Sequence-structure alignment using a statistical analysis of core models and dynamic programming

Marcus Brunnert[1], Paul Fischer[2] and Wolfgang Urfer[1]

[1]Fachbereich Statistik, Universität Dortmund

[2]IMM, Danmarks Tekniske Universitet, DK-2800 Kgs. Lyngby

Abstract

The expanding availability of protein data enforces the application of empirical methods necessary to recognize protein structures. In this paper a sequence-structure alignment method is described and applied to various Ubiquitin-like folded Ras-binding domains. On the basis of two probability functions that evaluate similarities between the occurrence of amino-acids in the primary and secondary protein structure, different versions of simple scoring functions are proposed. The application of the program 'PLACER' that uses a dynamic programming approach enables the search for an optimal sequence-structure alignment and the prediction of the secondary structure.

Keywords: Sequence-structure alignment, core model, dynamic programming, secondary structure prediction.

1 Introduction

Besides experimental methods like the x-ray crystallography, the application of empirical methods are discussed very frequently in order to recognize protein structures (Lathrop et al., 1996, Thiele et al., 1999, Bienkowska et al., 2000). In order to achieve statistical inferences from

multivariate protein data containing sequence and structure information, we give an empirical method for the *sequence-structure alignment.* Alignment methods are commonly used in the field of bioinformatics (Durbin et al., 1998) and can imply new information about protein structures using the analysis of protein data from experiments or protein data banks like the PDB of the Research Collaboratory for Structural Bioinformatics (RCSB-PDB). The idea of deducing the protein structure from an already known protein structure template uses the nature of structural similarities that are likely to persist during protein evolution (Madej et al., 1995). *Cores*, consisting of structural elements that are conserved during protein evolution are used for the deduction of protein structures, even if there is no relevant homology on the protein sequence level. Structural models like these cores are used for the sequence-structure alignment of a protein sequence (Lathrop et al., 1996).
In this technical report, we give a statistical method of a sequence-structure alignment by using the data of a specified protein structure template called a 'core model'. The constructing scheme of our sequence-structure alignment method is described in Section 2. In Section 3, we develop empirical *scoring functions* in order to compare different sequence-structure alignments. Furthermore, we present in Section 4 an algorithm for solving this sequence-structure alignment problem. In Section 5, we present an application to data of Ubiquitin-like folded proteins (SCOP–databank–http://scop.mrc–lmb.cam.ac.uk/ scop/Murzin et al., 1996). We apply the developed program 'PLACER' that uses a dynamic programming approach to search for an optimal sequence-structure alignment. Finally, we discuss the proposed statistical method for a secondary structure prediction using different sequence-structure alignments.

2 Constructing alignments

Modelling the protein structure uses the principal concept of the primary structure (sequence of amino–acids), the secondary structure (α-helix or β-sheet structures) and the tertiary structure (folded secondary structures). Important spatial structures that are highly conserved during evolution can be described by protein cores. These cores are cut off the whole protein structure and are used as structural templates in the sequence–structure alignment.

Similarities between amino–acid sequences and known core models representing the core structure can imply information about the structure of a new amino–acid sequence with hardly known structural information. The concept of similarity in sequence–structure alignment is not unique. Empirical approaches to scoring functions that quantify similarities according to the sequence–structure alignment evaluate complex multivariate sequence and structure data. Besides sequence identities, the scoring of different structural interactions make use of appropriate statistical methods and effective algorithms. A proposal to the concept of similarity in the context of sequence–structure alignment will be described later by using two different probability functions for scoring alignments.
Similarly to the core definitions in White et al. (1994), we construct sequence–structure-alignments from structure templates that are called *core models.* Using the information about the placement of α-helix or β-sheet structures, we can assign a *core segment* to a specific secondary structure with its amino–acid sequence. Every specific amino–acid in a core segment is called a *core element.* A vector $\mathbf{a_k}$ of n_k core elements referring to a core segment *k, k=1,. . .,K* is denoted by

$$\mathbf{a_k} = [a_k(1), \dots, a_k(n_k)]', \tag{1}$$

where $a_k(i) \in A$ is the observed amino-acid using an amino-alphabet A like the one–letter-code of 20 amino–acids (cf. Kanehisa, 2000). The core elements of the K core segments according to the core model are denoted by $a_1(1)$,. . ., $a_1(n_1)$,. . ., $a_K(1)$,. . ., $a_K(n_K)$. The lengths of the core segments are denoted by l_1,. . ., l_K. For simplicity, the lengths can be set to the number of core elements n_k due to the kth segment. The set of vectors $\mathbf{a_1}$,. . .,$\mathbf{a_K}$ describes the whole *core model.* We can denote a vector $\mathbf{b_{l_k}} = [b_{l_k}(t), \dots, b_{l_k}(t + l_k - 1)]', b_{l_k}(j) \in b(1), \dots, b(n)$, for the alignment of a primary sequence $\mathbf{b}$=$b(1)$,. . .,b(n), b(i)$\in A$, to segment *k.* In aligning a sequence of core elements $a_1(1)$,. . ., $a_1(n_1)$,. . .,$a_K(1)$,. . ., $a_K(n_K)$ to a primary sequence $\mathbf{b}$, we consider the following conditions to an admissible starting position t of the aligned sequence

$$1 \leq t_k \leq n + 1 - \sum_{k' \geq k} l_{k'}, k = 1, \dots, K \tag{2}$$

$$t_{k-1} + l_{k-1} - 1 < t_k, k = 2, \dots, K. \tag{3}$$

These conditions guarantee the alignment of all core segments and the sequential ordering of the segments. According to all admissible vectors $\mathbf{b_{l_k}}$, fulfilling (2) and (3), we can denote a set $\mathcal{T}_k := \{[b_{l_k}(t), \ldots, b_{l_k}(t+l_k-1)]' | t$ is an admissible starting position$\}$ according to segment $k = 1, \ldots, K$.

3 Probability functions for sequence–structure alignments

With respect to the structure templates, some amino–acids cannot be assigned to a structure, these amino–acids are called here *gaps*. Therefore we describe scoring functions without considering gaps and with considering gaps.

