# EVALUATION OF SEQUENCE ALIGNMENTS OF DISTANTLY RELATED SEQUENCE PAIRS WITH RESPECT TO STRUCTURAL SIMILARITY

**AYSAM GUERLER** guerler@chemie.fu-berlin.de
**ERNST-WALTER KNAPP** knapp@chemie.fu-berlin.de

*Freie Universität Berlin, Institut für Chemie und Biochemie*
*Takustr. 6, 14195, Berlin-Dahlem, Germany*

We evaluate the performance of common substitution matrices with respect to structural similarities. For this purpose, we apply an all-versus-all pairwise sequence alignment on the ASTRAL40 dataset, consisting of 7290 entries with a pairwise sequence identity of at most 40%. Afterwards, we compare the 100 highest scoring sequence alignments to their corresponding structural alignments, which we obtain from our structure alignment database. Our database consists of about 18.6 million pairwise entries. We calculated these alignments by applying the current version of GANGSTA, our non-sequential structural alignment tool, on about 26 million pairs. The results illustrate the difficulty of homology based protein structure prediction in cases of low sequence similarity. Further, the large fraction of structurally similar proteins in the ASTRAL40 dataset is quantitatively measured. Thereby, this investigation yields a new perspective on the topic of sequence and structure relation. Hence, our finding is a large-scale quality measure for any sequence based method, which aims to detect structural similarities.

*Keywords*: sequence alignment, protein structure prediction, substitution matrix, database comparison

## 1. Introduction

Protein sequence alignment plays a key role in the investigation of protein functionality [4, 12]. The protein sequence determines the structure and through it the protein's function. Similar sequences often share similar structures. However, the opposite is not the case since similar structures can be encoded by dissimilar sequences [11]. Shakhnovich et al. analysed this issue in terms of a "free energy landscape" in sequence space. During evolution of a protein sequence, amino acid residues are deleted, inserted or replaced by others. This process of sequence altering can lead to cross "barriers" and to seed new local minima in sequence space. In some cases the new minima correspond to similar structures, which are conservative with respect to the protein's function. Here, the mutations in sequence do not cause an unsatisfactory structural change at functionally relevant protein sites. Hence, the structural conservation for specific sites is higher than the sequential conservation. These properties of sequence and structure coherence can lead to difficulties in the application of common sequence alignment methods. Current strategies are based on substitution matrices, which are applied for measuring sequence similarities [8, 9]. However, the most common substitution matrices like PAM (point

accepted mutations) [2] and BLOSUM (blocks substitution matrix) [3] are based on preliminary sequence alignments of mainly similar protein sequence sections. Therefore, they are biased towards sequentially conserved regions. Despite these difficulties, many protein structure prediction methods apply a preliminary homology search in sequence databases [13]. In general, this process consists of four steps. First, the sequence homologue for a known sequence but unknown structure is searched. Then, both sequences are aligned. Afterwards, the backbone positions of the known structure are transferred to the other, based on the residue pairing on sequence level. Finally, the sidechains are added to the model. Certainly, this is a very effective and promising approach in case of high sequence similarity. Unfortunately, this search for structural properties based on sequence analysis becomes questionable when applied on distantly related sequences.

Sauder et al. performed an analysis with the structural alignment tool CE [9], the sequence alignment tool BLAST [8] and others. The quality of these methods on distantly related sequences is not known, yet [13]. In contrast to the current work, they measured the sequence alignment performance on sequence, instead of structure level. Further, the employed dataset was smaller. Sitbon et al. also applied an integrated analysis on sequence and structure information to determine the conservation of residues with respect to secondary structure elements. They found that helices and turns are underrepresented in conserved regions, in contrast to sheets, which are overrepresented. With respect to loops, they detected similar amounts in conserved and unconserved regions [4]. Further, Domingues et al. set up a benchmark protocol for sequence alignment algorithms with respect to threading. Thereby, they differ between local and global sequence alignment approaches. They claim that the alignments constructed with a combination of sequence alignment, atom pair interactions and protein solvent interactions are the most accurate. They evaluated the alignment quality by comparing the residue pairings between structure and sequence alignment results. Thereby, the local and global alignments performed quite similar. Additionally, they claim that the amount of incorrectly aligned residues with respect to the structural alignments is high for all algorithms [12].

