Chapter 5

Non-sequential Structure Alignment

5.1 Introduction

In the last decades several structure alignment methods have been developed but most of them ignore the fact that structurally similar proteins can share the same spatial arrangement of SSEs but differ in the underlying polypeptide chain connectivity (non-sequential SSE connectivity), i.e., they share the same architecture but differ in their overall topology. Good examples are proteins sharing the Rossmann fold motif whose SSE arrangements have been found in different sequential orderings [245].

GANGSTA (Genetic Algorithm for Non-sequential, Gapped protein STructure Alignment) [130] is a protein structure alignment method using a two-level hierarchical approach. On the first level, pairwise contacts and relative orientations between SSEs are maximized using a genetic algorithm (GA) and protein graph representation. On the second level pairwise residue contact maps resulting from the best SSE alignments are optimized. GANGSTA was developed to produce high quality global protein-structure alignments independent of SSE connectivity by optimizing the contact map overlap. The method can be used for pairwise protein-structure alignment or fast database searches and is available through a web server[1]. For the case of pairwise structure alignment, GANGSTA provides a statistical significance related to the GANGSTA similarity measure in the form of a P-value.

The pairwise protein-structure alignment problem can be defined as the task of identifying maximal common substructures of two proteins according to a given similarity measure. Algorithms solving this problem use different representations of protein structures. GRATH [98], SSM [134], TOP [149], TOPS [81], MATRAS [121]], PROTEP [162] and VAST [79] work on protein secondary structure level only. Such secondary structure representation was also used for index-based database searches [34, 49] that are able to search very fast for similarities but do not provide an alignment. Gille *et al.* [82] and Frömmel *et al.* [75] searched for conserved substructures in interfaces of SSEs [188]. DALI [103], CE [210], SSAP [227], FASE [232] and SCALI [245] work on the residue level

[1] http://gangsta.chemie.fu-berlin.de

or a combination of secondary structure and residue level. Another approach employs methods derived from computer vision to compare 3D models [173]. TOPSCAN [155] uses topology string representations for fast structural motif searches.

Biological meaningful comparison of protein structures require a structure similarity score that is transferable to biological and chemical classifications reflecting different protein architectures. Several measures for protein-structure similarity have been proposed (see Section 4.2). Contact map overlap (CMO, Definition 17) is based on the notion of contacts between two residues. A contact map captures a 3D structure in condensed form, representing the 3D protein conformation as a *Boolean* matrix of contacts. CMO-based structure alignment was introduced by Godzik and Skolnick [85] and was proved to be NP-hard by Goldman *et al.* [88]. However, Caprara *et al.* [36] succeeded with integer programming to get solutions for CMO in reasonable CPU times. Nevertheless, the protein structure alignment problem is computationally hard to solve.

To reduce the computational burden of protein structure alignment connected with direct use of residue contact maps, we developed for GANGSTA a hierarchical approach. On the first level of the hierarchy, an alignment of SSEs is performed. On the second level, solutions for the CMO are searched on the residue level. In analogy to protein sequence alignment, structure alignment methods can work with either a global or a local strategy. Global strategies start from whole structures and remove poorly matched parts of the structure. In contrast, local strategies start from small matching units and attempt to enlarge and merge these. Since GANGSTA searches for maximal common substructures, it uses a global strategy.

Protein architectures are essentially defined by the spatial arrangement of helices and strands. These SSEs generally form the core of protein structures, while loop, turn and coil structures connecting these SSEs are more irregular and preferentially localized on the protein surfaces. Furthermore, the composition and arrangement of SSEs are evolutionary more conserved. GANGSTA considers only regularly structured SSEs, which greatly reduces the complexity of the protein structure alignment problem and facilitates structure alignments with non-sequential SSE connectivity.

It is a widely assumed that similar protein structures can be aligned while the SSE connectivity in the polypeptide chain (sequential SSE connectivity) is conserved. Nevertheless, a considerable number of proteins possess different SSE connectivity but share the same architecture (see Yuan et al. [245] for a detailed list). It has been shown that permuted SSE alignments, i.e., alignments with non-sequential SSE connectivity, occur often [209]. Using protein representations in terms of graph-theory on the secondary-structure level, i.e., defining protein graphs (see Sections 2.7.2 and 3.2), the structural alignment of protein graphs on secondary structure level can be transformed into the problem of searching the maximum common subgraph of two protein graphs [12,129,162], a problem that is known to be NP-complete [125] (see next chapter for a detailed description). Therefore, we decided to use a genetic algorithm (GA) to perform connectivity-independent alignments on the SSE level, since evolutionary algorithms provide reasonable strategies to solve NP-complete problems [57]. GAs have been applied to a wide range of chemical and biological computational applications including matching chemical 2D compounds [32], conformational analysis of DNA [151], protein folding simulations [54] and molecular

recognition [182]. GAs have been used previously for protein structure alignment [38,156,218,219] and for detecting appropriate structure templates in homology modeling [51]. Only few structural alignment methods, such as SARF [2], K2 [218, 219], MASS [62, 63] or SCALI [245], can align protein-structure fragments in non-sequential order but these methods are using different protein representations and are searching rather for local similarities or motifs than for global alignments. Only the program PROTEP [162] also optimizes the matching of protein graphs, but it optimizes only a vertex product graph instead of the more accurate, but computational more expensive edge product graph (see Chapter 6 for details). Furthermore, the program is not available anymore for non-commercial users.

In this chapter we describe the GANGSTA method for global structural alignments independent from the sequential ordering of the SSEs and optimizing the contact map overlap. The method is divided into two hierarchical stages. On the first stage, the algorithm works on secondary structure level that acts as a filter. A GA is used to perform the structural alignment between protein graphs. The second stage maximizes the contact map overlap by shifting the aligned cores using a residue level description. After the alignment procedure the superposition of the two protein structures using the Least-Squares fitting method from Kabsch [118] is done to compute for the given residue matching the transformation that minimizes the RMSD. The performance of GANGSTA was assessed in pairwise structure alignments and database scans with sequential and non-sequential SSE connectivity on various datasets that are commonly used in several evaluation tests [70,172].

5.2 The GANGSTA Method

The GANGSTA method for protein structure alignment is organized in two hierarchical levels. On the secondary structure level, a protein is represented by its SSEs, which are helices and strands, while loops and turns connecting these SSEs are ignored. For the pairwise protein structure alignment problem, we call the smaller of the two protein structures the *source* structure and the larger the *target* structure. To increase flexibility of structure alignment we allow, in analogy with sequence alignment, gaps in the *source* structure. Thus, not all SSEs of the *source* structure are explicitly aligned. Gaps in the *target* structure occur naturally and are not subject to a penalty, since at most the number of SSEs in the *source* structure can be aligned. Note that no gaps are allowed within SSEs.

5.2.1 Protein Graph Representation

In GANGSTA, the secondary structure arrangement of a polypeptide chain is modeled as an attributed, undirected protein graph (see Definition 4 for the general protein graph):

Definition 18 (GANGSTA Protein Graph)**.** *The GANGSTA protein graph is defined as a 6-tuple* $PG = (V, E, f_T, f_L, f_C, f_O)$ *where* V *is the finite set of vertices,* $E \subseteq V \times V$ *the set of edges,* f_T *and* f_L *functions assigning labels to the vertices, and* f_C *and* f_O *functions assigning labels to the edges.*

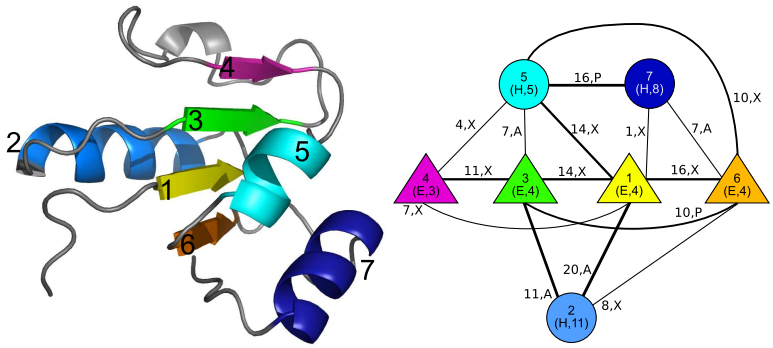


Figure 5.1: **Structure and protein graph for** $2uagA1$**.** Left: protein structure of $2uagA1$ (with Pymol [58]). Right: corresponding protein graph of $2uagA1$ (TOPS-like) [81].

The vertex set V *is an ordered set* $V = (v_1, .., v_m)$ *of vertices representing the SSEs numbered sequentially from the* N*- to the* C*-terminus. Each SSE* v_i *can be represented as a continuous set of residues of size* m*:* $v_i = (v_1^i, \ldots, v_m^i)$*. Vertices are labeled by two distinct attributes:*

$$f_T : V \to \{H, E\} \tag{5.1}$$

assigns a SSE type (helix or strand) to each vertex, and

$$f_L : V \to \mathbb{N}^+ \tag{5.2}$$

assigns the length, i.e., the number of residues per SSE, to each vertex.

The edge set E *represents spatial adjacencies between SSEs. An SSE contact between two SSEs* v_i *and* v_j *exists if there exists a residue contact between any residue* $r_k \in v_i$ *(*$k = (1, \ldots, m_i)$*) and* $r_l \in v_j$ *(*$l = (1, \ldots, m_j)$*). A residue contact exists if*

$$\forall k \in v_i, l \in v_j \;\; \exists \;\; dist\,(C\alpha(k), C\alpha(l)) < 11\mathring{A} \tag{5.3}$$

where $C\alpha(i)$ *is the coordinate vector of the* $C\alpha$*-atom of residue* i *and dist defines the Euclidean distance between two points.*

The edges are labeled by the following attributes:

$$f_C : E \to \mathbb{N}_0^+ \tag{5.4}$$

assigns the number of pairwise residue contacts between corresponding SSEs to each edge, and

$$f_O : E \to \{P, A, X\} \tag{5.5}$$

maps the relative orientation between two SSEs according to Definition 7. The following three conformations are distinguished: antiparallel (A*), parallel (*P*), and neither parallel nor antiparallel (crossed,* X*).*

An example protein graph for the SCOP [169] protein domain D-Glutamate ligase ($2uagA1$) is shown in Figure 5.1 (right). The vertices are colored correspondingly to the structure in Figure 5.1 (left). The same structure together

with its original TOPS diagram has been presented in Figure 3.2. In Figure 5.1 only the SSEs are colored and numbered that are building the protein core of the domain. These SSEs are then represented in a TOPS-like diagram with additional vertex and edge labels according to Definition 18. Here, we can see one problem defining the orientation between vertices (f_O), because the four strands, represented as triangles in the TOPS-like diagram, have clearly parallel orientation but our edge labeling denotes 'mixed' orientations, because the strands are slightly twisted in the original structure.

