Secondary structure classification of amino-acid sequences using state-space modeling

Marcus Brunnert, Tillmann Krahnke and Wolfgang Urfer
Department of Statistics, University of Dortmund

Abstract

The secondary structure classification of amino acid sequences can be carried out by a statistical analysis of sequence and structure data using state-space models. Aiming at this classification, a modified filter algorithm programmed in S is applied to data of three proteins. The application leads to correct classifications of two proteins even when using relatively simple estimation methods for the parameters of the state-space models. Furthermore, it has been shown that the assumed initial distribution strongly influences the classification results referring to two proteins.

Keywords: Secondary structure classification, discrete state-space models, filtering.

1. Introduction

In molecular biology the analysis of proteins according to their structure and function is essential to understand any organism. Proteins are built as long linear chains of several chemical components. Among these are the 20 amino acids. The structure of a protein can be described by the principal concept of the *primary structure* (amino acid sequence), the *secondary structure* (a-helix or β-sheet structures) and the *tertiary structure* (folded secondary structure). In this context, the classification of proteins into different secondary and tertiary structures enables the molecular biologist to draw conclusions about the function of the protein.

Because of the fast expanding availability of sequence and structure data, the empirical methods required to classify protein structures get more and more important. The two approaches commonly used for the empirical classification are the *hidden Markov models* and *discrete state-space-models* (Bienkowska *et al.*, 2000, White *et al.*, 1994). These statistical approaches are also important tools in many other molecular biological problems like the protein fold recognition problem (Bienkowska et al., 2000) or DNA sequence analysis (Urfer, 2001).

In the project we describe here, the discrete state space model was applied to a secondary structure classification. The statistical model is described in **Section 2**, followed by the description of the likelihood computation used as a classification criterion in **Section 3**. The following **Section 4** contains the stochastic modeling of three proteins that belong to the Ubiquitin-like folded family (Nassar *et al.*, 1995). Empirical results of an application to these proteins are presented in **Section 5**. The software package S-plus (MathSoft, 2000) was used to implement the modified filter algorithm. Finally, **Section 6** discusses this statistical method and the results of the application.

2. State-space model

Considering the primary sequence of amino acid residues, the correspondence between the amino acids on the primary sequence and the amino acids on the secondary structure can be thought of as a Markov chain. As a result of this, the stochastic output of an observable amino acid y_t at the residue position $t = 1,2,3,\ldots$ depends on the unobservable current state x_t at this residue position. The state describes the underlying secondary structure at this residue position. Additionally, the current state depends on the realized past state x_{t-1}. Assuming these stochastic characteristics hold true, a discrete state-space model can be defined as follows:

$$\boldsymbol{y_t} = \boldsymbol{H}\,\boldsymbol{x_t} \tag{1}$$

$$\boldsymbol{x_{t+1}} = \Phi\,\boldsymbol{x_t}\,. \tag{2}$$

In contrast to a Gaussian state-space model, the output variables y_t and the state variables x_t, $t = 1,2,3,\ldots$, of this categorical state-space model are only discrete random variables taking values in finite sets. Moreover, the output probability matrix $\boldsymbol{H}$ and the state transition matrix Φ are only arrays of conditional probability distributions. Each column of the matrices is conditioned on a specific state. Let denote a state space by $S = \{1, 2, \ldots, n\}$, where $x_t \in S$ and an output set $A = \{1, 2, \ldots, m\}$, where $y_t \in A$. The elements of the $(m \times n)$-output probability matrix $\boldsymbol{H}$ are

$$H(k, j) = P(y_t = k | x_t = j), \qquad k = 1, \ldots, m, \quad j = 1, \ldots, n \tag{3}$$

and the elements of the $(n \times n)$-state transition matrix are

$$\Phi(i, j) = P(x_{t+1} = i | x_t = j), \; i, j = 1, \ldots, n \,. \tag{4}$$

The distribution of one state x_t to a residue position t is denoted by

$$\boldsymbol{x_t} = \begin{pmatrix} P(x_t = 1) \\ P(x_t = 2) \\ \vdots \\ P(x_t = n) \end{pmatrix} \tag{5}$$

and the output probability distribution at residue position t is denoted by

$$\boldsymbol{y_t} = \begin{pmatrix} P(y_t = 1) \\ P(y_t = 2) \\ \vdots \\ P(y_t = m) \end{pmatrix}. \tag{6}$$

The sequence $(x_t)_{t=1,2,3,\ldots}$ can be interpreted as a Markov chain of order 1. The sequence $(y_t)_{t=1,2,3,\ldots}$ can be interpreted as a semi-Markov process, where the transition from y_t to y_{t+1} depends only on x_t and not on y_t. If $(x_t)_{t \in S}$ is not observable, the stochastic processes $((x_t)_{t=1,2,3,\ldots}, (y_t)_{t=1,2,3,\ldots})$ can be interpreted as a hidden Markov model (Rabiner, 1989 and Durbin *et al.*, 1998).

Summing up, a state-space model for a specific protein structure with an initial state distribution $\boldsymbol{x_1}$ is completely described by five parameters: $\mathscr{M} = (m, n, \Phi, \boldsymbol{H}, \boldsymbol{x_1})$.

3. Statistical analysis of protein data with a state-space model

State-space models can be used for different statistical analyses aiming at the filtering of an optimal primary sequence, classification of a protein structure, and estimation of a secondary structure for a given optimal sequence. The following procedures can be found in the literature:

(i.) Equations (1) and (2) enable the filtering of an optimal observation sequence generated by the underlying model $\mathscr{M}$. An algorithm is derived in White (1988) in order to filter an optimal primary sequence.

(ii.) By constructing q state-space models for different protein structures, it is possible to select the most probable model for a new primary sequence. A recursive algorithm is derived in White *et al.* (1994) for the calculation of the likelihood of a specific model $\mathscr{M}_l,\ l = 1,\ldots,q,$ generating the new primary sequence. Assuming an appropriate prior probability for each model, the posterior probability for a model given a primary sequence can be calculated on the basis of the likelihood. With these posterior probabilities for all q models the most probable protein structure can be selected.

