Structural analysis of effectors of the oncogenic Ras proteins

Marcus Brunnert

Department of Statistics, SFB 475
University of Dortmund

TIES Conference 2002,
Genova

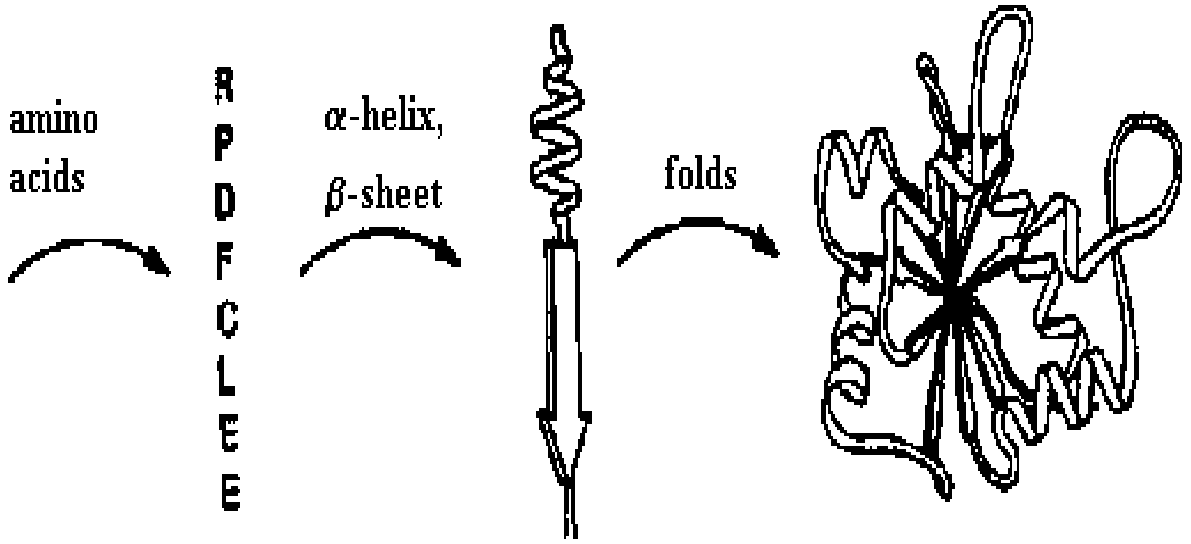
Outline

- Underlying molecular genetic problem.