5.2.2 Structure Alignment on SSE Level

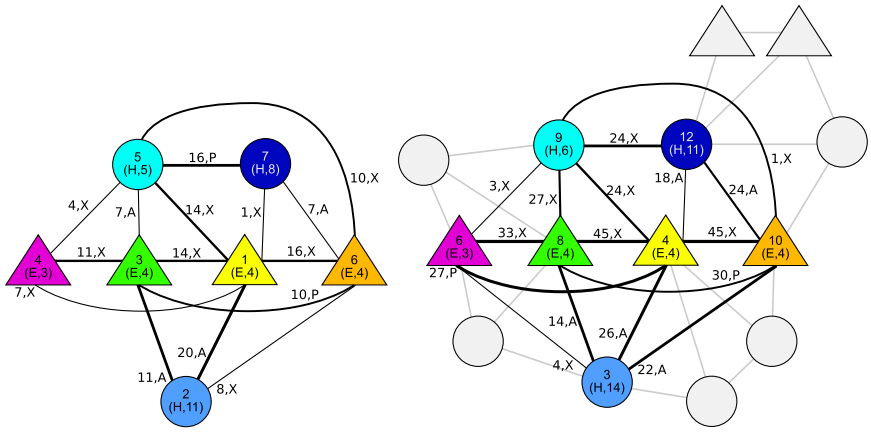


Figure 5.2: **Protein graphs** $2uagA1$ **and** $1gkuB1$**.** left: $2uagA1$, right: subgraph of $1gkuB1$

We assume that the *source* protein structure (*src*) will be aligned with the *target* structure (*target*) of equal or larger size. The structural alignment between the two protein graphs PG_{src} and PG_{targ} can be solved by finding the maximum common subgraph PG' applying

$$PG' \subseteq PG_{src}, PG_{targ} \ .$$

The task is to maximize the number of matching vertices. To search for the maximum common subgraph we need to apply a graph monomorphism $g : PG_{src} \to PG_{targ}$ composed of two mapping functions: $g_V : V_{src} \to V_{targ}$ and $g_E : E_{src} \to E_{targ}$ that are both relating structural details between the two protein structures. A detailed introduction into graph isomorphism as well as into the maximum common subgraph problem is given in the next chapter.

There are two constraining conditions that must be fulfilled for a valid structure alignment, the *SSE type criterion* and the *SSE length criterion*:

Definition 19 (SSE Type Criterion)**.** *The SSE type criterion guarantees the matching of SSEs of the same type only:*

$$f_{T_{src}}(v) = f_{T_{targ}}(g_v(v)), \quad \textit{considered } v \in V_{src} \ . \tag{5.6}$$

Definition 20 (SSE Length Criterion)**.** *The SSE length criterion ensures that the matched SSEs have similar length:*

$$\left|f_{L_{src}}(v) - f_{L_{targ}}\left(g_v(v)\right)\right| \leq LD, \quad \textit{considered} \quad v \in V_{src} \ . \tag{5.7}$$

LD is here the maximal allowed length difference between two SSEs.

These two edge conditions must hold only for SSEs that are explicitly considered in the structure alignment. If gaps are introduced, some SSEs in the *source* structure are ignored.

The two vertex constraints just guarantee a valid structure alignment, but to find the optimal structure alignment we have to add two additional criteria that have to be optimized (here minimized). First, the *contact number criterion* ensuring that only SSE interactions with similar number of residue interactions are aligned:

Definition 21 (Contact Number Criterion)**.** *The contact number criterion (CNC) is defined as the difference of the number of contacts in* source *and* target *protein:*

$$CNC := \sum_{e \in E_{src}} \left|f_{C_{src}}(e) - f_{C_{targ}}\left(g_E(e)\right)\right| \ . \tag{5.8}$$

Second, the *minimal orientation mismatch criterion* (MOMC) ensuring that the SSE matching conserves the relative position of the SSEs in the *target* protein, i.e., two parallel orientated SSEs in the *source* protein should be mapped on two parallel SSEs in the *target* protein:

Definition 22 (Minimal Orientation Mismatch Criterion)**.**

$$MOMC := \sum_{e \in E_{src}} \left|f_{O_{src}}(e) \ominus f_{O_{targ}}\left(g_E(e)\right)\right| \tag{5.9}$$

with the operator $\ominus$ *to compare the relative orientation*

$$x \ominus y := \begin{cases} 1 & x = y \\ 0.5 & (x \neq y) \wedge ((x = X) \vee (y = X)) \\ 0 & \text{else} \end{cases} \tag{5.10}$$

and $x, y \in \{P, A, X\}$.

An example alignment yielding a maximal common subgraph for the two domains is shown in Figure 5.2. A detailed discussion of the two domains and the resulting alignment is given in the Section 5.3.2. The total protein graph of $2uagA1$ could be aligned with a subgraph in $1gkuB1$. Both structures are represented in a TOPS-like diagram as described above, and only the aligned SSEs are colored and labeled.

5.2.3 The Genetic Algorithm

A genetic algorithm (GA) is a heuristic method for solving difficult optimization problems. It uses principles of evolution, e.g., genetic exchange and variation, to create a set of *individuals* and to let them evolve from *generation* to generation

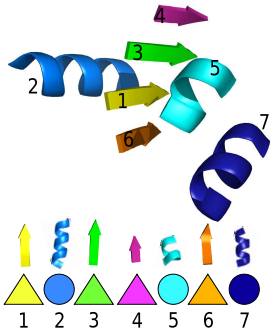


Figure 5.3: **GA encoding for** $2uagA1$**.** Composition of the genetic vector $\vec{g}$ for $2uagA1$ (helices: triangles, strands: circles, coloring according to SSE ordering from N- to C-terminus)

using specific *genetic operators*. Individuals are possible solutions (generally sub-optimal) of the optimization problem. Single individuals are evaluated using an *objective function* that has to be optimized. A new generation evolves by gene exchange and mutations applied to individuals to find improved solutions with better objective function values. The newly generated children and the fittest parents form the next generation. This procedure is repeated until the optimum is found or a suitable stop criterion is reached. There are three important aspects when creating a GA for a specific problem domain: First, a suitable representation of the problem and possible solutions have to be found, second, a computationally fast objective function is required to evaluate the solutions, and third, adequate evolutionary operators have to be designed. The basic design of the GA used in this work is that of a *steady-state-with-duplicates-GA* as described in [55].

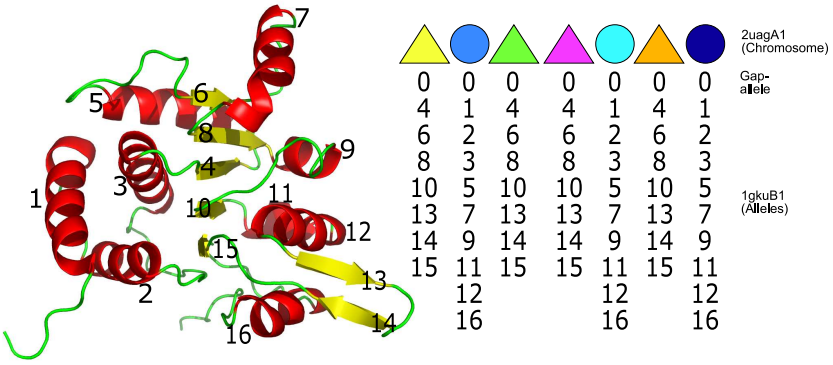


Figure 5.4: **SSE alignment.** Left: 3D structure of the $1gkuB1$ in cartoon description [58] (helices in red, strands in yellow, numbering from N- to C-terminus). Right: alleles for the structural alignment between $2uagA1$ and $1gkuB1$.

Encoding the Problem

The graph monomorphism is represented by a vector or *chromosome* $\vec{g}$ of dimension n probing the similarity between the two protein structures. A certain chromosome represents a single individual. Here, n is the number of SSEs of the *source* protein. The possible values for the single vector elements or *genes*, called *alleles*, are the SSEs from the *target* protein that hold for the same type and similar length criteria f_T and f_L, respectively. The alleles of the chromosome are given by:

$$\begin{aligned} alleles(v) = \quad & \{v' \in V_{targ} \mid \quad f_{T_{src}}(v) = f_{T_{targ}}(v') \wedge \\ & \mid f_{L_{src}}(v) - f_{L_{targ}}(v') \mid \leq LD\} \cup \{0\} \ . \end{aligned}$$

A vector element g_i represents the i-th gene and the alleles of the i-th gene (or SSE) of the *source* protein represent the possible matching SSEs of the *target* protein. In the following g_i and v_i are used synonymous differing just by their concrete representation as vector element or graph vertex.

Figure 5.3 shows the composition of the chromosome $\vec{g}$ for the protein $2uagA1$. The alleles for the structural alignment with SCOP protein domain $1gkuB1$ are shown in Figure 5.4. This example is used throughout the rest of this section.

By searching for the maximum common subgraph our approach includes the possibility to introduce gaps, i.e., to leave out SSEs from the *source* protein. Therefore every allele set for a certain gene contains additionally the value 0 or '-'. Using these gaps we can ignore SSEs from the *source* protein, so that there is no mapping onto SSEs of the *target* protein. Therefore, we can reformulate the general alignment definition for non-sequential alignments given in Definition 15 into the graph-based definition for GANGSTA protein graphs:

Definition 23 (GANGSTA-SSE-Alignment)**.** *A* GANGSTA-SSE-alignment *between a* source *and* target *protein structure represented as GANGSTA protein graphs* $PG_{src}=(V_{src}, E_{src})$ *and* $PG_{targ}=(V_{targ}, E_{targ})$ *is defined by a mapping* $m_v : V_{src} \rightarrow V_{targ} \cup \{'-'\}$ *if there exist for all* $v_i^{targ} \in V_{targ}$ *no two vertices* $v_i^{src}, v_j^{src} \in V_{src}$ *with* $i \neq j$ *such that* $m_v(v_i^{src}) = m_v(v_j^{src}) = v_i^{targ}$*. There are at most* $n_{gap} < \|V_{src}\|$ *mappings allowed such that* $m_v(v_i^{src}) = \ '-'$ *with* n_{gap} *as the maximal number of gaps allowed for a valid GANGSTA-SSE-alignment.*

Objective Function

To evaluate the quality of a given structural alignment represented by an individual $\vec{g}$, we use the following objective function:

$$\begin{aligned} obj(\vec{g}) \quad = \quad & w_C \left(1 - \frac{\sum\limits_{e \in E_{src}} \left| f_{C_{src}}(e) - f_{C_{targ}}\left(g_E(e)\right)\right|}{\sum\limits_{e \in E_{src}} f_{C_{src}}(e) + \sum\limits_{e \in E_{src}} f_{C_{targ}}\left(g_E(e)\right)} \right) + \\ & w_O \left(\frac{\sum\limits_{e \in E_{src}} |f_{O_{src}}(e) \ominus f_{O_{targ}}\left(g_E(e)\right)|}{|\{e \in E_{src} | f_{O_{src}}(e) \neq 0\}|} \right) \\ & - \ L(\vec{g}) - GP + Seq(\vec{g}) \ . \end{aligned} \tag{5.11}$$

The first term in the objective function measures the structural similarity between *source* and *target* protein by comparing the number of contacts between aligned SSEs. It is normalized to yield unity for contact identity (each contact in the *source* structure can be mapped on the *target* structure) and zero for no common contacts. The second term considers similarity in the relative orientation of SSE pairs in *source* and *target* structures, again normalized to yield unity for a perfect match and zero, if none of the orientations agree. These two terms are tuned by the weights w_C and w_O.

