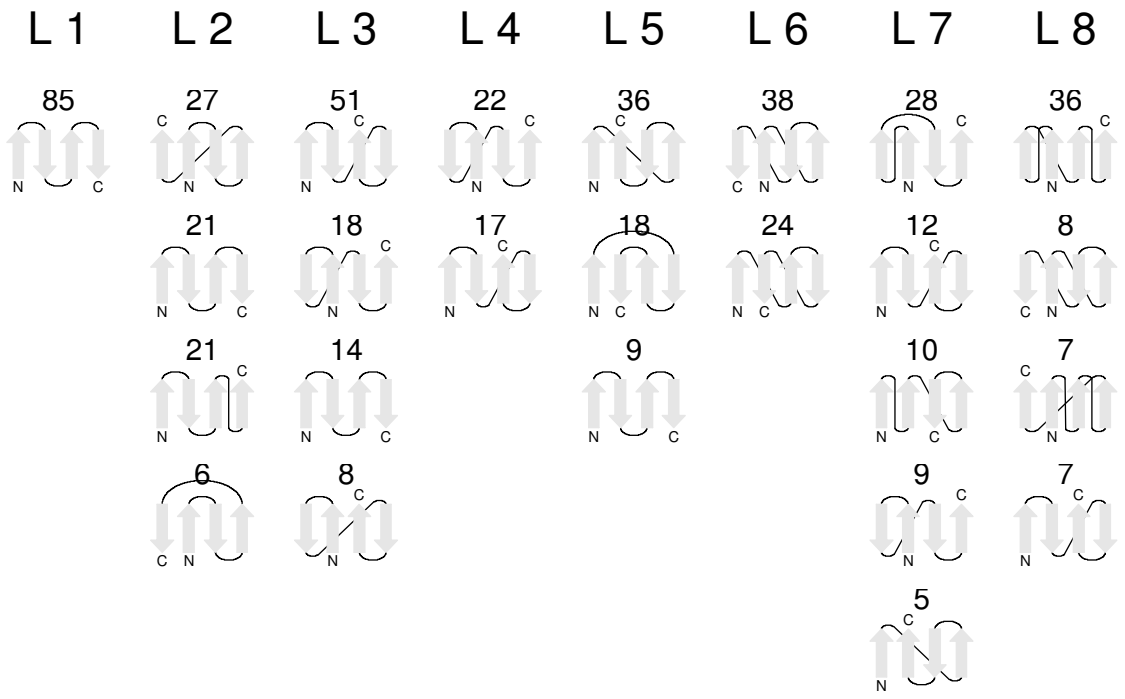
Sheets with Four Strands

Although there are 96 possible motifs for four-stranded sheets, we observed only 52 of those in the database. Among those, 18 motifs were observed only once. We saw 872 four-stranded sheets in the database, but less than 20 motifs were observed ten times or more.

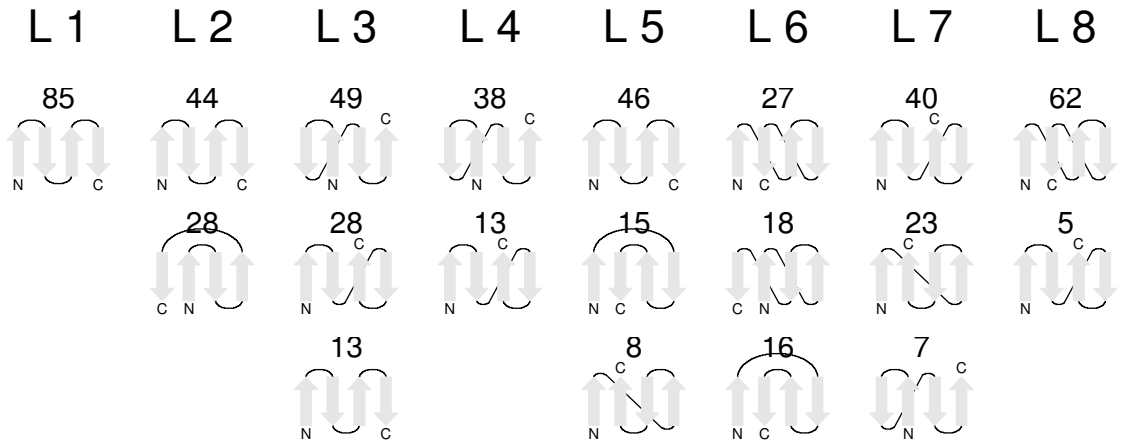
There are eight classes for the loop lengths between the four strands, labeled as follows:

L_1	short	short	short
L_2	short	short	long
L_3	short	long	short
L_4	short	long	long
L_5	long	short	short
L_6	long	short	long
L_7	long	long	short
L_8	long	long	long

As for the three-stranded motifs, we used χ^2-tests to determine which bins to collapse across helical status, and using pseudo-counts we fit the motif probabilities of the four-stranded sheets. Displaying all fitted probabilities would be rather confusing (96 motifs, 8 length classes, and 2 classes for the helical status equals 1536 bins!). Instead we show the motifs that have a probability of 5% or more in their respective bins in Figure 9.9.



(a) The most common four-stranded motifs for helical proteins.



(b) The most common four-stranded motifs for non-helical proteins.

Figure 9.9: Four-stranded motifs with probabilities bigger than 5%. The actual probabilities (rounded, in percent) are plotted above the motifs.