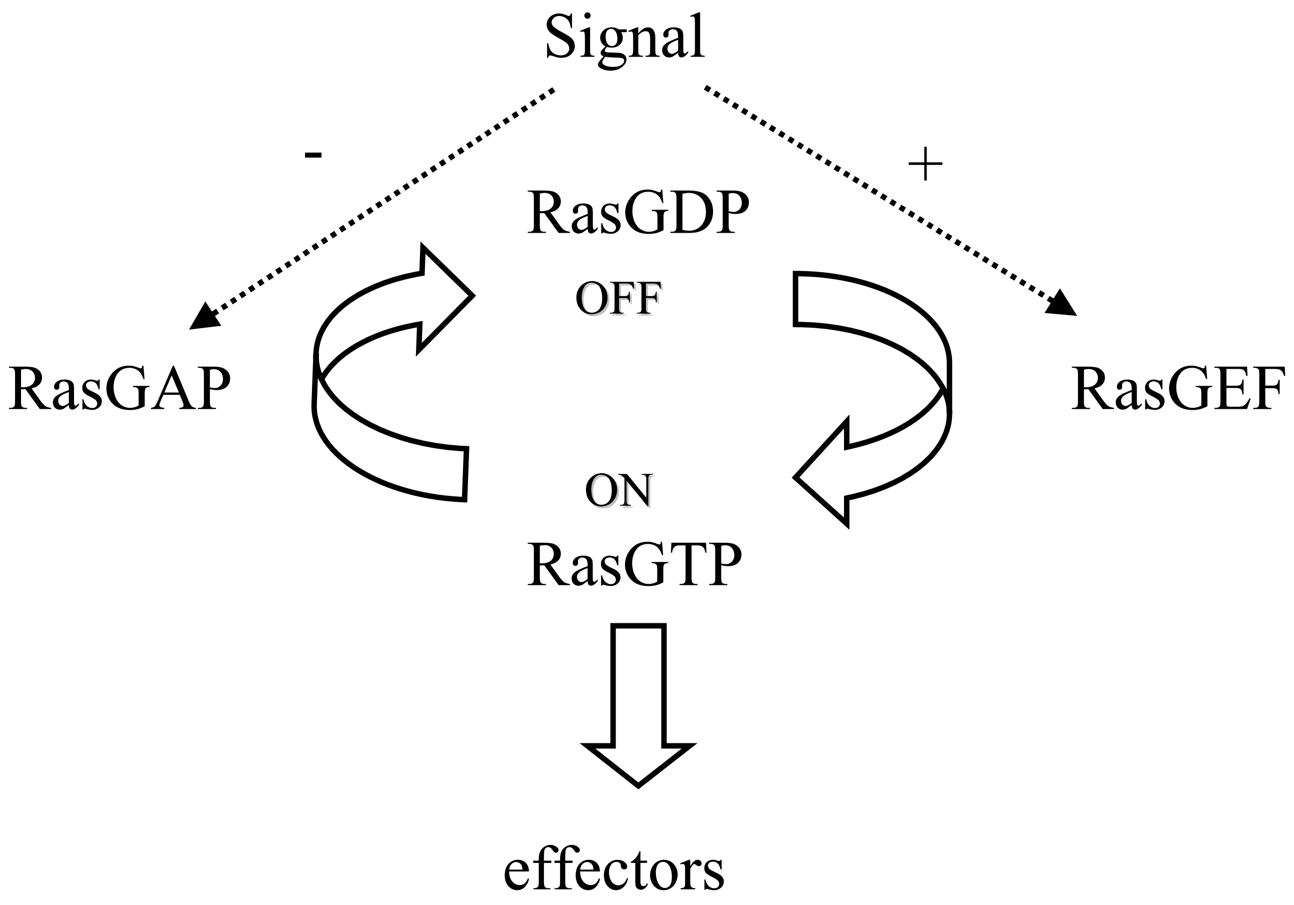
- Empirical protein structure prediction to sequence and structure data.

