TURUN YLIOPISTON JULKAISUJA
ANNALES UNIVERSITATIS TURKUENSIS

SARJA - SER. AII OSA – TOM. 228

BIOLOGICA - GEOGRAPHICA - GEOLOGICA

# Statistical Methods for Conservation and Alignment Quality in Proteins

**by**

**Virpi Ahola**

TURUN YLIOPISTO
Turku 2008

From the Department of Statistics
University of Turku
Turku, Finland

**Supervised by**

Professor Esa Uusipaikka
Department of Statistics
University of Turku
Turku, Finland

Professor Mauno Vihinen
Institute of Medical Technology
University of Tampere
Tampere, Finland

**Reviewed by**

Professor Timo Koski
Department of Mathematics
Royal Institute of Technology
Stockholm, Sweden

Chief Research Scientist Jaakko Hollmén
Department of Information and Computer Science
Helsinki University of Technology
Espoo, Finland

**Opponent**

Professor Arne Elofsson
Stockholm Bioinformatics Center / Center for Biomembrane Research
Stockholm University
Stockholm, Sweden

ISBN 978-951-29-3725-7 (PRINT)
ISBN 978-951-29-3726-4 (PDF)
ISSN 0082-6979

Painosalama Oy - Turku Finland 2008

To my parents

# Abstract

Construction of multiple sequence alignments is a fundamental task in Bioinformatics. Multiple sequence alignments are used as a prerequisite in many Bioinformatics methods, and subsequently the quality of such methods can be critically dependent on the quality of the alignment. However, automatic construction of a multiple sequence alignment for a set of remotely related sequences does not always provide biologically relevant alignments. Therefore, there is a need for an objective approach for evaluating the quality of automatically aligned sequences.

The profile hidden Markov model is a powerful approach in comparative genomics. In the profile hidden Markov model, the symbol probabilities are estimated at each conserved alignment position. This can increase the dimension of parameter space and cause an overfitting problem.

These two research problems are both related to conservation. We have developed statistical measures for quantifying the conservation of multiple sequence alignments. Two types of methods are considered, those identifying conserved residues in an alignment position, and those calculating positional conservation scores. The positional conservation score was exploited in a statistical prediction model for assessing the quality of multiple sequence alignments. The residue conservation score was used as part of the emission probability estimation method proposed for profile hidden Markov models.

The results of the predicted alignment quality score highly correlated with the correct alignment quality scores, indicating that our method is reliable for assessing the quality of any multiple sequence alignment. The comparison of the emission probability estimation method with the maximum likelihood method showed that the number of estimated parameters in the model was dramatically decreased, while the same level of accuracy was maintained.

To conclude, we have shown that conservation can be successfully used in the statistical model for alignment quality assessment and in the estimation of emission probabilities in the profile hidden Markov models.

# List of original publications

This thesis is based on an introduction part and the following original publications, referred to in the text by their Roman numerals.

**Publication I:** Ahola V., Aittokallio T., Uusipaikka E. and Vihinen M. (2003) Efficient estimation of emission probabilities in profile hidden Markov models. *Bioinformatics* Vol. **19**, No. 18, p. 2359-2368.

**Publication II:** Ahola V., Aittokallio T., Uusipaikka E. and Vihinen M. (2004) Statistical methods for identifying conserved residues in multiple sequence alignment. *Statistical Applications in Genetics and Molecular Biology*, Vol. **3**, Iss. 1, article 28.

**Publication III:** Ahola V., Aittokallio T., Vihinen M. and Uusipaikka E. (2006) A statistical score for assessing the quality of multiple sequence alignments. *BMC Bioinformatics* Vol **7**, article 484.

**Publication IV:** Ahola V., Aittokallio T., Vihinen M. and Uusipaikka E. (2008) Model-based prediction of sequence alignment quality. *Bioinformatics* Vol. **24**, No. 19, p. 2165-2171.

# Abbreviations

3D - 3-dimensional

APDB - Analyze PDB

AQ - Alignment quality

BAliBASE - Benchmark alignment database

Blast - Basic local alignment search tool

Blosum - Blocks of amino acid substitution matrix

ConsAA - Proportion of conserved residues

CS - Column score

EEP - Efficient emission probability

ET - Evolutionary trace

$f_m$ - Modeler's viewpoint

$f_d$ - Developer's viewpoint

FDR - False discovery rate

FSSP - Families of structurally similar proteins

FWE - Family-wise error rate

GDT - Global distance test

HMM - Hidden Markov model

Homstrad - Homologous structure alignment database

IC - Information content

IS - Importance sampling

maxZ - Maximum Z

MD - Mean distance

ML - Maximum likelihood

MOS - Multiple overlap score

MSA - Multiple sequence alignment

MSE - Mean square error

NiRMSD - Normalized iRMSD

PAM - Point-accepted-mutation

PDB - Protein data bank

PPI - Protein-protein interaction

RMSD - Root mean square deviation

SABmark - Sequence alignment benchmark

SH2 - Src homology 2

SMART - Simple modular architecture research tool

SP - Sum of pairs

norMD - Normalized mean distance

TIM - Triosephosphate isomerase

TM - Template modeling

UI - Union-intersection

WOOF - Wood-oriented objective function

# Contents

# Part I

# Background and summary of thesis

# Chapter 1

# Introduction

## 1.1 Comparative genomics

During the last few years, the whole genome sequencing projects of human and several animals, plants and microbe genomes have produced large amounts of raw nucleotide sequence data (Lander *et al.*, 2001; Venter *et al.*, 2001). Similarly, three-dimensional structures have been solved for thousands of proteins every year. In order to understand how this genetic information leads to observable traits and behaviors, advanced bioinformatics methods are needed. Comparative genomics has become a very important research area in the molecular genomics after the whole genome sequencing projects (Saccone and Pesole, 2003). Comparative genomics compares genomes of different species or strains to infer how selection has acted on genomes. The major principle of comparative genomics is that the DNA responsible for common features of organisms has been conserved among the species (Hardison, 2003). Under the neutral theory of molecular evolution, in the absence of selective constraints, the mutation rate is high, while in the presence of functional or structural constraints, the mutation rate is low, imposed by purifying or negative selection (Jukes and Kimura, 1984). Comparative genomics uses this theory conversely: the degree of conservation is used to find functional and structural constraints that have acted on a particular site of the genome.

## 1.2 Multiple sequence alignments

The multiple sequence alignments (MSA) are a core of comparative genomics (Batzoglou, 2005). A sequence alignment maps the residues of one sequence onto the residues of the other sequences. Gaps are inserted between the residues, and thereby, residues with identical or similar characters are aligned in the same column. The alignment columns represent nucleotide or amino acid residues having evolved from the same position of the common ancestor, superposable structures or sequence motifs having a common function

(Edgar and Batzoglou, 2006).

MSAs help to identify regions of similarity and dissimilarity, which is essential in understanding structural, functional or evolutionary relationships of DNA or protein sequences among different organisms. MSAs are useful in annotation of sequences, gene finding and finding new family members of distantly related proteins by sequence database search. They are valuable in locating and visualizing conserved domains and sequence patterns within a protein family. The evolutionary relationships for a set of sequences can be studied by phylogenetic methods, which are usually based on MSA. Protein structural analyses, such as homology modeling and protein secondary and tertiary structure prediction, are based on sequence alignments.

It has been recognized that the automatic construction of MSAs for a set of remotely related sequences can be a very demanding task. Nowadays, most algorithms aligning protein sequences originate from the pioneering work of Needleman and Wunsch for finding global pairwise alignments Needleman and Wunsch (1970). The Needleman–Wunsch's method allows matches, mismatches, insertions and deletions to occur in the alignment. It applies the dynamic programming algorithm to build up the best global alignment by using optimal alignments of smaller subsequences. Smith and Waterman (1981) modified the Needleman–Wunsch's algorithm to find an optimal local pairwise alignment. Globally optimal pairwise sequence alignments can be solved in $O(L^2)$ time (Gotoh, 1982) and $O(L)$ space (Myers and Miller, 1988). Most MSA methods rely on a sum-of-pairs scoring function (Carrillo and Lipman, 1988). The sum-of-pairs score can be optimized for MSA by the dynamic programming with time and space complexity $O(L^N)$. Thus, an optimal solution can only be found for a very few sequences in exponential time (Carrillo and Lipman, 1988; Wang and Jiang, 1994). Several powerful alignment algorithms have, however, been developed for multiple sequences. Since many methods of comparative genomics can be critically dependent on the quality of a given alignment, there is a need for an objective approach to evaluate the quality of automatically aligned sequences. The procedures developed for assessing the quality of MSAs are described in more detail in section 2.2.

## 1.3 Conservation

Protein conservation can be quantified from a MSA of homologous sequences. A wide variety of methods has been developed for calculating the degree of conservation at an alignment position (Valdar, 2002). Different approaches are described in section 2.1.

Conservation measures can be used to study evolutionary sequence conservation in relation to structural and functional properties of a given protein. To be more specific, many measures have been developed for functional and structural annotation to predict functionally or structurally important sites of protein families, such as catalytic and lig-

and binding residues (Mirny and Shakhnovich, 1999; Magliery and Regan, 2005; Capra and Singh, 2007; Fischer *et al.*, 2008). Extensive efforts have also been made to apply conservation scores for predicting residues involved in protein-protein interaction (Valdar and Thornton, 2001; Caffrey *et al.*, 2004; Bordner and Abagyan, 2005). A very popular application has been to detect positions responsible for functional and structural differences between subgroups, i.e. functional specificity of proteins (Lichtarge *et al.*, 1996; Hannenhalli and Russell, 2000; del Sol Mesa *et al.*, 2003; Kalinina *et al.*, 2004a; Pei *et al.*, 2006; Marttinen *et al.*, 2006). Finally, conservation measures have been effectively used for refining automatically produced MSAs, by detecting misaligned regions or unreliable aligned sequences (Sadreyev and Grishin, 2004; Thompson *et al.*, 2001; Castresana, 2000), for visualization (Schneider and Stephens, 1990; Thompson *et al.*, 1997) and for assessing the quality of MSAs (Pei and Grishin, 2001; Thompson *et al.*, 2001).

## 1.4 Profile analysis

Traditional profile analysis and profile hidden Markov models (HMM) have been proved to be very powerful methods in comparative genomics. The essence of the traditional profile analysis is that the information about sequence or structural aligned probe sequences is incorporated into a position-specific scoring table, a profile, whereby the analysis is able to detect structural similarities and remote homologies to the sequence family (Gribskov *et al.*, 1987). The profile includes information about conservation of residues, changes allowed at each position and penalties for insertion or deletion.

The profile HMMs originate from the profile analysis (Krogh *et al.*, 1994; Eddy, 1998). The underlying idea and the objective of the profile HMMs is exactly the same as in the profile analysis, the difference being that the HMM is a well-formulated probability model. The conserved positions of the alignment are modeled by match states, while other positions are modeled by either insertion or deletion states. Match states emit a residue according to the estimated probability distribution, which corresponds to the substitution score matrix in the traditional profile analysis. The gap penalties for insertions and deletions, by which positions of conserved regions are controlled, are provided by transition probabilities from/to insert and delete states.

In the profile HMMs, emission probabilities of all 20 amino acids are estimated in all emitting states, and thus, the number of estimated parameters can be enormous. However, the majority of estimated emission parameters are actually 'noise', that is, probabilities of uninteresting or unconserved residues. The phenomenon is related to overfitting, which occurs when there are not enough data to obtain good estimates for the model parameters, and consequently, the model will not generalize adequately to new data

The traditional profile analysis can be used to build local sequence-profile or profile-profile alignments and assess the statistical significance of the alignments (Altschul *et al.*,

1997; Sadreyev and Grishin, 2003; Yona and Levitt, 2002). The profile HMMs are useful in detecting remote homologues to the protein family, for instance. The profile analysis and conservation scores are, hence, differently motivated. A more general difference to the conservation scores is that the profile analysis provides a (complicated) model for describing a whole alignment, whereas the conservation score provides a single, decisive positional statistic (Valdar, 2002).

## 1.5 Aims of the thesis

The aim of the thesis was to develop a both biologically and statistically relevant method for measuring conservation of multiple protein sequence alignments, and to use this measure for assessing the quality of alignments (Publications III and IV).

The secondary aim of the thesis was to develop methods for identifying conserved residues in an alignment position, and by means of these scores to solve the overfitting problem in estimating emission probabilities in profile HMMs (Publications I and II).

## 1.6 Scientific novelty of the thesis

The thesis contributes to quantifying the conservation and assessing the quality of MSAs. The thesis focuses on developing both biologically and statistically sound scores for measuring conservation on two different levels: identifying conserved residues in one alignment position and scoring the conservation of an alignment position. The practical examples show that the scores are able to identify structurally and/or functionally important residues and alignment positions. The improvement over the earlier developed methods is that the scores have a strong statistical background and the significance of the conservation score can be reliably estimated.