Matching SSEs with length differences above a threshold are penalized depending on SSE type by the function L. The SSE penalty L is calculated according to the following expressions:

$$\begin{aligned} \triangle L(v,\vec{g}) &= f_L(v) - f_L(g_v(v)) \\ LP(v,\vec{g}) &= \begin{cases} 0.03 & \text{if } \triangle L(v,\vec{g}) > 3 \\ 0.01 & \text{if } 0 > \triangle L(v,\vec{g}) >= -3 \\ 0.1 & \text{if } -3 > \triangle L(v,\vec{g}) \\ 0 & \text{otherwise or if } v \text{ is a gap} \end{cases} \\ L(\vec{g}) &= \sum_{v \in V_{str}} LP(v,\vec{g}) \end{aligned} \tag{5.12}$$

with $\triangle L$ giving the length difference between matched SSEs. Since we want to align the *target* protein onto the *source* protein, the difference is not symmetric. LP is the penalty for each aligned SSE pair depending on $\triangle L$, and L is then the sum of penalties over all SSEs.

A small number of SSEs from the *source* structure may not be considered for structure alignment. Those gaps are penalized by the gap penalty factor GP to ensure that the GA tries to find the maximum common subgraph instead of an arbitrary, small subgraph.

The structure alignment can be performed with a user-specified bias preferring sequential or non-sequential connectivity alignments, tuned by the term Seq in the objective function in Equation 5.11. The function Seq is used to favor or disfavor sequential alignments:

$$Seq(\vec{g}) = SB \frac{\sum_{i=1}^{|V_{src}|-1} br(v_i, v_{i+1})}{|V_{str}| - 1} \tag{5.13}$$

where V_{str} is the set of ordered vertices in the *source* structure. SB is a parameter controlling the strength of the bias. Positive SB values favor structures aligned with sequential SSE connectivity while negative SB values disfavor such alignments. The function br denotes, how often the connectivity of the aligned *target* structure differs from the connectivity of the *source* structure:

$$br(v_i, v_j) = \begin{cases} 1, & \text{if for } v_{k_1} = g_V(v_i) \text{ and } v_{k_2} = g_V(v_j) \text{ follows } k_1 > k_2 \\ 0, & \text{else .} \end{cases} \tag{5.14}$$

For example, the structural alignment $\{1 \mapsto 2, 2 \mapsto 4, 3 \mapsto 3\}$ has exactly one ordering violation: $br(v_1, v_2) = br(2, 3) = 1$, since $g_V(2) = v_4$ and $g_V(3) = v_3$ with $4 > 3$.

The parameters w_C, w_O and the penalty factors L, GP, Seq in Equation 5.11 were optimized to yield maximum separation of structure pairs belonging to the same SCOP [169] superfamily from those belonging to different SCOP superfamilies referring to the GANGSTA *score*, Equation 5.16.

Genetic Operators

In the literature several genetic operators [55,87], here *crossover* and *mutation*, for chromosome strings of permutations of integers are known, which can be applied to the maximal common subgraph problem. Some of the operators create children that do not agree with our constraints (for instance, a duplicate use of one SSE in the same individual violating the injectivity of the monomorphism). Those 'lethal' children are discarded.

We use the following crossover operators in our GA:

1. **Shuffle Crossover** (Figure 5.5 a)
   A random number of randomly selected genes are exchanged.

2. **2-point Crossover** (Figure 5.5 b)
   Two genes are randomly selected. All genes in between those two genes, i.e., within the crossover region, are exchanged.

3. **Helix Crossover** (Figure 5.5 c)
   A crossover mechanism, where only genes with a helix type ($f_T(v) = H$) are exchanged. The idea is to conserve good solutions for one SSE type. The helix crossover operator can not generate lethal children.

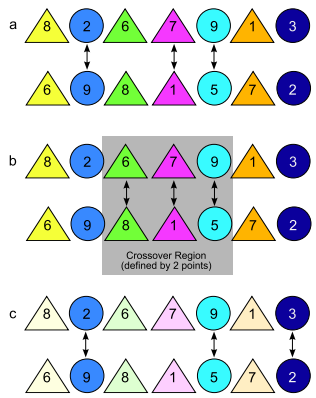


Figure 5.5: **Crossover operators.** a) Shuffle Crossover b) 2-point Crossover c) Helix Crossover.

We use the following mutation operators:

1. **Random Mutation**
   A small, random number of genes are set to a randomly selected, non-lethal allele (so, the type and length conditions hold).

2. **Exchange Mutation**
   Two randomly selected genes with matching SSE type and length, i.e.,

$$f_T(v_1) = f_T(v_2)) \wedge (|f_L(v_1) - f_L(v_2)| \leq LD \ ,$$

   are exchanged, if the exchange is non-lethal.

3. **Greedy Mutation**
   This is a newly developed operator, that employs a single local search, so that it can be called a memetic operator [38, 166]. For all alleles of a random gene all possible non-lethal children are evaluated and the one leading to the best objective function value is selected: $\max_{g_i \in alleles(v_i)}(obj(\vec{g}))$.

Population Restrictions

In order to prevent convergence to a suboptimal solution the population is managed to ensure that it contains diverse and meaningful individuals. No lethal or zero-fitness individuals are allowed. For generating the initial generation we use various initializing strategies in addition to randomized initialization:

- **identity**: if possible an individual is included that directly represents the *source* structure holding $\vec{g} = identity(V_{str})$ with

$$identity : \forall v_i \in V_{str} : g_v(v_i) = v_i^t \ \text{ with } \ v_i^t \in V_{targ} \ .$$

- **sequential**: a number of different individuals are generated containing no sequential breaks and holding $\vec{g} = sequential(V_{str})$ with

$$sequential : \forall v_i \in V_{str} \wedge i \in [1, \ldots, |V_{str}|] : br(v_i, v_{i+1}) = 0 \ .$$

- **inverse sequential**: like the sequential individuals, but inverting the sequential ordering: $\vec{g} = inverse(V_{str})$ with

$$inverse : \forall v_i \in V_{str} \wedge i \in [1, \ldots, |V_{str}|] : br(v_i, v_{i+1}) = 1 \ .$$

- **sequential with breaks**: if it is not possible to generate complete sequential individuals individuals with the least possible number of breaks are generated holding: $\vec{g} = minimize(V_{str})$ with

$$minimize : \min \left(\sum_{i=1}^{|V_{src}|-1} br(v_i, v_{i+1}) \right) \ .$$

In most cases optimal protein structure alignments are sequential. Thus, enhancing the initial generation with a restricted number of sequential structural alignments is very useful, because the algorithm converges faster against the optimal solution. However, the algorithm does will also find the optimal solutions if we omit this constraint but may be need more generations. A new generation is build from the best children and the best individuals from the parent generation. Individuals have a maximum lifetime ensuring that surviving only the best parents and children does not impoverish the gene pool. Additionally, we

are adding a small amount of randomly generated individuals to every generation. The population size is restricted to 100 alignments by default. During the execution time of the GA a list of the 10 best alignments so far and the best objective function seen at any point are maintained.

Termination Condition

The GA continues until an individual with

$$obj(\vec{g}) = w_C + w_O + SB$$

is found (see Equations 5.11 and 5.13). Since for the majority of alignments such a solution does not exist, the method terminates if a pre-defined number of iterations (generations) have been performed or the fitness of the best individual does not change over a restricted number of generations.

5.2.4 Structure Alignment on Residue Level

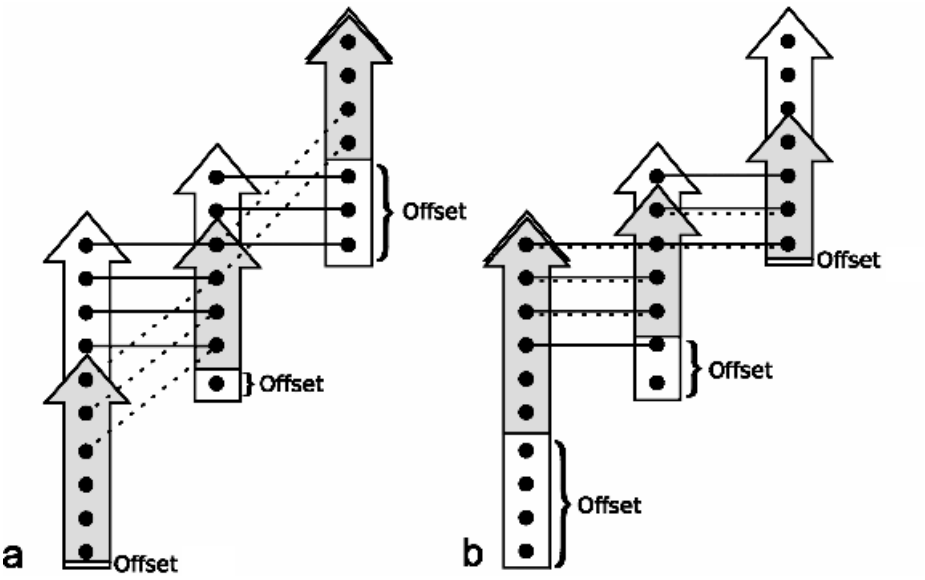


Figure 5.6: **Possible residue shifts for an aligned pair of SSEs from two different protein structures.** Two proteins (white: *source* protein, gray: *target* protein) consisting of three SSEs each. Each bold dot within a SSE represents a residue. Lines connect residue pairs of different SSEs that are in contact. Solid lines refer to contacts in the *source* protein; dotted lines refer to contacts in the *target* protein. For conserved contacts, residue pairs are connected by horizontal solid and dotted lines, simultaneously. a) No contacts of *source* and *target* proteins are conserved ($q = 0$). b) A maximum of five contacts from a total number of seven contacts are conserved ($q = 5/7$).