3. Classification method to secondary sequence and structure data.

1. Protein structures



Ras- a molecular switch



Signal

-

+

RasGDP

OFF

RasGAP

RasGEF

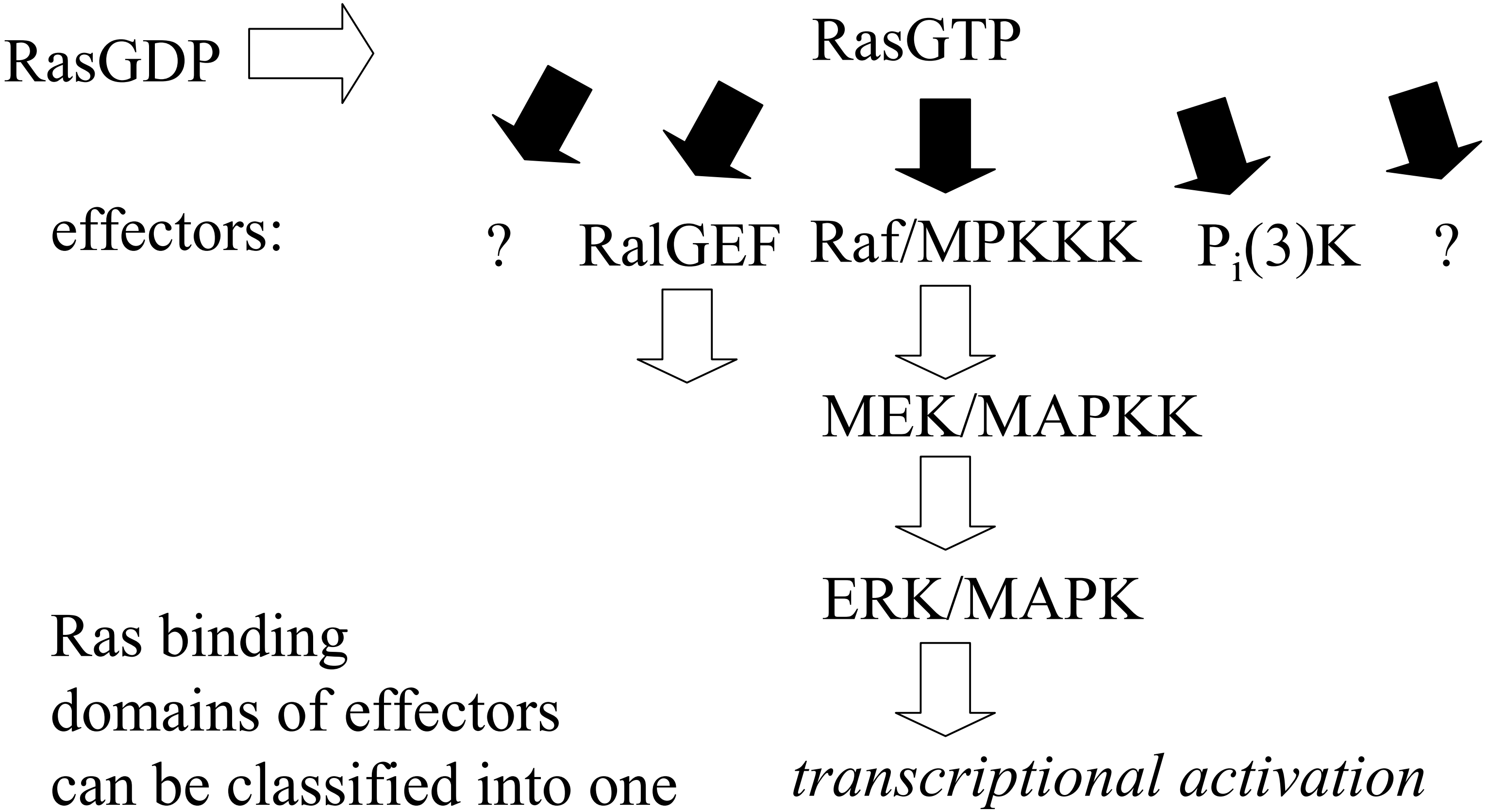
ON

RasGTP

effectors

Wittinghofer and Waldmann (2000)

More signal transduction pathways



RasGDP RasGTP

effectors: ? RalGEF Raf/MPKKK $P_i(3)K$?

MEK/MAPKK

ERK/MAPK

transcriptional activation

Ras binding domains of effectors can be classified into one protein structure family