The positional conservation score has been extended to define the quality of the whole alignment. The whole alignment quality score has been used as a key factor in a model-based alignment quality method. The results suggest that the novel method for assessing the quality of alignments can confidently predict the quality of any MSAs. The improvement over the existing scores is that our measure can be used without reference alignment or other additional information. The residue level conservation score has been applied in the emission probability estimation method developed for the profile HMMs. The emission probability estimation method dramatically decreases the number of estimated parameters and thus, solves the overfitting problem in the estimation of emission probabilities in profile HMMs without losing accuracy. The new method provides an alternative approach to Bayesian methods for profile HMMs.

# Chapter 2

# Conservation and alignment quality

## 2.1 Scoring residue conservation

This section presents several approaches for quantifying positional conservation, discusses statistical inference on the conservation scores and the performance of the scoring methods in predicting functional and structural sites of proteins.

### 2.1.1 Positional conservation scores

Conservation scores are important for predicting functionally important sites in protein sequences. A good conservation score should give biologically relevant results and fulfill the following criteria. The score should be a simple mathematical mapping to continuous and bounded space, should take into account relative frequencies and stereochemical properties of amino acids, should penalize for gaps and weight sequences against redundancy (Valdar, 2002). Additional criteria require that the maximally unconserved position obtains a minimum score of zero, and that an invariant position always obtains a maximum score and does not depend on an invariant amino acid (Fischer *et al.*, 2008). This section introduces different measures used for quantifying positional conservation during the last 40 years.

The variability of amino acid positions was quantified already in 1970 (Wu and Kabat, 1970). Their measure divided the number of different amino acids by the relative frequency of the most common amino acid at a given position. The later scores have taken into account relative frequencies of all amino acids at a given position and background distribution (Pei and Grishin, 2001). Lockless and Ranganathan (1999), for instance, assumed that the lack of evolutionary constraint should cause the distribution of amino acids to approach the mean distribution of the same set of proteins. They defined conservation as a root mean square deviation (RMSD) of residue relative frequencies from their background probabilities.

Probably the most popular conservation scores have been based on the Shannon en-

tropy, which was originally introduced for information theory (Shannon, 1948). For nucleotide sequences, the Shannon entropy and relative entropy were first described by Schneider *et al.* (1986) and used by several authors thereafter, for instance, to discover regulatory sites in co-regulated genes (Stormo and Fields, 1998). The entropy-based sequence logo method has become popular for visualizing sequence alignments (Schneider and Stephens, 1990). For protein sequences, Shenkin *et al.* (1991) used the Shannon entropy,

$$S = -\sum_{j=1}^{J} p_j \log_2 p_j, \tag{2.1}$$

where $p_j$ is the relative frequency of amino acid $j$, as a measure of variability in alignment positions. In the same year, Sander and Schneider (1991) proposed a similar measure. These measures were called the information content (IC). These and many other modifications of the Shannon entropy, relative entropy or mutual information have become very popular in scoring positional conservation (Pei and Grishin, 2001; del Sol Mesa *et al.*, 2003; Pirovano *et al.*, 2006), although, as such, they do not account for different stereochemical properties of amino acids. Furthermore, they can only be used for ungapped alignments.

Another often used approach for scoring residue conservation is based on substitution matrices, such as PAM, Blosum or Gonnet (Dayhoff *et al.*, 1978; Henikoff and Henikoff, 1993; Benner *et al.*, 1994). These matrices are used to quantify the similarity of amino acids in an aligned position. An often used score is a sum-of-pairs score, which is the sum of pairwise alignment scores, where the alignment score can be any residue comparison matrix normalized so that the diagonal values are always one (Karlin and Brocchieri, 1996). The idea of the sum-of-pairs score was originally introduced by Carrillo and Lipman (1988) for the alignment problem, and has been subsequently used in many MSA programs and conservation scores (Pilpel and Lancet, 1999; Armon *et al.*, 2001; Pei and Grishin, 2001; Valdar and Thornton, 2001). The original sum-of-pairs score has usually been supplemented by gap opening and extension penalties for introducing insertions and deletions into alignment. The mean distance (MD) score, implemented into the ClustalX program (Thompson *et al.*, 1997), does not directly use a sum-of-pairs score, but is based on the concept of continuous sequence space (Vingron and Sibbald, 1993). The MD score calculates the mean distance of all residues from the consensus point (Thompson *et al.*, 2001). The normalized MD score (norMD) is comparable among different alignments, and also penalizes for gaps. Another group of scores accounting for amino acid propensities uses stereochemical groups of amino acids. These scores calculate the minimal set of physiochemical properties that represent any group of amino acids in an aligned column (Taylor, 1986; Zvelebil *et al.*, 1987; Livingstone and Barton, 1993).

Many authors have proposed scores which incorporate physicochemical properties of amino acids into the Shannon, relative or mutual entropy scores. Mirny and Shakhnovich

(1999, 2001) incorporated grouping of amino acids into the Shannon entropy. The amino acids were grouped according to their physicochemical properties into six classes. Earlier Williamson (1995) had used relative entropy for groups of amino acids. These scores combining entropy and amino acid groups, however, ignore relative frequencies within the partition (Valdar, 2002).

Later measures have also incorporated sequence similarity into the entropy-based scores. Caffrey *et al.* (2004) introduced the use of the von Neumann entropy (Lifshitz and Pitaevskii, 1980) to test the conservation of protein interfaces. The von Neumann entropy uses the density matrix, whose components are obtained by multiplying residue relative frequencies by their similarities. Jensen-Shannon divergence measures the deviation between the relative frequencies of amino acids and background distribution using relative entropy (Capra and Singh, 2007; Lin, 1991). Kalinina *et al.* (2004a,b) used a modified mutual information score to predict residues which determine the functional specificity of predefined groups. Their method, SDPpred, uses smoothed residue relative frequencies, in which amino acid substitutions have been accounted for. Other modifications use amino acid similarities (e.g. Blosum matrix) instead of background probabilities in relative entropy or relative joint entropy-based scores (Capra and Singh, 2007; Mihalek *et al.*, 2007).

The evolutionary trace (ET) method, originally introduced by Lichtarge *et al.* (1996), was a first attempt to account for the evolutionary history of a protein family in the conservation scores. The ET method was developed to identify active sites and functional protein interfaces when the structure of a protein is known. The ET method consists of the following steps Lichtarge *et al.* (1996):

1. Construct a phylogenetic tree from a MSA.
2. Assemble a consensus sequence for each branch of the tree.
3. Align consensus sequences.
4. Compute an evolutionary trace by assigning each position as neutral, conserved (invariant) or class-specific.
5. Map a status of each site by color coding onto the 3D structure of the protein.

After the development of the original ET method, developers of Consurf and other authors have used different methods for calculating the phylogenetic tree (1) and conservation of consensus sequences (4) (Armon *et al.*, 2001). The first version of Consurf applied a conservation score, where physicochemical distances between each pair of amino acid were taken into account (Armon *et al.*, 2001). Later versions have used evolutionary rate as a conservation score. The evolutionary rate indicates how fast a site has evolved in relation to the average site (Landau *et al.*, 2005). In the Rate4site method, the evolutionary rate can be estimated using the likelihood of the data given the tree and the

evolutionary rate. The estimation can be carried out by the ML method (Pupko *et al.*, 2002) or alternatively by the empirical Bayesian method (Mayrose *et al.*, 2004). The Bayesian approach calculates a posterior distribution of the evolutionary rate assuming a Gamma prior distribution (Mayrose *et al.*, 2004). The evolutionary rate is estimated by the expected value of the posterior distribution.

A good conservation score should take into account sequence redundancy in a given MSA. However, among a considerable number of different conservation scores, only a few have originally used sequence weighting. Principally, however, sequence weighting could be added to most of the presented scores. One could apply average distance between sequences (Vingron and Argos, 1989), an entropy-based measure (Henikoff and Henikoff, 1994), Voronoi weights (Sibbald and Argos, 1990), or other methods using a phylogenetic tree, for instance (Durbin *et al.*, 1998; Valdar, 2002). The authors using Vingron and Agros -type weighting have used the mutation matrix (Valdar and Thornton, 2001) or one minus percentage identity to calculate the evolutionary distances between sequences (Sander and Schneider, 1991; Wass and Sternberg, 2008; Thompson *et al.*, 2001). Pei and Grishin (2001) proposed the option of using weighted relative frequencies of amino acids instead of plain relative frequencies in their three scores, and many others have followed them (Capra and Singh, 2007).

### 2.1.2 Inference on conservation scores

The interpretation of conservation scores have been traditionally made by dividing the scores as conserved or unconserved using a predefined threshold level. Only recently, have methods of statistical inference been used to interpret the scores and answer the questions: 'what is the expected probability of conservation' or 'what is the statistical significance of an alignment position'. A simple approach is to compare the average conservation of positions of interest with that of the other positions in order to compute the probability that the variability of positions of interest has been obtained by chance. This kind of approach has been used for studying the conservation of a folding nucleus (Mirny and Shakhnovich, 2001) and the conservation of protein-protein interfaces (Valdar and Thornton, 2001).

Several conservation measures use the Z score as a test statistic. The simplest way is to calculate the mean and standard deviation of the conservation score, which are needed in the Z score, over all positions in the alignment, and use a predefined threshold to determine the status of the position (Hannenhalli and Russell, 2000; Wass and Sternberg, 2008). A more sophisticated way is to calculate the mean and standard deviation from the random model, which is as close as possible to the original data, but does not include the functional constraints of the original data (Pei *et al.*, 2006). Mirny and Gelfand (2002) introduced two ways to generate a random sample: random shuffling of the original position and simulating data with the help of a phylogenetic tree. Pei *et al.* (2006) generated

a random model as a combination of random shuffling and evolutionary simulation. Fischer *et al.* (2008) directly calculated the probability that a given position is functional (catalytic or ligand binding). They defined the posterior probability of a position given a conservation score, relative frequencies of residues, and other parameters predicted for the local environment.

### 2.1.3 Performance of conservation scores and recent advances

Many authors have evaluated the performance of different scoring methods in predicting catalytic, ligand-binding and protein-protein interaction (PPI) sites in protein sequence alignments (Panchenko *et al.*, 2004; Capra and Singh, 2007; Fischer *et al.*, 2008). The results have shown that the prediction accuracy is highest in catalytic sites and next highest in ligand-binding sites, while that of PPI sites is much lower, indicating that the conservation cannot be solely used to predict the PPI sites. Clustering of neighboring positions seems to improve the accuracy of prediction of functional sites. Panchenko *et al.* (2004) scored clusters of residues in contact, while sequential neighbors have been exploited by others (Capra and Singh, 2007; Fischer *et al.*, 2008). Recent advances in prediction of protein functional sites have been made by combining different sequence and structural information, such as relative frequencies of amino acids, identity, solvent accessibility, secondary structure, relative position on protein surface, along with conservation, into the machine learning framework (Gutteridge *et al.*, 2003; Panchenko *et al.*, 2004; Petrova and Wu, 2006).

To conclude, conservation scores have been successfully used for finding and verifying many important structural and functional sites in proteins. Although recent methods often use additional sequence and structural information in the prediction, conservation has been proved to be the most important feature of the prediction (Petrova and Wu, 2006). In his review of conservation methods, Valdar (2002) concluded that no method was both biologically and statistically rigorous at that time. Since Valdar's review, new conservation scores have better taken into account his criteria for a good conservation score. Improvements could be made, however, by estimating the significance or posterior probability of conservation in an alignment position. Furthermore, the uncertainty of the prediction should be estimated by calculating confidence or probability intervals for the conservation.

## 2.2 Methods for assessing alignment quality

This section discusses factors having an impact on alignment quality, and describes methods developed for assessing the quality of multiple sequence alignments.

### 2.2.1 Factors affecting alignment quality

The major factors contributing to alignment difficulty and quality are evolutionary distance between sequences, sequence length and number of sequences (Sauder *et al.*, 2000; Griffiths-Jones and Bateman, 2002). The increase of evolutionary distance (decrease of sequence identity) complicates the alignment procedure and decreases the alignment quality. The increased sequence length has an opposite effect on alignment quality, especially in alignments with high evolutionary distances (Lassmann and Sonnhammer, 2002). Most of the alignment algorithms are capable of producing biologically plausible alignments, when pairwise identity of all sequences is more than 40 %. When identity drops below 20-25 %, the accuracy of most alignment methods decreases dramatically (Thompson *et al.*, 1999; Jaroszewski *et al.*, 2000). This is because local changes in structure between distantly related sequences can be remarkable: conserved regions such as hydrophobic core residues or key catalytic amino acids are usually correctly aligned, but alignment quality tends to decrease in more variable loop regions or other regions exposed to solvent. The sequence identity is, however, not alone a sufficient measure for validating sequence alignments (Pei and Grishin, 2006, 2007).

### 2.2.2 Reference-based alignment quality scores

Cline *et al.* (2002) defined three criteria for a good alignment quality measure. Firstly, the measure should be scaled so that a higher score implies higher alignment quality, secondly, it should be optimizable, so that the quality increases when badly aligned regions are removed, and thirdly, it should penalize for *over-alignment*, i.e. aligning pairs which are structurally not alignable, for *under-alignment*, i.e. not aligning structurally alignable parts, and for *misalignment*.