(iii.) A third statistical analysis which uses a smoothing algorithm (Stultz *et al.*, 1993) or the Viterbi algorithm (Durbin *et al.*, 1998, Rabiner, 1989) can determine the most probable state sequence according to a new primary sequence. This state sequence can be interpreted as an optimal secondary structure of the new primary sequence.

3.1 Filtering algorithm

Filtering algorithms are important statistical methods used in analysing state space models.
The Kalman filter (Kalman, 1960) is applied to continuous data in Schmitz and Urfer (1997).
For the secondary structure classification problem a categorical filter has to be applied.
Aiming at the filtering of an optimal primary sequence, the model parameters have to be known or they have to be estimated according to data on the protein structure. A re-estimation method for these parameters using the EM –algorithm (Dempster *et al.*, 1977) is described in Sousa *et al.* (2001).
The following filtering algorithm is adapted from White *et al.* (1994):

Input : Model $\mathscr{M} = (m, n, \Phi, \boldsymbol{H}, \boldsymbol{x_1})$ and observed primary sequence $Y_d = (y_1, y_2, \ldots, y_d)$.

Initialization : $\boldsymbol{x}_1^- = \boldsymbol{x_1}$. (7)

Recursion for $1 \le t \le d$: $\boldsymbol{y}_t^- = \boldsymbol{H}\,\boldsymbol{x}_t^-$ (8)

State update

$$\boldsymbol{v}_t = H[y_t = k]^T * \boldsymbol{x}_t^-$$

$$l = \sum_{j=1}^{n} v_t(j) \tag{9}$$

$$\boldsymbol{x}_t^+ = \frac{\boldsymbol{v}_t}{l} .$$

State Propagate $\boldsymbol{x}_{t+1}^- = \Phi \boldsymbol{x}_t^+$. (10)

Termination $t = d$: The conditional output probability distribution at each residue position $t = 1, \ldots, d$ according to the given sequence $Y_d = (y_1, y_2, \ldots, y_d)$.

The notation * denotes the Hadamard (elementwise) product of two matrices. In addition, $H[y_t = k]$ is the vector of the *k*th row of the $(m \times n)$-output probability matrix $\boldsymbol{H}$, while $\boldsymbol{y}_t^-$ is the vector of conditional distribution of the current observation $y_t = k,\ k \in A,$ at the sequence position t, given all past observations $Y_{t-1} = (y_1, \ldots, y_{t-1})$, i.e.,

$$\boldsymbol{y}_t^- = \begin{pmatrix} P(y_t = 1 | Y_{t-1}) \\ P(y_t = 2 | Y_{t-1}) \\ \vdots \\ P(y_t = m | Y_{t-1}) \end{pmatrix}. \tag{11}$$

For the current state x_t, the vector $\boldsymbol{x}_t^-$ is analogously defined as

$$\boldsymbol{x}_t^- = \begin{pmatrix} P(x_t = 1 | Y_{t-1}) \\ P(x_t = 2 | Y_{t-1}) \\ \vdots \\ P(x_t = n | Y_{t-1}) \end{pmatrix}. \tag{12}$$

$\boldsymbol{x}_t^+$ denotes the vector of the conditional distribution of the current state x_t, given all past and present observations $Y_t = (y_1, \ldots, y_t)$:

$$\boldsymbol{x}_t^+ = \begin{pmatrix} P(x_t = 1 | Y_t) \\ P(x_t = 2 | Y_t) \\ \vdots \\ P(x_t = n | Y_t) \end{pmatrix}. \tag{13}$$

The output of the filter is the filtering of the probability distribution of an optimal sequence.

3.2 Likelihood filtering algorithm

The classification of a protein structure can be connected with the search for the most probable model for a primary sequence. Given a primary sequence Y_d and a model $\mathscr{M}$, the computation of the probability that the sequence Y_d is generated by the model $\mathscr{M}$ enables the search for the most probable model. For that reason, the computation of the likelihood $L(Y_d|\mathscr{M}_l)$ due to several models $\mathscr{M}_1,\ldots,\mathscr{M}_q$ enables the classification of the primary sequence to q protein structures. Here, the likelihood $L(Y_d|\mathscr{M}_l)$ is as follows:

$$L(Y_d|\mathscr{M}_l) = P(Y_d|\mathscr{M}_l) = P(y_1)\prod_{t=2}^{d} P(y_t|Y_{t-1}). \tag{14}$$

Considering the output of the filtering algorithm in **Section 3.1**, the log-likelihood can be calculated recursively as follows:

$$\begin{aligned} &\log L(0) = 0 \ \ and \\ &\log L(t) = \log L(t-1) + \log P(y_t|y_{t-1}), \quad \text{for } t = 1,\ldots,d. \end{aligned} \tag{15}$$

In order to select the most probable model from q models, the posterior probability for model l can be calculated by assuming prior probabilities p_l and the q model likelihoods calculated from (15):

$$p(l|Y_d) = \frac{L(Y_d|\mathscr{M}_l)\,p_l}{c}, \tag{16}$$

where $c = \sum_{l=1}^{q} L(Y_d|\mathscr{M}_l)\,p_l$.

With the posterior probability at hand, it is possible to classify a protein structure optimally with respect to the structural hypotheses $\{\mathscr{M}_l\}_{l=1}^{q}$.