The result from the GA is a structure alignment on the SSE level. Often there are length differences among pairs of matched SSEs. In this case, the shorter

SSE is shifted along the longer SSE to find an optimal arrangement with respect to residue pair contacts. Two methods are used to solve the problem:

- An optimal search where all possible combinations of residue assignments for each pair of SSEs from the structure alignment are considered to find the most similar residue pair contact map.
- An estimated search that is described in Section 5.2.7.

Gaps in an individual SSE on the residue level would result in an SSE consisting rather of two instead of one SSE (if the gap is close to the center of the SSE) or in an effectively shorter SSE (if the gap is introduced on the edge of the SSE). These situations are considered on the SSE level as two independent SSEs or as a shorter SSE, respectively. Hence, no gaps in SSEs need to be considered.

The residue contact map overlap q, which is a measure for the residue pair contacts that are conserved in a structure alignment, is defined according to Definition 17 and [21]:

$$q = \frac{\sum_{i,j} C^{str}_{i,j} C^{targ}_{map(i),map(j)}}{\max\left[\sum_{i,j} C^{str}_{i,j}, \sum_{i,j} C^{targ}_{map(i),map(j)}\right]} \tag{5.15}$$

where C^{src} and C^{targ} are Cα-atom contact maps of the *source* and the *target* protein structures, respectively. The SSE alignment ($\vec{g}$) results in a residue map, which assigns residue j of the *source* protein to residue $map(j)$ of the *target* protein. Here, a residue contact is established, if the Cα-atoms of two residues are separated by less than 11Å, a value optimized empirically for protein-structure recognition by Bastolla *et al.* [21]. Note, that only residues within the aligned SSEs are mapped. All other contacts are ignored, because there exist no alignment. The objective of the second level of hierarchy is to maximize the residue contact overlap, Equation 5.15. In Figure 5.6 two examples for the same SSE mapping are shown with different residue shifts.

5.2.5 The GANGSTA Score

The last step in the GANGSTA procedure is the superpositioning of *source* and *target* protein structure with the best contact map overlap q (Equation 5.15) minimizing the RMSD (Equation 4.1) of the aligned Cα-atoms using the Kabsch algorithm [118]. To rank the quality of multiple pairwise structure alignments RMSD is not good similarity measure, as discussed in Chapter 4. The value of the objective function, Equation 5.11, is also only a crude method working on the SSE level, designed for fast screening of many individuals occurring in the GA. The residue contact map overlap q works on the residue level, but focuses on short distances only. In absence of chain connectivity, as is the case for structure alignment of SSEs, a short distance criterion alone is not sufficiently accurate to characterize global topologies of protein structures. Therefore, we have introduced a more detailed measure of protein structure similarity that considers simultaneously $RMSD$ (Å), number of not aligned SSEs in the *source*

protein N_{gap}, residue contact map overlap q, and relative difference in SSE pair distances $\triangle SSE$ between *source* and *target* structure given as

$$score = \frac{RMSD + 2 * N_{gap}}{N_{alnRes} * q * (1 - \triangle SSE) + \epsilon} \quad . \tag{5.16}$$

This GANGSTA *score* is normalized by the number of aligned residues N_{alnRes} and a small $\epsilon = 10^{-5}$ is added in the denominator to avoid division by zero. The smaller the GANGSTA *score* is, the larger is the structural agreement between the considered pair of proteins. $\triangle SSE$ is defined as

$$\triangle SSE = \frac{\sum_{k=1}^{N^{SSE}} |d_k^s - d_k^t|}{\max\left(\sum_{k=1}^{N^{SSE}} d_k^s, \sum_{k=1}^{N^{SSE}} d_k^t\right)} \tag{5.17}$$

where the sums run over the number of SSE pairs N_{SSE} considered for the structure alignment. The Euclidean distances d_k^s and d_k^t in Equation 5.17 refer to the Cα-atoms in the SSE centers of the corresponding pairs of SSEs in *source* and *target* structures, respectively. A pair of aligned proteins with evanescent GANGSTA *score* represents structures that are identical on the employed resolution level of Cα-atom coordinates.

5.2.6 Statistical Significance

To assess the quality of pairwise protein-structure alignments we use a method described by Ortiz *et al.* [179] and Vesterstroem *et al.* [232] following the work of Levitt and Gerstein [146] and Abagyan and Batalov [1]. To estimate the statistical significance of GANGSTA *scores*, Equation 5.16, we calculate a P-value describing the probability to get a better GANGSTA *score* than observed when aligning unrelated structures. This P-value can be obtained by fitting a Type I extreme value distribution function (Gumbel distribution) on the GANGSTA *score* distribution resulting from pairwise structure alignments of unrelated proteins. The Gumbel distribution possesses the probability density function [94]

$$f_G(x) = \frac{1}{b} exp\left(\frac{-(x-a)}{b}\right) exp\left(-exp\left(\frac{a-x}{b}\right)\right) \tag{5.18}$$

with parameters a for location and b for width of the density function, respectively. To fit the GANGSTA *score* distribution with the Gumbel probability density function the parameters a and b in Equation 5.18 need to be determined. Since the GANGSTA *score* assigns protein structure alignments of higher quality lower scores, i.e., a *score* of 0 is identity, the part with lower GANGSTA *scores* within the Gumbel distribution is more relevant for the fit than the tail at larger GANGSTA *scores* [179]. Therefore, we evaluated the probability to obtain GANGSTA *scores* t lower than a threshold x. The corresponding expression of the Gumbel distribution reads

$$P_G(t \leq x) = \int_0^x f_G(t)dt = exp\left(-exp\left(\frac{a-x}{b}\right)\right) \quad . \tag{5.19}$$

Equation 5.19 can be transformed into a linear expression by applying the natural logarithm function twice yielding

$$ln\left(-ln\left(P_G(t \leq x)\right)\right) = \frac{a}{b} - \frac{1}{b}x \ . \tag{5.20}$$

The parameters a and b can now easily be estimated by a linear fit between the probability of GANGSTA for *scores* $t \leq x$ obtained from structure alignments between unrelated proteins and the corresponding probability $P_G(t \leq x)$ from the Gumbel distribution. Once we have determined a and b, we can calculate the mean

$$\mu = a + \gamma b \ , \tag{5.21}$$

where $\gamma = 0.5772$ is the Euler-Mascheroni constant, and the standard deviation

$$\sigma = \frac{\pi}{\sqrt{6}} b \tag{5.22}$$

of this distribution. Using the linear transformation

$$z = \frac{x - \mu}{\sigma}$$

the P-value in Equation 5.19 can be obtained as function of the z-score:

$$P_G(Z < z) = exp\left(-exp\left(\frac{\pi}{\sqrt{6}} z + \gamma\right)\right) \ . \tag{5.23}$$

5.2.7 Database Search

For a database scan a *reference* structure is aligned against all *sample* structures in the database. In most applications the *reference* structure is also the *source* structure, i.e., the *reference* structure is smaller than the *sample* structure from the database. However, the *reference* structure can also be the *target* structure if the *sample* structure from the database is smaller than the *reference* structure. To speed up database searches a pre-filter is applied to limit the search for proteins that match certain criteria. These involve the number of SSEs, the structure diameter, i.e., the maximum distance between any pair of SSEs measured between Cα-toms in the SSE geometric centers, and the number of SSEs in contact based on Cα-atom distances. A protein structure from the database (*sample* structure) is only considered for structure alignment if the corresponding pair of *source* and *target* structures fulfills the following three basic criteria:

1. The *target* structure has at most one helix or one strand less than the *source* structure.

2. The structure diameter of the *source* structure should be at most twice as large as the diameter of the *target* structure.

3. The *source* structure should have no more than twice as many helix or strand pairs in contact as compared to the *target* structure.

Additionally, for the computationally demanding second level of the method, the residue-based structure alignment step, a rough estimate for the contact map optimization is used. To estimate the contact overlap value q, Equation 5.15, we use a greedy strategy which starts by finding the optimal offset (see Figure 5.6) for the considered SSE pair yielding the largest number of contacts. Then the algorithm continues by finding the optimal offset for the pair having the second largest number of contacts and so forth. While the problem of finding a global optimal residue alignment cannot be solved with such a local strategy, the estimated overlap values are in good agreement with optimal results. However, this estimate is sometimes up to $10,000$ times faster than the method used for finding optimal structure alignments on the residue level as described above. Since we are using an estimated contact overlap q, Equation 5.15, the reported P-value for database scans is only an upper bound of the P-value for pairwise alignments.

5.3 Results

5.3.1 Implementation

The GANGSTA structure-alignment method is implemented in C++ in a first version only for UNIX systems. As command line tool the user can choose between two methods for SSE assignment: DSSP [119] or Stride [74], and five different contact type definitions, as defined in Section 2.7.3: three distance-based contacts between Cα-, Cβ- or between all atoms, and contacts defined using Voronoi tessellation or overlapping van-der-Waals radii. It can be executed in the pairwise mode or against a list of structures.

Additionally, GANGSTA is available as web application[2]. The user can perform pairwise structure alignments or database searches against a non-redundant database of 3D structures, the ASTRAL Scop40 dataset (see Appendix D.1). The assignment of secondary structure can be done with DSSP, Stride, or according to the HELIX/SHEETS records in PDB [22] files. Here, only Cα-atom contacts are considered.

In Table 5.1 the runtimes for some exemplary pairwise structure alignments and database searches are shown. The database scan were performed using the estimated contact map overlap. All calculations were done on a Linux AMD Opteron 242 system, using one thread for the entire program including all initializations. GANGSTA is able to perform pairwise alignments within seconds and whole database searches against more than 5000 structures within minutes.

The GA used a default population of size 100. The mutation rate was 0.10 per individual and the crossover probability was set to 0.8. Fitness proportional selection was used to select the mating pool. These parameters were chosen after an initial assessment of parameter values on few structural alignments. The parameters of the objective function *obj* (Equation 5.11) are given in Table G.4 in the Appendix.