Let us denote the probabilities of an occurrence of a specific amino–acid at the sequence position j by $P(b_{l_k}(j) = w)$, $w \in A$ and the probability of an occurrence of a specific amino–acid pair in sequential adjacent positions $(j, j+1)$ by $P(b_{l_k}(j) = v, b_{l_k}(j+1) = w)$, $v, w \in A$. Besides the one-letter code of 20 amino–acids, we use a binary code, that takes into account the hydrophobicity of amino-acids. Then, we have to consider $A = \{hydrophobic, hydrophilic\}$. By neglecting stochastic dependencies between aligned amino–acids we can propose the following probability function for the alignment of the kth segment,

$$\begin{aligned} p_k \quad &: \quad \mathcal{T}_k \rightarrow [0,1], \\ p_k(\mathbf{b_{l_k}}) \quad &= \quad \prod_{j=t}^{t+l_k-1} P(b_{l_k}(j)) \prod_{j=t}^{t+l_k-2} P(b_{l_k}(j), b_{l_k}(j+1)) \end{aligned} \tag{4}$$

Taking into account the various segment lengths, we divide $p_k(\mathbf{b_{l_k}})$ by the length l_k. This ratio serves as a score for each aligned sequence.

Alternatively, we propose a second probability function for the sequence–structure alignment. Now, we consider stochastic dependencies within the sequence of amino-acids by applying the Markov property as follows

$$\begin{aligned} \tilde{p}_k \quad &: \quad \mathcal{T}_k \rightarrow [0,1], \\ \tilde{p}_k(\mathbf{b_{l_k}}) \quad &= \quad P(b_{l_k}(t)) \prod_{j=t}^{t+l_k-2} P(b_{j+1}|b_j) \end{aligned} \tag{5}$$

In order to recognize the various segment lengths, the ratio $\frac{\tilde{p}_k(\mathbf{b}_{\mathbf{l_k}})}{\tilde{p}_k(\mathbf{a_k})}$ serves as a score for each aligned sequence, too. Analogously to (5), $\tilde{p}_k(\mathbf{a_k})$ is the corresponding probability of the amino–acid sequence of the kth core segment.
To compute the scores, we need estimates for the probabilities. We estimate the probabilities by using the frequencies of the occurrences of the amino-acids and amino-acid pairs. In order to avoid zero values we add so called pseudo counts (Durbin et al., 1998). Let $c \geq 0$ be the pseudo count (we use $c = 1$ in the application). Then our estimators are:

$$\hat{P}_k(v) \quad := \quad \begin{cases} \frac{c+1}{|A|c+1} & \text{if } a_k(t) = v \\ \frac{c}{|A|c+1} & \text{otherwise} \end{cases} \tag{6}$$

$$\hat{P}_k(w|v) \quad := \quad \frac{c + |\{t \mid t \geq 2 \wedge a_k(t) = w \wedge a_k(t-1) = v\}|}{|\{t \mid t \geq 2 \wedge a_k(t) = w\}| + |A|\, c} \tag{7}$$

Finally, we add penalty terms to our scoring functions. In this context, we penalize false gap lengths between aligned core segments. The model gap length can be determined by the observed core gaps from data or it can be determined by estimating the gap length. For example, the arithmetic mean of all calculated gap lengths from the alignments can be an estimate. Given the model gap length and the resulting gap length of the alignment, we multiply the corresponding difference by a penalty parameter.

4 A Dynamic Programming Algorithm

In this section we present an abstract formulation of the problem and an algorithm solving it.

Let n be the length of the primary sequence and $\ell_0, \ldots, \ell_{k-1}$ be the lengths of the core sequences. We assume that indexing starts with 0. Let a score matrix $Sc(i,j)$, $i = 0, \ldots n-1$, $j = 0, \ldots, k-1$ with real entries be given. Proposals to the calculation of Sc are given in the previous Section. Then we want to solve the following problem: find integers $i_0, \ldots, i_{k-1}$ such that:

$$0 \leq i_0 < i_1 < \cdots < i_{k-1} \leq n - \ell_{k-1} \tag{8}$$

$$i_j - i_{j-1} \geq \ell_{j-1}, \quad j = 1, \ldots, k-1 \tag{9}$$

$$\sum_{j=0}^{k-1} Sc(i_j, j) \text{ is maximal.} \tag{10}$$

The conditions guarantee that order of the core sequences is maintained, that they do not overlap and are optimally placed.

The dynamic programming algorithms first computes the optimal value $Sc_{opt}(l, r, j)$ of $Sc(i, j)$ for every interval $[l, r]$, $0 \leq l \leq i \leq r - \ell_j + 1$ as well as the (left most) position where this assumed. Informally, $Sc_{opt}(l, r, j)$ is the optimal start position of an alignment of core sequence $\mathbf{a}_j$ with $\mathbf{b}$ such that it totally inside the interval $[l, r]$.

$$\begin{aligned} Sc_{opt}(l, r, j) &:= \max \{Sc(i, j) \mid l \leq i \leq r - \ell_j + 1\} \\ pos_{opt}(l, r, j) &:= \operatorname{argmax} \{Sc(i, j) \mid l \leq i \leq r - \ell_j + 1\} \end{aligned}$$

Then the optimal solutions are iteratively computed for the first two, three etc. core sequences. Let $Sol(i, j)$ denote the (summed) score for the optimal choice of $i_0, \cdots, i_j$ which respects conditions (8) – (10) and $i_j + \ell_j - 1 \leq i$, i.e. the last core sequence $\mathbf{a}_j$ does not exceed i. Obviously we have for $j = 0$:

$$Sol(i, 0) = Sc_{opt}(0, i, 0) \tag{11}$$

For convenience let $L_{j-1} := \sum_{t=0}^{j-1} \ell_t$ and $R_j(i) = i - \ell_j + 1$ be the first resp. last possible choice for i_j. Then the computation rule for $j > 0$ is

$$Sol(i, j) = \max_{L_{j-1} \leq t \leq R_j(i)} \{Sol(t, j - 1) + Sc_{opt}(t, i, j)\} \tag{12}$$

That is, for every choice t of i_j which respects conditions (8) – (10) we add the optimum solution for the first $j-1$ sequences before t and the optimal placement of the j-th segment after t but not exceeding i and take the maximum. Then an overall maximum is $Sol(0, n-1)$. In order to trace the positions of an optimal choice of the i_j a solution should also maintain a list $P(i, j)$ of the optimal positions for $i_1, i_2, \ldots, i_j$, where $i_j \leq i$, computed so far. In the computation step (12) the list $P(t, j - 1)$ is extended by the optimal choice if i_j to become $P(i, j)$.