The traditional method for validating sequence alignments has been to compare them with the corresponding structural alignments. This approach can, however, be widely applied only for pairwise sequence alignments. The standard quality scores for pairwise alignments are called modeler's ($f_m$) and developer's ($f_d$) viewpoint scores (Sauder *et al.*, 2000). The $f_m$ score measures the proportion of correctly aligned residues in the sequence alignment, whereas the $f_d$ score describes which part of the structural alignment is correctly represented in the sequence alignment. The $f_m$ and $f_d$ scores measure specificity and sensitivity, which control *over-alignment* and *under-alignment*, respectively. Hence, together they fulfill Cline's three criteria for a good alignment quality score, except for the last part of the third criterion. These two scores have been frequently reported together to assess the overall quality of pairwise alignments. A shift score calculates how many residues apart the residue in the sequence alignment is from the corresponding residue in the structural alignment (Domingues *et al.*, 2000; Cline *et al.*, 2002). The validation of the shift score shows that it effectively addresses all the criteria for a good alignment quality score (Cline *et al.*, 2002).

Similar measures have been widely used for evaluating MSAs. The most commonly used alignment quality measures, when reference alignment is available, are the sum of pairs (SP) score and the column score (CS) (Thompson *et al.*, 1999). The CS score measures the proportion of identical columns between the reference and test alignments: it gives a rough overall estimate of the alignment quality, but has the drawback that even one misaligned residue in a position reduces the score of that position to zero, and one misaligned sequence is enough to result in a zero CS score for the whole alignment. The SP score calculates the proportion of identically aligned residue pairs in the reference and test alignments. Karplus and Hu (2001) modified the SP score to account for gaps by weighting the identically aligned residues in the reference and test alignments by 2, and the residues aligned with a gap in both alignments by 1. The CS and SP scores are both measures of sensitivity. They have frequently been used for benchmarking the MSA methods in different reference alignment databases.

### 2.2.3 Reference-independent alignment quality scores

If no reference alignment is available, the alignment quality can be assessed by reference-independent methods. Conservation has traditionally been used as a measure of alignment quality (Pei and Grishin, 2001). Principally, any column-based conservation score presented in a section 2.1 could be used as a quality score by summing up the positional scores over the entire length of the alignment. One such measure is the sum-of-pairs measure of Carrillo and Lipman (1988), from which many other scores have obtained their inspiration. The benefit of the scores based on the sum-of-pairs is that the similarity of amino acids has been taken into account. It should be noted that this sum-of-pairs measure is different from the SP score presented in the previous section.

Pei and Grishin (2001) introduced several positional conservation measures, which are based on the use of the sum-of-pairs score, the Shannon entropy or the variance of relative frequencies of residues. The authors also used these positional scores to assess the quality of whole alignments. They compared the conservation of manually curated SMART (Schultz *et al.*, 1998) and FSSP structural alignments (Holm and Sander, 1996) with that of ClustalW alignments (Thompson *et al.*, 1994). The entropy-based measure was superior to the other methods for rather similar sequences (SMART), whereas the sum-of-pairs-based measure outperformed the others for very divergent sequences (FSSP). These results indicate that conservation-based measures are valuable tools for assessing the quality of MSAs.

Thompson *et al.* (2001) have used the norMD conservation score for measuring the overall quality of alignments and for the detection of badly aligned regions or unrelated sequences in MSAs. Their comparison with the other column-based methods, sum-of-pairs, IC and MD scores in the BAliBASE reference databases (Bahr *et al.*, 2001) shows that the norMD is superior to the other methods.

Beiko *et al.* (2005) presented a word-oriented objective function (WOOF) for alignment validation. Their method is not column-based, but relies on the scoring of regions of conserved amino acids. The WOOF generates patterns from each pair of sequences and calculates a weighted proportion of correctly aligned residues. The results of Beiko *et al.* (2005) show that both the WOOF score with exact pattern matches and the IC score assign very high scores to BAliBASE reference alignments compared to the automatically generated alignments. The norMD score, on the contrary, did not favor reference alignments over automatically produced alignments. This result is somewhat inconsistent with the results of Thompson *et al.* (2001).

In their article, Hertz and Stormo (1999) described two procedures for calculating the significance of IC for a whole MSA, and used these measures to identify optimal alignments. The first procedure used large-scale deviation statistics, while the other was based on pure numerical calculations. Later, Nagarajan *et al.* (2005) and Keich (2005) used modified fast Fourier transformation to estimate the $p$ value for an IC. Recently, Tomovic and Oakeley (2007) applied Bayes factors and posterior probabilities to distinguish random alignments from biologically relevant ones. Their results showed that the Bayesian method had higher specificity compared to the two methods proposed by Nagarajan *et al.* (2005) and Keich (2005). All these approaches calculating the significance of the IC are, however, only applicable for local ungapped alignments, and they have mostly been tested for short DNA sequences. Furthermore, like all positional quality scores, they assume that alignment positions are independent from each other, which is too strict an assumption.

Another approach to assess alignment quality is to use known protein structures. This is especially beneficial with divergent sequences, since the best alignments of remote homologues have been built using structural information (Jaroszewski *et al.*, 2000; Menke *et al.*, 2008). The iRMSD (Armougom *et al.*, 2006) and its previous version, APDB (O'Sullivan *et al.*, 2003), are based on the RMSD, which measures the distance between equivalent alpha carbons of two superposed structures. The iRMSD score has been improved so that the score is independent of any structural superposition methods. The iRMSD score assumes that if two residue pairs, say $A_1B_1$ and $A_2B_2$, are correctly aligned, then the distances of two residues, $d(A_1, A_2)$ and $d(B_1, B_2)$, within both sequences must be roughly equal. The normalized iRMSD (NiRMSD) takes into account alignment lengths, and can be used to compare alternative alignments. Armougom *et al.* (2006) have shown that in the BAliBASE database, the performance of the NiRMSD is 90% consistent with that of the SP score, when the SP score was calculated for the core blocks only. The drawback of the NiRMSD score is that it can only be used to compare the relative accuracy of two alignments and, more importantly, the structures of at least two sequences of alignment must be available.

If the structures of at least two sequences of alignment are available, another possibility is to use structural similarity scores, such as the DALI Z score (Holm and Sander, 1998), the TM score (Zhang and Skolnick, 2004), the GDT-TS score (Zemla *et al.*, 1999) or the

3D score (Rychlewski *et al.*, 2003), for the evaluation of alignment accuracy (Pei and Grishin, 2006, 2007). Pei and Grishin (2006; 2007) obtained good correlation between the reference-independent structural similarity and reference-dependent scores. The result indicates that the structural similarity scores could be helpful in benchmarking alignment programs, for instance, especially when the similarity of sequences is low. It should be remembered, however, that such evaluation is based only on sequences whose structure is available.

The concept of consistency was first introduced for constructing pairwise alignments by Gotoh (1990) and Vingron and Argos (1991). Currently, it is used in the best strategies for constructing MSAs (Edgar, 2004; Katoh *et al.*, 2002, 2005; Do *et al.*, 2005), and it has also been used to assess alignment quality. The use of consistency in measuring alignment quality originates from Mevissen and Vingron (1996), who presented a reliability score for every residue pair of optimal pairwise alignment using sub-optimal alignments. Their measure is based on the assumption of consistency, that is, that the positions which are consistent among several suboptimal alignments are usually highly conserved, and have been proved in many studies to be reliably aligned. Lassmann and Sonnhammer (2005) also used consistency as a core idea in their alignment quality method. Their MOS score is based on aligning the same set of sequences with several MSA programs and comparing the results. The alignment quality is determined by the proportion of similarly aligned residue pairs among all pairs of aligned residues. Their comparisons show that the MOS score clearly outperforms the norMD, average identity and the scores proposed by Pei and Grishin (2001). The authors report that the MOS score might be sensible for the choice and number of test alignments. Landan and Graur (2007, 2008) also relied on consistency. As distinct from Lassmann and Sonnhammer's method, they used several suboptimal alignments and the same alignments in reversed residue order. Vingron (1996) pointed out that consistency-based scores cannot recognize the real relation of sequences and should, therefore, only be applied to similar sequences, since even unrelated sequences could be similarly aligned.

In conclusion, the measures quantifying alignment quality can be divided into those relying on conservation, consistency or structural similarity. The evaluation of the scores is difficult, since no comprehensive comparison has been made. The performance of the new scores has usually been compared with that of the SP score. This comparison gives a test for sensitivity, whereas a test for specificity has often been ignored. Scores for measuring specificity, such as the $f_m$ score (Sauder *et al.*, 2000) or the shift score (Cline *et al.*, 2002) for pairwise alignments, should also be developed for the evaluation of MSA. Almost all measures are meaningful in the sense that the absolute scores of different alignments with different characteristics can be compared. The statistical inference of quality scores, such as significance tests, confidence intervals or posterior probabilities, has, however, usually been ignored.

# Chapter 3

# Methods

## 3.1 Multiple sequence alignment

### 3.1.1 Formulation of MSA

In the following, we present a mathematical formulation for the MSA (Koski, 1999; Waterman, 1995). Suppose, we have $N$ amino acid sequences

$$\boldsymbol{o^l} = (o^l_{i_1}, o^l_{i_2}, \ldots, o^l_{i_{m(l)}}), \qquad \text{for } l = 1, 2, \ldots, N, \tag{3.1}$$

of length $m(l)$. The sequences consist of symbols $o_i$ from the 20 letter alphabet $O = \{A, C, \ldots, T, Y\}$ of amino acids. The MSA of the sequences $\boldsymbol{o^l}$ is a two-dimensional array of $N$ rows and $L$ columns, and can be designated as

$$\mathcal{A}: \begin{matrix} o^1_1, & o^1_2, & \ldots, & o^1_L \\ o^2_1, & o^2_2, & \ldots, & o^2_L \\ \vdots & \vdots & \ddots & \vdots \\ o^N_1, & o^N_2, & \ldots, & o^N_L. \end{matrix}$$

Each row represents one sequence and each column the residues, which have been assumed to be evolved from the same position of the common ancestor. The residue mismatches within a column are interpreted as point mutations. Technically, we need to add a gap $'-'$ to the alphabets to describe insertion or deletion indels, which usually cannot be distinguished from each other. Pairwise sequence alignments are special cases of MSAs where $l = 2$.

### 3.1.2 Probabilistic model for MSA

In this section, we formulate a statistical model for MSA. The model assumes that the positions of MSA are independent. We assume that amino acids have fixed probabilities $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{20})$ to occur in the sequences. Let us introduce $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_{20})$ as

a vector of random variables defining the number of times each amino acid occurs in an alignment position. Then, the observations $(n_1, n_2, \ldots, n_{20})$ from the random variables $(Y_1, Y_2, \ldots, Y_{20})$ follow a multinomial distribution

$$\mathbb{P}(Y_1 = n_1, Y_2 = n_2, \ldots, Y_{20} = n_{20}) = \frac{n!}{\prod_{i=1}^{20} n_i!} \prod_{i=1}^{20} \beta_i^{n_i}, \tag{3.2}$$

where $n$ is the number of amino acids in an alignment position.

The probabilities of amino acids are unknown, but the maximum likelihood (ML) estimators of $\boldsymbol{\beta}$ are given by the relative frequencies of amino acids

$$b_i = \frac{n_i}{n}, \qquad \text{for } i = 1, 2, \cdots, 20. \tag{3.3}$$

The vector of background probabilities $\boldsymbol{\beta^0} = (\beta_1^0, \beta_2^0, \ldots, \beta_{20}^0)$ describes a random distribution of amino acids. As a background, we used the distribution of the amino acid composition of all proteins in the SWISS-PROT database (Boeckmann *et al.*, 2003) (Publications I and II) or the amino acid distribution of the whole MSA of interest (Publications III and IV). The latter background distribution was chosen to obtain a better estimate of the underlying amino acid composition of the protein family of interest.

The mean and variance of number of an amino acid $i$ are $\mathbb{E}(n_i) = n\beta_i$ and $\mathbb{V}\text{ar}(n_i) = n_i\beta_i(1 - \beta_i)$, from which it follows that the expectation vector and covariance matrix of $\boldsymbol{b} = (b_1, b_2, \ldots, b_{20})$ are given by

$$\mathbb{E}(\boldsymbol{b}) = \boldsymbol{\beta} \qquad \text{and} \qquad \mathbb{C}\text{ov}(\boldsymbol{b}) = \boldsymbol{\Sigma}, \tag{3.4}$$

where

$$\Sigma_{ij} = \frac{\beta_i(\delta_{ij} - \beta_j)}{n}, \qquad \text{for } i, j = 1, 2, \ldots, 20, \tag{3.5}$$

and the Kronecker's delta function is defined by $\delta_{jj} = 1$ and $\delta_{ij} = 0$ for all $i \neq j$.

### 3.1.3 Statistical hypotheses for MSA

This section presents the statistical hypotheses which were used to identify conserved residues and conserved positions in MSA. The following hypotheses were postulated for testing whether the occurrences of residues follow an underlying background distribution:

$$H_{0i} : \beta_i = \beta_i^o, \text{ for } i = 1, 2, \cdots, 20 \tag{3.6}$$

versus

$$H_{Ai} : \beta_i \geq \beta_i^o, \text{ for } i = 1, 2, \cdots, 20, \tag{3.7}$$

where at least one inequality is proper.