Table 5.1: **GANGSTA performance.** Runtimes for pairwise protein structure alignments and database scans against 5,397 protein domains from the ASTRAL SCOP40 dataset [41]. CPU time is including time for I/O.

source protein	*target* protein	CPU time
$2uagA1$	$1gkuB1$	0.579s
$2uagA1$	$1dhs$	0.707s
$2uagA1$	$1cjcA2$	0.575s
$1ziwA$	$2a0zA$	19.06s
$2uagA1$	database scan	13m32s

5.3.2 Example for a Non-sequential Alignment

To demonstrate the capability of GANGSTA to find protein structures with different SSE connectivity exemplarily, we consider the structure alignment of the two SCOP [169] domains $2uagA1$ and $1gkuB1$. In CATH [176] these protein domains correspond to $2uagA01$ and $1gkuB02$, respectively. For the naming convention of protein domains in SCOP and CATH see Appendix A. Both domains share the same protein structure class (α/β) but belong to different fold and superfamily categories in SCOP (see last chapter for details). Both structures have an incomplete Rossmann structure motif [199] (see also Section 3.5.3) in common. The Rossmann fold motif is ubiquitous in the universe of protein structures. It occurs with different SSE connectivity and usually comprises four helices and four strands. In the incomplete Rossmann fold motif one dangling helix is missing. Generally, it serves as a device for binding functionally relevant cofactors, such as nucleotide di-(tri-)phosphates and flavins. In the SCOP classification scheme, $2uagA1$ and $1gkuB1$ belong to the folds 'MurCD N-terminal domain' and 'P-loop containing nucleoside triphosphate hydrolase', respectively. In CATH [46], they are classified in the homologous superfamilies 'NAD(P)-binding with Rossmann-like domain' and 'P-loop containing nucleoside triphosphate hydrolase', respectively. Both proteins share the same level of CATH topology 'Rossmann-fold'. In the pairwise structure alignment mode of GANGSTA the smaller protein structure (*source*) is superimposed on the larger protein structure (*target*). In the *target* structure only the SSEs useful for the alignment are considered, while the omission of an SSE in the *source* structure (introducing a gap) is penalized (see Section 5.2.3). Figure 5.7 shows the result of the pairwise alignment for these two domains represented as the superposition of aligned SSEs using Stride [74] for SSE assignment. Table 5.2 summarizes results obtained from the pairwise structure alignment of the complete set of SSEs of *source* structure $2uagA1$ on the *target* structure $1gkuB1$ for two different SSE assignment methods, Stride and DSSP [119]. Although the two protein domains possess different SSE connectivity, GANGSTA was able to align them with a significant P-values (below 0.05 corresponding to a confidence level of 95%, see next Section) considering all SSEs of the *source* structure, i.e., no SSE gaps were introduced. For both SSE assignment methods GANGSTA produced the same SSE alignment but with different scores. Using Stride the

[2] http://gangsta.chemie.fu-berlin.de

color code of aligned SSEs							
SSE number [a] in source structure 2agA1	**1**	**2**	**3**	**4**	**5**	**6**	**7**
SSE number [a] in target structure 1gkuB1	**4**	**3**	**8**	**6**	**9**	**10**	**11**

Figure 5.7: **Structural superposition of** $2uagA1$ **and** $1gkuB1$**.** The superposed structures are displayed in two different orientations. Aligned SSEs of *source* ($2uagA1$) and *target* ($1gkuB1$) structures have the same color. The SSEs connecting loops and SSEs not considered for the alignment are displayed in light gray in both structures. Color coding and SSE numbering are given below. [a]In both structures the SSEs are numbered from N- to C-terminus. SSE assignment by Stride [74]. The figure was created with PyMol [58].

alignment had a $RMSD$ of 3.526 (0.0 is identity), a GANGSTA *score* of 0.0746 (0.0 is identity), and a contact map overlap of 0.7829 (1.0 is identity and 0.0 is minimum). Using DSSP the alignment had a better GANGSTA *score* but inferior $RMSD$ and contact map overlap values.

We also tried to reproduce this alignment with other structural alignment programs like DALI [103], CE [210] and LGA [247], three of the community-wide mostly used structure alignments methods, but none of these methods was able to reproduce our non-sequential alignment or produce an alignment with a comparable good $RMSD$.

Additionally, we tested for the same pairwise alignment the five different contact criteria as defined in Section 2.7.3. The results are shown in Table 5.3. Using the *ca* contact definition resulted in the best alignment according to GANGSTA *score* and contact map overlap q. Though the *vor* contact type produced the same SSE alignment more residues were aligned using the *ca* contacts leading to a better $RMSD$ value. For the *ca* contact type the contact map overlap is significantly better despite less residues were aligned. The reason for that is the flexible distance cutoff of 11Å used for the *ca* contact type: the more residue contacts are defined the more contacts between aligned residues can be conserved in the contact map alignment. All five alignments were non-sequential. Totally, three different SSE alignments were produced indicating that Rossmann folds give multiple possibilities to align its SSEs resulting all in

Table 5.2: **Summary of non-sequential structure alignment of** $2uagA1$ **and** $1gkuB1$**.** Structure alignment between $2uagA1$ (*source*) and $1gkuB1$ (*target*) using Cα contacts. q: contact map overlap (Equation 5.15), RMSD (Equation 4.1), GANGSTA *score* (Equation 5.16), N_{gap}: number of SSEs ignored in the *source* structure, and N_{alnRes} the number of aligned residues. SSE assignment by Stride [74] and DSSP [119].

quantity	value STRIDE	value DSSP	comments and details
q	0.7829	0.7225	1 is identity, 0 is minimum
$RMSD$ [Å]	3.526	3.548	0.0 is identity
GANGSTA *score*	0.0746	0.0594	0.0 is identity
P-value	0.0085	0.0059	< 0.01 is significant
N_{alnRes}	42	37	number of aligned residues
N_{gap}	0	0	number of gapped SSEs

Table 5.3: **Structure alignment of** $2uagA1$ **and** $1gkuB1$**with different contact definitions.** Contacts are defined according to Section 2.7.3. SSE assignment by Stride [74]. q: contact map overlap (Equation 5.15), RMSD (Equation 4.1), GANGSTA *score* (Equation 5.16), and N_{alnRes} the number of aligned residues.

contact type	q	GANGSTA *score*	$RMSD$ [Å]	N_{alnRes}	alignment
ca	0.7225	0.0594	3.548	37	3, 8, 9,14, 2, 7, 4, 5
cb	0.3953	0.0761	4.022	42	3, 2, 7, 5, 8, 9,11,14
all	0.6094	0.0651	4.022	42	3, 2, 7, 5, 8, 9,11,14
vor	0.6000	0.0616	3.394	42	3, 8, 9,14, 2, 7, 4, 5
vdW	0.5000	0.0949	4.554	36	3,10, 9,14, 2, 7, 4, 5

good $RMSD$ values with a similar number of aligned residues.

5.3.3 Statistical Significance

As described in Chapter 4, one of the most important applications of protein structure alignment methods is to find out whether a pair of proteins is structurally or evolutionarily related. The SCOP [169] or CATH [176] databases are often used for such a classification task. Whether the similarity measure employed in GANGSTA (the GANGSTA *score*, Equation 5.16) is suitable to assign two protein structures to the same SCOP superfamily was tested by a statistical study similar to the one described by Vesterstroem and Taylor [232]. For that purpose, we performed structure alignments of $5,000$ protein domain pairs belonging both to the same SCOP superfamily (dataset SAME40) and $90,000$ structure alignments of domains pairs belonging to different SCOP superfamilies (dataset DIFF40). The two datasets are subsets of the ASTRAL Scop40 dataset [41] consisting of globular domains with at most 40% pairwise sequence similarity. All datasets are explained in more detail in the Appendix D.

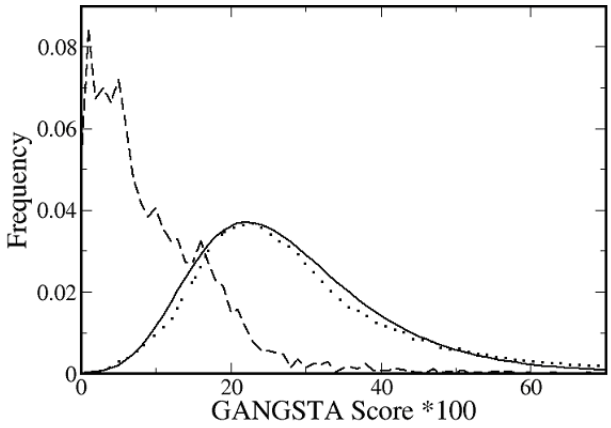


Figure 5.8: **The distribution of GANGSTA** *scores***.** The distribution of the GANGSTA *score*, Equation 5.18, for aligned protein pairs of the same (dashed line) and of different (dotted line) SCOP [169] superfamilies. The Gumbel distribution, Equation 5.18, $f(score * 100)$ (solid line) was fitted with $a = 22.2013$ and $b = 9.9384$. For more details see Method section above.

For the protein structure alignments from both datasets the distributions of GANGSTA *scores* are shown in Figure 5.8. A Gumbel distribution was fitted to the GANGSTA *score* distribution of the DIFF40 dataset with mean $\mu = 27.938$ and standard deviation $\sigma = 12.746$ (see Equations 5.21 and 5.22), as described in Section 5.2.6. Similar score distributions were reported by Levitt and Gerstein [146], MAMMOTH [179], and FASE [232] all using different measures of structural similarity and different optimization algorithms. According to Figure 5.8, the distributions of GANGSTA *scores* of the two datasets overlap partially. Hence, it is not possible to conclude reliably from the similarity of two protein structures that they belong to the same superfamily of proteins. But the ability of the GANGSTA *score* to discriminate between related and non-related protein structures can be illustrated as coverage-versus-error-rate plot as shown in Figure 5.9, as done before in [232, 244]. In short, the *coverage* is the ratio of true-positives at a given P-value threshold, while the *error-rate* defines the number of false-positives for that threshold. True-positives are defined as an alignment of two domains from the same SCOP superfamily with a P-value better than a given P-value threshold, and false-positives as an alignment from two domains from different SCOP superfamily with a P-value better than a given P-value threshold. The plots were calculated as described by Ortiz *et al.* [179]:

1. For each pairwise alignment the P-value is determined as computed by Equation 5.19, and it is noted whether the pair of domains is a true-positive or true-negative.

2. All alignments are sorted by increasing P-value.

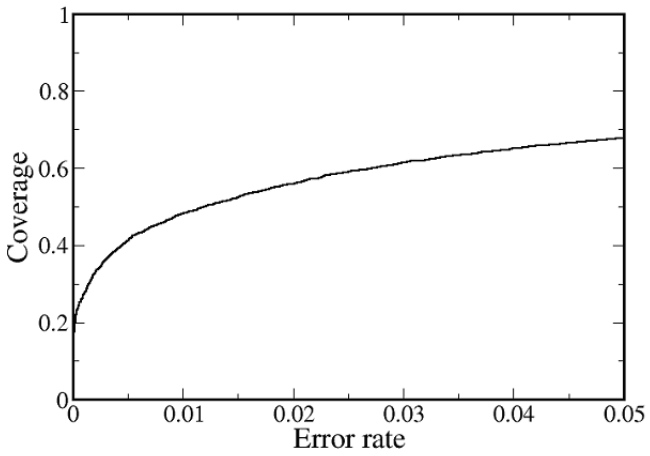


Figure 5.9: **Coverage-versus-error-rate plot for GANGSTA** *scores***.** For a given P-value threshold, we calculated the percentage of true-positives, i.e., the percentage of domain pairs from the same SCOP [169] superfamily that could be aligned with a GANGSTA *score* better than a given error-rate threshold.