Let us look at the running time of the algorithm. The computation of the initial values $Sc_{opt}(l, r, j)$ can be performed in time $O(kn^2)$ because for fixed j and r the value $Sc_{opt}(l, r, j)$ can be computed in constant time from $Sc_{opt}(l, r - 1, j)$. The update formula has to be computed for $O(kn)$ many pairs (i, j) every such update means computation the maximum over at most n values. Hence we have

Theorem 4.1 *The placement problem can be solved in time $O(kn^2)$.*

5 Application

This sequence-structure alignment method has been applied to data of protein domains. From the primary structure and the secondary structure information of the RCSB-PDB concerning the protein Ubiquitin (RCSB-PDB ID: 1ubi) as well as the Ras-binding domains of *RalGEF* (1lxd), *Raf* (1guaB), *Rgl* (1ef5), *Rlf* (1rfl), $P_i(3)$-kinase (1he8B) the core models were constructed. In the SCOP-databank, *Ral*, *Raf*, *Rgl*, *Rlf* and $P_i(3)$-kinase are classified into the superfamily of *Ubiquitin*-like folded proteins. Optimal sequence-structure alignment are computed with different versions of the program PLACER with respect to different scoring functions. The different versions of PLACER were programmed in JAVA. We use the binary code for A in all PLACER versions. First of all, we implemented the scoring function according to the probability function p_k from (4) in the PLACER versions 2, 3a and 3b. In PLACER version 2 we consider no gaps. Linear gap penalties are implemented in PLACER version 3a and 3b. In Placer version 3a the gap penalty parameter is estimated by using the data of a single protein. Contrary to this, in PLACER version 3b, the gap penalty parameter is estimated by the data of all proteins. Finally, the one-letter-code of the 20 amino–acids is implemented in PLACER version M alternatively to the binary code using the probability function in (5).

Due to the fact that the score matrices are based on simple probability estimates, some sequence-structure alignments have identical scores and therefore a set of possible optimal alignments have to be considered. In this application, we present the results of the program PLACER that choose the optimal alignment of a segment placed leftmost on the primary sequence. In the Appendix, examples of the calculated secondary structure according to the optimal sequence-structure alignments are included. The helices are denoted by 'H', the beta sheets are denoted by 'S' and the gaps are denoted by '-'.

5.1 Comparison of sequence-structure alignment methods

We compare the results of the PLACER versions by analysing the sequence-structure alignments calculated from sequence data of the five domains and *Ubiquitin* and their corresponding observed cores.

Figure 1: Frequencies of correctly aligned core elements referring to the $P_i(3)$-kinase sequence.

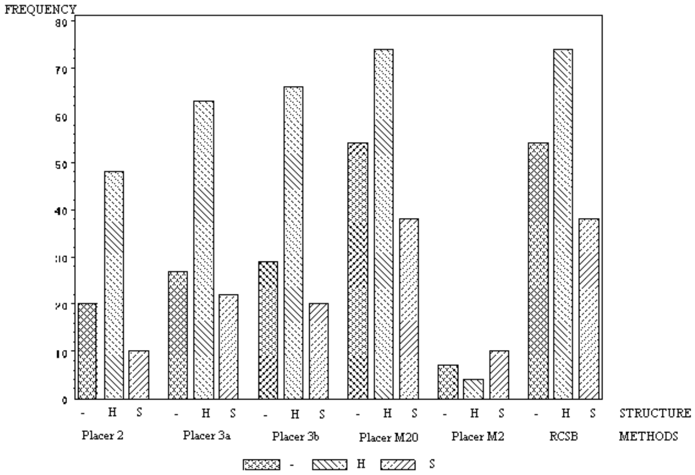


Figure 2: Frequencies of correctly aligned core elements referring to the Raf sequence.

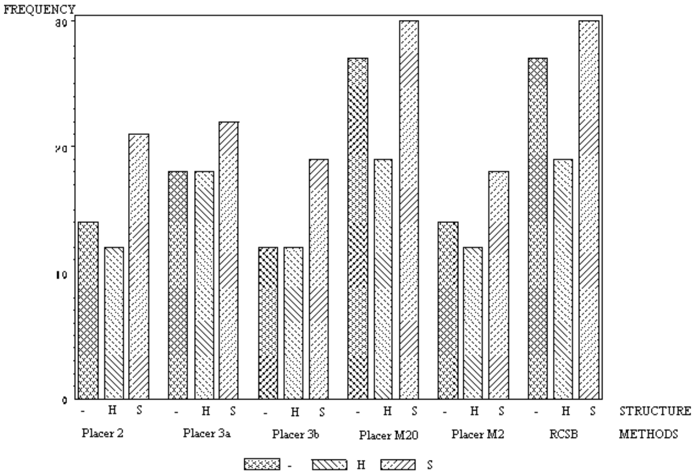


Figure 3: Frequencies of correctly aligned core elements referring to the *Ral* sequence.

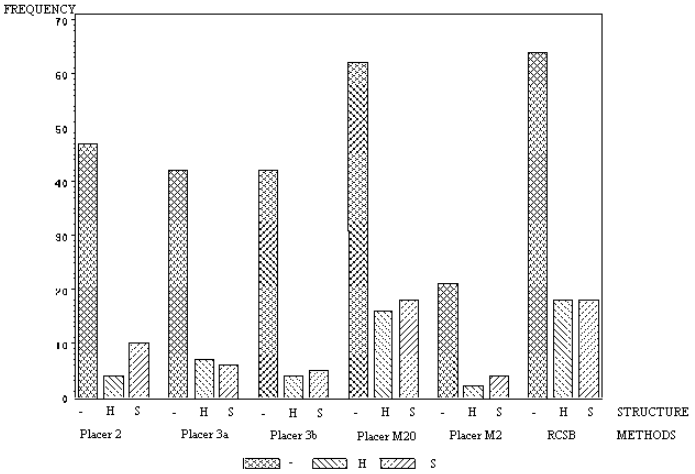


Figure 4: Frequencies of correctly aligned core elements referring to the *Rgl* sequence.

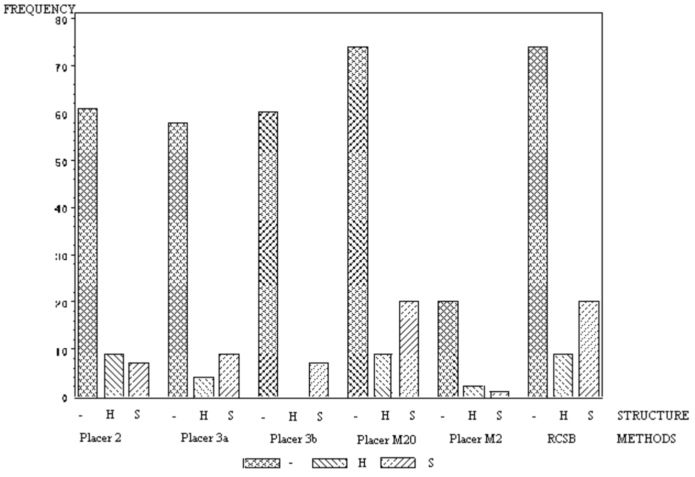


Figure 5: Frequencies of correctly aligned core elements referring to the *Rlf* sequence.