Since the hypotheses can be considered as a family of 20 hypotheses, we can express the null hypothesis as an intersection of hypotheses and the alternative as a union of hypotheses (Roy, 1953):

$$H_0 : \beta_i = \bigcap_{i=1}^{20} \beta_i^o \qquad \text{and} \qquad H_A : \beta_i = \bigcup_{i=1}^{20} \beta_i^o. \tag{3.8}$$

## 3.2 Methods for scoring conservation

This section introduces three types of conservation scoring methods. The residue conservation scores divide individual residues into conserved and unconserved in a given MSA position (Publications I and II). The positional conservation score calculates conservation for each alignment position (Publication III). Finally, the whole alignment conservation score defines conservation for the entire alignment (Publication III). In the following text, the terms "residue conservation" and "positional conservation" have been used to distinguish the first two of these approaches.

### 3.2.1 Scoring residue conservation

In this section, we present three methods, Dunn-Sidak, Bauer and Iterative scores, for identifying conserved residues at each MSA position. The Dunn-Sidak and Bauer scores have been originally presented in Publication II, and the Iterative score in Publication I. The Dunn-Sidak and Bauer scores test the $H_{0i}$ hypothesis (3.6) against an alternative (3.7) in order to decide which of the residues at one MSA position occur more often than would be expected from the background probability of that residue. Both scores use

$$Z_i = \frac{b_i - \mathbb{E}(b_i)}{\sqrt{\mathbb{V}\mathrm{ar}(b_i)}}, \qquad \text{for } i = 1, 2, \cdots, 20 \tag{3.9}$$

as a test statistic for each individual residue $i$. The $Z_i$ statistics are assumed to be independent and identically distributed $N(0, 1)$ random variables. For testing the null hypothesis $H_{0i}$, the rejection region for the $Z$ test is in the form

$$\max_{1 \le i \le 20} Z_i > \xi, \tag{3.10}$$

where $\xi$ is a critical value. The critical value can be defined from the Union-Intersection (UI) test for $H_0$ (3.8) (Roy, 1953). The rejection region is defined as a union of rejection regions, i.e. $H_0$ is rejected only if at least one $H_{0i}$ is rejected. In order to contol the family-wise error rate (FWE), that is, the risk of a false decision at some predefined level $\alpha$, the $\xi$ must be chosen so that under $H_0$

$$\mathbb{P}(\max_{1 \le i \le 20} Z_i > \xi) = \alpha. \tag{3.11}$$

Under $H_0$ $Z_1, Z_2, \ldots, Z_{20}$ have a 20-variate normal distribution and, hence, the critical value $\xi$ is an upper $\alpha$ point of the standard normal distribution, which is denoted as $Z_{20}^{\alpha}$.

For the Dunn-Sidak score, we specified a critical value $\xi$, so that it would better account for the multiple tests made at each alignment position (Publication II). To obtain a less conservative approximation for the critical value, we used the Dunn-Sidak procedure (Dunn, 1958; Hochberg and Tamhane, 1987). From Sidak's inequality (Sidak, 1967; Hochberg and Tamhane, 1987)

$$\mathbb{P}(\max_{1\leq i\leq 20} Z_i \leq \xi) \geq \prod_{1\leq i\leq 20} \mathbb{P}(Z_i \leq \xi) \tag{3.12}$$

it follows that if each test $Z_i$ is of size

$$\mathbb{P}(Z_i \leq \xi) = (1-\alpha)^{1/20} \tag{3.13}$$

then

$$\mathbb{P}(\max_{1\leq i\leq 20} Z_i \leq \xi) \geq 1-\alpha. \tag{3.14}$$

The approximation to the upper bound of $\xi$ results in $Z_{20}^{1-(1-\alpha)^{\frac{1}{20}}}$, which is an upper $1-(1-\alpha)^{\frac{1}{20}}$ point of the standard normal distribution.

Let us assume that under $H_0$ $\boldsymbol{b}$ has an asymptotically multivariate normal distribution. The expectation and variance of distibution of $\boldsymbol{b}$ are obtained from the equations (3.4) by replacing the unknown $\boldsymbol{\beta}$ by the corresponding value of the background distribution $\boldsymbol{\beta^o}$. The **Dunn-Sidak conservation score** can now be expressed as

$$I_{DS} = \{i : \frac{b_i - \beta_i^o}{\sqrt{\frac{\beta_i^o(1-\beta_i^o)}{n}}} > Z_{20}^{1-(1-\alpha)^{\frac{1}{20}}}, 0<\alpha<1\}. \tag{3.15}$$

The score determines the residues whose $Z$ score under the $H_0$ is over the critical point $Z_{20}^{1-(1-\alpha)^{\frac{1}{20}}}$ as conserved and the other residues as unconserved. Let us define the sets of conserved and unconserved residues as $J_1$ and $J_2$, respectively, and $J = J_1 \cup J_2$.

In Publication II, we also introduced the **Bauer conservation score**. The Bauer score is assigned as

$$I_B = \{i : n_i = n \text{ or } \frac{b_i - \beta_i^o}{\sqrt{\frac{b_i(1-b_i)}{n}}} > n^{\frac{1}{c}},\ c>2,\ n_i>0\}, \tag{3.16}$$

where $c$ is a fixed threshold value. Here $n^{\frac{1}{c}} \to \infty$ and $n^{\frac{1}{c}}\sigma_{in} \to 0$ for $n \to \infty$ and all $0 < \beta_i < 1$ (Bauer *et al.*, 1988). The Bauer score differed from the Dunn-Sidak mainly in the way it calculates the variance. While in the Bauer score, $\beta_i$ is replaced by its ML estimate $b_i$, in the Dunn-Sidak score, the variance is calculated under the null hypothesis. Another difference between the methods is that in the Dunn-Sidak score, the

significance level $\alpha$ is fixed, meaning that, even if the sample size tends to infinity, the risk of misclassification of unconserved residues is always $\alpha$. In the Bauer score, on the contrary, the significance level converges to zero, i.e. $c_i(n) \to \infty$. Therefore, as the sample size increases, the number of misclassifications tends to zero, although very slowly.

In Publication I, we introduced an algorithm to calculate an **Iterative conservation score**. In that publication, conserved and unconserved residues have been called effective and ineffective, respectively. The iterative conservation score ($I_{IT}$) is based on comparing the proportion of residues in one alignment position with respect to their background probabilities. The residue which determines the largest ratio between the ML estimates and background probabilities exceeding a fixed threshold value $c > 1$ is chosen as conserved. The rest of the ML estimates and background probabilities of the residues are renormalized to sum to one. The renormalization ensures that the residues with low and high background probabilities are handled equally. The iteration is continued until the largest ratio does not exceed the threshold value. The algorithm for calculating the $I_{IT}$ score has been elaborated in Publication I.

### 3.2.2 Scoring positional conservation

For scoring positional conservation, the $Z$ statistic (3.9) was calculated for the profile instead of the ML estimates of residue probabilities (Publication III). The profile for amino acid $i$ is expressed as

$$f_i = \sum_{j=1}^{20} b_j c_{ij} = \boldsymbol{c}_i^T \boldsymbol{b}, \qquad \text{for } i = 1, 2 \cdots, 20, \tag{3.17}$$

where $c_{ij}$ denotes one component $(i, j)$ of the whole substitution or similarity matrix $\boldsymbol{C}$. The expectation vector and covariance matrix for the profile under $H_0$ (3.6) are defined as

$$\mathbb{E}(\boldsymbol{f}) = \boldsymbol{C}\boldsymbol{\beta}^0 \qquad \text{and} \qquad \mathbb{C}\text{ov}(\boldsymbol{f}) = \boldsymbol{C}\boldsymbol{\Sigma}^0\boldsymbol{C}^T, \tag{3.18}$$

where the entries of $\boldsymbol{\Sigma}^0$ are defined as in (3.5) but $\beta_i$ and $\beta_j$ has been replaced with $\beta_i^0$ and $\beta_j^0$. Applying the expectation and covariance to the $Z$ statistic (3.9) gives us a $Z$ statistic for the profile (3.17)

$$Z_i = \frac{\boldsymbol{c}_i^T(\boldsymbol{b} - \boldsymbol{\beta}^0)}{\sqrt{\boldsymbol{c}_i^T \boldsymbol{\Sigma}^0 \boldsymbol{c}_i}}, \qquad i = 1, 2, \ldots, 20. \tag{3.19}$$

This statistic differs from the $Z$ statistic used for scoring individual residues by taking into account the similarities or other criteria describing the stereochemical relationships between amino acids. For scoring alignment positions, we used as a test statistic the maximal $Z_i$ value, maxZ. Hence, we avoided carrying out multiple tests within one alignment position.

The significance of the maxZ statistic was calculated by testing the null hypothesis (3.6) against the alternative (3.7) using an importance sampling (IS) method (Rubin, 1988). The low observed significance level indicates that the observed value of maxZ is significantly larger than that which would be likely to arise under $H_0$ due to random variation. The alignment position with a low significance level was defined as conserved.

In order to apply the IS method, we defined the IS distribution as a mixture of multinomial distributions for 20 amino acid frequencies (Publication III)

$$\begin{aligned} g^* &= \mathbb{P}(n_1, n_2, \ldots, n_{20} | \beta_1^0, \beta_2^0, \ldots, \beta_{20}^0, \alpha, \epsilon) \\ &= \alpha(\tfrac{n}{n_{1,0} \ldots n_{20,0}}) \prod_{i=1}^{20} \beta_{i,0}^{0 n_{i,0}} \\ &+ \tfrac{1-\alpha}{K} \sum_{k=1}^{K} (\tfrac{n}{n_{1,k} \ldots n_{20,k}}) \prod_{i=1}^{20} \beta_{i,k}^{n_{i,k}}, \end{aligned} \tag{3.20}$$

where $\alpha > 0, \epsilon < 1$, $K+1$ denotes a number of mixture components, which is here the number of amino acids plus one, $\beta_{i,0}^0$ is the background probability of the $i$th amino acid in the 0th mixture, $n_{i,k}$ is the $i$th amino acid frequency in the $k$th mixture and

$$\beta_{i,k} = \begin{cases} \epsilon + (1-\epsilon)\frac{\beta_i^0}{K}, & i = k, \\ (1-\epsilon)\frac{\beta_i^0}{K}, & i \neq k. \end{cases}$$

This IS distribution has one mixture component for the background distribution and one for each amino acid. The $\alpha$ (mixture) parameter determines which parts of the samples are drawn from the background distribution (first mixture) and which from the other $K$ mixtures. The $\epsilon$ (shape) parameter approximates the probability of the highest amino acid, while the probability of the other amino acids are proportional to their background distribution.

Using this IS distribution, the IS procedure converges rather rapidly. This is because by using $K+1$ mixtures we can obtain a good coverage in the 20-dimensional parameter space, but at the same time, by using large $\epsilon$ values, we can obtain extreme samples from the parameter space, and hence, obtain more exact significance levels for extreme observations. The entire IS procedure for calculating significance levels for each alignment position has been elaborated in Publication III.

### 3.2.3 Scoring whole alignment conservation

After calculating the significance tests for each alignment position by IS sampling, we corrected the false-positive error rate of multiple tests by controling the false discovery rate (FDR), i.e. the expected proportion of erroneously rejected null hypotheses (Benjamini and Hochberg, 1995). The FDR was chosen since our main interest is not to increase the statistical power of single tests, but to find the set of conserved alignment positions. Furthermore, we expected to find many conserved positions in the alignment, and it has been proven that the FDR methods are most applicable for this kind of approach (Dudbridge and Koeleman, 2004). We applied a step-up procedure, which takes the

dependency of test statistics into account (Benjamini and Yekutieli, 2001). With the given FDR and the length of MSA, the procedure chooses which of the hypotheses $H_{0i}$ (3.6) are not supported, i.e. which of the alignment positions can be considered as conserved. The full step-up procedure has been described in Publication III.

Next, we defined a quality score for the whole alignment using the FDR-corrected significance tests. The whole alignment score measures the conservation level of MSA. The score was defined by the proportion of residues located in the conserved alignment positions, hence, it also takes the number of gaps into account. The whole alignment conservation score ConsAA is designed as

$$\text{ConsAA} = \frac{\sum_{j \in J^*} n_j}{\sum_{j=1}^{L} n_j}, \tag{3.21}$$

where $n_j$ is the number of residues at position $j$, and $J^*$ is a set of conserved alignment positions. The ConsAA will be used as a core for building the alignment quality scores in the next section.

## 3.3 Applications of conservation scores

This section first describes how the methods identifying conserved and unconserved residues in an alignment position can be used to solve the overfitting problem in profile HMMs (Publication I). Then, it introduces two measures for quantifying alignment quality. Both of these measures determine quality in terms of the whole alignment conservation score ConsAA. The first score assumes the true reference alignment to be known (Publication III), while in the second score, this assumption has been relaxed (Publication IV).