3. Count down the list from the best P-value to the worst P-value and at each point in the list the number of false-positives is counted, i.e., with increasing P-values the number of false-positives also increases, because it is summed up over all P-values before defining the error-rate.

4. Compute in the same fashion the fraction of true-positives that are more significant than the P-value defining the coverage.

In the above application, GANGSTA was able to detect 48% and 67% true-positives of the SCOP superfamily relationships at a confidence level of 99% and 95% (see Figure 5.9), respectively. The discrimination between structurally related and non-related proteins is comparable with other methods. At a confidence level of 99% PrISM [244] reported 54% and MAMMOTH 50% true-positives. At a confidence level of 95% MAMMOTH reported 60% and FASE 72% true-positives. In contrast to these studies GANGSTA reports the P-value for SCOP superfamily classification instead of SCOP fold classification.

5.3.4 Comparison with other Methods

Most programs or web servers for protein-structure alignment deal with sequential structure alignments only and most of the known curated structure alignments or benchmark sets for structure alignment are constructed to test methods preserving the sequential SSE connectivity. To obtain a more representative comparison with other alignment methods we tested the performance of GANGSTA for structure alignments with exclusively sequential SSE connectivity. The two structure alignment tasks we conducted here complement the evaluation of web-based programs and servers for structure alignment applied

Table 5.4: **Results Novotny dataset [172].** Comparison of different structure alignment methods for three structure classes according to CATH [176]. Except for GANGSTA all data were taken from literature [37, 172]. Average performances differ slightly, since structures with low secondary structure content were omitted. The 53 proteins of the Novotny dataset (see Appendix D.8) were aligned against the SCOP40 reference database (see Appendix D.1). For the GANGSTA evaluation the assignment of a *reference* structure was successful, if at least one *sample* structure with appropriate CATH topology was found among the top 100 ranked protein domains.

Program/Server	Mainly-α	Mainly-β	Mixed-α-β	Average performance (%)
Total Number[a]	19	19	15	
CE [210]	17	19	13	93
DALI [103]	14	19	14	89
DEJAVU [124, 153]	14	19	9	79
GANGSTA	18[b]	19	15	98
LOCK [211]	0	14	11	47
MATRAS [121]	11	19	14	83
PRIDE [40]	14	14	7	66
SSM [134]	5	13	10	53
TOP [149]	14	18	12	83
TOPS [81]	2	15	14	59
TOPSCAN [155]	15	12	9	68
VAST [79]	12	17	15	83
YAKUSA [37]	17	19	14	94

[a] Number of *sample* protein structures belonging to the specified CATH class that are used for assignment to the appropriate CATH topology.

[b] Since protein 1*c*3*u* was moved to another topology class in more recent CATH versions, 18 is the maximum number of correct structure alignments achievable.

in recent performance tests conducted by Novotny *et al.* [70, 172]. All methods used in this evaluation test are described and listed in Appendix E. The authors identified structural related protein structures by using the CATH classification scheme [176], as described in Section 4.4.2. True-positives were defined on the CATH topology level. Each *reference* structure was submitted to all servers evaluated in the Novotny study [172], and it was determined whether any of the structures from the same CATH topology level, other than the *reference* structure, was found as a true-positive hit. Since the various servers and methods all use different databases and scoring systems, the simple counting of true-positives was not feasible. Therefore, they used a simple binary scoring system: at least one true-positive either was or was not found in the list of significant hits. For servers that did not indicate the significance of the hits, up to 100 hits were examined. To have a similar test scenario, we decided to reproduce these structure alignment task using the database scan version of GANGSTA. All database scans were performed using DSSP [119] for SSE assignment. We used the GANGSTA *score* (Equation 5.16) to rank the resulting structure alignments.

However, the P-value was not used, because for database scans GANGSTA calculates only an estimated contact map overlap q (Equation 5.15) to increase the computational performance (see Method section).

The first task was based on a selection of protein domains (Novotny dataset, see Appendix D.8) belonging to four different CATH classes (mainly-α, mainly-β, mixed α-β, few SSEs) as used in [172]. Proteins from the fourth CATH class (few SSEs) have only low secondary structure content and thus few SSE contacts. Since GANGSTA considers helices and strands only, we tested it only on those proteins of the Novotny dataset (reduced Novotny dataset) belonging to CATH classes mainly-α, mainly-β, mixed α-β. This resulted in 53 *reference* protein structures (see Appendix Table D.3). The results of the structure alignments performed with GANGSTA and 11 other methods are shown in Table 5.4. Except for the data obtained with GANGSTA all data were taken from the literature [37, 172]. Average performances differ slightly from the literature values, since the structures with low secondary structure content were omitted. In analogy to the preceding investigations on the Novotny dataset the assignment of a *reference* structure was successful with GANGSTA, if at least one *sample* structure with appropriate CATH topology was found among the top 100 ranked protein domains. GANGSTA was able to detect true-positives for 52 of all 53 *reference* structures (98%) of the reduced Novotny dataset except for the mainly-α protein $1c3u$. This protein had been moved to another topology in more recent CATH versions (Table 5.4 and D.3 in the Appendix), so we could not compare the GANGSTA results directly to results listed for other methods. Hence, GANGSTA reaches the best result possible for the reduced Novotny dataset.

The second structure alignment task considers a database search with eleven pairs of structures from the DIFFAL dataset [70] (see Appendix D.7 for details) that were considered as difficult structure alignment cases [79] before. For each pair Fischer defined a *reference* and a *target* structure. According to Novotny *et al.* [172], a search was considered to be successful, if for a *reference* structure the defined *target* structure or a homologous structure according to CATH was found. For the 11 *reference* structures true-positives were searched among the best 100 ranked structures from a database scan against the Scop40 dataset. GANGSTA was able to find appropriate true-positive structures for each of the eleven protein pairs (see Table 5.5 for more details). Seven results were found at top 1 position, eight within the top 10, and all within the top 50 ranked structures. Hence, in this test GANGSTA outperforms DALI and CE, which both found only ten out of eleven possible structure pairs [172].

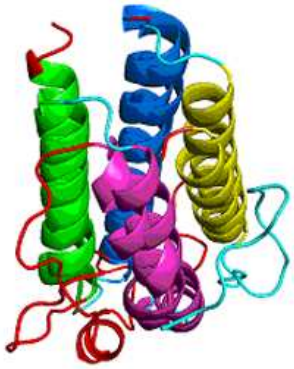
5.3.5 Non-sequential Structure Alignments

We studied the performance of GANGSTA for alignments of protein structures with non-sequential SSE connectivity that are known from literature. We show examples for four-helix-bundles or the TRAF-domain-like fold studied in [62, 63]. Additionally, we show significant alignments of protein structures with non-sequential SSE connectivity involving the Rossmann and Rossmann-like structural motifs according to classifications in SCOP [169] or CATH [176] and circular permuted proteins. All comparisons were done in the pairwise structure alignment mode using Stride [74] for SSE assignment.

Table 5.5: **Results for DIFFAL dataset.** For all 11 pairs of the DIFFAL dataset the structures from the ASTRAL Scop40 (Appendix D.1) dataset are given, which are most similar to the specified target structure, together with their rank from the GANGSTA database search and their CATH [176] hierarchy levels. CATH hierarchy levels are: *H* same homologous superfamily and *S* same sequence family.

protein pair		successful matches		
reference structure	*target* structure	rank	PDB code	CATH level
1*bgeB*	2*gmfA*	1	1*bgc*	*H*
		2	1*alu*	*H*
		6	1*lki*	*H*
1*cewI*	1*molA*	49	1*eqkA*	*H*
		66	1*stfI*	*H*
1*cid*01	2*rhe*	25	1*eajA*	*H*
		35	1*ojaE*1	*H*
1*crl*	1*ede*	1	1*llfA*	*S*
1*fxiA*	1*ubq*	14	1*m*94*A*	*H*
		23	1*c*1*yB*	*H*
		26	1*lm*8*B*	*H*
		40	1*lfdA*	*H*
1*ten*	3*hhrB*	2	1*fnf*02	*H*
		4	1*f*6*fB*2	*H*
		5	1*fhyB*2	*H*
		8	1*cd*9*B*2	*H*
1*tie*	4*fgf*	1	1*avwB*	*H*
		2	1*wba*	*H*
		6	1*jlxA*1	*H*
		8	1*md*6*A*	*H*
		12	1*q*1*uA*	*H*
2*azaA*	1*paz*	1	1*qhqA*	*H*
		2	1*jzgA*	*H*
		3	1*sdfA*	*H*
		4	1*plc*	*H*
		8	1*jw*0*A*3	*H*
2*sim*	1*nsbA*	1	3*sil*	*H*
		14	1*usrA*	*H*
3*hlaB*	2*rhe*	1	1*k*5*nB*	*H*
		4	1*fp*5*A*2	*H*
		13	1*mjaH*	*H*
		15	1*ojae*2	*H*
1*g*61	1*jdw*	1	1*jdw*	*H*
		2	1*g*62*A*	*H*
		54	1*bwdA*	*S*

Four-Helix-Bundles



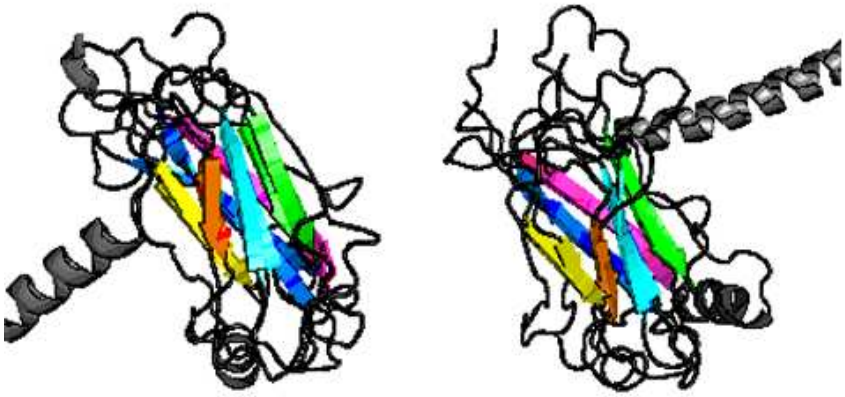
reference structure	four-helix-bundle structures								
2hmzA	2ccyA	256a	3inkC	1rcb	1bgeB	1bbhA	1flx	1le2A	1aep
H1 [a]	H3	H1	H7	H6	gap	H3	H4	H7	gap
H2 [a]	H4	H2	H4	H3	H4	H4	gap	H4	H5
H3 [a]	H1	H3	H5	H4	H3	H1	H2	H3	H4
H4 [a]	H2	H4	H1	H1	H5	H2	H3	gap	gap
SSE connectivity [b]	-	+	-	-	-	-	-	-	-
Score, eq. (10)	0.0296	0.0283	0.0360	0.0405	0.0446	0.0446	0.0830	0.0906	0.1030
P-value	0.0009	0.0009	0.0015	0.0020	0.0026	0.0026	0.0174	0.0235	0.0364
RMSD [Å]	2.952	2.993	3.093	3.792	1.890	3.760	2.973	3.496	1.927