In this paper, we evaluate the performance of common substitution matrices in detecting structural similarities. Therefore, we employ the ASTRAL40 dataset. The set consists of 7290 protein chains, which share less than 40% sequence identity. The sequences and the structures are available online [6]. In a first step, we align the sequences of each ASTRAL40 entry on the complete sequence set with FASTA [7]. Thereby, we retrieve the list of the 100 highest ranked protein pairs for each entry (as SCOP 1.69 codes [6]). Then, we select the corresponding structural scores (SC) of these pairs from our structure alignment database (SD). This procedure is applied in combination with BLOSUM50, BLOSUM62 and PAM120. The resulting structural scores (SC) are plotted in figure 6. Additionally, the 100 highest structural scores (SC) for each ASTRAL40 entry are selected from our structure alignment database and plotted as reference, respectively as upper performance limit. Since, our structure alignment method

is able to detect non-sequential similarities between two protein structures we additionally plotted the sequential structure alignments separately.

## 2. Methods

### 2.1. *Sequence alignment*

Currently, the most popular sequence alignment tools are FASTA [7] and BLAST [8]. Both employ a set of substitution matrices to score the sequence alignment results. The most commonly used matrices are PAM and BLOSUM. Both matrix types are calculated on the basis of prior gapless sequence alignments. Initially the observed substitution frequencies $q_{ij}$ are obtained by counting all of the aligned amino acid pairs *ij*. Further, the occurrence frequency $p_i$ of each amino acid *i* is calculated. Finally, the log-odds ratio of the substitution frequencies against the background distribution of the amino acids is evaluated for each pair. The score $s_{ij}$ is then written as

$$s_{ij} = (\ln \frac{q_{ij}}{p_i p_j}) / \lambda \tag{1}$$

with lambda [5] the scaling parameter. This procedure yields a symmetrical 20x20 substitution matrix. Sequence alignments are scored as summation of the $s_{ij}$ values, corresponding to the aligned amino acid pairs *ij*. Since the scores employ a logarithmic scale, this is equivalent to the multiplication of amino acid occurrence probabilities against the background distribution under the independence assumption [5].

### 2.2. *Structure alignment scoring*

The basis of the structure alignment evaluation is the structure alignment score (SAS), which has been proposed by Kolodny et al. [10]. This score weights the RMSD of the C-alpha atoms by the number of structurally aligned residues $N_{aligned}$ (see equation 2).

$$SAS = \frac{RMSD * 100}{N_{aligned}} \tag{2}$$

Linear scaling yields the structural score (SC), which we define in this investigation to evaluate the structural similarity between two proteins. The structural score is defined as

$$SC = 100 - 2 * SAS \tag{3}$$

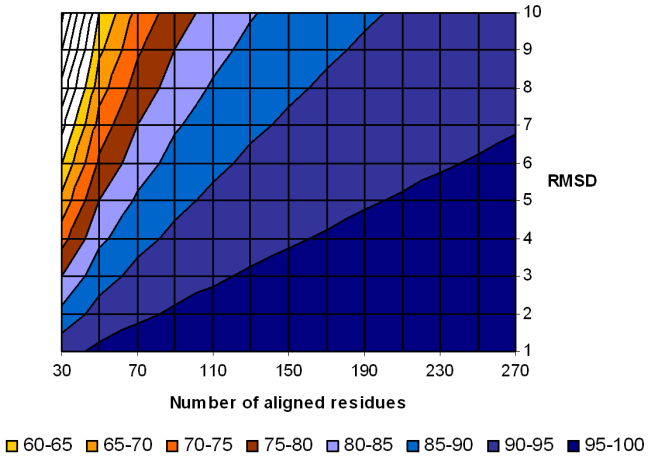


**Figure 1** The plot shows the range of structural scores (SC) as function of RMSD and the number of aligned residues (structural scoring scheme).

Figure 1 shows the range of the structural score versus the RMSD and the amount of aligned residues.

### 2.3. *Structure alignment database*

Setting up the structure alignment database (SD) involved the evaluation of all ASTRAL40 (7290 entries) pairs, which leads to about 26 million structural alignments. These have been calculated with our non-sequential structure alignment method based on maximizing the GANGSTA score [1]. In contrast to sequence alignment methods the structural alignment does not incorporate amino acid identities, but crystallographic protein details. Our method is designed to ignore the sequential order of secondary structure elements in protein chains. Additionally, the method ensures that alignments are always topologically correct, such that only secondary structure elements of the same type are aligned on each other. Thereby, we attempt to capture the biologically relevant similarities between two proteins more accurately.