### 3.3.1 Emission probability estimation method for profile HMMs

In the profile HMMs, the conserved residues of MSA are modeled by emissions in the match states (Krogh *et al.*, 1994; Eddy, 1998). The emission probability distribution of match states is usually estimated by the ML or an alternative Bayesian method. When using the ML method, the emission probabilities are calculated for each amino acid at each alignment position. Since the profile HMMs can be very long, the number of estimated emission parameters can be enormous, and this may cause the overfitting problem. In the Bayesian framework, this problem has been solved by incorporating some underlying characteristics of alignment environments into the model using the Dirichlet prior distribution (Sjölander *et al.*, 1996). The use of the 1-component Dirichlet distribution is a probabilistic way to add simple pseudocounts to the count of each residue in the alignment (Durbin *et al.*, 1998), whereas the use of the Dirichlet mixture prior distribution corresponds to the use of linear combination of pseudocounts (Sjölander *et al.*, 1996).

We have developed an efficient emission probability (EEP) estimation method to reduce the parameter space in the emission probability estimation of profile HMMs, and hence, overcome the overfitting problem (Publication I). The method is based on the idea of choosing which of the residues in the alignment are conserved and unconserved, and then incorporating this information into the ML estimation. The method assumes that unconserved residues follow the background distribution.

First, the three constraints have been determined for the log-likelihood function of the multinomial distribution function

$$l = \sum_{j \in J} n_j \log b_j. \tag{3.22}$$

The constraints enable the reduction of the parameter space. The first constraint forces the mutual ratios between the residue relative frequencies and their background probabilities to be equal in the set of unconserved residues. The second condition is needed to give a small, non-zero estimate for unconserved residues when only conserved residue(s) appear in that position. The third condition ensures that the emission probabilities always sum to one.

The optimisation problem has then been solved with the Lagrange multipliers method (Luenberger, 1984). In the first solution, the EEP estimated probabilities $b_j^*$ are given by

$$\begin{aligned} b_j^* &= \frac{n_j}{\sum_{j \in J} n_j}, && \text{for } j \in J_1, \quad \text{and} \\ b_j^* &= \frac{b_j^o}{\sum_{j \in J_2} b_j^o} \frac{\sum_{j \in J_2} n_j}{\sum_{j \in J} n_j}, && \text{for } j \in J_2. \end{aligned}$$

The EEP estimators for conserved residues are ML estimators, but the estimates of unconserved residues are obtained by dividing the sum of the probability of unconserved residues in proportion to the background probabilities. If a position has been occupied with conserved residues only, the EEP emission probability estimates take the forms

$$\begin{aligned} b_j^* &= \frac{c \sum_{j \in J_1} b_j^o}{c \sum_{j \in J_1} b_j^o + \sum_{j \in J_2} b_j^o} \frac{n_j}{\sum_{j \in J_1} n_j}, && \text{for } j \in J_1, \quad \text{and} \\ b_j^* &= \frac{b_j^o}{c \sum_{j \in J_1} b_j^o + \sum_{j \in J_2} b_j^o}, && \text{for } j \in J_2. \end{aligned}$$

The latter solutions are an alternative to the pseudocount methods, ensuring that the emission probabilities are always non-zero.

The EEP method has two advantages over the ML method: the first advantage is that the number of parameters needed for emission probability estimation is dramatically reduced. While in the ML estimation, 20 parameters are needed for each match state, in the EEP estimation, the number of parameters is the number of conserved residues plus one. Hence, only one parameter is needed to estimate the emission probabilities of all unconserved residues. The second advantage is that the EEP method always produces non-zero emission probability estimates.

### 3.3.2 Reference-based alignment quality score

When reference alignment is available, the quality of the alignment can be assessed by explicitly comparing the proportion of conserved residues, ConsAA, in reference and test alignments. The alignment quality (AQ) score, originally presented in Publication III, can be expressed as

$$AQ = [1 - (|\mathrm{ConsAA}_{\mathrm{ref}} - \mathrm{ConsAA}_{\mathrm{test}}|/\mathrm{ConsAA}_{\mathrm{ref}})] * 100. \quad (3.23)$$

The $AQ$ score measures what percentage of conservation of reference alignment is included in a test alignment. The $AQ$ score can be used in benchmarking MSA programs, for instance. Compared with the $SP$ score, which calculates the proportion of correctly aligned residues and does not distinguish between divergence and homologous regions, the $AQ$ is focused on comparing the total conservation of alignments.

### 3.3.3 Reference-independent alignment quality score

When the "true" reference alignment is not known, the quality of alignments cannot be assessed by the score presented in the previous section, but a reference-independent validation method is needed. This section describes a model-based quality score, which can be used when only primary multiple sequence alignment is available. The score does not need structural information or several alternative alignments, but is based on the conservation of the whole MSA and additional *ab initio* sequence information. The model-based quality score was originally presented in Publication IV. An overview of the building and use of the quality score is illustrated in Figure 3.1.

For building a model-based quality score, one has to choose one or several reference alignment databases, which are used as a source of conservation information about known proteins from the whole protein fold space. The database should include multiple alignments of reference protein sequences, and the number of aligned sequences should be large enough for statistical analysis. The reference alignments should be biologically as correct as possible, for example, structural alignments. This is a huge demand, since the construction of multiple structural alignments, especially for remote homologues, is often a difficult or even impossible task. We used two thirds of the Homstrad database as a reference (Mizuguchi *et al.*, 1998).

Next, one or several MSA programs have to be chosen. The programs should be proven to be accurate and fast enough to align sequences in one or several reference alignment databases. We used three alignment programs, Mafft (L-INS-i mode), Muscle and Probcons (Katoh *et al.*, 2002, 2005; Edgar, 2004; Do *et al.*, 2005). The sets of reference sequences were re-aligned using the three methods, and the reference-based quality scores were calculated to measure how far the automatically aligned sequences are from the reference alignments. We chose the $SP$ score as a reference quality score, but principally

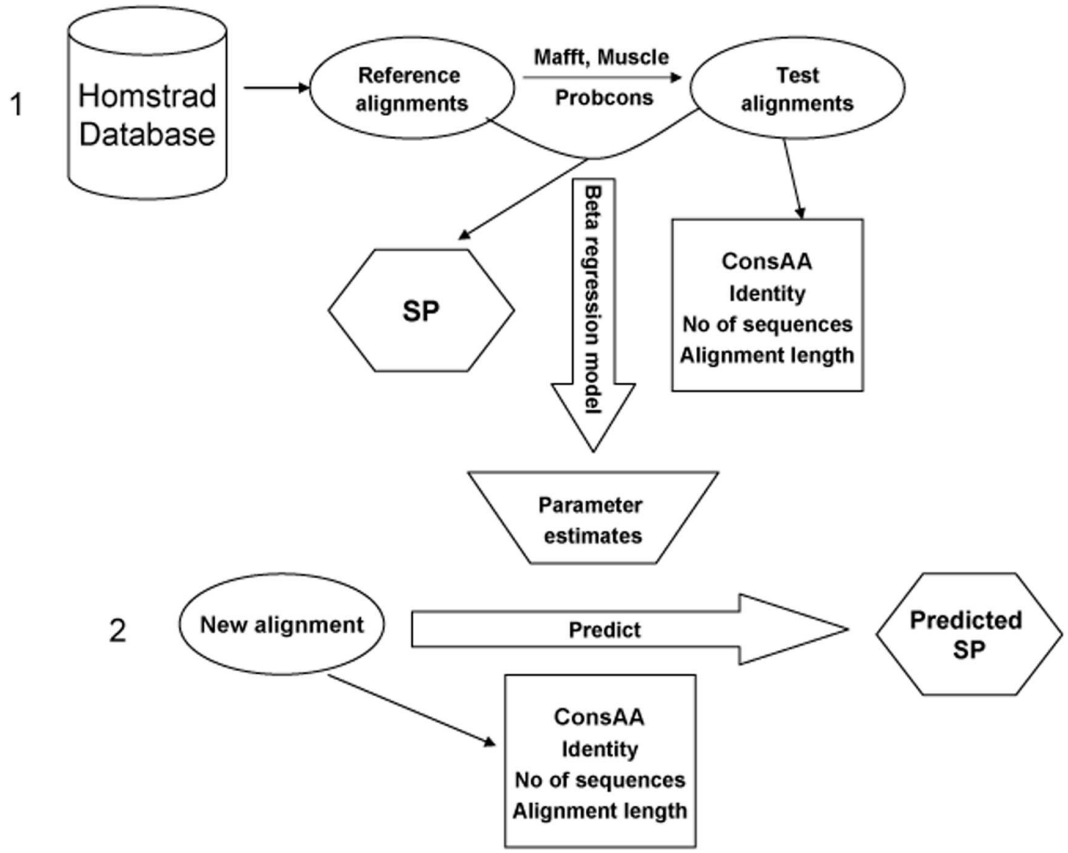


Figure 3.1: Diagram for building and using the alignment quality model. Stage 1 illustrates the steps for calculating the beta regression parameter estimates, and stage 2 their use in predicting the SP score for an unknown alignment.

any other quality score could have been used. The conservation score and other *ab initio* alignment characteristics were used as predictors in the statistical model. We used the ConsAA as the conservation measure, and average of pairwise sequence identity, number of sequences and alignment length as predictors.

The $SP$ score can obtain values on the bounded unit interval [0,1], and hence, the $SP$ score can be assumed to follow a beta distribution. The beta distribution is very flexible; it can have many different shapes on the open unit interval (0,1). Since the $SP$ score can also be zero or one, the endpoints of the interval have to be transformed. We added $\frac{1}{2M}$ to the zero observations and subtracted $\frac{1}{2M}$ when the $SP$ score was equal to one, where $M$ denotes the number of alignments (McMillian and Creelman, 2005).

Now let us assume that the transformed $SP$ score follows a beta distribution

$$f(\widetilde{SP};\omega,\tau) = \frac{\Gamma(\omega+\tau)}{\Gamma(\omega)\Gamma(\tau)}\widetilde{SP}^{\omega-1}(1-\widetilde{SP})^{\tau-1},$$

where $\widetilde{SP} \in (0,1)$, $\Gamma(.)$ is the gamma function and $\omega, \tau > 0$ are the shape parameters of the beta distribution. Let the mean and variance of $\widetilde{SP}$ be

$$\mathbb{E}(\widetilde{SP}) = \frac{\omega}{\omega+\tau} \qquad \text{and} \qquad \mathbb{V}\text{ar}(\widetilde{SP}) = \frac{\omega\tau}{(\omega+\tau)^2(\omega+\tau+1)}. \tag{3.24}$$

In order to transform the beta regression model to the form of the generalized linear model, the parameters $\omega$ and $\tau$ are often transformed as (Ferrari and Cribari-Neto, 2004)

$$\begin{cases} \mu &= \frac{\omega}{\omega+\tau} \\ \\ \phi &= \omega+\tau \end{cases} \quad \Leftrightarrow \quad \begin{aligned} \omega &= \mu\phi \\ \\ \tau &= \phi - \mu\phi. \end{aligned}$$

The mean and variance of $\widetilde{SP}$ can now be written in the form

$$\mathbb{E}(\widetilde{SP}) = \mu \qquad \text{and} \qquad \mathbb{V}\text{ar}(\widetilde{SP}) = \frac{\mu(1-\mu)}{1+\phi},$$

where the new parameters $\mu$ and $\phi$ can be interpreted as location and precision parameters. The location and precision parameters of the transformed beta distribution can now be estimated for alignment $i$ in the same way as in the generalized linear models for many other distributions using the linear equations for alignment $i$

$$\begin{aligned} g(\mu_i) &= \textstyle\sum_{j=0}^{k} x_{ij}\theta_j \qquad \text{and} \\ h(\phi_i) &= -\textstyle\sum_{j=0}^{t} w_{ij}\gamma_j \end{aligned}$$

for the location and precision of $\widetilde{SP}$. Adding a minus sign to the equation of precision turns the interpretation of precision to dispersion (Smithson and Verkuilen, 2006). Thus, $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_k)$ and $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_t)$ are vectors of unknown regression parameters for location and dispersion, and $\boldsymbol{x_i} = (x_{i0}, \ldots, x_{ik})$ and $\boldsymbol{w_i} = (w_{i0}, \ldots, w_{it})$ are known predictors for the location and dispersion models, respectively. Link functions $g$ and $h$ have to be chosen so that the mean can have any real value between zero and one and the variance is always positive. We used a logit and log functions as link functions. In Publication IV, the models for the location and dispersion have been determined by

$$\begin{aligned} g(\mu_i) &= \theta_0 + \theta_1 * \text{ConsAA}_i + \theta_2 * \text{identity}_i \\ &\quad + \theta_3 * \text{number of sequences}_i + \theta_4 * \text{alignment length}_i \\ h(\phi_i) &= -\gamma_0 - \gamma_1 * \text{ConsAA}_i - \gamma_2 * \text{identity}_i \\ &\quad - \gamma_3 * \text{number of sequences}_i - \gamma_4 * \text{alignment length}_i. \end{aligned}$$

The alignment quality score, the predicted average $SP$ score and its dispersion are obtained by replacing $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ with their estimated parameter values, and predictors $\boldsymbol{x_i}$ and $\boldsymbol{w_i}$ with their observed values and by applying the equations

$$\begin{aligned} \mu_i &= \frac{\exp(\sum_{j=0}^{k} x_{ij}\theta_j)}{1+\exp(\sum_{j=0}^{k} x_{ij}\theta_j)} \\ \phi_i &= \exp(-\textstyle\sum_{j=0}^{t} w_{ij}\gamma_j). \end{aligned}$$

The uncertainty of the predicted $SP$ score can be estimated by calculating the prediction intervals for the location parameter

$$\mu_i \pm t_{1-\alpha}\hat{\sigma},$$

where $t_{1-\alpha}$ is the $1-\alpha$'s quantile of the Student's $t$ distribution and $\hat{\sigma}$ the standard error of $\mu_i$, which can be calculated using the delta method (Cox, 1998).

