Statistical analysis of sequence-structure alignment scores

Marcus Brunnert [1], Ralf Thiele [2], Heinz-Theodor Mevissen [3] and Wolfgang Urfer [1]

[1]Department of Statistics, University of Dortmund, D-44221 Dortmund, Germany

[2]Bioinformatics, Bayer AG, D-42096 Wuppertal, Germany

[3]Fraunhofer-Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, D-53754 Sankt Augustin, Germany

Abstract

The structural analysis of proteins is fundamental to the analysis of protein functions. In this context, sequence-structure alignment methods are important among the different empirical methods. In order to assess the quality of sequence-structure alignments, a statistical method using a Bayesian approach proposed by Lathrop et al. (1998) will be presented. Finally, the results of a developed statistical analysis of scores of RDP(recursive dynamic programming)-sequence-structure alignments (Thiele et al., 1999) according to data of six proteins will be described.

Keywords: Sequence-structure-alignment, statistically significant scores, Combined Bayesian approach and RDP-alignment.

1 Introduction

Proteins are composed of amino-acid chains. The amino-acid chain can be described by using the one-letter code of amino acids (Kanehisa, 2000). The resulting amino-acid sequence can be aligned to a protein structure due to the chemical and physical characteristics of the amino acids. Furthermore, the sequence and structure similarities between evolutionary or functionally related proteins are used in sequence-structure-alignments. In Thiele et al. (1999) a method for the calculation of sequence-structure alignments on the basis of a recursive dynamic programming (RDP) approach is presented. By combining a Bayesian approach proposed by Lathrop et al. (1998), we statistically assess the optimal sequence-structure alignments of six proteins due to this RDP-method. In Section 2 we present the Bayesian approach. The statistical model for the calculation of confidence intervals according to scores of a sequence-structure alignment is described in Section 3. With respect to data of six proteins belonging to the Ubiquitin like-folded family (classification from the SCOP-databank, http://scop.mrc-lmb.cam.ac.uk/uk/scop/, Murzin et al., 1995), the results of the statistical analysis will be presented in Section 4. Finally, the results will be discussed.

2 Bayesian approach to sequence-structure alignments

A scoring function that gives values (scores) to hypothetical sequence-structure alignments enables the search for an optimal sequence-structure alignment. In this context, the optimality depends on the chemical and physical characteristics of amino acids that are used for the definition of a scoring function. For our purposes here, we assume a scoring function according to a commonly used sequence-structure alignment method.
Let us denote the scoring function of a sequence-structure alignment (alignment) by f. This function is mapping into the set of scores of all admissible alignments. Following the notation of Lathrop et al. (1998) the score of an alignment $\mathbf{t}$ depends on the sequence $\mathbf{a}$ of length n and the structure C. Assuming a score calculated from a common scoring function, the score can be transformed into a probability,

$$P(\mathbf{a} \mid C, n, \mathbf{t}) = \frac{f(\mathbf{a}, C, \mathbf{t})}{\sum_{\mathbf{b} \in A^n} f(\mathbf{b}, C, \mathbf{t})}. \tag{1}$$

In this context the scores are standardized referring to the set A^n of all possible sequences of length n. Then the Bayesian approach of Lathrop et al. (1998) can be applied to these probabilities in terms of a statistical analysis of alignments. An alignment can be selected by using the posterior probabilities of an alignment given the structure C and the sequence $\mathbf{a}$, i.e.,

$$P(\mathbf{t} \mid C, \mathbf{a}, n) = \frac{P(\mathbf{a} \mid C, \mathbf{t}, n) P(\mathbf{t} \mid C, n)}{P(\mathbf{a} \mid C, n)}, \tag{2}$$

where $P(\mathbf{t} \mid C, n)$ and $P(\mathbf{a} \mid C, n)$ are prior probabilities due to the distributions of an alignment $\mathbf{t}$ and an amino-acid sequence $\mathbf{a}$.
Besides the selection of probable alignments, the selection of an appropriate structure leads to a second statistical problem. In Lathrop et al. (1998) this statistical problem is also faced with a Bayesian approach. The main idea of this structure selection can be explained by the following ratio of means,

$$\frac{\mu_t}{\mu_{\mathbf{a}}}, \tag{3}$$

where the mean μ_t is defined as,

$$\mu_t = \sum_{x \in \mathcal{T}[C,n]} f(\mathbf{a}, C, \mathbf{t}) P(\mathbf{x} \mid C, n) \tag{4}$$

and the mean $\mu_{\mathbf{a}}$ is defined as,

$$\mu_{\mathbf{a}} = \sum_{\mathbf{b} \in A^n} f(\mathbf{b}, C, \mathbf{t}) P(\mathbf{b} \mid C, n). \tag{5}$$