After evaluation, we kept the highest scoring alignment of each pair with a structural score (SC) above 30 and at least 50% of the secondary structure elements in the smaller of both proteins aligned. This amounts to about 18.6 million protein pairs. From them, about 450.000 pairs have a structural score above 90 (SC). Thus on average, each ASTRAL40 entry shares very high structural similarities with about 60 other proteins. About 7.15 million pairs score above 80 (SC), which indicates significant structural similarities between each ASTRAL40 entry and 980 other proteins in average (about 13% of the ASTRAL40 set). Figure 2 shows the distribution of structural scores for the

structure alignment database. About 7% (in total about 1.2 million) of all alignments are sequential, such that the secondary structure elements are aligned in sequence direction.

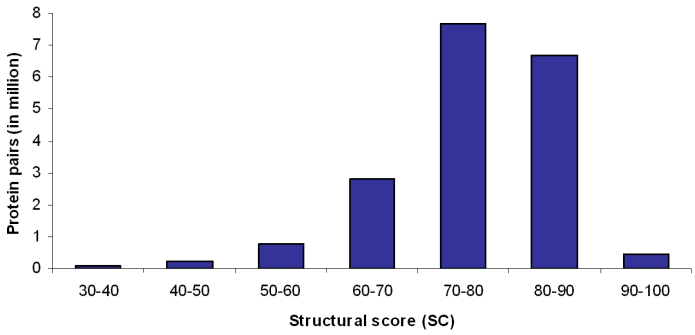


**Figure 2** Structural score (SC) histogram in our structure alignment database (SD).

The highest scoring pairs are **1mdah** (= SCOP code) with **2bbkh_** (*SC* = 0.99, *RMSD* = 0.50 Å, $N_{aligned}$ = 337) in the sequential and **1fw8a_** with **1v6sa_** in the non-sequential entries (*SC* = 0.99, *RMSD* = 1.27 Å, $N_{aligned}$ = 323) (see figure 3). The highest amount of residues has been aligned in sequence direction between **1ogya2** and **2napa2** (*SC* = 0.99, *RMSD* = 1.07 Å, $N_{aligned}$ = 512). Figure 4 illustrates a case of non-sequential alignment by **1erja_** and **1m1xa4** (*SC* = 0.98, *RMSD* = 1.59 Å, $N_{aligned}$ = 180). About half of the secondary structure elements are aligned non-sequentially.

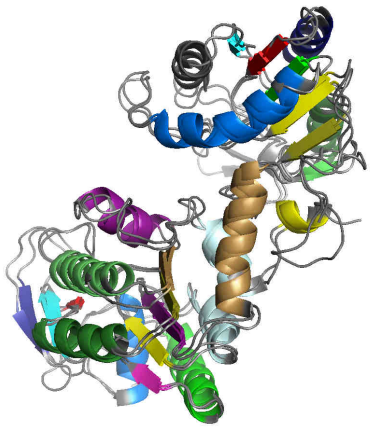

**Figure 3** Non-sequential alignment between 323 residues from **1fw8a_** and **1v6sa_** with a RMSD of 1.27Å. With respect to sequence direction, the initial three secondary structure elements (SSE) of 1v6sa_ are aligned on the last three elements of 1fw8a_. Secondary structure elements in dark, loops in light grey.

**Figure 4** Non-sequential structure alignment between **1erja_** and **1m1xa4** with 180 residues at 1.59Å. About half of the secondary structure elements are not aligned in sequence direction. Secondary structure elements in dark, loops in light grey.

## 3. Results

Initially, all-versus-all sequence alignments are performed on the ASTRAL40 dataset with FASTA. The highest ranking 100 sequences are kept for each entry. This yields 7290 sets of 100 sequentially high scoring entry pairs (=< 729000). Then, we select the structural scores (SC) for each of these pairs from our structure alignment database. Figure 5 illustrates this data acquisition process and figure 6 shows the distribution of the corresponding structural scores plotted for FASTA with BLOSUM50. This evaluation has also been done with BLOSUM62 and PAM120. Since this gave almost identical results, only the BLOSUM50 plot is shown. Additionally, we plotted the 100 highest structural scores available for each entry from our database as reference. The reference plot is an upper performance limit for the sequence alignment. Since, the secondary structure elements can be disordered in terms of sequence direction (non-sequential alignments), we plotted the highest structural scores of the in-sequential structural alignment entries separately. The distribution of sequential entry scores has its mode at 85 (SC).