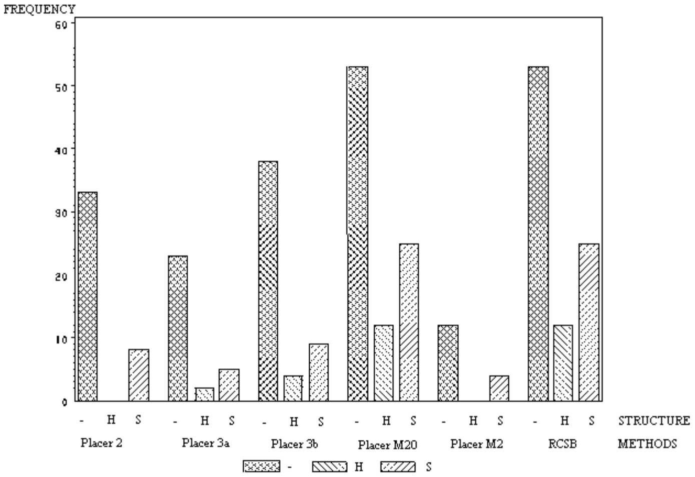
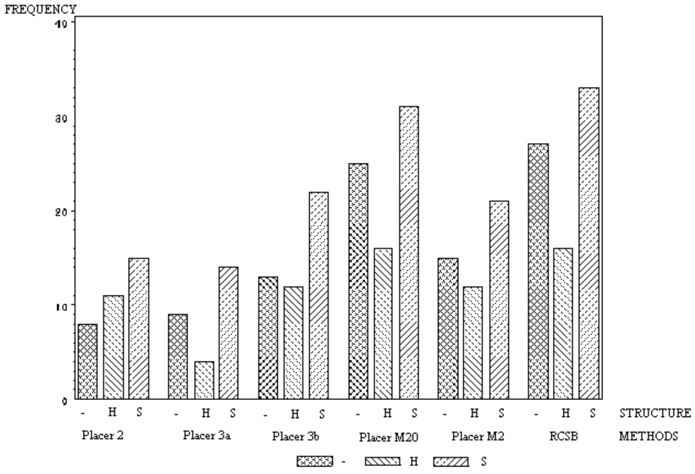


Figure 6: Frequencies of correctly aligned core elements referring to the *Ubiquitin* sequence.



In Figure (1)-(6) the frequencies of correct helix-, β-sheet- and gap-elements are shown. Additionally, in these Figures the reference frequencies of helix-, β-sheet- and gap-elements with respect to the RCSB-PDB are shown. Only the PLACER version M20 yield an alignment with the correct assignments of secondary structure elements (core elements). Analogously, we calculated the optimal sequence-structure alignments from the results of the secondary structure prediction method DSSP (Kabsch and Sander, 1983). Again, PLACER version M20 yielded the best results (Figures not shown).

5.2 Empirical prediction of the secondary structure

In this Section, the empirical prediction of the secondary structure using the results of optimal sequence-structure alignments will be presented. The sequence of the protein $Byr2$ (1i35) is aligned with different versions of PLACER to one of the five core model from Section 5.1. With respect to the differences in the sequence lengths of $P_i(3)$-kinase and $Byr2$ we divided the core of the $P_i(3)$-kinase in two halves of 5 and 6 segments. Therefore the relevant case that the core is not known and has to be predicted from related cores is examined now. With the set of optimal sequence-structure alignments the empirical distribution at each residue position for the three structure types (helix, β-sheet and gap) can be calculated. Aiming at the empirical prediction of the secondary structure, the maximal probable core element (or the structure type) can be selected. In comparison to the known secondary structures of the RCSB-PDB data bank a validation of this prediction method can be done.

In Figure (7) it is shown that the sequence–structure alignments referring to the core of $P_i(3)$-kinase yield the best prediction of the secondary structure of $Byr2$. More than 50% of the core elements are correctly aligned. The worst result yield the prediction on the basis of the *Ubiquitin* core model. Comparing the scoring methods according to the different PLACER versions, Figure (8) indicates the pooling of all Placer results as the best prediction method. Moreover, the usage of the sequence– structure alignments results according to the binary code versions 2 and M2 yields less correct aligned core elements than the usage of the one–letter–code versions. Finally, the comparison of the versions based on the probability functions leads to the conclusion that the consideration of stochastic dependencies (denoted by p_2 in

Figure 7: Frequencies of correctly aligned core elements referring to the $Byr2$ sequence and methods.

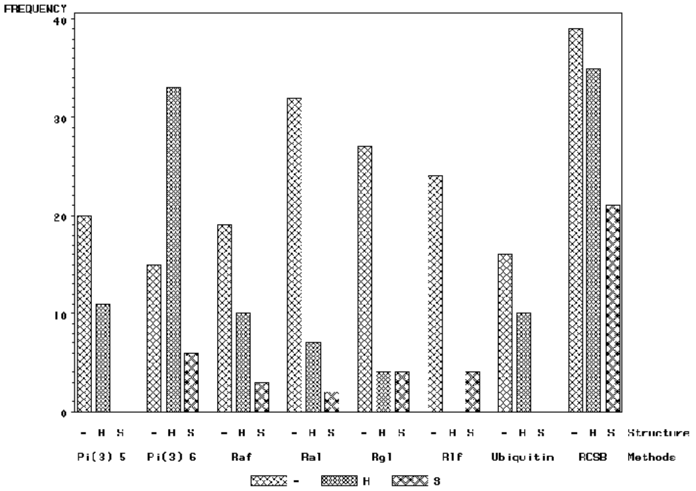


Figure 8: Frequencies of correctly aligned core elements referring to the $Byr2$ sequence and core models.