This ratio enables the comparison of the mean of scores according to a fixed structure given all possible alignments of a sequence **a** belonging to the set $\mathcal{T}[C, n]$ and the mean of scores according to a fixed structure given all sequences of length n. Therefore, different structures can be compared by using the corresponding sequence permutations of a sequence **a** and corresponding alignments. In Lathrop et al. (1998) this ratio in (3) was deduced from a Bayesian approach. Both statistical problems are not solvable without the other. On the one hand the alignment problem implies an appropriate structure selection and on the other hand the structure problem implies the calculation of alignment scores. For that reason, besides a consecutive solving of these two problems the simultaneous problem solving is of interest. In Lathrop et al. (1998) a combined method is proposed in order to solve both problems. Again, a Bayesian approach was applied to calculate posterior probabilities of the pairs $\langle C, \mathbf{t} \rangle$,

$$\begin{aligned} P(C, \mathbf{t} \mid C, \mathbf{a}, n) &= P(\mathbf{t} \mid C, \mathbf{a}, n) P(C \mid \mathbf{a}, n) \\ &= \frac{P(\mathbf{a} \mid C, \mathbf{t}, n) P(\mathbf{t} \mid C, \mathbf{a}, n) P(C \mid n)}{P(\mathbf{a} \mid n)}. \end{aligned} \tag{6}$$

With appropriate prior probabilities, this simultaneous selection of $\langle C, \mathbf{t} \rangle$ enables statistical evaluations of core and alignment scores.

3 Statistical model

With respect to the distribution of the alignment scores, different assumptions for the statistical model can be made. Aiming at the combined Bayesian approach, statistical models for the alignment scores and the prior probabilities will be proposed next.
Regarding a sample of alignment scores $x_1, \ldots, x_m$, the following statistical model can be assumed.

1. The scores X_i, $i=1,\ldots,m$ are normally distributed.
2. The mean μ and the variance σ^2 of all scores are constant.
3. The normally distributed random variables X_i, $i=1,\ldots,m$ are stochastically independent.

Later on, these assumptions have to be examined by analysing the empirical distributions of the scores. Following these assumptions, the common probability distribution of a sample $x_1, \ldots, x_m$ is denoted by

$$p(x_1, \ldots, x_m \mid \mu, \sigma^2) = \prod_{i=1}^{m} \frac{1}{\sigma\sqrt{2\pi}} exp[-\frac{1}{2\sigma^2}(x_i - \mu)^2]. \tag{7}$$

Then the likelihood of the parameters given the scores can be deduced. The deduction of the standardized likelihood function is shown in Mood et al. (1974). Considering the Bayesian approach in (6), the likelihood function is

$$p(\mathbf{t} \mid C, \mathbf{a}, n) = l(\mu \mid \bar{x}, \sigma^2) = \frac{1}{\sqrt{2\pi\frac{\sigma^2}{n}}} exp[-\frac{1}{2}(\sqrt{n}\frac{\mu - \bar{x}}{\sigma})^2]. \tag{8}$$

Critical to this likelihood function is the assumed known variance $\sigma^2 \ / \ n$. In the case of an unknown variance, another likelihood function (not presented here) has to be considered (see for example Mood et al., 1974).
Furthermore, in order to enable the combined Bayesian approach, prior distributions have to be assumed. In this context, simple uninformative uniform distributions are proposed by Lathrop et al. (1998). Therefore, the following ratio can be set to a constant,

$$\frac{P(\mathbf{t} \mid C, \mathbf{a}, n)P(C \mid n)}{P(\mathbf{a} \mid n)} = const. \tag{9}$$

This constant completes the Bayesian approach in (6).
Summing up, the posterior distribution is approximately the product of this constant in (9) and the likelihood function in (8). According to this normally distributed posterior, confidence intervals for the mean of the distribution of $\langle C, \mathbf{t} \rangle$ can be calculated. Finally, the statistical significance on the basis of these confidence intervals can be assessed for every pair $\langle C, \mathbf{t} \rangle$.