Most of the reference scores are above 80 (SC) and the mode (about 17%) is at about 92 (SC). As mentioned in section 2.3., this indicates significant structural similarities among the ASTRAL40 entries. The sequence alignment with FASTA was able to determine the structurally most similar protein pairs (SC >= 98). Furthermore, in most of these cases the corresponding structure alignment is arranged in sequence direction, more precisely these are sequential structure alignments (see dashed line in figure 6). However, only a small fraction of protein pairs scores in the range between 93 and 98 (SC). The mode (~ 4%) of accepted scores is at about 81 (SC). Unfortunately, for about 25% of the

high ranking protein sequence pairs only very little structural similarity (SC < 30) could be detected by our structure alignment method.

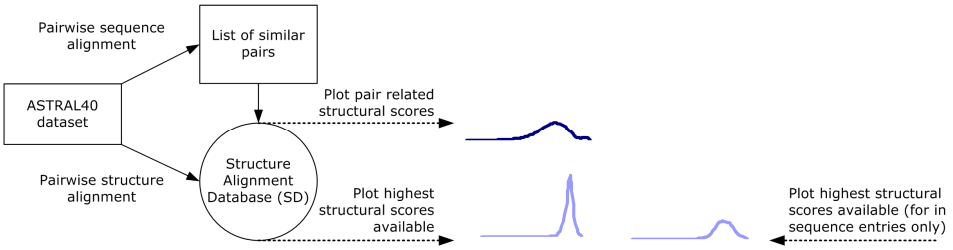


**Figure 5** This figure illustrates the data acquisition process by usage of sequence (dark) and structure (light) alignments. As result the structural score distributions, according to the structural alignment database (SD), are plotted. Additionally, the sequential structure alignment entries are plotted separately.

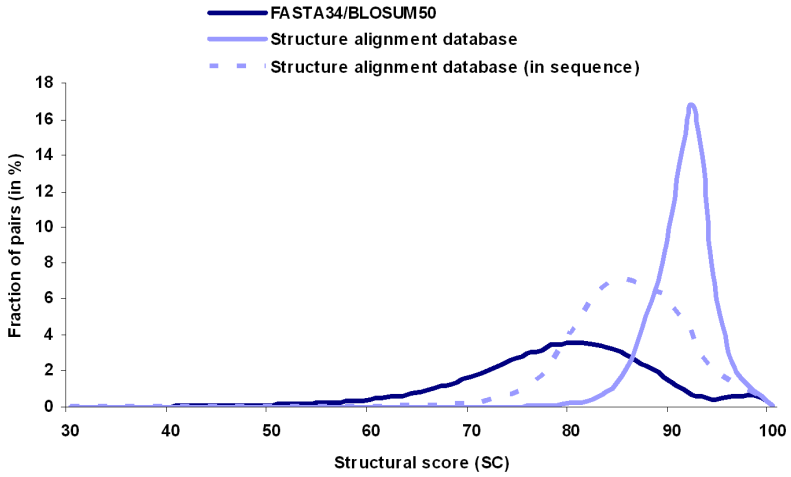


**Figure 6** Structural score distribution for similar protein pairs with respect to sequence (dark) and structure (light). The dashed line is related to the sequential structure alignments, in which the secondary structure elements of two proteins are aligned in sequence direction.

## 4. Discussion

The application of sequence alignment methods in protein science aims to reproduce structural similarities. Therefore, structure alignment methods, incorporating crystallographic details, are applied as a “gold standard” with respect to protein sequence alignment methods [14]. Since in many cases no crystal structure is known, sequence alignment is a promising and essential approach for the first step in protein structure prediction.