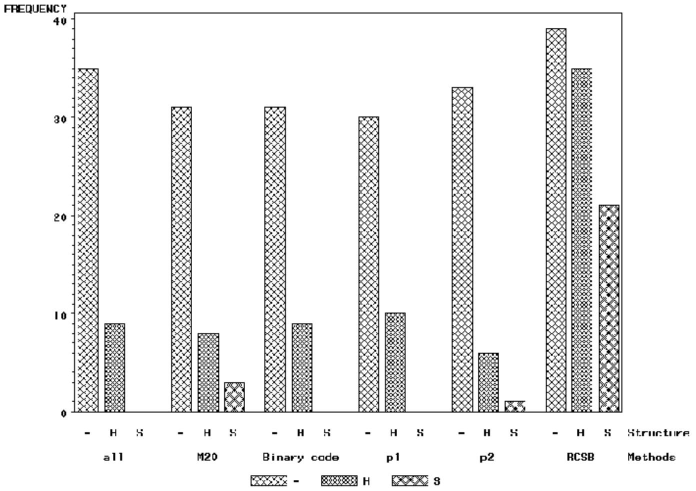


Figure (8)) is more appropriate for the prediction.

6 Discussion and outlook

The computations of the score matrices described above require the multiplication of many potentially small real numbers. This leads to numerical problems on computers working with a fixed precision. Adding the logarithms of the factors often helps to achieve greater numerical stability. This can, however, lead to different results of the dynamic programming algorithm because there are $a, b, c, d > 0$ such that $a + b < c + d$ and $\log(a) + \log(b) > \log(c) + \log(d)$.

Another way to overcome the problem is to use implementations of real numbers which allow arbitrary precision. These are present in most modern programming languages (e.g., the class `BigDecimal` in Java) or are available as libraries.

Arbitrary precision arithmetic can become quite slow. Thus one might want to represent the the model parameters as rationals, i.e., two integers. This is possible because they are derived from empirical frequencies. When computing the score matrix one first tries to find cancellations before multiplying. Nevertheless one might be forced to use integer arithmetic with arbitrary precision which is also supported by many modern programming languages but slows down the computations.

In this paper, four sequence-structure alignment methods were proposed and applied by using a dynamic programming approach. In the application to a small protein data set of *Ubiquitin*-like folds proteins we were able to choose the best alignment method among the proposed. Furthermore, this application showed that the consideration of gap penalties improves the sequence-structure alignment results. Nevertheless, the usage of the one-letter-code of the 20 amino-acids without gap consideration yielded the best alignment method. But the validation of the empirical prediction method using the optimal sequence-structure alignments of the PLACER version M showed that only almost the half of the $Byr2$ secondary structure can be correctly predicted. Therefore, this prediction methods failed in the commonly used case of unknown core structure but known related core structures. The consideration of multivariate sequence data that contain the information of more than one amino-acid alphabet may be an improvement of the sequence-structure alignment and the empirical prediction method.

Acknowledgement

We would like to thank Dr. Ingrid Vetter from the Max-Planck-Institute for Molecular Physiology in Dortmund for the helpful discussions and comments. This report has been supported by the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity for multivariate data structures").

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.*

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998), Biological sequence analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press.

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

Kanehisa, M. (2000), Post-Genome Informatics. Oxford University Press.

Karlin, S. and Ghandour, G. (1985), "Multiple-alphabet amino acid sequence comparisons of the immunoglobulin -chain constant domain", *Proc. Natl. Acad. Sci USA, 82, 8597–8601.*

Lathrop, R. H., and Smith, T.F. (1996), "Global optimum protein threading with gapped alignment and empirical pair score function", *Journal of Molecular Biology, 255, 641–665.*

Madej, T., Gibrat, J.F. and Bryant, S.H. (1995), "Threading a database of protein cores", *Proteins: Structure, Function and Genetics, 23, 256–369.*

Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C. (1995), "SCOP: A Structural Classification of Protein Databases for the

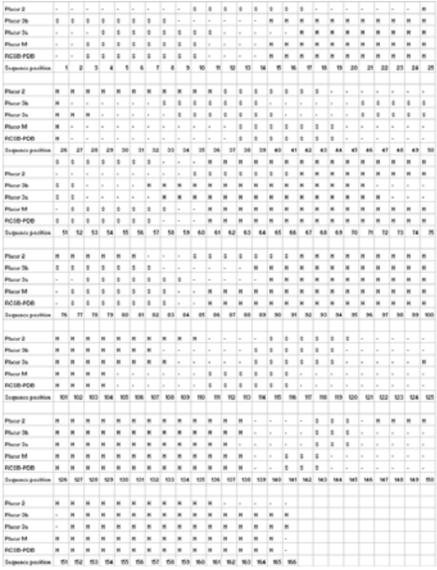
Investigation of Sequences and Structures", *Journal of Molecular Biology, 247(4), 536–540.*

Thiele, R., Zimmer, R. and Lengauer, T. (1999), "Protein threading by recursive dynamic programming", *Journal of Molecular Biology, 290, 757–779.*

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

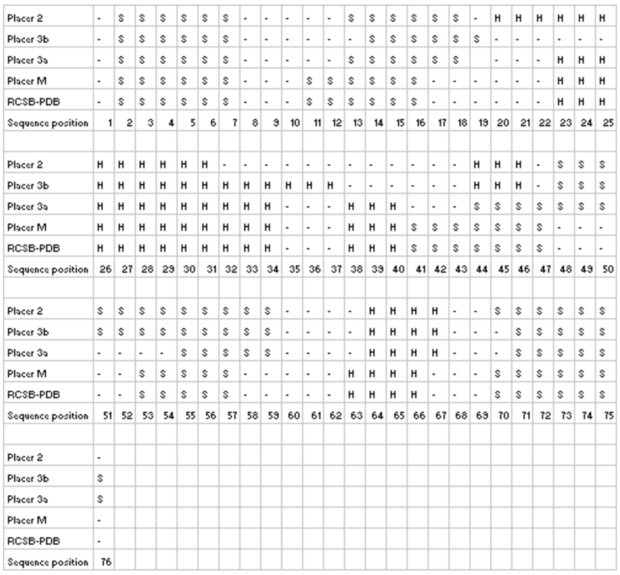
Appendix

Figure 9: Optimal sequence-structure alignments of the $P_i(3)$-kinase sequence.