Sheets with more than Four Strands

At the beginning of this chapter we established that for a n-stranded sheet there are $n! \times 2^{n-2}$ possible motifs. In our probability model we also have to take into account the knowledge we have about the helical status of the protein (2 classes) and the loop lengths between the strands (2^{n-1} classes). To model the probability distribution of the motifs for a n-stranded sheet we therefore have to consider $n! \times 2^{n-2} \times 2 \times 2^{n-1} = n! \times 2^{2n-2}$ bins. What this means in actual numbers is shown in Table 9.4.

Table 9.4: The number of possible sequences (seq), orientations (or), motifs, length bins (L), helical bins (H) number of bins we condition on (cond), and overall number of bins for a sheet with n strands.

n	seq	or	motifs	L	H	cond	bins
2	1	2	2	2	2	4	8
3	3	4	12	4	2	8	96
4	12	8	96	8	2	16	1536
5	60	16	960	16	2	32	30720
6	360	32	11520	32	2	64	737280
7	2520	64	161280	64	2	128	20643840
8	20160	128	2580480	128	2	256	660602880
9	181440	256	46448640	256	2	512	23781703680
10	1814400	512	928972800	512	2	1024	951268147200

For up to four strands the counts from the database were sufficient to model the probability distribution of the motifs, using the raw counts for each motif. Considering how many motifs we have for sheets of size five or bigger, this is not feasible anymore, especially in light of the declining counts of larger sheets from the database, shown in Figure 9.10.

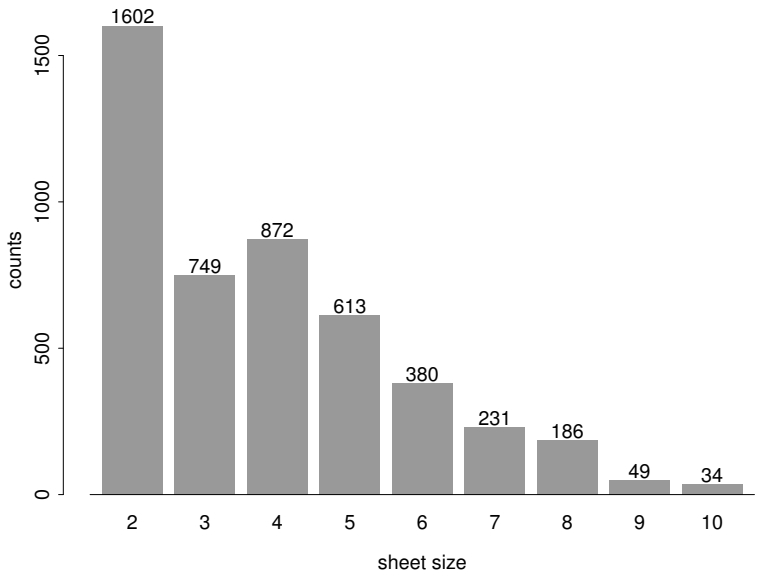


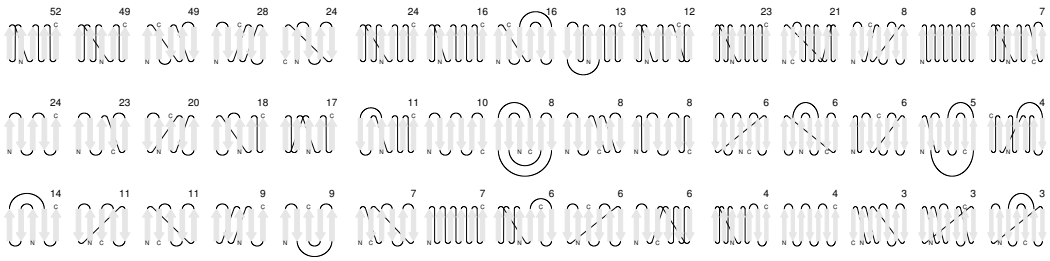
Figure 9.10: The counts of sheets of sizes $2-10$ observed in the database.

In the case of four-stranded sheets, there are 96 possible motifs, but only about twenty occur fairly frequently (ten times or more) in the database. The majority of these motifs looked very similar though, in the sense that in most motifs all neighboring strands were either parallel or anti-parallel. This observation is even more obvious in the most common motifs for sheets of size five or larger, shown in Figure 9.11.

To proceed with the model of the sheet configuration term, we make the assumption that the likelihood of in individual motif can be modeled by some global features, such as the number of parallel pairs and the positioning of the first strand in the motif. We use the following abbreviations:

P_p Number of parallel neighbor strands in a motif.

P_p^s Number of parallel neighbor strands in a motif with a short loop in between.



(a) The most common 5-stranded motifs.

(b) The most common 6-stranded motifs.

(c) The most common 7-stranded motifs.

Figure 9.11: Larger sheets frequently observed in the database. The sheets clearly have common "patterns".

J Number of strand pairs adjacent in sequence that are not neighbors in the sheet. We also refer to this feature as "jump".

J^s Number of jumps with a short loop between the strand pair.

F The position of the first strand in the motif.

Since we try to model the configuration of individual sheets, we can drop the sheet decomposition in the term $P(\text{SC}^i|\text{SD}, n_\text{S}, H, L)$, and for simplification we write $P(\text{SC}|n, H, L)$ when referring to the conditional probability of the configuration SC^i of n_S-stranded sheet i.