The strength of the use of the beta regression model is that it takes heteroscedasticity into account in the dispersion model. This means that the association of different predictor values with the changes in variability of SP scores can be taken into account in the dispersion model.

# Chapter 4

# Summary of the Publications

The thesis is based on the introduction part and four original publications. This chapter summarizes the contents of the original publications.

## Publication I: Efficient estimation of emission probabilities in profile hidden Markov models

Publication I presents a likelihood-based approach for the estimation of mission probabilities in the profile HMMs. The method overcomes the overfitting problem in the emission probability estimation by explicitly taking into account conservation of alignment.

First, we presented an iterative classification algorithm to divide residues into conserved and unconserved (effective and ineffective) at each alignment position. Then, we introduced the EEP estimation method for the profile HMMs. The underlying assumption of the EEP method is that the unconserved residues follow a background distribution. The Lagrange multipliers method was used to discover the estimation formulae for the emission probabilities. In most alignment positions, the emission probability estimates of conserved residues are ML estimates, whereas the probabilities of unconserved residues are obtained by dividing the remaining probability in proportion to their background probability. In the absence of unconserved residues, the estimates of the EEP method are proportional to their background distribution, and hence, all residues obtain non-zero emission probability estimates.

The performance of the EEP and ML methods was compared in simulations and in a database search of 20 protein families. The results of the database search showed a dramatic reduction in the parameter space when the EEP was used instead of the ML method. The sensitivities were 98% and 97% in the EEP and ML methods, respectively, whereas the specificities were 100% in both methods. The result indicates that the accuracy of the EEP method was slightly better than that of the ML method, although the number of parameters was reduced to 15 % of the original. The performance of the EEP

method was further compared to that of the ML, HMMER and Blast in the database search of the triosephosphate isomerase (TIM) family. The results of this small example show that all three profile HMMs were comparable with each other when the same log-odds score threshold value was used. The specificity of the EEP and ML methods was 100%, whereas the Blast had some false positive findings. The sensitivity of the EEP and ML methods was somewhat lower than that of the Blast.

To conclude, the EEP method flexibly incorporates residue conservation scores into the profile HMMs. The results suggest that this combination provides a potent method for a database search.

## Publication II: Statistical methods for identifying conserved residues in multiple sequence alignment

Publication II continues the work of Publication I in developing measures for identifying conserved residues in MSA. The residue conservation measures predict single conserved residues at aligned position. The predicted residues are assumed to be under strong evolutionary constraints and, hence, to be important for maintaining the 3D structure or function of a protein.

This article introduces two scores for measuring residue conservation, the Bauer and Dunn-Sidak conservation scores. The proposed scores measure whether the estimated residue probabilities differ from their background probabilities. The scores have two major advantages: firstly, they are based on statistical multiple comparison methodology, and hence, the decision for selecting conserved residues is made simultaneously, and secondly, they account for the variability of residue estimates. The two scores differ mainly in the way they incorporate variances into the scores.

The new test procedures were compared with the iterative approach, presented in Publication I, and traditional background- and entropy-based methods in an extensive simulation study. The simulations were used to determine the threshold levels for the conservation scores and to compare the number of false positive (type I error) and negative (type II error) classifications in the different procedures. Additionally, an assessment of the effect of number of sequences on the estimation was made. The practical performance of the methods was evaluated in the alignments of the Src homology 2 (SH2) domain, and three other protein families: globins, ras-like proteins and serine proteases.

The results of the simulation study showed that the false positive rate was very low, especially in the multiple comparison based methods (<3%). The false negative rate was greatest when the conservation level was 20%, but decreased rapidly as the level increased, being less than 1 % with the conservation level of 30%. The classification of conserved residues was heavily dependent on the background distribution. When the conservation level was near to 20%, the false negative rate of rarely occurring amino acids was very

low, but the rate was increased in the more often occurring amino acids. The impact of number of sequences appeared when the number of sequences was decreased from 30 to 20. If the alignment had 30 or more sequences, then the error rates always remained low (<10%).

The conservation analysis of the SH2 domain was performed by comparing the experimental, functional and structural information of the domain with the results of the conservation scores. Functionally, the most important sites of the SH2 domain are those involving phospotyrosine binding and those forming the binding pockets for phosphotyrosine-following residues. All the scores classified as conserved the most important amino acids, whose mutation would be crucial for the protein. The advantage of the multiple comparison methods was that they mostly identified only the functionally important residues as conserved, whereas the other methods also suggested that some additional residues were conserved.

The performance of the Dunn-Sidak score and Sequence logo was further studied using the alignments and entropy/variability classifications of globins, ras-like proteins and serine proteases (Oliveira *et al.*, 2003). The results showed that both the scores detected the most important positions and highly conserved residues (Box1). When the entropy and/or variability were decreased, both the scores identified several residues as conserved. In the highly variable positions with low entropy, only a few residues were identified as conserved (Box33). These findings might be false positives, or they might indicate functional specificity of the protein subfamily (Oliveira *et al.*, 2003).

To summarize, we have presented two multiple comparison -based residue conservation scores. The scores, especially the Dunn-Sidak score, could be used to predict functionally and structurally important residues in the MSAs or used, in conjunction with the EEP method, to estimate emission probabilities in the profile HMMs.

# Publication III: A statistical score for assessing the quality of multiple sequence alignments

Publication III has two objectives. Firstly, it introduces a new conservation score for quantifying the degree of conservation at each alignment position, and secondly, it introduces the AQ measure based on this conservation score.

The positional conservation score is based on a modified Z-score for the sequence profile. The Z-score includes the background distribution of amino acids and the covariance structure of residue probability estimates, but also considers physicochemical properties of residues. The statistical significance of the maxZ score was estimated at each alignment position using the IS method. The novel IS distribution was introduced for this particular problem. The ability of the maxZ score to predict functionally and/or structurally important regions in a given MSA was studied in the SH2 domain, ras-like proteins, peptidase

M13, subtilase and $\beta$-lactamase families. The effect of different scoring matrices on the results was also studied. The results of the maxZ score was compared with that of the IC and MD conservation scores.

The AQ score was derived from the results of the positional conservation score. The positional significance tests were adjusted using a step-up procedure for controling a predefined FDR. The conservation of the whole alignment, ConsAA, was defined as a proportion of conserved residues in MSA, i.e. the proportion of residues occurring in conserved positions. The AQ score quantifies the divergence of the whole alignment conservation between the reference and test alignments. The performance of the AQ score was evaluated by comparing the scorings of seven alignment methods with that of the SP and CS scores in the BAliBASE database. A comprehensive comparison of the alignment methods is reported.

The performance of different scoring schemas in the use of the maxZ score was compared in the SH2 domain and ras-like proteins. The comparison matrices included the Blosum62, Gonnet250 and PAM250 substitution matrices, the identity matrix and a 6-class grouping of amino acids. All the scoring matrices provided similar conservation scores. With the Blosum62, Gonnet250 and PAM250 substitution matrices, the scoring of the highly conserved positions was, however, heavily dependent on the mutability of the most conserved amino acid, i.e. the diagonal value of the substitution matrix.

The conservation scoring of the maxZ score was compared with that of the IC and MD scores in the five protein families. The three scores mostly predicted the functionally or structurally important sites as highly or moderately conserved. The result clearly shows three main differences between the scores. Firstly, the scorings of the maxZ are dependent on the average substitution rate of amino acids, i.e. the diagonal values of the scoring matrices. Therefore, the maxZ score gives different scorings for invariant positions with different amino acids. Secondly, the maxZ score considers only the most conserved residue at the position, and therefore, is not affected by the distribution of other residues. The MD score, on the contrary, might fail to detect some important positions, in which, for instance, subgroups are conserved on different amino acids. Thirdly, the IC score can only be used for ungapped alignments. Nor does the IC score give equal scores for invariant positions.

The quality of the seven alignment methods, Clustal, Dialign, Mafft (L-INS-l and FFT-NS-2 modes), Muscle, Probcons and TCoffee, was assessed using the BAliBASE reference sequence alignment database. The performance of the presented AQ score was compared with that of the frequently used SP and CS scores. The AQ and SP scores were moderately correlated (r=0.53-0.67) in the alignments of the BAliBASE database. Moreover, the median scorings in the six reference sets were very similar between the AQ and SP scores, while those of the CS were considerably lower. The results show clear difference between the AQ and SP scores. Using the AQ score, the L-INS-i strategy of Mafft obtained the best overall result, being the best method in four reference sets,

while the SP score ranked the Probcons as the best overall method. In the results of the AQ score, the Probcons outperformed the other methods in two reference sets. The differences to the Muscle, TCoffee and Clustal were, however, negligible, whereas the FFT-NS-2 strategy of Mafft and Dialign usually scored significantly more poorly than the other methods.

To summarize, the third article introduces a novel approach for quantifying the positional conservation of a given MSA, and uses this score as a core function to formulate an alignment quality measure. The study of the five protein families shows that the positional conservation score is able to predict the functionally and structurally important sites of a protein. An alignment quality assessment of seven alignment methods and comparison with the other reference-based scoring methods suggest that the presented method is reliable for assessing the quality of MSAs, when the reference alignment is available.

# Publication IV: Model-based prediction of sequence alignment quality

Publication IV generalizes the AQ score presented in Publication III for the situations where no reference alignment is available. The novel score is a model-based prediction of the alignment quality and is based on measuring the conservation of reference alignments using the whole alignment conservation score, ConsAA, presented in Publication III.

The reference MSAs with different similarities, alignment lengths and number of sequences were obtained from the Homstrad database. The reference sequences were realigned with Mafft, Muscle and Probcons alignment programs and their quality was measured by the SP score. The beta regression model was fitted for the SP score using conservation level, identity, number of sequences and alignment length as predictors in the model. The new quality measure, called the predicted SP score, uses the parameter estimates of the beta regression model to predict the quality of a given global MSA. We tested the novel quality score on the structural alignments in the test sets of the Homstrad and SABmark databases by comparing the predicted SP with the $CS$ and correct $SP$ scores in the Homstrad and median $f_d$ and $f_m$ scores in the SABmark database. Additionally, we compared the performance of the predicted SP score with that of the MOS, NiRMSD and NorMD quality scores.

The results suggest that the predicted SP was highly correlated with the correct quality scores in the test set of the Homstsrad database ($r_{SP}$=0.65 and $r_{CS} = 0.60$, mean of Mafft, Muscle and Probcons alignments) and in the SABmark database ($r_{f_d} = 0.73$ and $r_{f_m} = 0.72$). Among the other quality scores, the MOS score had a very strong relationship with the correct alignment quality scores ($r_{SP} = 0.87$ and $r_{CS} = 079$ in Homstrad and $r_{f_d} = r_{f_m} = 0.83$ in SABmark databases). The NiRMSD and NorMD scored slightly more poorly than the other two methods in the Homstrad database, while the results of all the

approaches were very similar in the SABmark database. In the Homstrad database, the agreement of the correct and predicted SP scores and the MOS score was measured by the mean square error (MSE) rate. All three scores were within the 1 % mean square difference from each other, indicating very low divergence among the three scores. The ability of the quality scores to distinquish correct alignments from alignments with badly aligned or unrelated sequences was studied by adding random sequences to the Homstrad alignments with 5 to 10 sequences. The results suggest that the predicted SP score decreases in the same proportion as the number of added random sequences, whereas the MOS score overcorrects the influence of unrelated or badly aligned sequencs.

To conclude, Publication IV uses the statistical prediction model, together with the conservation scoring method, for the prediction of alignment quality. The results suggest that the quality of any global MSA can be evaluated by the novel model-based approach.

# Contribution of the author

The ideas of Publications I, II and III were joint work. In Publications II and III, the author participated in developing the theoretical approaches. The design of computational experiments was planned by the author with the help of the other co-authors. The author implemented the methods and carried out all the simulations, computational experiments and statistical analyses. Publications I, II and III were mostly written by the author. The co-authors commented and helped by revising the manuscript. Publication IV was initiated by the author. The author implemented the methods, and planned and carried out the computational experiments and statistical analyses. The author was the principal writer of Publication IV. The co-authors contributed by commenting on and revising the manuscript.