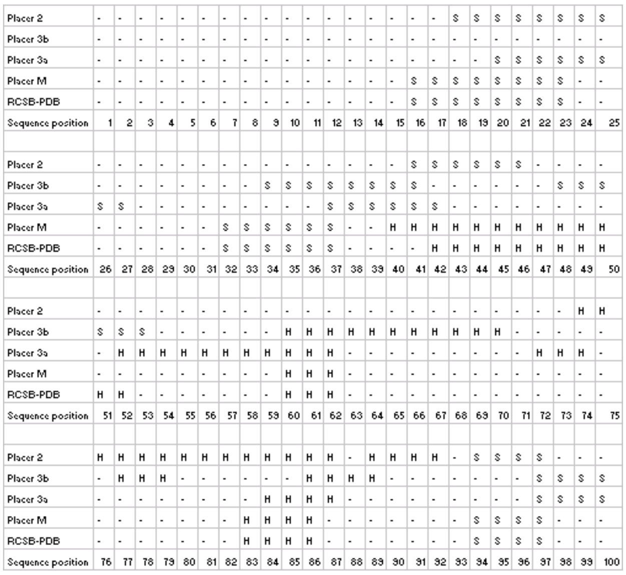
Phscor 2	-	-	-	-	-	-	-	-	-	S	S	S	S	S	S	S	S	-	-	-	-	-	-	-	H
Phscor 3b	S	S	S	S	S	S	S	S	-	-	-	-	-	-	H	H	H	H	H	H	H	H	H	H	H
Phscor 3a	-	-	-	S	S	S	S	S	S	S	S	-	-	-	-	-	H	H	H	H	H	H	H	H	H
Phscor M	-	-	S	S	S	S	S	S	S	S	-	-	-	-	H	H	H	H	H	H	H	H	H	H	H
RCSB-PDB	-	-	S	S	S	S	S	S	S	S	-	-	-	-	H	H	H	H	H	H	H	H	H	H	H
Sequence position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Phscor 2	H	H	H	H	H	H	H	H	H	H	H	S	S	S	S	S	S	S	-	-	-	-	-	-	-
Phscor 3b	H	-	-	-	-	-	-	S	S	S	S	S	S	S	-	-	-	-	-	-	S	S	S	S	S
Phscor 3a	H	H	H	-	-	-	-	-	S	S	S	S	S	S	S	-	-	-	-	-	S	S	S	S	S
Phscor M	H	-	-	-	-	-	-	-	-	-	-	-	S	S	S	S	S	S	S	-	-	-	-	-	-
RCSB-PDB	H	-	-	-	-	-	-	-	-	-	-	-	S	S	S	S	S	S	S	-	-	-	-	-	-
Sequence position	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
	S	S	S	S	S	S	S	-	-	-	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
Phscor 2	-	-	-	-	-	-	-	-	-	S	S	S	S	S	S	S	H	H	H	H	H	H	H	H	H
Phscor 3b	S	S	-	-	-	-	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	-	-	-	-
Phscor 3a	S	S	-	-	-	-	-	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	-	-	-
Phscor M	-	S	S	S	S	S	S	S	-	-	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
RCSB-PDB	S	S	S	S	S	S	S	-	-	-	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
Sequence position	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
Phscor 2	H	H	H	H	H	H	-	-	-	S	S	S	S	S	S	S	H	H	H	H	H	H	H	H	H
Phscor 3b	S	S	S	S	S	S	S	-	-	-	-	-	-	H	H	H	H	H	H	H	H	H	H	H	H
Phscor 3a	-	-	S	S	S	S	S	S	S	-	-	-	-	-	H	H	H	H	H	H	H	H	H	H	H
Phscor M	-	S	S	S	S	S	S	S	-	-	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
RCSB-PDB	-	S	S	S	S	S	S	S	-	-	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
Sequence position	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
Phscor 2	H	H	H	H	H	H	H	H	H	H	-	-	-	-	S	S	S	S	S	S	-	-	-	-	-
Phscor 3b	H	H	H	H	H	H	H	-	-	-	-	-	-	S	S	S	S	S	S	-	-	-	-	-	-
Phscor 3a	H	H	H	H	H	H	H	H	-	-	-	-	-	S	S	S	S	S	S	-	-	-	-	-	H
Phscor M	H	H	H	H	-	-	-	-	-	-	S	S	S	S	S	S	-	-	-	-	-	-	-	-	-
RCSB-PDB	H	H	H	H	-	-	-	-	-	-	S	S	S	S	S	S	-	-	-	-	-	-	-	-	-
Sequence position	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125
Phscor 2	H	H	H	H	H	H	H	H	H	H	H	H	H	-	-	-	-	S	S	S	-	H	H	H	H
Phscor 3b	H	H	H	H	H	H	H	H	H	H	H	H	H	-	-	-	-	S	S	S	-	-	-	-	-
Phscor 3a	H	H	H	H	H	H	H	H	H	H	H	H	-	-	-	-	-	S	S	S	-	-	-	-	-
Phscor M	H	H	H	H	H	H	H	H	H	H	H	H	H	-	-	S	S	S	-	-	-	-	-	-	-
RCSB-PDB	H	H	H	H	H	H	H	H	H	H	H	H	H	-	-	S	S	S	-	-	-	-	-	-	-
Sequence position	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150
Phscor 2	H	H	H	H	H	H	H	H	H	H	H	-	-	-	-	-									
Phscor 3b	-	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H									
Phscor 3a	-	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H									
Phscor M	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	-									
RCSB-PDB	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	-									
Sequence position	151	152	153	154	155	156	157	158	159	160	161	162	163	164	165	166									

Figure 10: Optimal sequence-structure alignments of the Raf sequence.