To not interrupt the flow of this chapter, we only state that after carrying out some exploratory data analysis, we found that the data support the assumption that the sheet configurations are characterized by their number of parallel pairs and how many of those have a short loop in between, their number jumps and how many of those have a short loop in between, plus the position of their first strand in the sheet. With this assumption we can

reduce the number of bins, assigning all sheet configurations with the same number of parallel pairs, jumps, etc in a single bin. The bins may contain different numbers of sheet configurations. For example, for an all-parallel 5-stranded motif of a protein with all loops longer than 10 residues starting at the first position (of either helical status) we have three possibilities if we allow one jump (Figure 9.12(a-c)), but only one possibility if we don't allow a jump (Figure 9.12(d)).

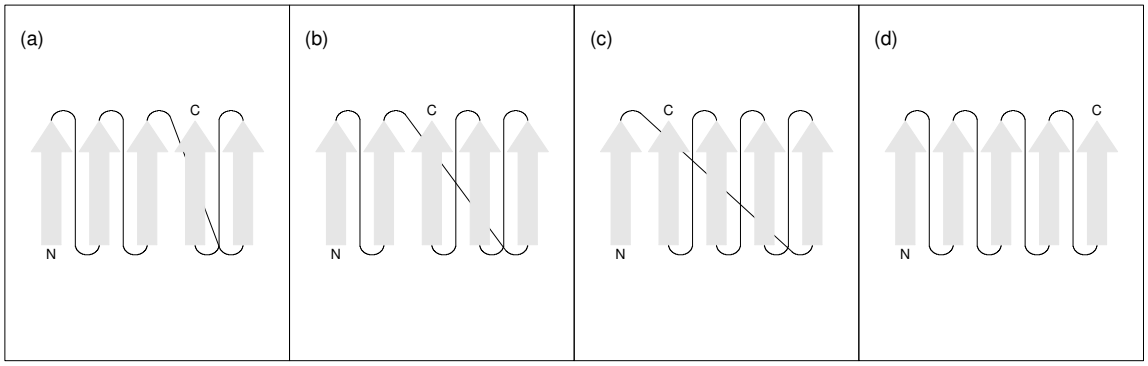


Figure 9.12: All parallel, five-stranded sheets starting at the first position.

Formally, we make the following assumption:

$$P(\mathrm{SC}|n,H,L) = \frac{P(P_p,P_p^s,J,J^s,F|n,H,L)}{k_{n,L}(P_p,P_p^s,J,J^s,F)}, \tag{9.18}$$

with $k_{n,L}(P_p,P_p^s,J,J^s,F)$ being the number of motifs with n strands and loop-lengths distribution L in the (P_p,P_p^s,J,J^s,F) bin, which is independent of the helical status of the protein under consideration. The term on the right-hand side in (9.18) looks even more complicated than the term on the left-hand side at first glance, but the description of the motifs by global features enables us to estimate the distribution of the sheet configurations in a meaningful way.

Using rules for conditional probabilities, we get

$$\begin{aligned} & P(P_p, P_p^s, J, J^s, F|n, H, L) \\ = \; & P(F|n, H, L) \times P(P_p, P_p^s, J, J^s|n, H, L, F) \\ = \; & P(F|n, H, L) \times \\ & P(P_p, J|n, H, L, F) \times P(P_p^s, J^s|n, H, L, F, P_p, J) \end{aligned} \tag{9.19}$$

In the following, we model each of the three above terms separately.

The term $\mathbf{P(F|n, H, L)}$ Analyzing the data, we found that

$$P(F|n, H, L) = P(F|n, H) \tag{9.20}$$

is a reasonable assumption. It seems possible that if for example all loops were short, the first strand of the sheet is slightly more likely to be in position 1 than in other positions, since having no long loops might prohibit jumps in certain configurations. However, we didn't have enough data to establish this, and hence accepted the assumption made in equation (9.20). Figure 9.13 shows histograms of F (scaled to probabilities) for sheets of size $5 - 10$ from helical and non-helical proteins.

For $n = 5$ and $n = 6$ we have plenty of data, and clearly there are differences between helical and non-helical proteins. Hence we estimated $P(F|n = 5, H)$ and $P(F|n = 6, H)$ directly from the data, for helical and non-helical proteins. For $n \geq 7$ the counts were really low, and we combined the counts of helical and non-helical proteins. For those, we see some differences in the distributions of the counts between sheets with even and odd numbers of strands. Below we describe briefly which assumptions were made to finish the modeling of the distribution. The modeling was done with the requirement in mind that the model should also be usable for sheets with more than ten strands.