3.3 Optimal secondary structure computation by smoothing

Using the results of the filtering, it is possible to compute an optimal probability prediction of the secondary structure given a primary sequence. From the filtering algorithm, we can calculate $P\overbrace{\left(x_t = j \middle| y_{t-1},\ldots,y_1\right)}^{from\ x_t^-}$, $j=1,\ldots,n$, $t=1,2,3,\ldots$. If we apply the filtering algorithm to the reverse sequence $y_d,\ldots,y_1$, we can calculate $P\left(x_t \middle| y_{t+1},\ldots,y_d\right)$. For each residue position t, we can calculate the smoothing distribution for a state $j=1,\ldots,n$ (Stultz *et al.*, 1993):

$$P\left(x_t = j \middle| y_1,\ldots,y_d\right) = C_k \frac{P\left(x_t = j \middle| y_{t-1},\ldots,y_1\right) H\left(y_t \middle| x_t = j\right) P\left(x_t = j \middle| y_{t+1},\ldots,y_d\right)}{f_t(j)}, \qquad (17)$$

where $f_{t+1}(j) = \sum_{j'} \Phi\left(x_{t+1} = j \middle| x_t = j'\right) f_t\left(x_t = j\right),\ f_1(j) = P\left(x_1 = j\right),\quad t=1,2,3,\ldots$

and C_k is chosen, so that $\sum_{j=1}^{n} P\left(x_t = j \middle| y_1,\ldots,y_d\right) = 1$.

This way of calculating the smoothing distribution is an alternative to the sequence-to-structure-alignment described in Brunnert *et al.* (2002).

4. Modeling the secondary structure

The protein domain is built in an hierarchical fashion using the three basic modeling structures: α-helix, β-strand and an ”other” or random structure that is neither an α-helix nor a β-strand. These basic structures are put together to larger secondary structure called plexes. Two basic structure are linked at ”junctions”. These junctions are silent states with no output. For simplicity, only one plex is used here: the β-sheet containing two β-strands and one additional state. This additional state contains only one residue and therefore no basic modeling is needed.

In **Figures 1-7,** the states are represented by circles and the junctions by squares. The transition probabilities are written above the arrows.

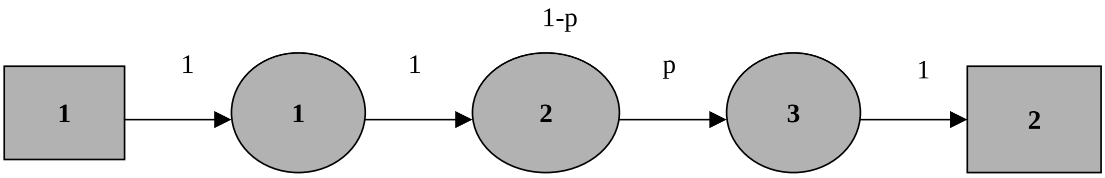
4.1 Basic modeling structures

The following sections describe the stochastic modeling of the three basic structures: α-*helix*, β-*strand* and *random structure.*

4.1.1 α-helix

The first basic structure described here is the α-helix.

Figure 1. Schematic modeling of the α-helix.



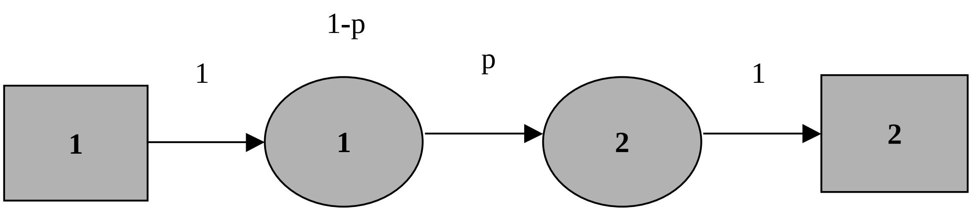
As we consider a minimal length of an α-helix of three residues, the transition matrices $\Phi_{ss} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1-p & 0 \\ 0 & p & 0 \end{pmatrix}$, $\Phi_{sj} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$, $\Phi_{js} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$ and $\Phi_{jj} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ are given, where $\Phi_{ss}, \Phi_{sj}, \Phi_{js}$ and Φ_{jj} denote the state-to-state transition matrix, the junction-to-state transition matrix, the state-to-junction matrix and the junction-to-junction matrix.

4.1.2 β-strand

Considering **Figure 2**, we get the following transition matrices:

$$\Phi_{ss} = \begin{pmatrix} 1-p & 0 \\ p & 0 \end{pmatrix},\ \Phi_{sj} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},\ \Phi_{js} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},\ \Phi_{jj} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

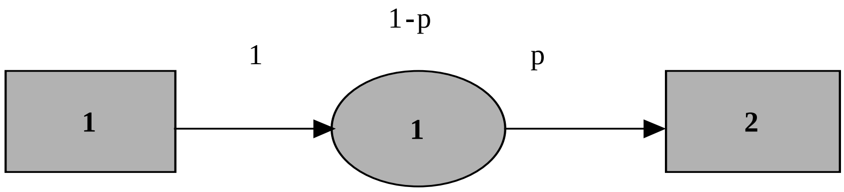
Figure 2. Schematic modeling of the β-strand.



4.1.3 ”Other” (random) structure

A stochastic modeling of a random structure is shown in **Figure 3**.

Figure 3. Schematic modeling of the random structure.



The transition matrices for a random structure with a minimal length of 1 is given by

$$\Phi_{jj} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},\ \Phi_{sj} = \begin{pmatrix} 1 & 0 \end{pmatrix},\ \Phi_{js} = \begin{pmatrix} 0 \\ p \end{pmatrix},\ \Phi_{ss} = 1 - p\ .$$

4.2 Plexes

The next hierarchical level in the state modeling describes combinations of basic structures.

4.2.1 β-sheet

The plex described here, links two β-*strands* at an additional state. Contrary to a junction state, the additional state is not silent and yields an output.

Figure 4. Schematic modeling of the β-sheet.