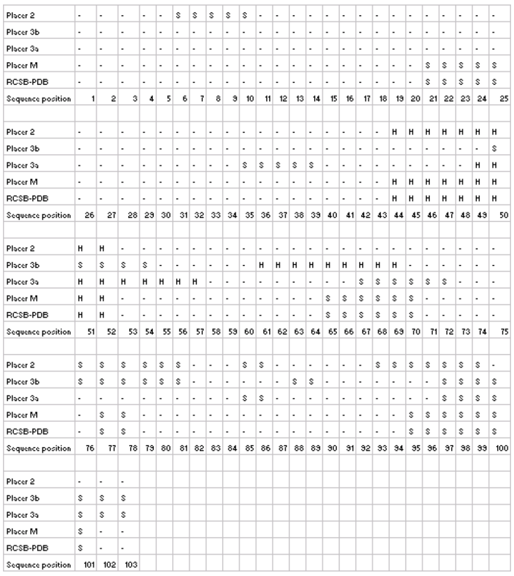
Placer 2	-	S	S	S	S	S	S	-	-	-	-	-	S	S	S	S	S	S	-	H	H	H	H	H	H
Placer 3b	-	S	S	S	S	S	S	-	-	-	-	-	-	S	S	S	S	S	S	-	-	-	-	-	-
Placer 3a	-	S	S	S	S	S	S	-	-	-	-	-	S	S	S	S	S	S		-	-	-	H	H	H
Placer M	-	S	S	S	S	S	S	-	-	-	S	S	S	S	S	S	-	-	-	-	-	-	H	H	H
RCSB-PDB	-	S	S	S	S	S	S	-	-	-	S	S	S	S	S	S	-	-	-	-	-	-	H	H	H
Sequence position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Placer 2	H	H	H	H	H	H	-	-	-	-	-	-	-	-	-	-	-	-	H	H	H	-	S	S	S
Placer 3b	H	H	H	H	H	H	H	H	H	H	H	H	-	-	-	-	-	-	H	H	H	-	S	S	S
Placer 3a	H	H	H	H	H	H	H	H	H	-	-	-	H	H	H	-	-	-	S	S	S	S	S	S	S
Placer M	H	H	H	H	H	H	H	H	H	-	-	-	H	H	H	S	S	S	S	S	S	S	-	-	-
RCSB-PDB	H	H	H	H	H	H	H	H	H	-	-	-	H	H	H	S	S	S	S	S	S	S	-	-	-
Sequence position	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
Placer 2	S	S	S	S	S	S	S	S	S	-	-	-	-	H	H	H	H	-	-	S	S	S	S	S	S
Placer 3b	S	S	S	S	S	S	S	S	S	-	-	-	-	H	H	H	H	-	-	-	S	S	S	S	S
Placer 3a	-	-	-	-	S	S	S	S	S	-	-	-	-	H	H	H	H	-	-	-	S	S	S	S	S
Placer M	-	-	S	S	S	S	S	-	-	-	-	-	H	H	H	H	-	-	-	S	S	S	S	S	S
RCSB-PDB	-	-	S	S	S	S	S	-	-	-	-	-	H	H	H	H	-	-	-	S	S	S	S	S	S
Sequence position	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
Placer 2	-																								
Placer 3b	S																								
Placer 3a	S																								
Placer M	-																								
RCSB-PDB	-																								
Sequence position	76																								

Figure 11: Optimal sequence-structure alignments of the Ral sequence.



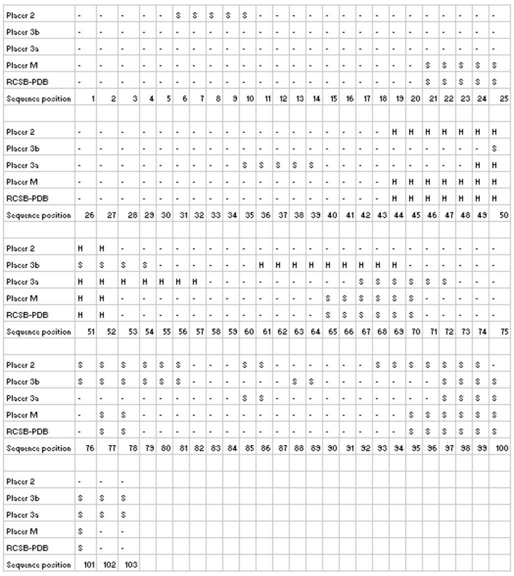
Placer 2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	S	S	S	S	S	S	S
Placer 3b	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Placer 3a	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	S	S	S	S	S
Placer M	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	S	S	S	S	S	S	S	-	-
RCSB-PDB	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	S	S	S	S	S	S	S	-	-
Sequence position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Placer 2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	S	S	S	S	S	-	-	-	-
Placer 3b	-	-	-	-	-	-	-	-	S	S	S	S	S	S	S	S	-	-	-	-	-	-	S	S	S
Placer 3a	S	S	-	-	-	-	-	-	-	-	-	S	S	S	S	S	S	-	-	-	-	-	-	-	-
Placer M	-	-	-	-	-	-	S	S	S	S	S	S	-	-	H	H	H	H	H	H	H	H	H	H	H
RCSB-PDB	-	-	-	-	-	-	S	S	S	S	S	S	-	-	-	-	H	H	H	H	H	H	H	H	H
Sequence position	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
Placer 2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	H	H
Placer 3b	S	S	S	-	-	-	-	-	-	H	H	H	H	H	H	H	H	H	H	H	-	-	-	-	-
Placer 3a	-	H	H	H	H	H	H	H	H	H	H	H	-	-	-	-	-	-	-	-	-	H	H	H	-
Placer M	-	-	-	-	-	-	-	-	-	H	H	H	-	-	-	-	-	-	-	-	-	-	-	-	-
RCSB-PDB	H	H	-	-	-	-	-	-	-	H	H	H	-	-	-	-	-	-	-	-	-	-	-	-	-
Sequence position	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
Placer 2	H	H	H	H	H	H	H	H	H	H	H	H	-	H	H	H	H	-	S	S	S	S	-	-	-
Placer 3b	-	H	H	H	-	-	-	-	-	-	H	H	H	H	-	-	-	-	-	-	-	S	S	S	S
Placer 3a	-	-	-	-	-	-	-	-	H	H	H	H	-	-	-	-	-	-	-	-	-	S	S	S	S
Placer M	-	-	-	-	-	-	-	H	H	H	H	-	-	-	-	-	-	-	S	S	S	S	-	-	-
RCSB-PDB	-	-	-	-	-	-	-	H	H	H	H	-	-	-	-	-	-	-	S	S	S	S	-	-	-
Sequence position	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100

Figure 12: Optimal sequence-structure alignments of the Rgl-kinase sequence.