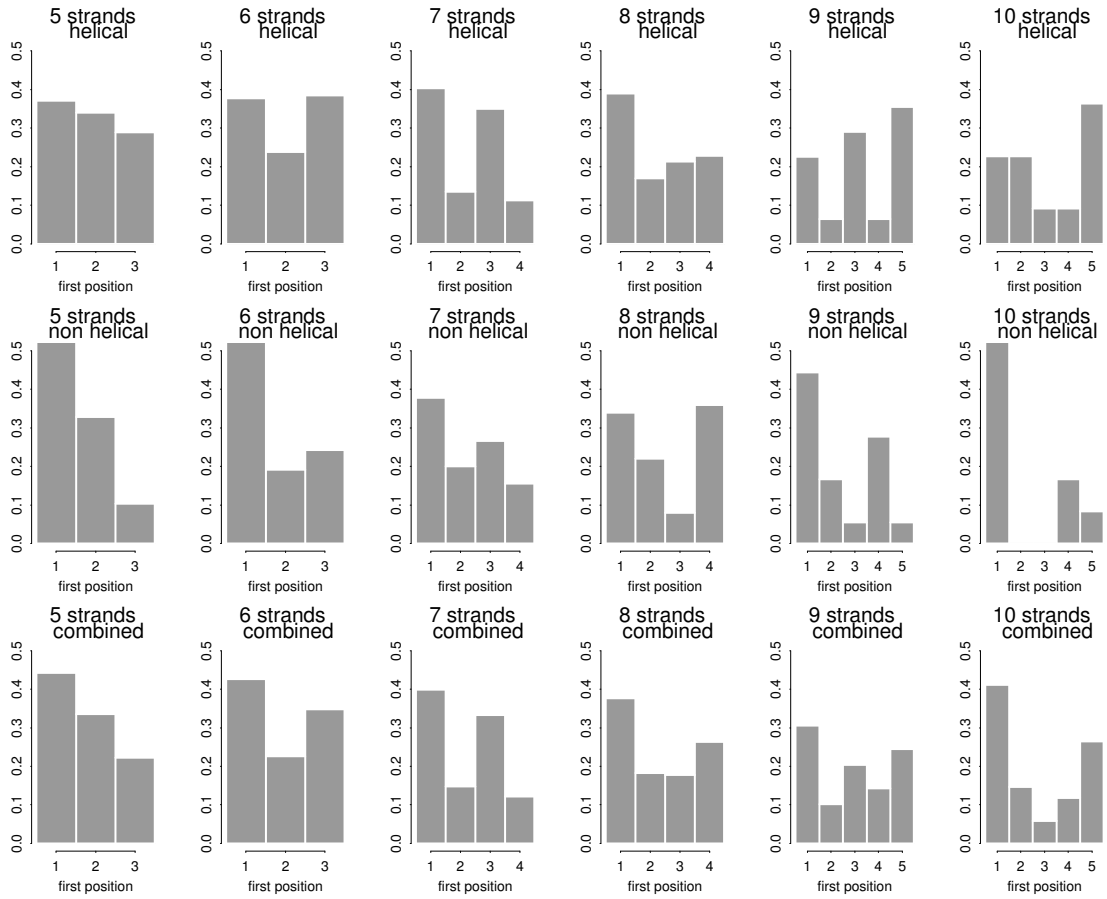


Figure 9.13: Probability distribution of the position of the first strand in the sheet, split by sheet size and helical status.

- Even number of strands ($n = 8, 10$):

  The pattern of the bars in Figure 9.13 for even n resembles a pattern such as

  $$ab \cdots bc \tag{9.21}$$

  Under this assumptions we estimated $\sum b = \frac{1}{3}$ and $a + c = \frac{2}{3}$ with $\frac{a}{c} = \frac{3}{2}$. Hence $a = \frac{6}{15}$ and $c = \frac{4}{15}$, regardless of the number of bs.

- Odd number of strands ($n = 7, 9$):

  The pattern of the bars in Figure 9.13 for odd n resembles a pattern such as

  $$abab \cdots a \tag{9.22}$$

  Under this assumptions we estimated $\sum a = \frac{3}{4}$ and $\sum b = \frac{1}{4}$.

We try to clarify the above "pattern analysis" and other calculations by showing the fitted probabilities of $P(F|n, H)$ for $H = 0, 1$ and $n = 5, \ldots, 10$ in Table 9.5.

Table 9.5: The fitted probabilities for the position of the first strand in the sheet, split by sheet size and helical status.

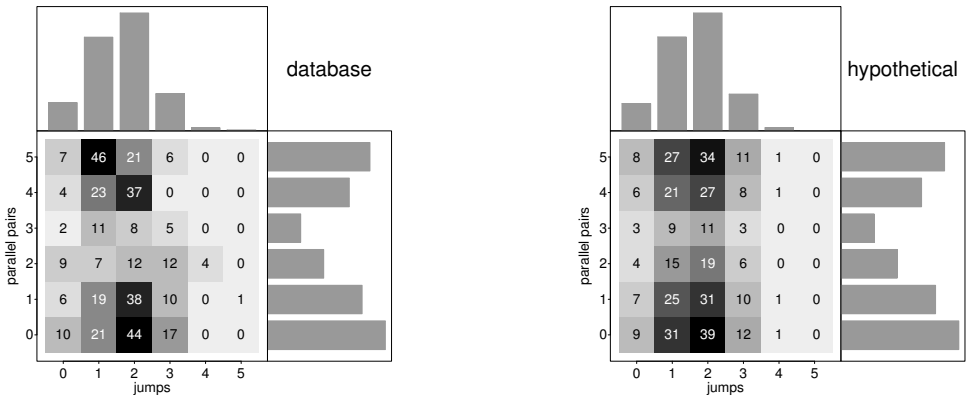
	helical					non-helical				
	1	2	3	4	5	1	2	3	4	5
5	0.370	0.340	0.290			0.570	0.330	0.100		
6	0.380	0.240	0.380			0.570	0.190	0.240		
7	0.375	0.125	0.375	0.125		0.375	0.125	0.375	0.125	
8	0.400	0.167	0.167	0.267		0.400	0.167	0.167	0.267	
9	0.250	0.125	0.250	0.125	0.250	0.250	0.125	0.250	0.125	0.250
10	0.400	0.111	0.111	0.111	0.267	0.400	0.111	0.111	0.111	0.267

The term $\mathbf{P(P_p, J|n, H, L, F)}$ For fixed n there are

n	possibilities for P_p
n	possibilities for J
2	possibilities for H
2^{n-1}	possibilities for L
$\left[\frac{n+1}{2}\right]$	possibilities for F

Hence there are roughly $2^{n-1} \times n^3$ bins in the term $P(P_p, J|n, H, L, F)$, and some simplification is needed. Below, we describe which assumptions we made and which bins we collapsed after examining the data.