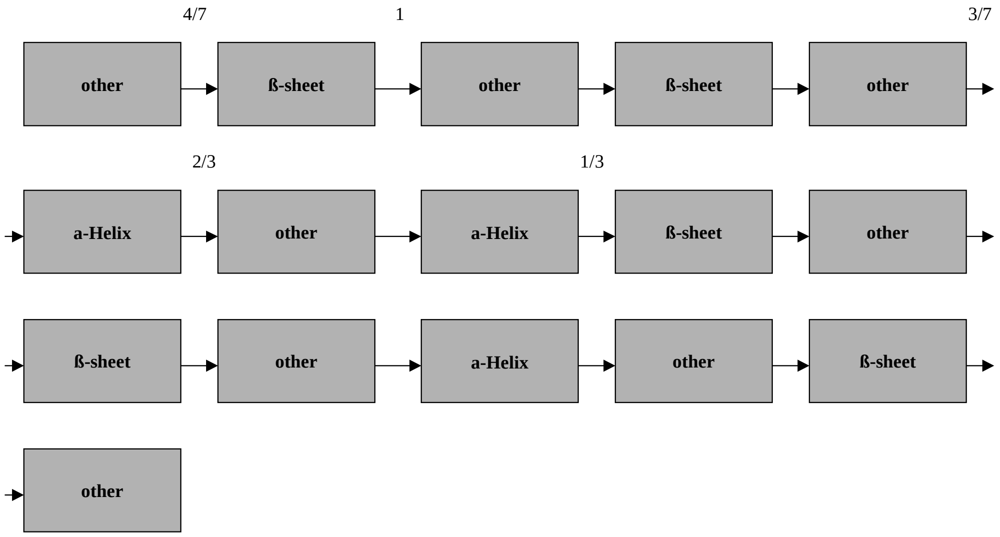
Connecting the two basic structures β-*strand* by an additional state (cf. **Figure 4**) with one output yields the following transition matrices:

$$\Phi_{ss} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix},\ \Phi_{sj} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix},\ \Phi_{js} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix},\ \Phi_{\mathrm{jj}} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

4.3 Modeling the domain

Now, the basic structures and plexes have to be put together to models for protein domains according to the application in **Section 5**. This will be done in an example in **Section 5**. The corresponding estimated transition probabilities are presented here for each state transition according to the observed secondary structures from the PDB (Protein Data Bank).

Figure 5. Modeling the observed Raf protein domain.



Considering the transition probabilities of the domain in **Figure 5**, one obtains the transition matrix $\vec{\Phi}_{Raf} = \begin{pmatrix} 0 & 0 & \frac{3}{7} \\ \frac{1}{3} & 0 & \frac{4}{7} \\ \frac{2}{3} & 1 & 0 \end{pmatrix}$, with only three states (α-helix, β-sheet and other).

Figure 6. Modeling the observed Ubiquitin protein.

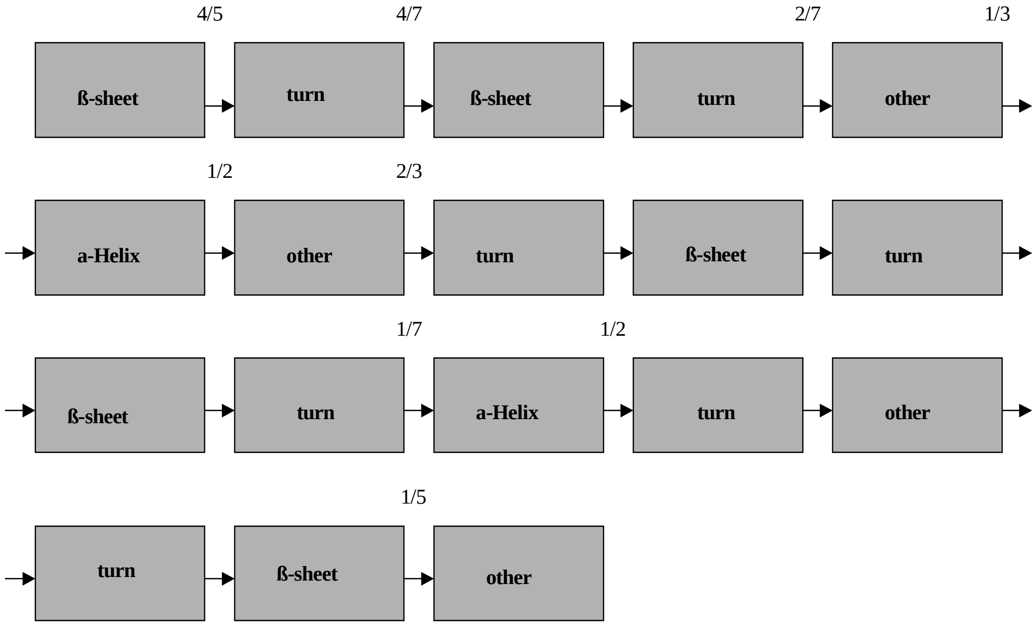
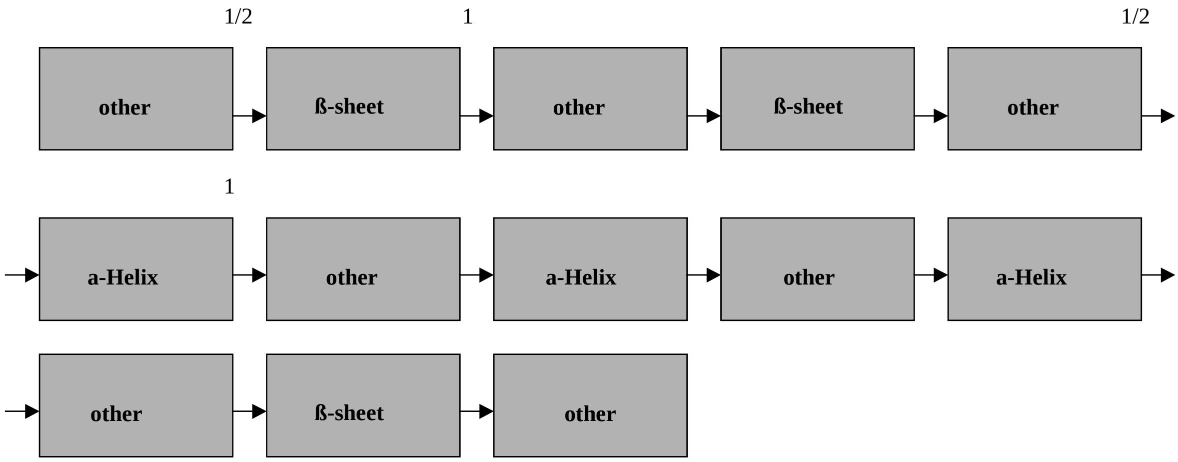


Figure 7. Modeling the observed RalGDS protein domain.