# Chapter 5

# Discussion

This study introduces methods for predicting conserved residues or positions in a given MSA. These measures were used as the core in developing further methods. The measures for quantifying alignment quality were derived from the positional conservation scores, whereas the estimation method for profile HMMs was based on the methods identifying individual residue conservation.

We have developed two statistical approaches for scoring residue conservation in an alignment position. We have shown that these scores are useful when they are used in conjunction with profile HMMs. The method proposes an alternative to the frequently used Bayesian approach (Sjölander *et al.*, 1996). It is a noteworthy solution to the question of how the overfitting problem could be handled in the profile HMMs, and more broadly speaking, how the explicit prediction of conserved residues could be used in the comparative genomics methods.

The positional conservation score uses as a test statistic a maximum Z value, which has been calculated for the sequence profile. The test statistic fulfills most of the criteria defined for a good conservation score (Valdar, 2002). The maxZ score is simple and maps an alignment position into the bounded interval of real numbers, it accounts for the relative frequencies and stereochemical properties of residues as well as gaps. The sequence weighting is the only requirement that was not fulfilled. The ability to normalize against redundancy is an important characteristic of a conservation score and should always be accounted for. A variety of weighting methods related to genetic distance between sequences or symbol entropy, for instance, have been proposed for conservation and profile analysis (Vingron and Argos, 1989; Durbin *et al.*, 1998; Valdar, 2002). The weighting of relative frequencies of residues in the maxZ score could be done by applying any appropriate weighting method in the same way as described by Pei and Grishin (2001).

Traditionally, many conservation scores are simple and have not been presented in the form of a test statistic or probability model, which would enable statistical inference, such as calculation of significance tests, confidence intervals or posterior probabilities. This might have arisen from the fact that they often require either complicated sampling

or asymptotical approximations, whereupon the methods might become time-consuming. During the last few years, more and more statistical inference has been incorporated into the conservation analysis (Mirny and Gelfand, 2002; Pupko *et al.*, 2002; Mayrose *et al.*, 2004; Marttinen *et al.*, 2006; Pei *et al.*, 2006; Fischer *et al.*, 2008). We have used the IS method to estimate the significance of the maxZ score. The method has been proved to provide reliable results, but calculation of $p$ values for large alignments might be time-consuming. The difference from another sampling method presented by Mirny and Gelfand (2002) is that our method does not use an evolutionary tree, but the sampling is controlled by the IS distribution. Recently, conservation has been calculated for clusters of alignment positions (Panchenko *et al.*, 2004; Capra and Singh, 2007; Fischer *et al.*, 2008). The results suggest that the prediction power of functional sites could be improved if the consecutive, interacting or co-evolving alignment positions were simultaneously considered.

Conservation has been proved to be a valuable measure for alignment quality assessment (Pei and Grishin, 2001; Thompson *et al.*, 2001). The reference-independent alignment quality measures have been traditionally formulated by summing over the positional conservation scores (Valdar, 2002). Our approach also uses conservation, but accounts for the effect of making multiple tests and the effect of gaps. Alternative approaches usually require structural information or that parallel alignments are available (Armougom *et al.*, 2006; Lassmann and Sonnhammer, 2005). Our model-based quality score does not need any additional information on tested sequences. It uses conservation and other *ab initio* sequence information available in the existing reference alignments to formulate a prediction model, and exploits this information in the prediction of the quality of a given alignment. The difference from the consistency-based methods is that in our prediction method, information on other alignments is computed only once, while in the consistency-based methods, the alternative alignments have to be constructed for each alignment quality assessment. Furthermore, the uncertainty of the prediction can be estimated using the statistical approach.

The model-based prediction assumes that the reference alignments are biologically 'correct', and that they are a comprehensive set of a type of alignments one would like to predict. The limitation inherent in the reference alignment databases is that different structural alignment methods may produce somewhat different alignments (Goldsmith-Fischman and Honig, 2003; Notredame, 2007). In the Homstrad database, which we used, the limitation was that the structural alignments have been supplemented using ClustalW, and hence, the alignments are not necessarily structurally correct. The model-based prediction method could be further developed by applying a more comprehensive reference alignment database. Furthermore, cross-checking of the effect of different structural alignment databases on quality assessment could be useful in further improving the prediction power.

Recent advances in protein functional annotation have been made by combining infor-

mation from different sources (Gutteridge *et al.*, 2003; Panchenko *et al.*, 2004; Petrova and Wu, 2006). Our prediction method is flexible for testing different predictors and conservation scoring methods. In future work, the alignment quality prediction could also take into account factors within the protein sequence, such as secondary structures or solvent accessibility. The performance of different conservation scores could also be tested.

# Chapter 6

# Conclusion

The thesis has introduced statistical measures for quantifying conservation in multiple protein sequence alignments. These measures have been used as a core in developing further methods for comparative genomics.

The main objective of this study was to develop both biologically and statistically relevant methods for measuring conservation of MSAs, and to use these measures for assessing the quality of protein sequence alignments. The positional conservation score uses a Z statistic of the sequence profile (Publication III). The Z statistic was used to test whether the amino acids follow the underlying background distribution. The hypothesis testing has been carried out by the IS procedure, which has been particularly tailored for calculating positional significance levels. The significance level is our new positional conservation score. The applicability of this score has been carefully tested in several protein families. The result indicates that the score detects key functional positions, such as the catalytic residues or ligand-binding sites of proteins.

The positional conservation score was exploited in the statistical prediction model for assessing the quality of MSAs (Publication IV). The key idea of our alignment quality score is that the conservation information available in the reference alignment databases was incorporated into the statistical model. In this way, the information about known proteins in the whole fold space was available for the prediction model. Conservation of alignments was defined using our positional conservation scores to calculate the proportion of conserved residues in the alignment. The reference alignments were realigned using three frequently used alignment programs, and the SP score was used to measure the quality of the realigned sequences. The beta regression model was built for the SP score using conservation level and other *ab initio* alignment characteristics as predictors. The estimated model parameters of the prediction model can be used to predict alignment quality. The comparisons of the predicted and correct quality scores show high correlation and low MSE between the two scores. The results suggest that our method is reliable for assessing the quality of any global MSA.

The secondary aim of the study was to develop residue conservation scores for iden-

tifying conserved residues in an alignment position and, by means of these scores, to develop a method for estimating emission probabilities in the profile HMMs (Publications I-II). We developed two residue conservation scores based on statistical hypothesis testing (Publication II). A careful examination of these scores in the SH2 domain shows that the scores are capable of identifying functionally and structurally important residues in MSA.

One of the constructed residue conservation scores was used as a preliminary stage in the emission probability estimation method for profile HMMs (Publication I). The EEP method developed overcomes the overfitting problem in the estimation of emission probabilities. The results of the EEP method in the database search indicate that the novel estimation method dramatically reduces the average number of estimated emission parameters, while the accuracy was maintained at the same level. The EEP method, in conjunction with some residue conservation score, provides a flexible method for detecting remote homologues to the protein families, for instance.

# Acknowledgements

The studies have been carried out at the Department of Statistics of the University of Turku and at the Biotechnology and Food research, MTT Agrifood Research Finland.

I am sincerely grateful to Professor Esa Uusipaikka for teaching me the secrets of statistics and for supporting and supervising the thesis. I am indebted to Professor Mauno Vihinen for introducing me to the fascinating field of bioinformatics, for guidance in biochemistry and for patiently supervising me throughout these years. I also thank him for giving me the opportunity to use the computer facilities of the Institute of Medical Technology. I owe special thanks to Docent Tero Aittokallio for his scientific guidance and cooperation. He has been easy to work with and he has helped me especially in scientific writing without sparing his time. I am very grateful to Pentti Riikonen, MS.c, who modified the MultiDisp software used in this thesis and made the illustration of the alignments easier. He always supported me in the use of the Linux cluster and gave his time unsparingly despite his other duties.

I wish to thank the staff of the Department of Statistics, with whom it was a pleasure to work. They introduced me to the field of statistics, made me part of the department and always encouraged me in my work. I owe special thanks to Kalle Lertola, Ph.Lis, for his help in Linux and other computational issues.

I would like to thank Eeva-Liisa Ryhänen, director of Biotechnology and Food Research, and team managers, Vesa Joutsjoki and Johanna Vilkki, for giving me the possibility to finalize this thesis at the MTT. They have always encouraged me with their positive attitude towards this thesis. I wish to express my deepest gratitude to my colleagues at the MTT. They have taught me what genomics research really means and have always patiently answered my questions. It has been a pleasure to work in this inspiring atmosphere with lots of fun.

I wish to thank to my official pre-examiners, Chief Research Scientist Jaakko Hollmén and Professor Timo Koski, for their positive statements, their constructive criticism and their suggestions to improve the manuscript of this thesis.

Very special thanks go to all my friends for discussions about all possible matters in the world. I owe special thanks to my sisters Sari and Marjo and their families for relaxing moments and support. My warmest thanks are due to my mother and father for caring for and supporting me in everything throughout the years. To them I also dedicate this

dissertation. Thank you for having confidence in me. My dearest thanks belong to Eero for his never failing support, encouragement and patience during many years of work. Thank you for keeping me sane.

The financial support from the Graduate School of Computational Biology, Bioinformatics and Biometry (ComBi) and The Finnish Academy is gratefully acknowledged.

Jokioinen, October 2008

Virpi Ahola

# Bibliography


Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389–3402.

Armon, A., Graur, D., and Ben-Tal, N. (2001). ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Biol*, **307**, 447–463.

Armougom, F., Moretti, S., Keduas, V., and Notredame, C. (2006). The iRMSD: a local measure of sequence alignment accuracy using structural information. *Bioinformatics*, **22**(14), e35–39.

Bahr, A., Thompson, J. D., Thierry, J. C., and Poch, O. (2001). BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Res*, **29**, 323–326.

Batzoglou, S. (2005). The many faces of sequence alignment. *Brief Bioinform*, **6**(1), 6–22.

Bauer, P., Ptscher, B. M., and Hackl, P. (1988). Model selection by multiple test procedures. *Statistics*, **19**, 39–44.

Beiko, R. G., Chan, C. X., and Ragan, M. A. (2005). A word-oriented approach to alignment validation. *Bioinformatics*, **21**(10), 2230–2239.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J R Stat Soc Ser B*, **57**, 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann Stat*, **29**, 1165–1188.

Benner, S. A., Cohen, M. A., and Gonnet, G. H. (1994). Amino-acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Eng*, **7**, 1323–1332.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*, **31**, 365–370.

Bordner, A. J. and Abagyan, R. (2005). Statistical analysis and prediction of protein-protein interfaces. *Proteins*, **60**, 353–366.

Caffrey, D. R., Somaroo, S., Hughes, J. D., Mintseris, J., and Huang, E. S. (2004). Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci*, **13**, 190–202.

Capra, J. A. and Singh, M. (2007). Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875–1882.

Carrillo, H. and Lipman, D. (1988). The multiple sequence alignment problem in biology. *SIAM J Appl Math*, **48**, 1073–1082.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*, **17**, 540–552.

Cline, M., Hughey, R., and Karplus, K. (2002). Predicting reliable regions in protein sequence alignments. *Bioinformatics*, **18**(2), 306–314.

Cox, C. (1998). Delta Method. In P. Armitage and T. Colton, editors, *Encyclopedia of Biostatistics*, pages 1125–1127. John Wiley & Sons, Inc., New York.

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978). *A model of evolutionary change in proteins*, volume 5 of *Atlas of protein sequence and structure*, chapter 22, pages 345–358. National biomedical research foundation, Washington DC.

del Sol Mesa, A., Pazos, F., and Valencia, A. (2003). Automatic methods for predicting functionally important residues. *J Mol Biol*, **326**, 1289–1302.

Do, C. B., Mahabhashyam, M. S., Brudno, M., and Batzoglou, S. (2005). ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res*, **15**, 330–340.

Domingues, F. S., Lackner, P., Andreeva, A., and Sippl, M. J. (2000). Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *J Mol Biol*, **297**, 1003–1013.

Dudbridge, F. and Koeleman, B. P. (2004). Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *Am J Hum Genet*, **75**, 424–435.

Dunn, O. J. (1958). Estimation of the means of dependent variables. *Ann Math Statist*, **29**, 1095–1111.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis; Probabilistic Models of Proteins and Nucleic*. Cambridge Univeristy Press, Cambridge.

Eddy, S. R. (1998). Profile hidden markov models. *Bioinformatics*, **14**, 755–763.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32**, 1792–1797.

Edgar, R. C. and Batzoglou, S. (2006). Multiple sequence alignment. *Curr Opin Struct Biol*, **16**, 368–373.

Ferrari, S. L. P. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *J Appl Stat*, **10**, 1–18.

Fischer, J. D., Mayer, C. E., and Soding, J. (2008). Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics*, **24**, 613–620.

Goldsmith-Fischman, S. and Honig, B. (2003). Structural genomics: Computational methods for structure analysis. *Protein Sci*, **12**, 1813–1821.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *J Mol Biol*, **162**(3), 705–708.

Gotoh, O. (1990). Consistency of optimal sequence alignments. *Bull Math Biol*, **52**, 509–525.