We first noted that we cannot assume conditional independence of P_p and J, which was formally tested using χ^2-tests. For illustration, we show the counts of P_p and J from six-stranded sheets in Figure 9.14 without taking H and L into account. These counts are shown in panel (a), together with their marginal distributions. If P_p and J were independent, then given the margins from panel (a), we would expect the counts to be roughly distributed as shown in panel (b), which is clearly not the case.



(a) Counts of jumps and parallel pairs derived from the database.

(b) Counts derived from the margins under the assumption of independence.

Figure 9.14: Counts of jumps and parallel pairs in six-stranded sheets in the database, plus their marginal distributions. Preferred are motifs with all or no parallel neighbors, and one or two jumps. Conditioning on the margins, hypothetical counts were calculated under the assumption of independence.

Having established this, we investigated which assumptions are reasonable that allow us to collapse bins. The data support the following decisions:

- For F, we only discriminate between starting at the first position versus starting at any other permissible position. F now takes on only two values: $F = 1$ when starting in the first position, and $F = 2$ otherwise.

- Certain motifs only happen when all loops between strands are more than ten residues long. We categorized $L = 1$ if all loops between strands are long, and $L = 0$ otherwise.

- The above stated fact was clearly observed in helical proteins. For non-helical proteins we did not have enough data to establish a loop length dependency of P_p and J at all. We omitted L from the probability term for non-helical proteins.

Using the above described binning, we have sufficient data to model the term $P(P_p, J|n, H, L, F)$ for 5 and 6-stranded sheets $(n = 5, 6)$ with $P \in \{0, \ldots, n-1\}$ and $J \in \{0, \ldots, n-1\}$. For sheets of 7 or more strands this is unfortunately not the case, and since the data prohibits more collapsing of variables we condition on, we cannot model every single level of P_p and J. However, it appears that for large sheets ($n \geq 7$), the patterns in the term $P(P_p, J|n, H, L, F)$ look somewhat similar (data not shown), although the number of levels are different for different sheet sizes ($n \times n$ for a sheet of n strands). For each $n \geq 7$ we binned the $n \times n$ table of counts observed for P_p and J, given n, H, L and F, into a 7×4 table. Given the number of strands n there are n possibilities for the number of parallel pairs, since $P \in \{0, \ldots, n-1\}$. The data suggest that we consider that we leave the counts for $0, 1, 2$ and $n-1, n-2, n-3$ and create a "middle bin", collapsing all counts between 3 and $n-4$ parallel pairs (see Table 9.6). The possible number of jumps is also n, as $J \in \{0, \ldots, n-1\}$. Since we condition on the first strand being at position 1 or not, we know that the number of jumps cannot be zero if the first strand is not in position 1. Hence we leave the bin for zero jumps as is. We split the remaining $n-1$ levels ($J \in \{1, \ldots, n-1\}$) into three bins (small medium and large number of jumps). The number of levels going into the small (medium) [large] bin are 2 (2) [2] for $n = 7$, they are

Table 9.6: The binning used to collapse the $n \times n$ table of counts of jumps and parallel pairs into a 7×4 table.

	P_p							J			
	1	2	3	4	5	6	7	1	2	3	4
7	0	1	2	3	4	5	6	0	1-2	3-4	5-6
8	0	1	2	3-4	5	6	7	0	1-3	4-5	6-7
9	0	1	2	3-5	6	7	8	0	1-3	4-6	7-8
10	0	1	2	3-6	7	8	9	0	1-3	4-6	7-9

3 (2) [2] for $n = 8$, 3 (3) [2] for $n = 9$, and 3 (3) [3] for $n = 10$ (data not shown). This binning allowed us to estimate $P(P_p, J|n, H, L, F)$ for $n \geq 7$.

The term $\mathbf{P(P^s_p, J^s|n, H, L, F, P_p, J)}$: Checking the data, we decided that it is reasonable to assume the following conditional independence:

$$\begin{aligned} & P(P^s_p, J^s|n, H, L, F, P_p, J) \\ = \;& P(P^s_p|n, H, L, F, P_p, J) \times P(J^s|n, H, L, F, P_p, J) \end{aligned} \tag{9.23}$$