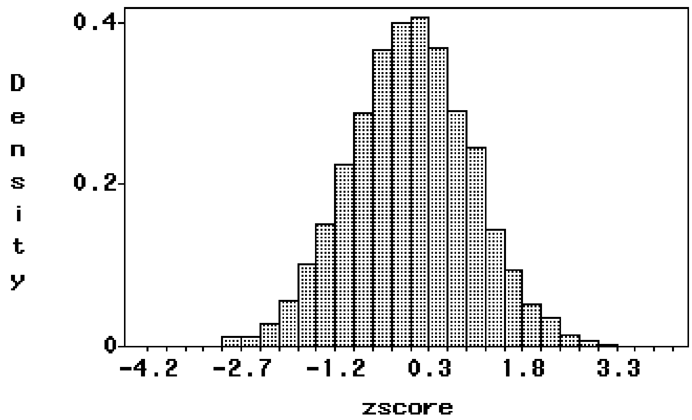
4 Application

In this application, the described Bayesian approach is applied to RDP-alignment scores (Thiele et al. (1999)). In contrast to the RDP sequence-structure approach, the Bayesian approach in Lathrop et al. (1998) evaluates scores of ungapped sequence-structure alignments based on so-called *cores*. For that reason, the assumption of constant mean is not fulfilled here. But the comparison of the results due to length dependent means and constant means yielded the same confidence intervals. Therefore, we only consider a standardization of the scores in order to fit a normal distribution.
The five data sets containing scores of alignments to the structures of the Ras-binding domains of Raf, Ral, Rgl, Rlf and $P_i(3)$-kinase were generated with the ToPLign-system (http://cartan.gmd.de/ToPLign.html, GMD, Gesellschaft für Mathematik und Datenverarbeitung). Each data set contained the scores of RDP-alignments of 1000 permutated sequences of the protein Ubiquitin aligned to one of these five structures. The scoring function was composed of so called Voronoi-tesselations (Zimmer et al., 1998).
With these five data sets (see Table 1), a descriptive analysis was carried out. The assumption of normally distributed scores was examined. This analysis gave no evidence for normally distributed scores. With respect to this result,

Table 1: Descriptive statistics and RDP-scores.

Alignment	Arith. mean	Median	Min.	Max.	Stand. dev.	Score
Ubiquitin-PI(3)	-378.651	-379.390	-456.080	-273.820	26.787	88.38
Ubiquitin-Raf	-303.970	-306.250	-396.600	-170.090	32.951	-88.92
Ubiquitin-Ral	-71.780	-69.805	-159.660	-22.410	21.004	-114.58
Ubiquitin-Rgl	-325.002	-325.235	-454.420	-207.270	33.129	-83.35
Ubiquitn-Rlf	-310.624	-310.460	-408.850	-212.830	32.878	-55.80

Figure 1: Histogram of the standardized and pooled data set.



the pooled data set was standardized due to the mean and the standard deviation of each single data set with respect to the different length of alignments.

The pooled data set of 5000 standardized scores yielded a p-value of 0.15 according to the Kolmogorov-Smirnov test (calculated in SAS).

Moreover, the histogram in Figure 1 confirms the normal distribution of the standardized scores (zscores).

Table 2: Confidence interval for the expectation of standardized scores.

Lower confidence limit	Upper confidence limit
-0.023	0.023

The 90% confidence interval for the expectation of standardized scores was calculated for the mean of the posterior distribution (see Table 2). The interval did not contain any standardized RDP score (see Table 3) of the Ubiquitin sequence aligned to each structure.

Table 3: Standardized RDP-scores.

Alignment	Standardized RDP-Score
Ubiquitin-PI(3)	17.4252
Ubiquitin-Raf	6.5264
Ubiquitn-Ral	-2.0377
Ubiquitin-Rgl	7.2943
Ubiquitn-Rlf	7.7506

5 Discussion and outlook

The Bayesian approach to sequence-structure analysis proposed in Lathrop et al. (1998) was applied to RDP-alignment scores. Besides the assessment of alignment quality by RMS (root mean square deviation), this Bayesian approach enables a statistical assessment of alignment scores. Considering the pooled data set of all alignment scores, the optimal alignment scores are statistically significant according to the calculated confidence interval. A normal distribution was used for fitting the score distribution. The descriptive analysis confirmed these assumed distributions.

The statistical models of this Bayesian approach used uninformative priors. This statistical validation method of RDP-alignments can be improved by using prior distributions due to the scores of suboptimal RDP-alignments. With these scores, the alignment distribution can be estimated in order to carry out the described Bayesian structure selection. Finally, an alternative posterior distribution of the structure and alignment pairs can be achieved by this estimated alignment distribution. The selection of an optimal structure and alignment pair can also be carried out by using this alternative posterior distribution.

Acknowledgement

We would like to thank the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") for financial support.

References

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

Lathrop, R.H., Rogers, R.G., Smith, T.F. and White, J.V. (1998), "A Bayes-optimal sequence-structure theory that unifies protein sequence-structure recognition and alignment", *Bulletin of Mathematical Biology*, 60, 1039–1071.

Mood, A.M., Graybill, F.A. and Boes, D.C. (1974), "Intoduction to the theory of statistics", Tokio.

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.

SAS/STAT-User's Guide, Version 6, Fourth Edition, SAS-Institute Inc.

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

Zimmer, R., Wöhler, M. and Thiele, R. (1998), "New scoring schemes for protein fold recognition based on Voronoi contacts", *Bioinformatics*, 14(3), 295–308.