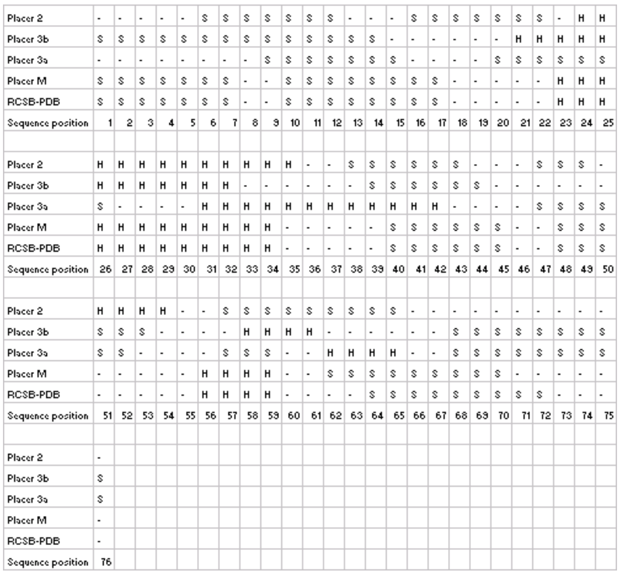
Placer 2	-	-	-	-	-	S	S	S	S	S	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Placer 3b	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Placer 3a	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Placer M	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	S	S	S	S
RCSB-PDB	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	S	S	S	S
Sequence position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Placer 2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	H	H	H	H	H	H	H
Placer 3b	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S
Placer 3a	-	-	-	-	-	-	-	-	-	S	S	S	S	S	-	-	-	-	-	-	-	-	-	H	H
Placer M	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	H	H	H	H	H	H	H
RCSB-PDB	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	H	H	H	H	H	H	H
Sequence position	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
Placer 2	H	H	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Placer 3b	S	S	S	S	-	-	-	-	-	-	H	H	H	H	H	H	H	H	H	-	-	-	-	-	-
Placer 3a	H	H	H	H	H	H	H	-	-	-	-	-	-	-	-	-	S	S	S	S	S	S	-	-	-
Placer M	H	H	-	-	-	-	-	-	-	-	-	-	-	-	S	S	S	S	S	S	-	-	-	-	-
RCSB-PDB	H	H	-	-	-	-	-	-	-	-	-	-	-	-	S	S	S	S	S	S	-	-	-	-	-
Sequence position	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
Placer 2	S	S	S	S	S	S	-	-	-	S	S	-	-	-	-	-	-	S	S	S	S	S	S	S	-
Placer 3b	S	S	S	S	S	S	-	-	-	-	-	-	S	S	-	-	-	-	-	-	-	S	S	S	S
Placer 3a	-	-	-	-	-	-	-	-	-	S	S	-	-	-	-	-	-	-	-	-	-	S	S	S	S
Placer M	-	S	S	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	S	S	S	S	S
RCSB-PDB	-	S	S	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	S	S	S	S	S
Sequence position	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
Placer 2	-	-	-																						
Placer 3b	S	S	S																						
Placer 3a	S	S	S																						
Placer M	S	-	-																						
RCSB-PDB	S	-	-																						
Sequence position	101	102	103																						

Figure 13: Optimal sequence-structure alignments of the Rlf sequence.



Placer 2	-	-	-	-	-	S	S	S	S	S	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Placer 3b	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Placer 3a	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Placer M	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	S	S	S	S
RCSB-PDB	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	S	S	S	S
Sequence position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Placer 2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	H	H	H	H	H	H	H
Placer 3b	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S
Placer 3a	-	-	-	-	-	-	-	-	-	S	S	S	S	S	-	-	-	-	-	-	-	-	-	H	H
Placer M	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	H	H	H	H	H	H	H
RCSB-PDB	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	H	H	H	H	H	H	H
Sequence position	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
Placer 2	H	H	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Placer 3b	S	S	S	S	-	-	-	-	-	-	H	H	H	H	H	H	H	H	H	-	-	-	-	-	-
Placer 3a	H	H	H	H	H	H	H	-	-	-	-	-	-	-	-	-	S	S	S	S	S	S	-	-	-
Placer M	H	H	-	-	-	-	-	-	-	-	-	-	-	-	S	S	S	S	S	S	-	-	-	-	-
RCSB-PDB	H	H	-	-	-	-	-	-	-	-	-	-	-	-	S	S	S	S	S	S	-	-	-	-	-
Sequence position	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
Placer 2	S	S	S	S	S	S	-	-	-	S	S	-	-	-	-	-	-	S	S	S	S	S	S	S	-
Placer 3b	S	S	S	S	S	S	-	-	-	-	-	-	S	S	-	-	-	-	-	-	-	S	S	S	S
Placer 3a	-	-	-	-	-	-	-	-	-	S	S	-	-	-	-	-	-	-	-	-	-	S	S	S	S
Placer M	-	S	S	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	S	S	S	S	S
RCSB-PDB	-	S	S	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	S	S	S	S	S
Sequence position	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
Placer 2	-	-	-																						
Placer 3b	S	S	S																						
Placer 3a	S	S	S																						
Placer M	S	-	-																						
RCSB-PDB	S	-	-																						
Sequence position	101	102	103																						

Figure 14: Optimal sequence-structure alignments of the Ubiquitin sequence.



Placer 2	-	-	-	-	-	S	S	S	S	S	S	S	-	-	-	S	S	S	S	S	S	S	-	H	H
Placer 3b	S	S	S	S	S	S	S	S	S	S	S	S	S	S	-	-	-	-	-	-	H	H	H	H	H
Placer 3a	-	-	-	-	-	-	-	-	S	S	S	S	S	S	S	-	-	-	-	S	S	S	S	S	S
Placer M	S	S	S	S	S	S	S	-	-	S	S	S	S	S	S	S	S	-	-	-	-	-	H	H	H
RCSB-PDB	S	S	S	S	S	S	S	-	-	S	S	S	S	S	S	S	S	-	-	-	-	-	H	H	H
Sequence position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Placer 2	H	H	H	H	H	H	H	H	H	H	-	-	S	S	S	S	S	S	-	-	-	S	S	S	-
Placer 3b	H	H	H	H	H	H	H	-	-	-	-	-	-	S	S	S	S	S	S	-	-	-	-	-	-
Placer 3a	S	-	-	-	-	H	H	H	H	H	H	H	H	H	H	H	H	-	-	-	-	S	S	S	S
Placer M	H	H	H	H	H	H	H	H	H	-	-	-	-	-	S	S	S	S	S	S	-	-	S	S	S
RCSB-PDB	H	H	H	H	H	H	H	H	H	-	-	-	-	-	S	S	S	S	S	S	-	-	S	S	S
Sequence position	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
Placer 2	H	H	H	H	-	-	S	S	S	S	S	S	S	S	S	-	-	-	-	-	-	-	-	-	-
Placer 3b	S	S	S	-	-	-	-	H	H	H	H	-	-	-	-	-	-	S	S	S	S	S	S	S	S
Placer 3a	S	S	-	-	-	-	S	S	S	-	-	H	H	H	H	-	-	S	S	S	S	S	S	S	S
Placer M	-	-	-	-	-	H	H	H	H	-	-	S	S	S	S	S	S	S	S	S	-	-	-	-	-
RCSB-PDB	-	-	-	-	-	H	H	H	H	-	-	-	-	S	S	S	S	S	S	S	S	S	-	-	-
Sequence position	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
Placer 2	-																								
Placer 3b	S																								
Placer 3a	S																								
Placer M	-																								
RCSB-PDB	-																								
Sequence position	76																								