Gribskov, M., Mclachlan, A. D., and Eisenberg, D. (1987). Profile analysis - detection of distantly related proteins. *Proc Natl Acad Sci U S A*, **84**, 4355–4358.

Griffiths-Jones, S. and Bateman, A. (2002). The use of structure information to increase alignment accuracy does not aid homologue detection with profile HMMs. *Bioinformatics*, **18**, 1243–1249.

Gutteridge, A., Bartlett, G. J., and Thornton, J. M. (2003). Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J Mol Biol*, **330**, 719–734.

Hannenhalli, S. S. and Russell, R. B. (2000). Analysis and prediction of functional subtypes from protein sequence alignments. *J Mol Biol*, **303**, 61–76.

Hardison, R. C. (2003). Comparative genomics. *PLoS Biol*, **1**, E58.

Henikoff, S. and Henikoff, J. G. (1993). Performance evaluation of amino-acid substitution matrices. *Proteins*, **17**, 49–61.

Henikoff, S. and Henikoff, J. G. (1994). Position-based sequence weights. *J Mol Biol*, **243**, 574–578.

Hertz, G. Z. and Stormo, G. D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. John Wiley & Sons, New York.

Holm, L. and Sander, C. (1996). Mapping the protein universe. *Science*, **273**, 595–603.

Holm, L. and Sander, C. (1998). Dictionary of recurrent domains in protein structures. *Proteins*, **33**, 88–96.

Jaroszewski, L., Rychlewski, L., and Godzik, A. (2000). Improving the quality of twilight-zone alignments. *Protein Sci*, **9**, 1487–1496.

Jukes, T. H. and Kimura, M. (1984). Evolutionary constraints and the neutral theory. *J Mol Evol*, **21**, 90–92.

Kalinina, O. V., Mironov, A. A., Gelfand, M. S., and Rakhmaninova, A. B. (2004a). Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein Sci*, **13**, 443–456.

Kalinina, O. V., Novichkov, P. S., Mironov, A. A., Gelfand, M. S., and Rakhmaninova, A. B. (2004b). SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucleic Acids Res*, **32**, W424–428.

Karlin, S. and Brocchieri, L. (1996). Evolutionary conservation of RecA genes in relation to protein structure and function. *J Bacteriol*, **178**, 1881–1894.

Karplus, K. and Hu, B. (2001). Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set. *Bioinformatics*, **17**, 713–720.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res*, **30**, 3059–3066.

Katoh, K., Kuma, K., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*, **33**, 511–518.

Keich, U. (2005). sFFT: a faster accurate computation of the p-value of the entropy score. *J Comput Biol*, **12**, 416–430.

Koski, T. (1999). Hidden Markov models and probabilistic learning with applications to bioinformatics. Volume c10 of lecture notes, University of Turku, Institute of Applied Mathematics.

Krogh, A., Brown, M., Mian, I. S., Sjölander, K., and Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol*, **235**, 1501–1531.

Landan, G. and Graur, D. (2007). Head or tails: a simple reliability check for multiple sequence alignments. *Mol Biol Evol*, **24**, 1380–1383.

Landan, G. and Graur, D. (2008). Local reliability measures from sets of co-optimal multiple sequence alignments. *Pac Symp Biocomput*, **13**, 15–24.

Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T., and Ben-Tal, N. (2005). ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res*, **33**, W299–302.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., and International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lassmann, T. and Sonnhammer, E. L. L. (2002). Quality assessment of multiple alignment programs. *FEBS Lett*, **529**, 126–130.

Lassmann, T. and Sonnhammer, E. L. L. (2005). Automatic assessment of alignment quality. *Nucleic Acids Res*, **33**, 7120–7128.

Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol*, **257**, 342–358.

Lifshitz, E. M. and Pitaevskii, L. P. (1980). *Statistical physics*. Pergamon Press, Oxford, UK.

Lin, J. (1991). Divergence measures based on Shannon entropy. *IEEE Trans Inf Theory*, **37**, 145–151.

Livingstone, C. D. and Barton, G. J. (1993). Protein-sequence alignments - a strategy for the hierarchical analysis of residue conservation. *Comput Appl Biosci*, **9**, 745–756.

Lockless, S. W. and Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295–299.

Luenberger, D. G. (1984). *Linear and Nonlinear Programming*. Addison-Wesley, Reading, MA, 2nd edn. edition.

Magliery, T. J. and Regan, L. (2005). Sequence variation in ligand binding sites in proteins. *BMC Bioinformatics*, **6**, 240.

Marttinen, P., Corander, J., Toronen, P., and Holm, L. (2006). Bayesian search of functionally divergent protein subgroups and their function specific residues. *Bioinformatics*, **22**, 2466–2474.

Mayrose, I., Graur, D., Ben-Tal, N., and Pupko, T. (2004). Comparison of site-specific rate-inference methods for protein sequences: empirical bayesian methods are superior. *Mol Biol Evol*, **21**.

McMillian, N. A. and Creelman, C. D. (2005). *Detection theory; a user's guide*. Lawrence Erlebaum Associates, Mahwah, NJ.

Menke, M., Berger, B., and Cowen, L. (2008). Matt: local flexibility aids protein multiple structure alignment. *PLoS Comput Biol*, **4**, e10.

Mevissen, H. T. and Vingron, M. (1996). Quantifying the local reliability of a sequence alignment. *Protein Eng*, **9**, 127–132.

Mihalek, I., Res, I., and Lichtarge, O. (2007). Background frequencies for residue variability estimates: BLOSUM revisited. *BMC Bioinformatics*, **8**, 488.

Mirny, L. and Shakhnovich, E. (2001). Evolutionary conservation of the folding nucleus. *J Mol Biol*, **308**, 123–129.

Mirny, L. A. and Gelfand, M. S. (2002). Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J Mol Biol*, **321**, 7–20.

Mirny, L. A. and Shakhnovich, E. I. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J Mol Biol*, **291**, 177–196.

Mizuguchi, K., Deane, C. M., Blundell, T. L., and Overington, J. P. (1998). HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci*, **7**, 2469–2471.

Myers, E. W. and Miller, W. (1988). Optimal alignments in linear space. *Comput Appl Biosci*, **4**(1), 11–17.

Nagarajan, N., Jones, N., and Keich, U. (2005). Computing the P-value of the information content from an alignment of multiple sequences. *Bioinformatics*, **21**, i311–318.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, **48**, 443–453.

Notredame, C. (2007). Recent evolutions of multiple sequence alignment algorithms. *PLoS Comput Biol*, **3**, e123.

Oliveira, L., Paiva, P. B., Paiva, A. C., and Vriend, G. (2003). Identification of functionally conserved residues with the use of entropy-variability plots. *Proteins*, **52**, 544–552.

O'Sullivan, O., Zehnder, M., Higgins, D., Bucher, P., Grosdidier, A., and Notredame, C. (2003). APDB: a novel measure for benchmarking sequence alignment methods without reference alignments. *Bioinformatics*, **19**, i215–221.

Panchenko, A. R., Kondrashov, F., and Bryant, S. (2004). Prediction of functional sites by analysis of sequence and structure conservation. *Protein Sci*, **13**, 884–892.

Pei, J. and Grishin, N. V. (2001). AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, **17**, 700–712.

Pei, J. and Grishin, N. V. (2006). MUMMALS: multiple sequence alignment improved by using hidden markov models with local structural information. *Nucleic Acids Res*, **34**(16), 4364–4374.

Pei, J. and Grishin, N. V. (2007). PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics*, **23**(7), 802–808.

Pei, J., Cai, W., Kinch, L. N., and Grishin, N. V. (2006). Prediction of functional specificity determinants from protein sequences using log-likelihood ratios. *Bioinformatics*, **22**, 164–171.

Petrova, N. V. and Wu, C. H. (2006). Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. *BMC Bioinformatics*, **7**, 312.

Pilpel, Y. and Lancet, D. (1999). The variable and conserved interfaces of modeled olfactory receptor proteins. *Protein Sci*, **8**, 969–977.

Pirovano, W., Feenstra, K. A., and Heringa, J. (2006). Sequence comparison by sequence harmony identifies subtype-specific functional sites. *Nucleic Acids Res*, **34**, 6540–6548.

Pupko, T., Bell, R. E., Mayrose, I., Glaser, F., and Ben-Tal, N. (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18 Suppl 1**, S71–7.

Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Ann Math Statist*, **24**, 220–238.

Rubin, D. B. (1988). Using the SIR algorithm to simulate posterior distributions. In M. H. Bernardo, K. M. an DeGroot, C. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics 3*, pages 395–402. Oxford University Press, Oxford UK.

Rychlewski, L., Fischer, D., and Elofsson, A. (2003). LiveBench-6: large-scale automated evaluation of protein structure prediction servers. *Proteins*, **53 Suppl 6**, 542–547.

Saccone, C. and Pesole, G. (2003). *Handbook of Comparative Genomics: Principles and Methodology*. John Wiley & Sons, Inc, Hoboken, New Jersey.

Sadreyev, R. and Grishin, N. (2003). COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J Mol Biol*, **326**, 317–336.

Sadreyev, R. I. and Grishin, N. V. (2004). Estimates of statistical significance for comparison of individual positions in multiple sequence alignments. *BMC Bioinformatics*, **5**, 106.

Sander, C. and Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.

Sauder, J. M., Arthur, J. W., and Dunbrack, R. L. (2000). Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins*, **40**, 6–22.

Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, **18**, 6097–6100.

Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *J Mol Biol*, **188**, 415–431.

Schultz, J., Milpetz, F., Bork, P., and Ponting, C. P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A*, **95**, 5857–5864.

Shannon, C. D. (1948). A mathematical theory of communication. Technical Report 27, Bell System Tech J.

Shenkin, P. S., Erman, B., and Mastrandrea, L. D. (1991). Information-theoretical entropy as a measure of sequence variability. *Proteins*, **11**, 297–313.

Sibbald, P. R. and Argos, P. (1990). Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *J Mol Biol*, **216**, 813–818.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *J Am Stat Assoc*, **62**, 626–633.

Sjölander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I. S., and Haussler, D. (1996). Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput Appl Biosci*, **12**, 327–345.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol*, **147**, 195–197.

Smithson, M. and Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychol Methods*, **11**, 54–71.

Stormo, G. D. and Fields, D. S. (1998). Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem Sci*, **23**, 109–113.

Taylor, W. R. (1986). The classification of amino-acid conservation. *J Theor Biol*, **119**, 205.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22**, 4673–4680.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*, **25**, 4876–4882.

Thompson, J. D., Plewniak, F., and Poch, O. (1999). A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res*, **27**, 2682–2690.

Thompson, J. D., Plewniak, F., Ripp, R., Thierry, J. C., and Poch, O. (2001). Towards a reliable objective function for multiple sequence alignments. *J Mol Biol*, **314**, 937–951.

Tomovic, A. and Oakeley, E. J. (2007). Quality estimation of multiple sequence alignments by Bayesian hypothesis testing. *Bioinformatics*, **23**, 2488–2490.

Valdar, W. S. and Thornton, J. M. (2001). Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins*, **42**, 108–124.

Valdar, W. S. J. (2002). Scoring residue conservation. *Proteins*, **48**, 227–241.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, Y. M., Evans, C. A., Holt, R. A., and *et al.* (2001). The sequence of the human genome. *Science*, **291**, 1304–1351.

Vingron, M. (1996). Near-optimal sequence alignment. *Curr Opin Struct Biol*, **6**, 346–352.

Vingron, M. and Argos, P. (1989). A fast and sensitive multiple sequence alignment algorithm. *Comput Appl Biosci*, **5**, 115–121.

Vingron, M. and Argos, P. (1991). Motif recognition and alignment for many sequences by comparison of dot-matrices. *J Mol Biol*, **218**, 33–43.

Vingron, M. and Sibbald, P. R. (1993). Weighting in sequence space: a comparison of methods in terms of generalized sequences. *Proc Natl Acad Sci U S A*, **90**, 8777–8781.

Wang, L. and Jiang, T. (1994). On the complexity of multiple sequence alignment. *J Comput Biol*, **1**, 337–348.

Wass, M. N. and Sternberg, M. J. (2008). ConFunc–functional annotation in the twilight zone. *Bioinformatics*, **24**, 798–806.

Waterman, M. S. (1995). *Introduction to Computational Biology: Maps, Sequences and Genomes; Interdisciplinary Statistics*. Chapman and Hall, London, UK.

Williamson, R. M. (1995). Information theory analysis of the relationship between primary sequence structure and ligand recognition among a class of facilitated transporters. *J Theor Biol*, **174**, 179–188.

Wu, T. T. and Kabat, E. A. (1970). An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J Exp Med*, **132**(2), 211–250.

Yona, G. and Levitt, M. (2002). Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J Mol Biol*, **315**, 1257–1275.

Zemla, A., Venclovas, C., Moult, J., and Fidelis, K. (1999). Processing and analysis of CASP3 protein structure predictions. *Proteins*, **Suppl 3**, 22–29.

Zhang, Y. and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.

Zvelebil, M. J., Barton, G. J., Taylor, W. R., and Sternberg, M. J. E. (1987). Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J Mol Biol*, **195**, 957–961.