2. Sequence-structure alignment

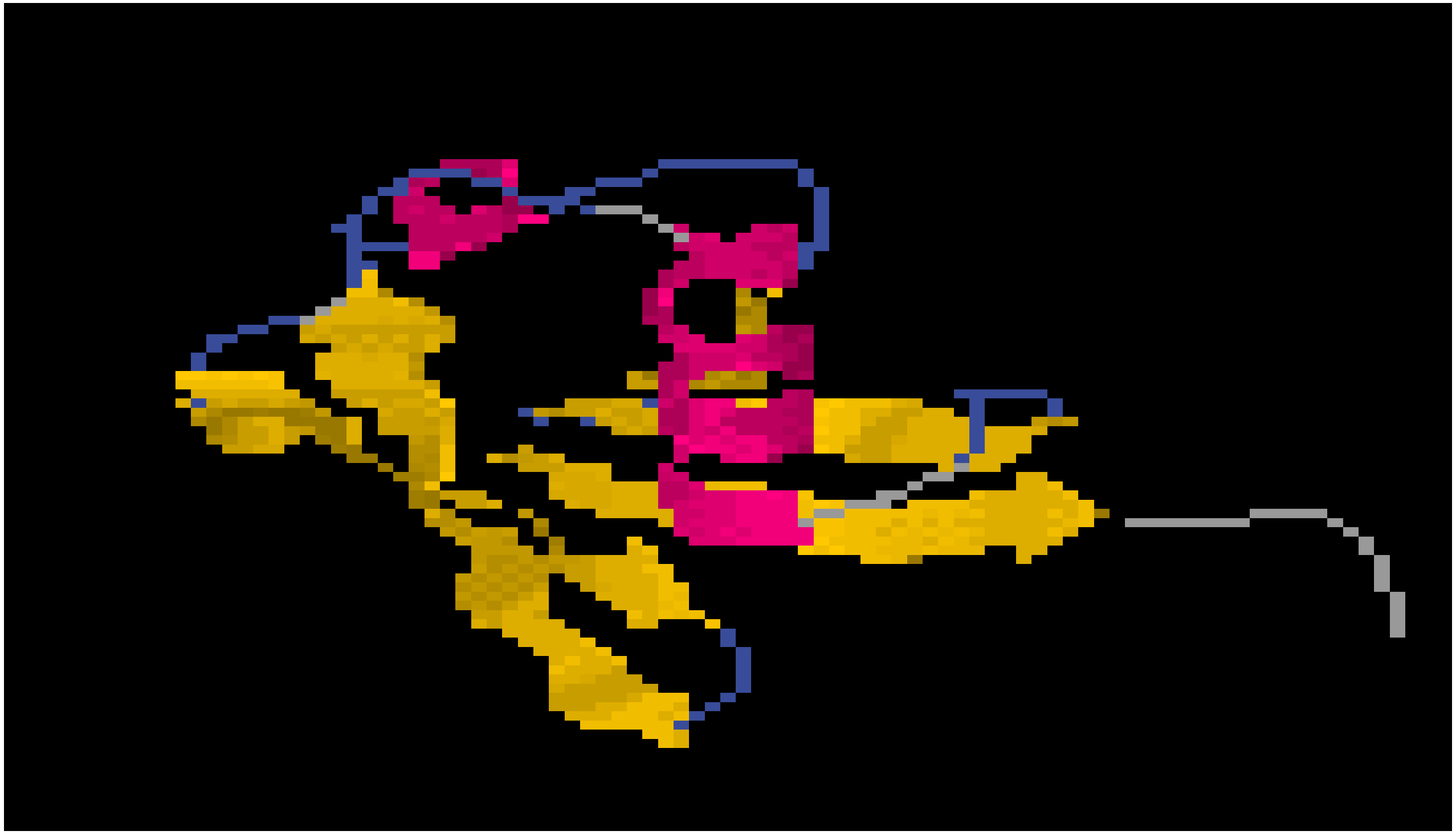
- Data of a protein core (protein domain)
- Proposal of a s scoring function
- Search algorithm for an optimal sequence-structure alignment
- Application
- Outlook

Data of a protein core

A protein core is composed of several quantitative and qualitative traits.

- **Core segments**
  - Information about the position of the secondary structures.
  - A segment is composed of a subsequence of the amino-acid sequence. The elements of this subsequence are called core elements.
  - ...
- **Properties of amino acids**
  - Hydrophobicity
  - …
- **Spatial neighbourhood of the segments**
  - Order of segments in the tertiary structure
  - Gaps between segments (amino acids not assigned to a secondary structure) are not considered in the core.
  - ...
- **...**