However, the results illustrate the difficulties of sequence alignment approaches in cases of low sequence similarity to already known protein structures. The sequence alignment method is able to reproduce the structurally most similar protein pairs, but in 25% of all high ranking FASTA results only very little structural similarity could be detected. This is related to the simplification of the model, since the sequence alignment method only incorporates the primary structure. Additionally, the sequence alignment method employs substitution matrices, which are biased towards conserved sequence segments. The structural alignment does not incorporate amino acid identities and the ASTRAL40 consists of distantly related sequences only. However, we applied the sequence alignment method only to produce pair lists of "similar" proteins. The evaluation of the similarities proceeded without taking any further information from the sequence alignment into account (e.g. score, residue assignment). Unfortunately, the recognition performance of structural similarities is low.

The fraction of sequential with respect to the non-sequential entries is at only about 7% (see details in 2.3.). Therefore, further investigations must be done to accurately measure the advantage of non-sequential versus sequential structure alignments. However, the results indicate a qualitative and quantitative gain through the non-sequential structure alignment approach. A reason for this can be the biochemical process of splicing. Furthermore, other genetic operations can reorder sequence segments [15]. Hence, our database incorporates relations between proteins and protein families, which are less constrained by these processes. Evaluating these relations can be useful to detect alternative structures and thereby support and improve protein structure prediction methods. Further, the database can be applied as reference for other sequence based approaches.

## Acknowledgements

We like to thank Jorge Numata, Stephan Lorenzen and Jonas Maaskola for their comments and support. Furthermore, this study has been supported by the International Research Training Group (IRTG) on "Genomics and Systems Biology of Molecular Networks" (GRK1360, Deutsche Forschungsgemeinschaft (DFG)).

## References

[1] B. Kolbeck, P. May, T. Schmidt-Goenner, T. Steinke and E.W. Knapp, Connectivity independent protein-structure alignment: a hierarchical approach, *BMC Bioinformatics*, 7:510, 2006

[2] M.O. Dayhoff, R.M. Schwartz, B.C. Orcutt, A model of evolutionary change in proteins, *Atlas of Protein Sequence and Structure*, 5(3):345 – 352, 1978

[3] S. Henikoff, J.G. Henikoff, Amino acid substitution matrices from protein blocks, *PNAS*, 89(22):10915-10919, 15, 1992
[4] E. Sitbon, S. Pietrokovski, Occurrence of protein structure elements in coserved sequence regions, *BMC Structural Biology*, 7:3, 2007
[5] S.F. Altschul, Amino acid substitution matrices from an informaition theoretic perspective, *JMB*, 219(3):555-565, 5, 1991
[6] J.M. Chandonia, G. Hon, N.S. Walker, L. Lo Conte, P. Koehl, M. Levitt, S.E. Brenner, The ASTRAL compendium in 2004, *Nucleic Acids Research*, 32:189-192, 2004
[7] W.R. Pearson, Rapid and Sensitive Sequence Comparison with FASTP and FASTA, *Methods in Enzymology*, 183:63 – 98, 1990
[8] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W., Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, 25:3389-3402, 1997
[9] I. N. Shindyalov, P. E. Bourne, A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm, *Nucleic Acids Research*, 29(1):228-229, 2001
[10] R. Kolodny, P. Koehl, M. Levitt, Comprehensive Evaluation of Protein Structure Alignment Methods: Scoring by Geometric Measures, *JMB*, 346, 1173-1188, 2005
[11] E. Shakhnovich, Protein Folding Thermodynamics and Dynamics: Where Physics, Chemistry and Biology Meet, *Chem Rev.*, 106 (5):1559-88, 2006
[12] F. S. Domingues, P. Lackner, M. J. Sippl, Structure Based Evaluation of Sequence Comparison and Fold Recognition Alignment Accuracy, *JMB*, 297, 1003-1013, 2000
[13] J.M. Sauder, J.W. Arthur, R.L. Dunbrack, Large-scale comparision of protein sequence alignment algorithms with structure alignments, *Proteins*, 40, 6-22, 2000
[14] P. Briffeuil, G. Baudoux, C. Lambert, X. De Bolle, C. Vinals, E. Feytmans, E. Depiereux, Comparative analysis of seven multiple protein sequence alignment servers: clues to enhance reliability of predictions, *Bioinformatics*, 14(4):357-366, 1998
[15] D.N. Cooper, E.V. Ball, M. Krawczak, The human gene mutation database, *Nucleic Acids Research Volume*, 26(1):285-287, 1998