Note, that the state "turn" is additionally observed in the Ubiquitin protein domain.

According to the **Figures 6 and 7**, we yield $\hat{\Phi}_{Ral} = \begin{pmatrix} 0 & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} \\ 1 & 1 & 0 \end{pmatrix}$ and $\hat{\Phi}_{Ubiquitin} = \begin{pmatrix} 0 & 0 & \frac{1}{7} & \frac{1}{3} \\ 0 & 0 & \frac{4}{7} & 0 \\ \frac{1}{2} & \frac{4}{5} & 0 & \frac{2}{3} \\ \frac{1}{2} & \frac{1}{5} & \frac{2}{7} & 0 \end{pmatrix}$.

5. Application

In an application, a modified filter algorithm is applied to the primary and secondary structure data of domains of the Ubiquitin-like folded proteins RalGDS, Raf and Ubiquitin. The state update in the filter algorithm is modified according to the admissible states in the modeled secondary structure (see **Figures 1-4**). For example, an inadmissible state transition in the $\alpha - helix$ structure (**see Figure 1**) is the transition from state 1 to state 3. Formally the state update (9) of **Section 3.1**

$$\boldsymbol{v}_t = H[y_t = k]^T * \boldsymbol{x}_t^-$$

$$l = \sum_{j=1}^{n} v_t(j)$$

$$\boldsymbol{x}_t^+ = \frac{\boldsymbol{v}_t}{l}$$

is modified as follows:

$$\boldsymbol{v}_t = H[y_t = k]^T * \boldsymbol{x}_t^-$$

$$\boldsymbol{v}_t = \boldsymbol{x}_t^- * \boldsymbol{e}_j,$$

$$l = \sum_{j=1}^{n} v_t(j)$$

$$\boldsymbol{x}_t^+ = \frac{\boldsymbol{v}_t}{l} \qquad \text{, where}$$

$\boldsymbol{e}_j$ is the $n \times 1 - unit\ vector$ and j is the state number according to the highest probability in $\boldsymbol{x}_t^-$. This modified filter algorithm has been implemented in S-plus.

In order to estimate the transition matrices of the three models, the structure information described in **Section 4** has been used. For the estimation of the probabilities p in the transition matrices, a geometric distribution for the length of the basic modeling structures has been assumed. Moreover, the uniform distribution for the initial states has been used.

Three methods for the estimation of the output probability matrices according to each of the three models have been applied. These are:

1. Computation of relative frequencies of the occurrence of amino acids given a state.
2. Consideration of pseudocounts using Laplace's rule (Durbin *et al.*, 1998).
3. Consideration of pseudocounts using the following estimator (Durbin *et al.*, 1998):

$$\hat{H}(k,j) = \frac{c_{k,j} + 20\, p_{k,j}}{\sum_{k'} c_{k',j} + 20}$$, where

$c_{k,j}$ is the number of occurences of amino acid k given state j. The output probability $p_{k,j}$ referring to amino acid k given state j is estimated by using the relative frequency of the occurrence of amino acid k given state j according to the data of all three domains (see **Appendix**).

The resulting state space models for the three protein domains can be described as follows:

Raf: $\mathscr{M}_1\left(20, 9, \hat{\Phi}_{Raf}, \hat{H}_{Raf}, \hat{x}_{1,Raf}\right)$,

RalGDS: $\mathscr{M}_2\left(20, 9, \hat{\Phi}_{Ral}, \hat{H}_{Ral}, \hat{x}_{1,Ral}\right)$ and

Ubiquitin: $\mathscr{M}_3\left(20, 10, \hat{\Phi}_{Ubiquitin}, \hat{H}_{Ubiquitin}, \hat{x}_{1,Ubiquitin}\right)$.

5.1 Results

According to the estimation methods (1-3) of the H matrices, the following three figures summarise the results of this application. Each figure demonstrates the course of differences of negative log likelihoods calculated from the filtering of the observed amino acid sequence, starting with the first observed residue. The minuend of the difference is always the negative log likelihood that is calculated from the filtering of the sequence due to its reference model. Therefore a positive difference indicates a less probable structure than the reference structure.

If we consider the positive values in the three figures belonging to the false structure models, all three sequences except for the third estimation method referring to the Ubiquitin sequence (**Figure 10**) are classified correctly with its modeled secondary structure model. After filtering the whole sequence, the estimation methods (1-3) according to the emission probabilities do not influence the resulting lowest logarithmic likelihood execept for the Ubiquitin sequence. It is remarkable that no sequence is classed with its modeled secondary structure very early except for the Raf sequence referring to the first estimation method **(Figure 8)**. It might be possible that the assumed uniform distribution in some cases yields disadvantaged initial states for some models. For that reason, the influence of the initial distribution $\boldsymbol{x_1}$ has been analysed for all sequences. A graphical analysis analogous to the **Figures 8-10** for all three models with a fixed initial state for all three models has been carried out (**Figures** not shown). In this graphical analysis the classification of the Ral sequence and the Ubiquitin sequence has been influenced by the choice of the initial state. The Raf sequence classification was not influenced by the choice of an initial state.

Figure 8: Differences of negative log likelihoods according to method 1.