The number of parallel pairs of strands in the sheet that are connected with a short loop of not more than ten residues depend on n, H, L, P, but given n, L, P, the number does not depend on F and J. Hence

$$P(P^s_p|n, H, L, F, P_p, J) = P(P^s_p|n, H, L, P), \tag{9.24}$$

and analogous

$$P(J^s|n, H, L, F, P_p, J) = P(J^s|n, H, L, J). \tag{9.25}$$

Let pp be the number of parallel pairs in the sheet and n_s the number of short loops. Since we have $n-1$ pairs of strands in the sheet, the lowest possible number of parallel pairs of strands in the sheet that are connected with a short loop (say n_{l_s}) is

$$l = \max(pp + n_s - (n-1)\ ,\ 0) \tag{9.26}$$

The maximum of parallel pairs of strands in the sheet that are connected with a short loop is

$$u = \min(pp, n_s) \tag{9.27}$$

Since

$$n_{l_s} \in \{l, \ldots, u\}, \tag{9.28}$$

we are interested in modeling the number of parallel pairs connected with a short loop in excess of l (say X). We have

$$X \in \{0, \ldots, u-l\}. \tag{9.29}$$

The data support the following model:

$$\begin{aligned} P(P_p^s = k+l|n, H, L, P) &= P(P_p^s = k+l|n, H, n_s, pp) \\ &= P(X = k) \end{aligned} \tag{9.30}$$

with

$$X \sim B(u-l, p_{\text{parpair}}(n, H)) \tag{9.31}$$

Analyzing the data, we concluded that for both helical and non-helical proteins the probability of the binomial term is not significantly different for different sheet sizes, and estimated

$$p_{\text{parpair}}(n, H) = p_{\text{parpair}}(H) = \begin{cases} 0.51 & \text{if } H = 0 \\ 0.24 & \text{if } H = 1 \end{cases} \tag{9.32}$$

For the number of jumps on a short loop we used exactly the same reasoning to derive our model. Let j be the number of jumps in the sheet and n_s the number of short loops. Let Y be the number of short jumps in excess of the lowest possible number of jumps on a short loop. We have

$$\begin{aligned} P(J^s = k + l | n, H, L, J) &= P(P_p^s = k + l | n, H, n_s, j) \\ &= P(Y = k) \end{aligned} \tag{9.33}$$

with

$$Y \sim B(u - l, p_{\text{jump}}(n, H)) \tag{9.34}$$

and

$$l = \max(j + n_s - (n - 1)\ ,\ 0) \tag{9.35}$$

$$u = \min(j, n_s) \tag{9.36}$$

However, in this case we found that the the binomial term does depend on the sheet size, and derived

$$p_{\text{jump}}(n, H) = \begin{cases} 0.25 & \text{if } H = 0 \text{ and } n = 5 \text{ or } 6 \\ 0.54 & \text{if } H = 1 \text{ and } n = 5 \text{ or } 6 \\ 0.18 & \text{if } H = 0 \text{ and } n > 7 \\ 0.28 & \text{if } H = 1 \text{ and } n > 7 \end{cases} \tag{9.37}$$

Summarizing the terms and the model we developed for sheets with more than four strands,

we have

$$\begin{aligned} & P(P_p, P_p^s, J, J^s, F|n, H, L) \\ = \; & P(F|n, H, L) \times P(P_p, P_p^s, J, J^s|n, H, L, F) \\ = \; & P(F|n, H, L) \times P(P_p, J|n, H, L, F) \times \\ & P(P_p^s, J^s|n, H, L, F, P_p, J) \\ = \; & P(F|H, n) \times P(P_p, J|n, H, L, F) \times \\ & P(P_p^s|n, H, L, P_p) \times P(J^s|n, H, L, J) \end{aligned} \tag{9.38}$$

with

$$P(\text{SC}|\text{SD}, H, L) = P(\text{SC}|n, H, L) = \frac{P(P_p, P_p^s, J, J^s, F|n, H, L)}{k_{n,L}(P_p, P_p^s, J, J^s, F)}, \tag{9.39}$$

Example: A fairly common motif for a five-stranded sheet in small proteins, such as the SH3 domain of spectrin (1aey), is shown in Figure 9.15.

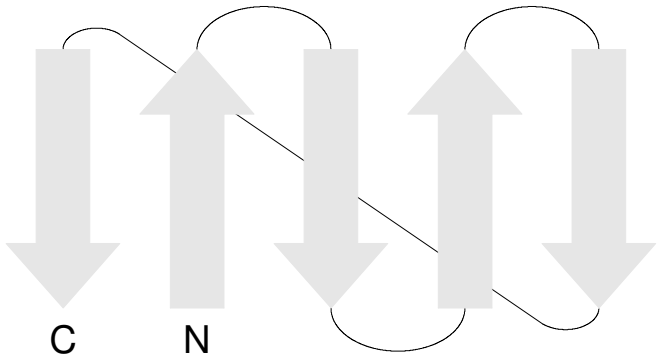


Figure 9.15: The sheet motif of the SH3 domain of spectrin.

The SH3 domain of spectrin is a non-helical protein with 62 residues. The loop lengths

in sequence are 18, 5, 2 and 3 residues. The strands along the sequence form anti-parallel neighbors in the sheet, except for strand 5, which jumps and forms an anti-parallel strand pair with the first strand. The sheet in the protein is quite warped, which allows the jump between strands 4 and 5 with a loop of only 3 residues.

We now score this motif, using the above explained model. In our notation, we have $n = 5, H = 0, L = (1, 0, 0, 0), F = 2, P_p = 0, J = 1, P_p^s = 0, J^s = 1$. Since this configuration is the only motif in the (P_p, P_p^s, J, J^s, F) bin, we have

$$\begin{array}{lcl} P(F|n,H) & = & 0.330 \\ P(P_p,J|n,H,L,F) & = & 0.396 \\ P(P_p^s|n,H,L,P_p) & = & 1.000 \\ P(J^s|n,H,L,J) & = & 0.250 \\ \hline & \prod & 0.033 \end{array}$$

The score of 3.3% doesn't seem to be very high, but given that there are 960 possible motifs, this is more than 31 times higher than "background frequency"!