Figure 5.10: **Four-Helix-Bundle alignments.** Top: superposition of the two aligned four-helix-bundle proteins $2hmzA$ (red line) and $3inkC$ (cyan line). Aligned SSEs have the same color coding. Below: structure alignments for *reference* structure $2hmzA$ against the four-Helix-Bundle dataset(Appendix D.2). For each structure alignment the SSE mappings, the GANGSTA *score*, the P-value, and the RMSD are given. Helices are numbered from N- to C-terminus according to SSE connectivity in the reference structure $2hmzA$. The structures are ordered by P-value. [a]Color code as in structure above. [b]Structure alignments with sequential or non-sequential SSE connectivity are denoted as '+' or '-', respectively.

Four-helix bundles are the most common α-helical motif that has been found in many α-domains with a range of diverse functions such as oxygen transport, nucleic acid binding, and electron transport (see also Section 3.5.1). We selected the protein domain $2hmzA$ as *reference* structure as representative structure for four-helix-bundles and aligned it pairwise with the nine other protein domains from the Four-Helix-Bundle dataset (see Appendix D.2). The results are comprised in Figure 5.10. For all pairwise alignments the SSE mapping (relative to the *reference* structure), the GANGSTA *score* (Equation 5.16), P-value (Equation 5.19), and RMSD (Equation 4.1) are listed. GANGSTA was able to align all structures within 95% confidence level. Only three protein domains ($1le2A$, $1aep$, $1flx$) were not aligned within the 99% confidence level but all contain alignment gaps, i.e., some SSEs of the *source* structure were not aligned. All

structure alignments were superimposed with an RMSD smaller than 3.5Å. It is noteworthy that only the alignment of $256a$ with the *reference* structure $2hmzA$ has sequential SSE connectivity. Figure 5.10 shows the structural superposition of the two protein domains $2hmzA$ and $3inkC$.

TRAF-Immunoglobin Dataset



1frtB	1k8iA	1czyA	1bmg	1kzzA	1lb4A	1igtA	1k2fA
E1 [a]	E7	E5	E1	Gap	E5	E1	Gap
E2 [a]	E8	E6	E2	E3	E4	E3	E9
E3 [a]	E9	E7	E3	E7	Gap	E4	E5
E4 [a]	E12	E9	E4	E11	E3	E7	E14
E5 [a]	E13	E10	E5	E6	E6	E8	E4
E6 [a]	E14	E2	E6	E5	E7	Gap	Gap
SSE connectivity [b]	+	-	+	-	-	+	-
Score	0.0134	0.0474	0.0610	0.0884	0.0955	0.1019	0.1973
P-value	0.0003	0.0030	0.0064	0.0216	0.0281	0.0351	0.2774
RMSD [Å]	0.980	2.696	0.468	2.126	2.555	2.501	4.363

Figure 5.11: **TRAF-Immunoglobin alignments.** Top: superposition of aligned structures of $1frtB$ with $1czyA$ (left) and with $1kzzA$ (right). Aligned SSEs have the same color coding. Below: structure alignments for reference *structure* 1frtB against the seven other structures of the TRAF dataset. The SSEs are numbered from N- to C-terminus according to SSE connectivity in the *reference* structure $1frtB$. The structures are ordered by P-value. [a]Color code as in structure above. [b]Structure alignments with sequential or non-sequential SSE connectivity are denoted as '+' or '-', respectively.

The TRAF dataset consists of eight proteins that belong to two different folds in the all-β class of the SCOP [169] database. Four proteins (PDB-IDs: $1czyA$, $1kzzA$, $1lb4$, $1k2fA$) belong to the 'TRAF (TNF Receptor Associated Factor) domain-like' fold, the other four proteins (PDB-IDs: $1bmg$, $1frtB$, $1igtA$, $1k8iA$) belong to the 'C1 set domains' family of the 'Immunoglobulin-like beta-sandwich' fold (see AppendixD.3). We aligned the *reference* structure $1frtB$ against all other seven domains. The results are shown in Figure 5.11. The

GANGSTA method was able to align six of the seven proteins within a 95% confidence threshold. Only protein domain $1k2fA$ could not be aligned with a significant P-value (0.2774). This protein could only be aligned to the *reference* structure if two gaps are introduced in the $1k2fA$ structure, resulting in a structure superposition with a RMSD of 4.3Å. For all other structures the corresponding RMSDs are smaller than 2.7Å with at most one gap introduced in the alignment. All structure alignments of $1frtB$ with proteins from different families possess different SSE connectivity ($1czyA$, $1kzzA$, $1lb4$, $1k2fA$). Only the alignments with members of the same family as the *reference* structure (1bmg, 1igtA, 1k8iA) preserve the same SSE connectivity. Additionally, Figure 5.11 shows the superposition of $1frtB$ with $1czyA$ (top left) and with 1kzzA (top right), two protein domains from different SCOP superfamilies than $1frtB$. Both alignments are non-sequential in SSE connectivity relative to $1frtB$.

Rossmann Structural Motif

reference structure	structures of the Rossmann fold dataset						
2uagA1	1f0kA	1f8yA	1rqlA	1dhs	1geeA	1dih_1	1cjcA2
E1 [a]	E1	E4	E11	E17	E1	E4	E1
H2 [a]	H2	H3H5	H12	H18	H6	H5	H7
E3 [a]	E3	H4E6	E13	E19	E7	E6	E8
E4 [a]	E5	H1E8	E16	E21	E9	E8	E9
H5 [a]	H9	H9	H17	H22	H2	H9	H2
E6 [a]	E8	E1	E1	E4	E3	E1	E3
H7 [a]	H11	H2	H7	H6	H4	H2	H5
SSE connectivity [b]	-	+	-	-	-	-	-
Score, eq. (10)	0.0449	0.0482	0.0494	0.0526	0.0573	0.0636	0.0700
P-value	0.0026	0.0032	0.0034	0.0041	0.0053	0.0072	0.0099
RMSD [Å]	2.963	3.254	3.156	3.424	3.846	4.220	4.054
CATH [c]	1f0ka01/02,R	1f8yA00, R	1rqlA01, R	1dhs000,-	1geeA00, R	1dhi001, R	1cjcA01, R
SCOP [c]	-	-	-	R-like	R	R	R-like

Figure 5.12: **Rossmann structural motif alignments.** Results of structure alignments of *reference* structure $2uagA1$ against the structures of the Rossmann-fold dataset. SSEs are numbered according to SSE connectivity of $2uagA1$ from N- to C-terminus. Structures are ordered by P-value. [a] Color code as in Figure 5.1. [b]Structure alignments with sequential or non-sequential SSE connectivity are denoted as '+' or '-', respectively. [c]R= Rossmann fold, R-like = Rossmann fold like.

Here, we consider a sufficiently complex and widespread structure motif, the Rossmann structure motif [199] that was first identified in dinucleotide-binding proteins (see also Section 3.5.3). We used the $2uagA1$ as *reference* structure and the Rossmann dataset (see Appendix D.5) as *sample* structures. Six of the seven proteins are classified as Rossmann-fold in the CATH topology level except $1dhs$, which is classified in SCOP as Rossmann-fold. The results are shown

in Figure 5.12. GANGSTA was able to align all proteins with the *reference* structure $2uagA1$ within the 99% confidence level. All alignments were non-sequential with respect to the SSE connectivity of the *reference* structure, and all superpositions could be made with RMSD smaller than 4.2Å.

Circular Permutations

The circular permutation (CP) of a protein is defined as a genetic operation in which part of the C-terminus of the protein is moved to its N-terminus [237]. The prevailing opinion about circular permuted proteins occurring in nature is, that most permutations are a result of gene duplication and subsequent deletion of unnecessary parts at the ends of the resulting tandem repeat. An example for this is swaposin, one of the first CPs reported [186]. Another method how nature can generated CPs was found in concanavalin A, where the CP is a result of posttranslational modification, namely the ligation of C- and N-terminus, and subsequent cleavage of the peptide chain [236]. In recent years numerous CPs found in native proteins have been reported [117]. In addition, artificial CPs were produced experimentally to study the effect of shifted C- and N-termini on fold stability and related subjects [97, 117].

Here, we want to test GANGSTA's ability to detect potential CPs on a set of CPs known from literature [117], observing whether exactly one break with the features described above is reported in the sequential order of the SSE alignment. We used Stride [74] for SSE assignment. The resulting alignments are shown in Table 5.6. The numbers represent SSE numbers of the *target* structure starting at the N-terminus. The position of the number equals the SSE number of the *source* protein. A 'G' stands for a gap denoting one SSE in the *source* structure not aligned with any SSE in the *target* structure. Additionally,

Table 5.6: **Sequential order of CP alignments.**. Pairs of protein structures that are known to represent CPs [117]. The GANGSTA SSE alignment is given as well as if the CP has been recognized.

CP-Pair	SSE Alignment by GANGSTA	CP found?
$1rin$-$2cna^{n}$	9 10 11 12 0 1 2 3 4	yes
$1nkl$-$1qdm^{n}$	2 3 0 1	yes
$1rsy$-$1qas^{n}$	G 28 21 22 24 G 25 26 27	yes
$1aqi$-$1boo^{n}$	G 4 5 7 8 G G 9 10 11 G G 0 1 2 3 G G	yes
$1onr$-$1fba^{*n}$	G 16 G 17 G 18 19 0 1 G G 2 3 6 7 8 10 11 12 13 14 15	yes
$1gbg$-$1ajk^{a}$	6 7 8 9 10 11 12 13 14 0 1 2 3 4 5	yes
$1avd$-$1swg^{a}$	6 7 8 0 1 2 3 5	yes

*found with GANGSTA, using a bonus for sequential SSE order.

^{n} CP found in nature.