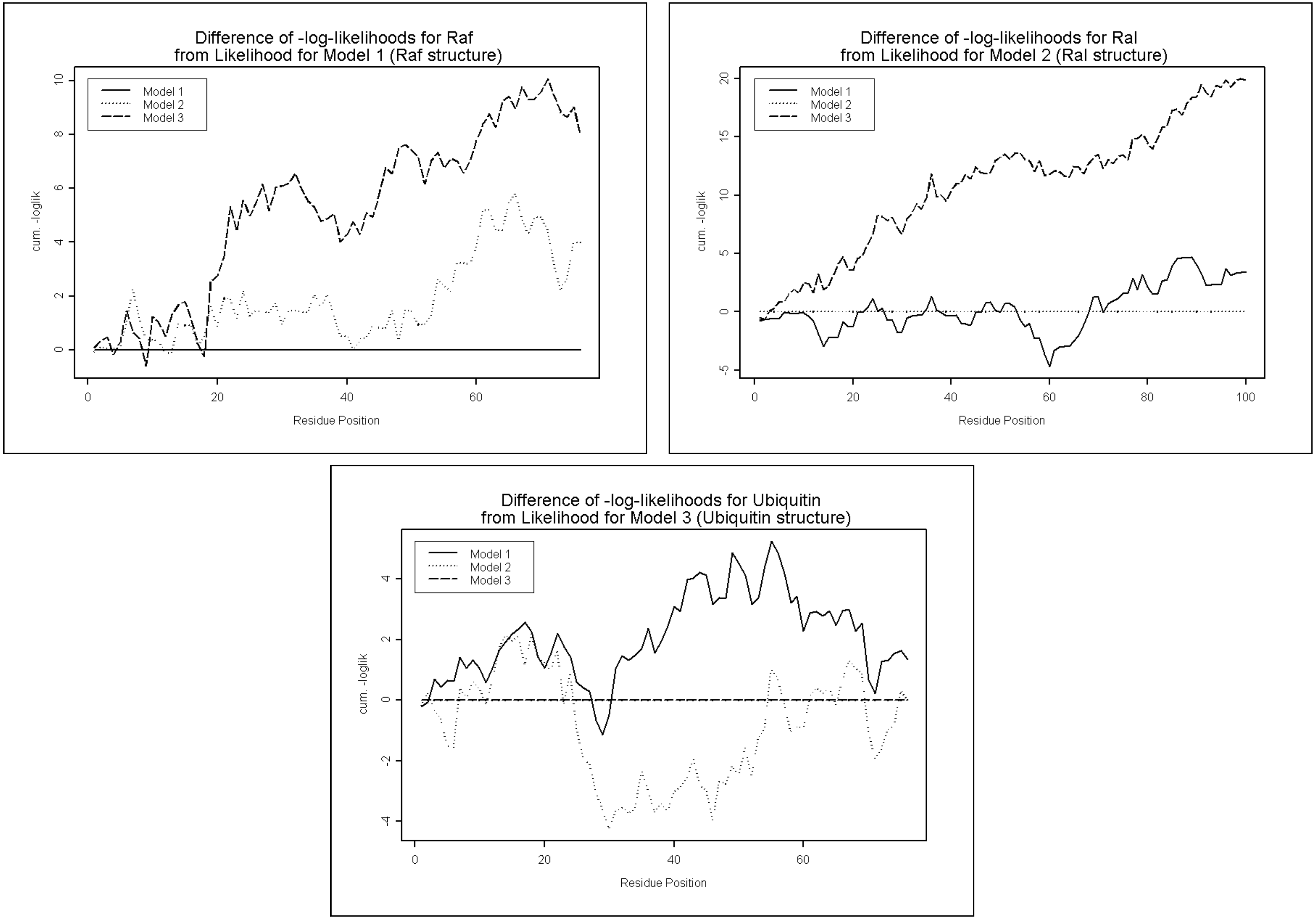


Figure 9: Differences of negative log likelihoods according to method 2.