Core of the protein Ubiquitin



M Q I F V K T L T G K T I T L G V G P S A T I G N V K A K I Q A K G G I P P

A Q Q R L I F A G K Q L G A G R T L S A Y N I Q K G S T L H L V L R L R G G

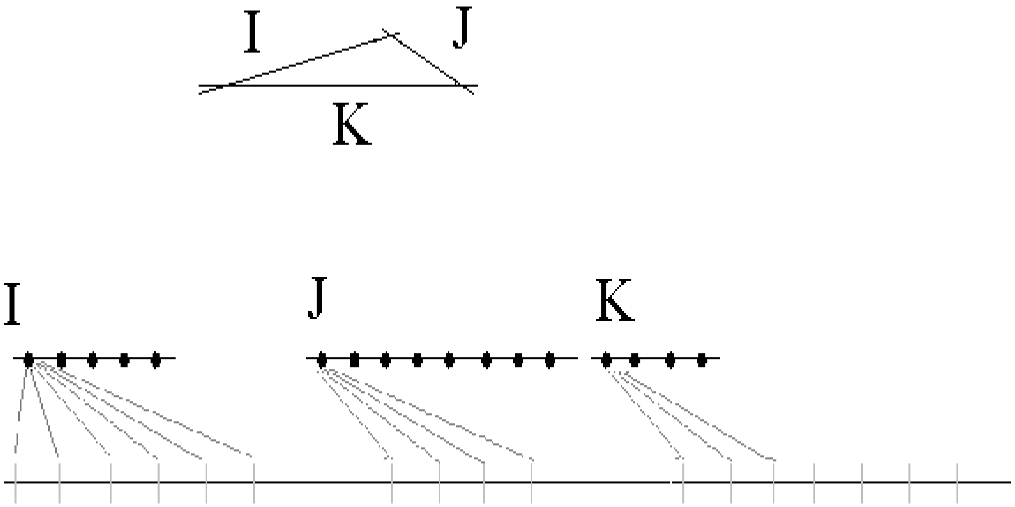
Core of the Ras binding domain of Raf

P S K T S N T I R V F L P N K Q R T V V N V R N G M S L H D -
C L M K A L K L V R G Q P G C C A V F R L L H G H K G K K
A R L D W N T D A A S L I G G G L

Core of the Ras binding domain of Ral-GEF

G S S S S L P L Y N Q Q V G D C C I I R V S L D V D N G N M
Y K S I L V T S Q D K A P T V I R K A M D K H N L D G D G P
G D Y G L L Q I I S G D H K L K I P G N A N V F Y A M N S A
A N Y D F I L K K R

Proposal of a scoring function

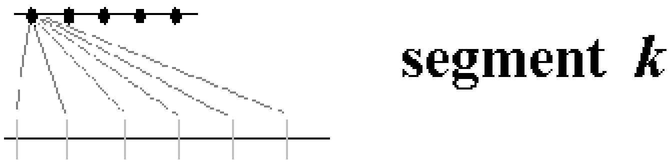


core segments

amino-acid sequence of core segments

aligned amino-acid sequence

Proposal of a scoring function



$$p_k :\ T_k \to [0,1],$$

$$p_k\left(\boldsymbol{b}_{l_k}^{(t)}\right)=\prod_{j=t}^{t+l_k-1}\mathrm{P}\left(b_{l_k}^{(t)}(j)\right)\prod_{j=t}^{t+l_k-2}\mathrm{P}\left(b_{l_k}^{(t)}(j),b_{l_k}^{(t)}(j+1)\right).$$

➢Score of a core segment: $S[k,t]$

Search algorithm

- Search for an optimal sequence-structure alignment

$\sum_{k=1}^{K} S[k, t_k]$ has to be maximized with respect to the constraints:

$$1 \leq t_k < n + 1 - \sum_{k' > k} l_{k'}, \; k = 1, \ldots, K \quad \wedge$$

$$t_{k-1} + l_{k-1} - 1 < t_k, \; k = 1, \ldots, K, \; t_0 = 0 \; and \; l_0 = 0.$$