9.2.3 Outlook

The probabilities we fit in this chapter reflect what we saw in the database. To assess the usefulness of our model for sheet configurations, we need to determine how much it improves the protein structure prediction, i. e. which impact it has on the "quality" of a set of decoys we generate. For example, if we use the model to select decoys from a larger set, we can compare the percentage of near-native proteins in the entire decoy set versus the percentage of near-native proteins in a subset of decoys selected by making use of this scoring function. We currently use the model in CASP4, and plan to thoroughly assess its usefulness after the structures of the targets have been published.

BIBLIOGRAPHY

[1] BAHAR, I., AND JERNIGAN, R. L. Coordination geometry of nonbonded residues in globular proteins. *Folding & Design 28* (1996), 357–370.

[2] BERNSTEIN, F. C., KOETZLE, T. F., WILLIAMS, G. J., MEYER, E. E., BRICE, M. D., RODGERS, J. R., KENNARD, O., SHIMANOUCHI, T., AND TASUMI, M. The protein data bank: a computer-based archival file for macromolecular structures. *J Mol Biol 112* (1977), 535–542.

[3] BONNEAU, R., AND BAKER, D. Ab initio protein structure prediction: progress and prospects. *manuscript in preparation*.

[4] BOWIE, J. U. Helix packing angle preferences. *Nature Struct. Biol. 4* (1997), 915–917.

[5] BOWIE, J. U., LUTHY, R., AND EISENBERG, D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science 253* (1991), 164–170.

[6] BREIMAN, L., FRIEDMAN, J., OLSHEN, R., AND STONE, C. *Classification and Regression Trees*. Chapman & Hall, 1984.

[7] BRYAN, R. N., MANOLIO, T. A., SCHERTZ, L. D., JUNGREIS, C., POIRIER, V. C., ELSTER, A. D., AND KRONMAL, R. A. A method for using mr to evaluate the effects of cardiovascular disease on the brain: the cardiovascular health study. *AJNR 15*, 9 (1994), 1625–1633.

[8] BYSTROFF, C., AND BAKER, D. Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol 281* (1998), 565–577.

[9] CERNÝ, V. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications 45*, 1 (1985), 41–51.

[10] CHIPMAN, H., GEORGE, E., AND MCCULLOUCH, E. Bayesian cart model search. *Journal of the American Statistical Association 93*, 443 (1998), 935–960.

[11] DANDEKAR, T., AND ARGOS, P. Identifying the tertiary fold of small proteins with different topologies from sequence and secondary structure using the genetic algorithm and extended criteria specific for strand regions. *J Mol Biol 256* (1996), 645–660.

[12] EFIMOV, A. V. A novel super-secondary structure of proteins and the relation between structure and the amino acid sequence. *FEBS 166* (1984), 33–38.

[13] FLEISHER, H., TAVEL, M., AND YEAGER, J. Exclusive-or representations of boolean functions. *IBM J. Res. Develop. 27*, 4 (1983), 412–416.

[14] FLEISS, J. *Statistical methods for rates and proportions*, 2 ed. Wiley, 1981.

[15] FRIED, L. P., BORHANI, N. O., ENRIGHT, P., FURBERG, C. D., GARDIN, J. M., KRONMAL, R. A., KULLER, L. H., MANOLIO, T. A., MITTELMARK, M. B., NEWMAN, A. B., O'LEARY, D. H., PSATY, B., RAUTAHARJU, P., TRACY, R. P., AND WEILER, P. G. The cardiovascular health study: Design and rationale. *Annals of Epidemiology 1*, 3 (1991), 263–276.

[16] GENEST, C., AND ZIDEK, J. Combining probability distributions: A critique and an annotated bibliography. *Statist. Science 1* (1986), 114–148.

[17] GILPIN, E., OLSHEN, R., HENNING, H., AND ROSS, J. Risk prediction after myocardial infarction. *Cardiology 70* (1983), 73–84.

[18] HARRIS, N. L., PRESNELL, S. R., AND COHEN, F. E. Four helix bundle diversity in globular proteins. *J Mol Biol 236* (1994), 1356–1368.

[19] HENNING, H., GILKPIN, E., COVELL, J., SWAN, E., O'ROURKE, R., AND ROSS, J. Prognosis after acute myocardial infarction: A multivariate analysis of mortality and survival. *Circulation 59*, 6 (1979), 1124–1136.

[20] HOBOHM, U., AND SANDER, C. Enlarged representative set of protein structures. *Protein Sci. 3* (1994), 522–524.

[21] HOLM, L., AND SANDER, C. Database algorithm for generating protein backbone and side-chain co-ordinates from a ca trace. *J Mol Biol 218* (1991), 183–194.

[22] HONG, S. J. R-mini: An iterative approach for generating minimal rules from examples. *IEEE Transactions on Knowledge and Data Engineering 9*, 5 (1997), 709–717.

[23] HUANG, E. S., SUBBIAH, S., TSAI, J., AND LEVITT, M. Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations. *J Mol Biol 257* (1996), 716–725.