^{a} artificially designed CP.

we compared GANGSTA structure alignments with the results of other structure alignment tools, here, MAMMOTH [179], Dali [103], K2 [218, 219], and SSAP [178], to check if the results for the CP alignments are correct and competitive in terms of structural similarity. The results are shown in Table 5.7.

The protein structure pairs have been ordered by the RMSD value computed by the Dali method that is often used as standard method for structure alignments. All protein structure alignments considered here were manually compared and

Table 5.7: **CP alignments using different methods.** Comparison of residue numbers (*no*) and RMSD (*rmsd*) [Å]; *mean* gives the mean value over all seven alignments per method.

	Dali		K2		Mammoth		SSAP		GANGSTA	
	no	*rmsd*	*no*	*rmsd*	*no*	*rmsd*	*no*	*rmsd*	*no*	*rmsd*
1*rin*-2*cna*^{n}	106	1.7	107	1.0	24	3.8	90	9.8	123	0.9
1*nkl*-1*qdm*^{n}	55	2.7	29	2.6	27	3.7	74	11.9	55	2.8
1*rsy*-1*qas*^{n}	109	3.7	89	1.1	74	2.9	127	14.2	82	1.0
1*aqi*-1*boo*^{n}	113	3.9	57	2.2	44	3.8	156	9.4	104	2.9
1*onr*-1*fba*^{*n}	198	4.1	81	2.3	54	3.8	224	11.5	147	4.0
1*gbg*-1*ajk*^{a}	123	1.2	116	0.4	119	1.4	125	1.7	194	0.5
1*avd*-1*swg*^{a}	74	1.7	68	1.0	57	2.8	77	2.9	85	0.7
mean	111	2.7	78	1.5	57	3.2	125	8,8	113	1.8

* found with GANGSTA, using a bonus for sequential SSE order.
^{n} CP found in nature.
^{a} artificially designed CP.

found to be in overall agreement for all methods, except the pair 1*onr*-1*fba* for which Dali and GANGSTA proposed the same overall alignment but none of the other methods agrees with them or each other. Since there is no exact alignment given in literature, each of the alignments could be right. However, the agreement between Dali and GANGSTA (see *mean* in Table 5.7) as well as the prediction of a CP in the GANGTSA alignment give credibility to these alignments. Only MAMMOTH produced better alignment RMSDs as GANGSTA, but MAMMOTH could only align 57 residues on average instead of GANGSTA 113. The SSAP method resulted in the largest number of aligned residues but accompanied by very large RMSD values resulting in overall different alignments for most of the naturally occurring CPs.

5.3.6 Different Contact Definitions

Here, we compared the quality of pairwise GANGSTA alignments using different contact type definitions. For all pairwise alignments we used Stride [74] for SSE assignment. As alignment task we used the Fischer dataset [70] as described in the Appendix D.9 that consist of 70 SCOP [169] *reference* protein structures that have to be aligned against a database of 333 structures. The Fischer dataset has been designed for evaluating sequence-to-structure alignments, also known as fold recognition or threading, and covers a wide range of different protein families and folds including protein pairs showing low sequence similarity. For every *reference* domain one or more *target* domains from the database of 333 structures are defined according to SCOP fold classification (see Table D.4). In the dataset exists no *reference-target* pair that has a sequence identity over 35%. The authors proposed that all *reference-target* alignments could be superimposed within a RMSD threshold of 3Å aligning at most half of the residues of the larger structure with residues of the smaller structure (alignments were

performed manually by the authors). A match occurs when one of the *reference-target* alignments were found on the top rank or within the top three ranks while ranking all 333 alignments per *reference* structure according to the GANGSTA *score* (Equation 5.16). In Table 5.8 the results for the five different contact

Table 5.8: **Different contact definitions for the Fischer dataset [70].** The Fischer dataset comprises overall 70 *reference-target* pairs (see Appendix D.9). The first column gives the type of contact definition used. Columns two and three give the number of *reference-target* alignments that were ranked according the GANGSTA *score* within the TOP1 and TOP3 best alignments, respectively. Columns four and five give the mean GANGSTA *score* for the overall best alignment and the best *reference-target* alignment. Column six and seven show the corresponding mean RMSD values.

contact	no hits		score[b]		RMSD[Å]	
type[a]	TOP1	TOP3	best	target	best	target
ca	66	70	0.033	0.036	1.68	1.79
cb	67	70	0.042	0.045	1.96	2.11
all	67	70	0.041	0.043	2.10	2.12
vdW	65	70	0.042	0.044	1.53	1.54
voronoi	66	70	0.041	0.043	2.05	2.10

[a] as defined in see Section 2.7.3.

[b] as defined in Equation 5.16, optimum is 0.

definitions, as defined in Section 2.7.3, are shown. For all contact definition types GANGSTA was able to find a correct *reference-target* alignment within the best three alignments using the GANGSTA *score* for ranking. The best *reference-target* pair has for all contact types a mean GANGSTA *score* of at most 0.045 what corresponds to a confidence level of 95%. Overall, the GANGSTA *score* is for Cα-atom contacts better than for all other contact types. Aligning two similar folds means aligning two similar contact maps. Therefore, the more contacts are defined between SSEs the better works the GA on secondary structure level. Additionally, the more residues are aligned the more contacts can be conserved on the residue level. Interestingly, alignments using van-der-Waals contacts result in significant better RMSD values than alignments using other contact types. This can be explained by the fact that this contact definition type is very well defined in terms of spatial proximity. GANGSTA is able to align these contacts, if present, with very low RMSD values. The drawback of this contact type definition is that only very similar contacts can be aligned explaining the lower GANGSTA *scores* resulting from lower contact map overlaps on the residue level.

5.4 Discussion

We have tested GANGSTA on different datasets to assess its performance for challenging tasks in protein structure alignment. These include

1. classifications of protein superfamilies,

2. searching for structure alignments with non-sequential SSE connectivity, and

3. comparisons with other methods considering datasets of protein structures that require sequential SSE connectivity.

It was shown that for structure alignments from different SCOP superfamilies the distribution of GANGSTA *scores* follows the well-known Gumbel distribution. Levitt and Gerstein [146], MAMMOTH [179], and FASE [232] reported the same distribution before. These methods use all different measures of structural similarity and different optimization algorithms. At confidence levels of 95% and 99% significance we found 67% and 48% true-positives, respectively. The discrimination between structurally related and non-related proteins, as pictured in the coverage-error plot in Figure 5.9, is comparable with that of other methods: At a confidence level of 99% PrISM [244] reported 54% and MAMMOTH 50% true-positives. At a confidence level of 95% MAMMOTH reported 60% and FASE 72% true-positives. In contrast to these studies GANGSTA reports the P-value for SCOP [169] superfamily classification instead of SCOP fold classification. This test is more demanding, since protein structures may share the same SCOP fold but belong to different SCOP superfamilies. Generally protein structure alignments are validated using classification schemes that discriminate according to specified criteria between related and unrelated structures. For this purpose most studies use the CATH [177] or SCOP database of classified proteins. However, these databases were also generated with specific classification criteria, which naturally may build in biases. This adds to the difficulties of fairly comparing different methods of protein structure alignment. Additionally, many methods, including GANGSTA, demonstrate the quality of their methods using receiver operating characteristic (ROC) curves. Kolodny *et* al. [131] showed that comparisons between different methods based on ROC curves are often unsatisfactory with respect to the quality of protein structure alignment. So far, the best insight into the quality of protein structure alignments can be obtained by visual inspection. This depends on the structural and functional features upon which the viewer focuses and is obviously subjective in nature.

Protein structure alignments from different SCOP families and superfamilies have demonstrated that GANGSTA is able to find reasonable structure alignments that may provide new insights for structure-function relationships of proteins and also for structural motifs that occur with different SSE connectivity. The results for the Rossmann dataset demonstrate that GANGSTA finds structural similarities for proteins that are known to have similar function but have no obvious structural or sequence similarity. The Rossmann structure motifs are ubiquitous, appearing in the large enzyme family of kinases [42] that catalyze the transfer of phosphate groups. In these proteins, the Rossmann structure motif constitutes just a small fraction of the whole structure, which can differ significantly in the remaining part of the structure. Hence, SCOP classifies these proteins in different superfamilies, such as MurCD N-terminal domain, FAD/NAD(P)-binding domain, HAD-like, NAD(P)-binding Rossmann-fold domains, DHS-like NAD/FAD-binding domain, UDP-Glycosyltransferase/glycogen phosphorylase, and Flavodoxin-like. These structural similarities hint at functional similarity in nucleotide binding. GANGSTA is able to detect the structural similarity of those proteins despite their topological differences with

respect to SSE connectivity. Protein structures with different SSE connectivity often exhibit large structural variations in terms of RMSD, but can simultaneously have large contact overlaps and a GANGSTA *score* (Equation 5.16) close to zero, corresponding to high quality structure alignment.

Another example for structural similar protein structures with non-trivial relationships are circular permuted proteins (CPs). In the case of CPs, one break in the sequential connectivity is inherently necessary. Most methods trying to identify CPs mostly work on sequence level. But sequence alignment methods are unsuitable for related proteins with low sequence identity. Therefore, structure alignment tools are necessary that are able to detect CPs despite of missing sequence similarity. Since most of the structure alignment methods preserve the sequentially order of the SSEs they are able to produce a good structural alignment mostly only for one of the two domains of CPs. Therefore, they are unable to detect CPs directly by aligning complete chains. GANGSTA, however, has been designed to generate alignments without sequential ordering constraints, and, therefore, it is not only able to align CPs, but also to detect the specific break in the sequential order of the SSEs. We showed that GANGSTA can find alignments that show the specific behavior predicted for CPs and that can in principle be used to detect possible unknown CPs by structure alignment only, an important task for the automated classification procedure of protein structures.

Although GANGSTA was designed and implemented specifically to find unusual protein-structure alignments with non-sequential SSE connectivity that are hard to detect, we could show that even for sequential SSE connectivity GANGSTA is able to compete with other established protein structure alignment methods like DALI [103], VAST [79], YAKUSA [37], and CE [210]. Regarding the number of aligned residues and the overall RMSD results individual pairwise protein-structure alignments with GANGSTA are generally somewhat inferior to the results obtained with other methods. But, for the more imprecise database scan method GANGSTA outperforms structure-alignment methods that consider sequential SSE connectivity only.

GANGSTA can be used with different contact type definitions for pairwise residue contacts. Although the total numbers of contacts are varying significantly for the different contact types, we could show that the performance of GANGSTA is not dependent on the used contact definitions. Moreover, it produces for all contact types highly significant structural alignments with low RMSD values as well as correct ranking using the GANGSTA *scores* for global ranking.