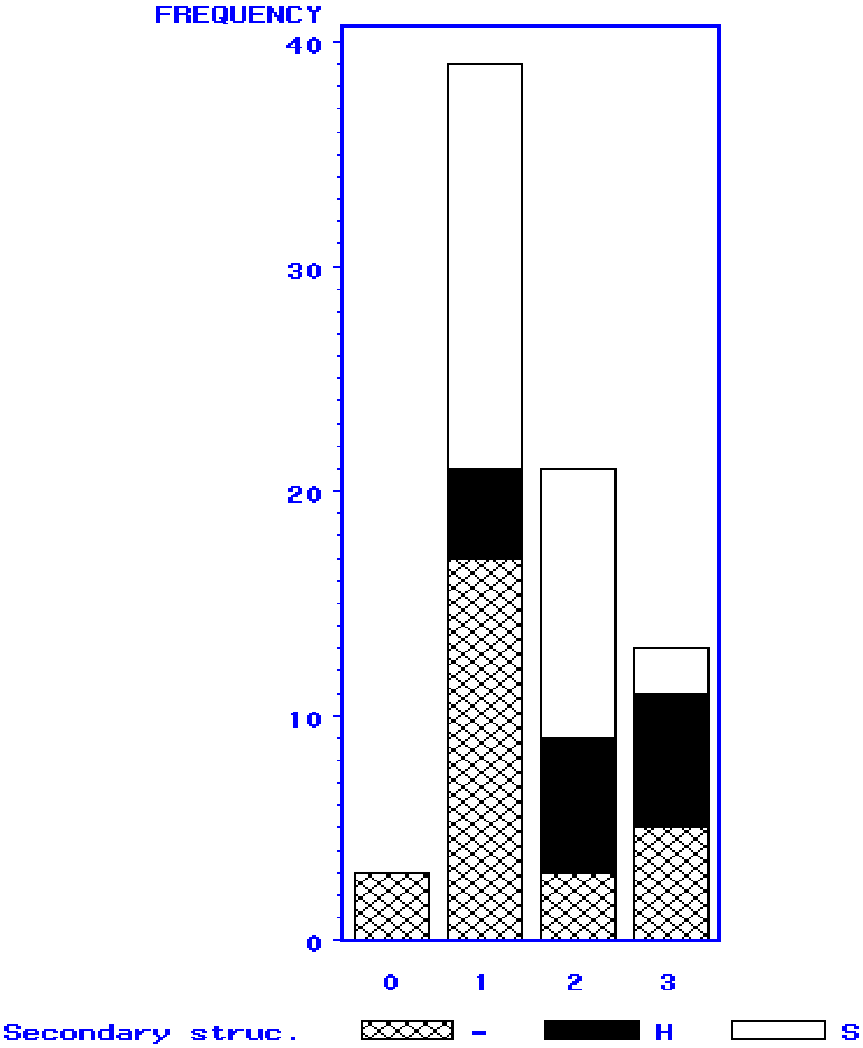
- Dynamic programming approach has been implemented in the program *Placer*.

Results of the application

Figure: Parts of the sequence-structure alignment of Ubiquitin

Core Raf	-	-	-	-	-	S	S	S	S	S	S	-	-	-	-	-	-	-	-	S	S	S	S	S	S
Core Ral	S	S	S	S	S	S	S	S	-	-	-	S	S	S	S	S	S	-	-	-	H	H	H	H	H
Core Ubiquitin	-	-	-	-	-	S	S	S	S	S	S	S	-	-	-	S	S	S	S	S	S	S	-	H	H
Original core	S	S	S	S	S	S	S	-	-	S	S	S	S	S	S	S	-	-	-	-	-	-	H	H	H
Identical structures	1	1	1	1	1	3	3	0	1	2	2	2	1	1	1	2	1	2	2	1	0	0	1	2	2
Sequence position	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5

Results of the application

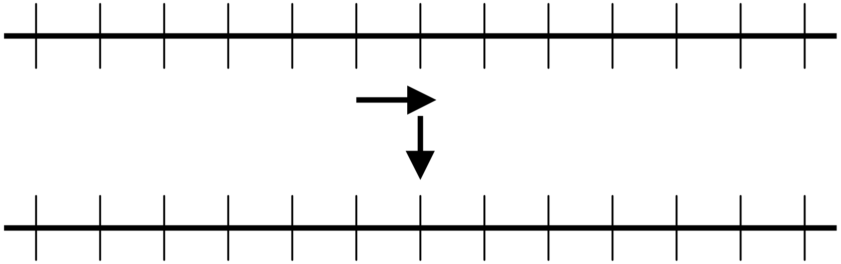


Outlook

- Consideration of gaps between segments.
- Improvement of the probability function on the basis of Markov random fields (MRF).
    - Definition of spatial neighbourhoods according to Voronoi contact relations (Voronoi tesselations).
    - Modeling spatial neighbourhoods in graphs.
    - Definition of a MRF on the graph.
    - Assuming this MRF, the probability of the occurrence of several neighbouring amino acids in the core can be used for scoring the core segments.