[24] KIRKPATRICK, S., GELATT, C. D., AND VECCHI, M. P. Optimization by simulated annealing. Tech. rep., IBM Research Report RC 9355, 1982.

[25] KOCHER, J. A., ROOMAN, M. J., AND WODAK, S. J. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J Mol Biol 235* (1994), 1598–1613.

[26] KOLINSKI, A., AND SKOLNICK, J. Monte carlo simulations of protein folding. i. lattice model and interaction scheme. *Proteins 18* (1994), 338–352.

[27] KOOPERBERG, C., BOSE, S., AND STONE, C. Polychotomous regression. *Journal of the American Statistical Association 92*, 437 (1997), 117–127.

[28] LUCEK, P. R., AND OTT, J. Neural network analysis of complex traits. *Genetic Epidemiology 14* (1997), 1101–1106.

[29] MCCLELLAND, R. *Regression Based Variable Clustering for Data Reduction*. PhD thesis, University of Washington, Seattle, WA 98195, June 2000.

[30] MCCULLAGH, P., AND NELDER, J. A. *Generalized Linear Models*, 2 ed. Chapman & Hall, 1989.

[31] MICHALSKI, R., MOZETIC, I., HONG, J., AND LAVRAC, N. The multi-purpose incremental learning system aq15 and its testing application to three medical domains. In *Proc. AAAI* (1986), pp. 1041–1045.

[32] MITCHELL, J. B. O., LASKOWSKI, R. A., AND THORNTON, J. M. Non-randomness in side-chain packing: the distribution of interplanar angles. *Proteins 29* (1996), 370–380.

[33] MIYAZAWA, S., AND JERNIGAN, R. L. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol 256* (1996), 623–644.

[34] MONGE, A., LATHROP, E. J., GUNN, J. R., SHENKIN, P. S., AND FRIESNER, R. A. Computer modeling of protein folding: conformational and energetic analysis of reduced and detailed protein models. *J Mol Biol 247* (1995), 995–1012.

[35] OTTEN, R. H., AND GINNEKEN, L. P. *The Annealing Algoritm*. Kluwer Academic Publishers, 1989.

[36] PAGALLO, G. Learning dnf by decision trees. In *Proc. 11th IJCAI* (1989), pp. 639–644.

[37] PARK, B. H., HUANG, E. S., AND LEVITT, M. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J Mol Biol 266* (1997), 831–846.

[38] PARK, B. H., AND LEVITT, M. Energy functions that discriminate x-ray and near-native folds from well-constructed decoys. *J Mol Biol 258* (1996), 367–392.

[39] REDDY, B. V. B., AND BLUNDELL, T. L. Packing of secondary structural elements in proteins. *J Mol Biol 233* (1993), 464–479.

[40] RICHARDSON, J. S. The anatomy and taxonomy of protein structure. *Adv Protein Chem 34* (1981), 167–339.

[41] SAMUDRALA, R., AND MOULT, J. An all-atom distance dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol 275* (1998), 895–916.

[42] SHORTLE, D., SIMONS, K. T., AND BAKER, D. Clustering of low-energy conformations near the native structures of small proteins. *Proc. Natl. Acad. Sci. 95*, 19 (1998), 11158–11162.

[43] SIMONS, K. T., BONNEAU, R., RUCZINSKI, I., AND BAKER, D. Ab initio protein structure prediction of casp iii targets using rosetta. *Proteins 37*, 3 (1999), 171–176.

[44] SIMONS, K. T., KOOPERBERG, C., HUANG, E., AND BAKER, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J Mol Biol 268* (1997), 209–225.

[45] SIMONS, K. T., RUCZINSKI, I., KOOPERBERG, C., FOX, B. A., BYSTROFF, C., AND BAKER, D. Improved recognintion of native-like protein structures using a combination of sequence dependent and sequence independent features of proteins. *Proteins 34* (1999), 82–95.

[46] SIPPL, M. J. Calculation of conformational ensembles from potentials of mean force. an approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol 213* (1990), 859–883.

[47] SRINIVASAN, R., AND ROSE, G. D. Linus: a hierarchic procedure to predict the fold of a protein. *Proteins 22* (1995), 81–99.

[48] SWAYNE, D. F., COOK, D., AND BUJA, A. Xgobi: Interactive dynamic graphics in the x window system with a link to s. In *Proceedings of the 1991 American Statistical Association Meetings* (ASA, Alexandria, VA, 1992).

[49] TENG, E. L., AND CHUI, H. C. The modified mini-mental state (3ms) examination. *Journal of Clinical Psychiatry 48*, 8 (1987), 314–318.

[50] VAN LAARHOVEN, P. J., AND AARTS, E. H. *Simulated Annealing: Theory and Applications*. Kluwer Academic Publishers, 1987.

[51] WALTHER, D., EISENHABER, F., AND ARGOS, P. Principles of helix-helix packing in proteins: the helical lattice superposition model. *J Mol Biol 255* (1996), 536–553.

[52] WEISS, S. M., AND INDURKHYA, N. Optimized rule induction. *IEEE Expert* (December 1993), 61–69.

[53] WENTWORTH, P. *Boolean Logic and Circuits*. Department of Computer Science, Rhodes University, http://diablo.cs.ru.ac.za/func/bool/.

[54] ZHANG, J., AND MICHALSKI, R. S. Rule optimization via sg-trunc method. *EWSL* (1989), 251–262.