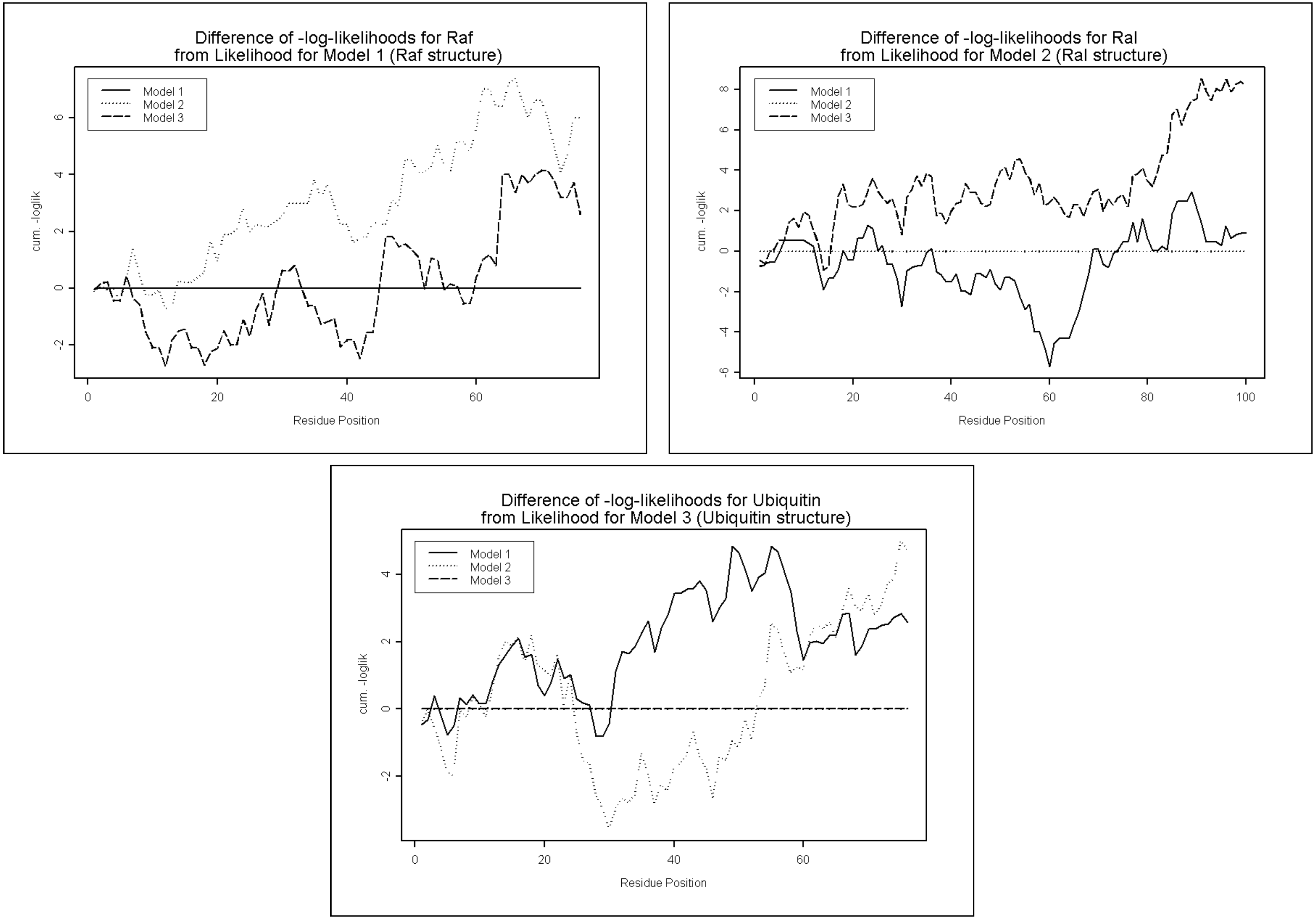
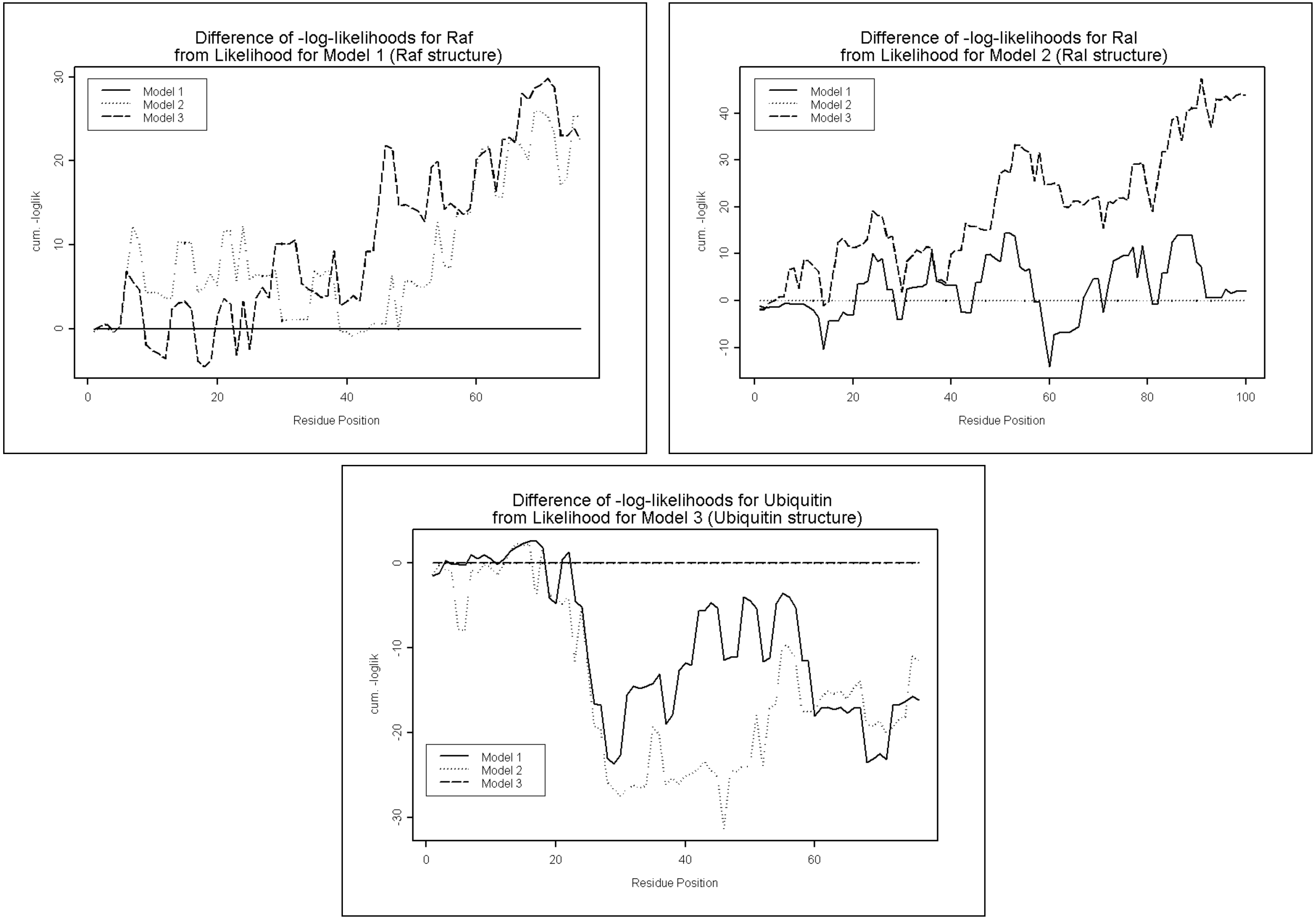


Figure 10: Differences of negative log likelihoods according to method 3.