3. Classification of amino-acid sequences

- Classification of an amino-acid sequence to a secondary structure.



Secondary structure

Primary structure, observed amino-acid sequence

- State-space model
  - Filtering algorithm
    - Likelihood calculation

State-space model

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

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

.

$$M = \left(m, n, \Phi, \boldsymbol{H}, \boldsymbol{x_1}\right)$$

$$\boldsymbol{x_t} = \begin{pmatrix} P(x_t = 1) \\ P(x_t = 2) \\ \vdots \\ P(x_t = n) \end{pmatrix} \qquad \boldsymbol{y_t} = \begin{pmatrix} P(y_t = 1) \\ P(y_t = 2) \\ \vdots \\ P(y_t = m) \end{pmatrix}$$

$$\left(x_t\right)_{t=1,2,3,\dots}$$

$$\left(y_t\right)_{t=1,2,3,\dots}$$

$$Y_d = \left(y_1, y_2, \dots, y_d\right)$$

Filtering algorithm

Input : Model $M \quad = \left(m, n, \Phi, \boldsymbol{H}, \boldsymbol{x_1}\right)$ and observed sequence

$Y_d = \left(y_1, y_2, \ldots, y_d\right)$

Initialisation : $\boldsymbol{x}_1^- = \boldsymbol{x_1}$

Recursion for *t*, $1 \leq t \leq d$: $\boldsymbol{y}_t^- = \boldsymbol{H}\boldsymbol{x}_t^-$

State update: $\boldsymbol{v}_t = H\left[y_t = k\right]^T * \boldsymbol{x}_t^-$

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

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

State propagate : $\boldsymbol{x}_{t+1}^- = \Phi\boldsymbol{x}_t^+$

Termination $t = d$

Likelihood calculation

$$M_1, \ldots, M_q$$

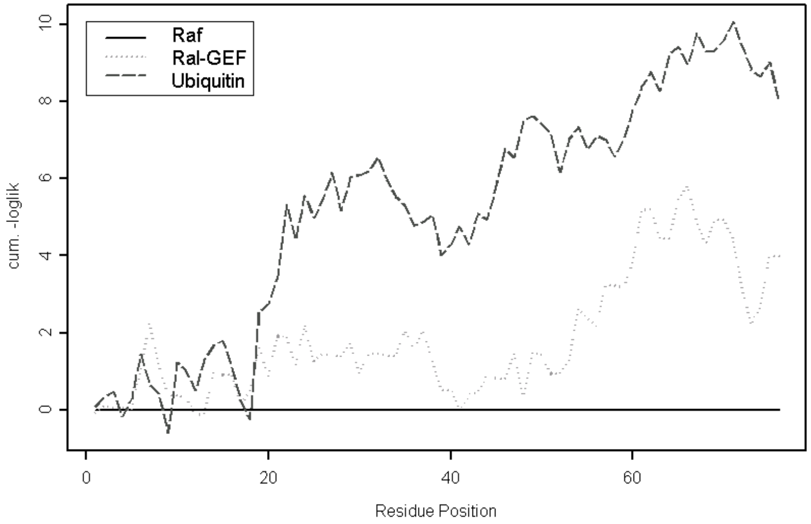
$$L(Y_d | M_l) = P(Y_d | M_l) = P(y_1) \prod_{t=2}^{d} P(y_t | Y_{t-1}).$$