6. Discussion and outlook

In this paper a modified filter algorithm was applied to sequence and structure protein data in order to classify the three amino acids sequences according to their secondary structures. This filter algorithm has been implemented in S-Plus. The application demonstrated that the choice of the initial distribution influenced the classification results. Further applications of this filter algorithm should consider this kind of influence in order to avoid false classifications. In this context, the iterative application of this filter algorithm is of interest and will be investigated in further analyses.

The estimation of the H and the Φ matrices was done by a simple computation of relative frequencies. These estimation methods can be improved by the application of the EM-algorithm. Nevertheless, the results for the Raf and the RalGDS sequence demonstrated that the state-space modeling is appropriate for the secondary structure classification even for a small number of observations and simple estimation methods. An application to whole protein sequence or structure families aim at a further validation of this classification method. For that purpose, the computation of posterior probabilities (see **Section 3.2**) extends the classification tools of the filtering approach. Besides this, the consideration of the Kullback-Leibler distance enables the assessment of the discrimination of the different models.

Moreover, the smoothing algorithm is an alternative to sequence-to-structure alignments and will be compared with the sequence-to-structure alignment described in Brunnert *et al.* (2002). In Bienkowska (2000) the state-space modeling has been applied successfully to 188 protein structures.

Acknowledgement

The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") is gratefully acknowledged.

Appendix

Table A1. Estimates of the conditional output probabilities.

Amino acid (three-letter-code, name)	$\alpha-$Helix	$\beta-$sheet	Turn	Other
Ala, Alanine	**0.0755**	**0.0182**	**0.0435**	**0.0707**
Arg, Arginine	**0.0566**	**0.0364**	**0.0435**	**0.0856**
Asn, Asparagine	**0.0566**	**0.0273**	**0.0435**	**0.0367**
Asp, Aspartic acid	**0.1132**	**0.0182**	**0.1089**	**0.0733**
Cys, Cysteine	**0.0189**	**0.0455**	**0.0217**	**0.0123**
Gln, Glutamine	**0.0189**	**0.0727**	**0.0435**	**0.0247**
Glu, Glutamic acid	**0.0566**	**0.0636**	**0.0869**	**0.0489**
Gly, Glycine	**0.0377**	**0.0364**	**0.0869**	**0.0732**
His, Histidine	**0.0189**	**0.0455**	**0.0217**	**0.0123**
Ile, Isoleucine	**0.0566**	**0.100**	**0.0217**	**0.0489**
Leu, Leucine	**0.0566**	**0.1364**	**0.0653**	**0.1099**
Lys, Lysine	**0.1132**	**0.0909**	**0.0653**	**0.0733**
Met, Methionine	**0.0566**	**0.0182**	**0.0217**	**0.0123**
Phe, Phenylalanine	**0.0189**	**0.0455**	**0.0217**	**0.0367**
Pro, Proline	**0.0377**	**0.0273**	**0.0869**	**0.0610**
Ser, Serine	**0.0377**	**0.0727**	**0.0869**	**0.0489**
Thr, Threonine	**0.0377**	**0.0727**	**0.0653**	**0.0610**
Trp, Tryptophan	**0.0189**	**0.0090**	**0.0217**	**0.0123**
Tyr, Tyrosine	**0.0377**	**0.0090**	**0.0217**	**0.0247**
Val, Valine	**0.0755**	**0.0545**	**0.0217**	**0.0733**

References

Bienkowska, J.R., Yu, L., Zarakhovich, S., Rogers, R.G. and Smith, T.F. (2000), “Protein fold recognition by total alignment probability”, *Proteins: Structure, Function and Genetics*, 40, 451-462.

Brunnert, M., Fischer, P., Vetter, I. and Urfer, W. (2002), “Sequence-to-structure alignment using a statistical analysis of core models and dynamic programming”, *in preparation*.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977), “Maximum likelihood from incomplete data via the EM algorithm”, *Journal of the Royal Statistical Society, Series B*, 39, 1-38.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998), *Biological sequence analysis*, Cambridge University Press, Cambridge.

Kalman, R.E. (1960), “A new approach to filtering and prediction problems”, *Journal of Basic Engineering*, 82, 35-45.

MathSoft (2000), *S-PLUS 2000*, MathSoft Inc., Seattle, WA.

Nassar, N., Horn, G., Herrmann, C., Scherer, A., McCormick, F. and Wittinghofer, A. (1995), “The 2,2 $\dot{A}$ crystal structure of the Ras-binding domain of the serine/threonine kinase c-Raf1 in complex with Rap1A and GTP analogue”, *Nature*, 375, 554-560.

Rabiner, L.R. (1989), “A tutorial on hidden Markov models and selected applications in speech recognition”, *IEEE*, 77, 257-286.

Schmitz, N. and Urfer, W. (1997), “State dependent time series models for heart rate dynamics data and their application to psychophysiology”, Informatik, Biometrie und Epidemiologie in Medizin und Biologie, 28, 169-184.

Sousa, L., Santos, M., Turkman, M.A.A. and Urfer, W. (2001), “Bayesian Analysis of Protein Sequence Data”, *Technical Report 48/2001, SFB 475, University of Dortmund*.

Stultz, C.M., White, J.V. and Smith, T.F. (1993), “Structural analysis based on state-space Modeling”, *Protein Science*, 2, 305-314.

Urfer, W. (2001), “Statistical tools for extracting information from DNA sequence data”, In: Mathematical Statistics with Applications in Biometry: Festschrift in Honour of Siegfried Schach. J. Kunert and G. Trenkler (eds.), Eul Verlag, Köln, 103-112.

White, J.V. (1988), “Modeling and filtering for discretely valued time series”, in J.C. Spall, ed., chapter 10, *Bayesian analysis of time series and dynamic models*, Dekker, New York.

White, J.V., Stultz, C.M. and Smith, T.F. (1994), “Protein classification by stochastic modeling and optimal filtering of amino-acid sequencing”, *Mathematical Biosciences*, 119, 35-75.