$$\log L(0) = 0 \;\; and$$

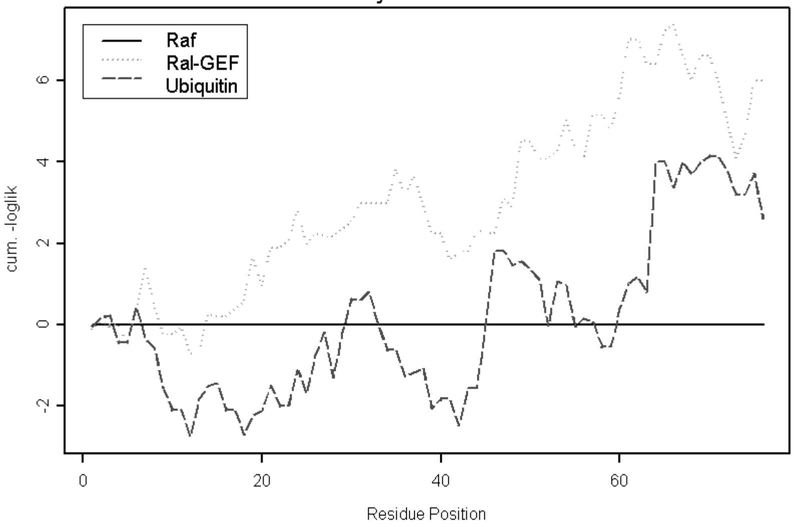
$$\log L(t) = \log L(t-1) + \log P(y_t | y_{t-1}), \; t = 1, \ldots, d.$$

Results

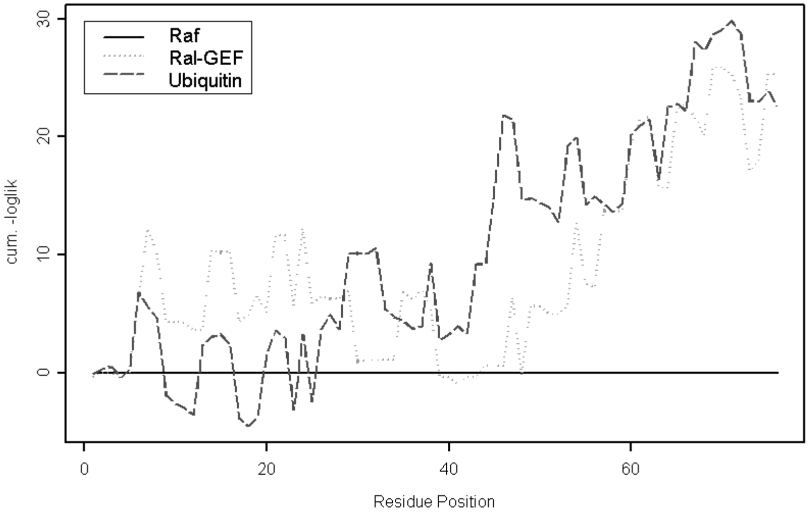
Difference of negative log-likelihoods of the Raf-sequence
from its reference negative log-likelihood.
Parameters estimated by method 1.



Difference of negative log-likelihoods of the Raf-sequence
from its reference negative log-likelihood.
Parameters estimated by method 2.



Difference of negative log-likelihoods of the Raf-sequence
from its reference negative log-likelihood.
Parameters estimated by method 3.



Summary and outlook

- Two empirical methods were applied to known protein structures.
- Improvement of the sequence-structure alignment:
  - Other scoring function.
- Improvement of the classification method:
  - Smoothing.
- Combination of both methods.

References

Brunnert, M., Krahnke, T. and Urfer, W. (2001), “Secondary structure classification of amino-acid sequences using state-space models”, *Technical Report 49/01*, SFB 475, University of Dortmund.

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.

White, J. V., Muchnik, I., and Smith, T.F. (1994), “Modeling protein cores with Markov random fields”, *Mathematical Biosciences*, 124, 149-179.

Wittinghofer, A. and Waldmann, H. (2000), “Ras-A Molecular Switch Involved in Tumor Formation”, *Angewandte Chemie*, 39/23, 4192-4214.