Chapter 7

Conclusion and Future Work

In this thesis, we have made progress toward analyzing evolutionary and functional important aspects of proteins. No problem is more central to the workings of the cell than understanding protein structure, as proteins are the building blocks of the most important cellular processes. More generally, determining the role each protein plays in the cell is one of the basic steps in making sense of cellular networks. Due to the more and more increasing data coming from high-throughput experiments the assignment of function to proteins and the understanding of their evolution can only be achieved, if similarities between protein structures can automatically be determined using suitable structure alignment methods. A problem that is also relevant for the classification of protein structures. We could show that our methods are able to detect biological meaningful similarities that are not detectable for sequence-based or other state-of-the-art structure alignment methods, and thus provide new insights into evolutionary and functional relationships of protein structures.

Biologically useful structure alignment methods need adequate representations. The representation of protein structures as residue contact maps and at the secondary structure level has two major advantages, efficiency and accuracy: efficiency, because the average number of SSEs in a globular protein is smaller by 10-fold compared to the average number of residues, and the average number of residues is smaller by 10-fold compared to the average number of atoms. Protein structures can be treated more easily and significant improvement in computation can be achieved, especially when many structures are analyzed. Accuracy due to the high atom density in protein structure, because it is possible to superimpose any random pair of proteins so that many of their atoms are aligned. However, such an alignment is most probably biologically irrelevant, because it would not reflect any evolutionary or functional relatedness. Additionally, the definition of protein graphs, either for the representation of protein topologies (see Chapter 3) or for structural alignment (see Chapter 5), has the advantage to describe protein structures without reference to atom coordinates. Rather, only contact relationships are considered. This allows for a comparison of protein structures that do not superimpose well but show similar spatial arrangements of their SSEs. This is in contrast to most other methods (see

Chapter 4) that base their protein representation on atomic level. It enables us to detect similarities of protein structures in a wider range than methods based on geometric descriptions only. Since we are using the contact information between SSEs, essential information on sequential relationships within the protein structure is encoded without necessarily fixing the sequential order of the SSEs.

In order to define a unique description of protein topology, we described the secondary structure topology of a protein by methods of applied graph theory. We defined the secondary structure topology of a protein as an undirected labeled graph on three description levels of its SSEs: the Alpha graph, the Beta graph, and the Alpha-Beta graph. For each graph type exist four linear notations and corresponding graphical representations. We developed the PTGL database that enables the user to search for the specific proteins or for certain topologies or sub-topologies, and for sequence similarity in SSEs. The database can be used for any kind of theoretical protein structure analysis, protein structure prediction, and protein function prediction. Additionally, we have developed a system, which supports fast pattern searching over PTGL linear notations. The search engine is based on simple regular-expressions to search for the most common structural motifs. Users can search on motifs from a library or define their own search patterns. We are now in the process of enhancing the database system to permit users to compare the topology of a given structure with all the other structures in a database using graph-theoretical methods. Additionally, we want to provide more accurate descriptions of structural motifs.

We developed a hierarchical method for protein structure alignment, called GANGSTA. The first stage of this method is a maximal common subgraph (MCS) search of protein graphs based on secondary structure representation. Therefore, a genetic algorithm (GA) was developed to search for maximal common substructures. Additionally, we determined the exact graph-theoretical algorithm (ExactGANGSTA) for this problem. The experiments in Section 6.4 showed that the GA method is an adequate search strategy to solve the MCS problem. Although the exact solution produces better results in terms of the used objective function, the quality of SSE alignments produced from the GA method are comparable to the exact solution in most cases. The second stage of the GANGSTA method maximizes the contact map overlap using a residue level description. After the alignment procedure the superposition of the two protein structures is performed to compute the transformation that minimizes the RMSD. GANGSTA is able to find protein structure alignments independent of the SSE connectivity. Such a capability is essential for detecting structural similarities that exist due to convergent evolution, but with no fold homology. In certain cases, where order dependency is preferred, there is also an option in GANGSTA to consider the order of the SSEs. This option can be used to cluster topologically similar proteins or to obtain a structure-based sequence alignment. We could show that functionally related protein domains can have large structural variations in terms of RMSD. The contact map overlap and the newly introduced GANGSTA score can identify structures with different SSE connectivity not detectable by sequence-based alignment methods or methods maintaining the SSE connectivity. Structure alignment methods considering the geometry of loops that connect the regularly structured SSEs in a protein have a strong bias for sequential SSE connectivity. Hence, these methods have difficulty finding structural alignments that are non-sequential in SSE connectivity. Even if a protein fold cannot be aligned to another protein structure

while maintaining the SSE connectivity, structural similarity may still exist for different SSE connectivity despite large RMSD values. GANGSTA tends to align large fold motifs regardless of the SSE connectivity. This is due to the following features:

- GANGSTA does not optimize distances between residue pairs, but maximizes the number of residue pair contacts.
- The number of gaps, i.e., the number of not aligned SSEs in the smaller structure, is restricted to make sure that a maximum number of SSEs and consequently also of residues are aligned.
- GANGSTA ignores loop structures, which helps to find structure alignments that are non-sequential in SSE connectivity.
- GANGSTA is able to construct decoy structures (alternative alignments).
- GANGSTA is robust against different contact type definitions, i.e., independently of the contact type used it detects the correct arrangement of SSEs.

Our method is able to detect functional important sequential and non-sequential structural similarities. The quality of sequential alignments is comparable to other state-of-the-art structure alignment methods.

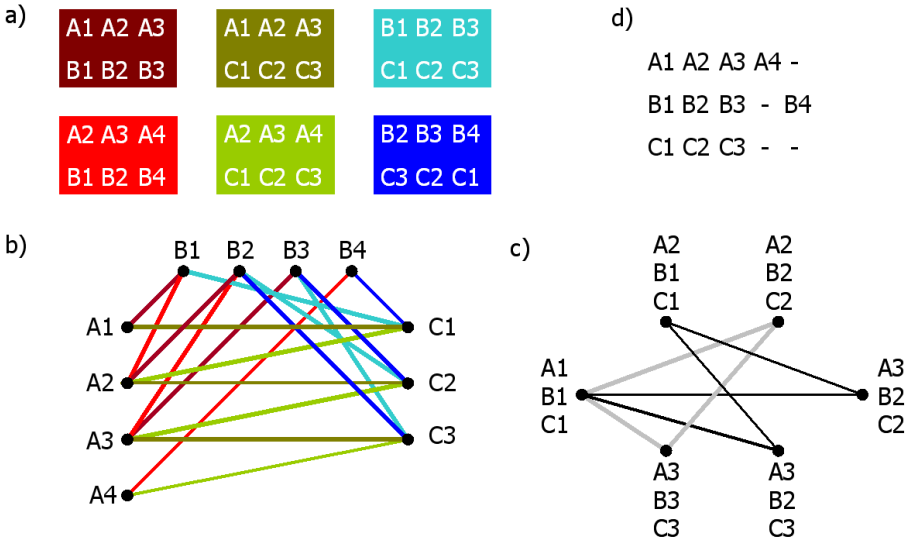

Figure 7.1: **MultipleGANGSTA.** a) Pairwise SSE alignments between three protein graphs (A,B,C). For each protein pair two alternative alignments are shown. SSE numbering from N- to C-terminus. Every alignment is highlighted with a different color. b) SSE-alignment graph: cliques represent possible columns in a multiple alignment. Edges are colored according to the SSE-alignment. Cliques are shown with bold edges. c) Column-consistency graph. The maximum clique is shown in gray. d) Multiple structure alignment built from the maximum clique.

As discussed in the previous chapters, the alignment of proteins is crucial for many purposes in biology. Pairwise sequence alignment is unreliable if the

proteins diverge on the sequence level. Although multiple sequence alignments are more accurate, even they are inadequate when dealing with distantly related proteins sharing little sequence similarity. Many methods have been developed to address the pairwise structural alignment task (see Chapter 4). In contrast, only a few methods are available for aligning multiple structures. However, it seems to be clear that multiple structure alignment gives more insight in evolutionary relatedness and is thus a much more powerful method. Most of the currently available methods for multiple structural alignment are based on pairwise structure alignments. They find common substructures through a series of comparisons between pairs of molecules. These methods combine a pairwise structural alignment and a heuristic to merge pairwise alignments into a multiple alignment. Well-known methods of this type are SSAPm [226], PrISM [244], STAMP [202], or MUSTANG [132]. The pairwise-based methods have the limitation that in each pairwise alignment the only available information is about the two molecules involved. Thus, alignments optimal for the whole input set might be missed, if they are not also optimal for every pair [64]. There are other methods, like MASS [63], Escalier *et al.* [66], MUSTA [139] and MultiProt [207], that are considering all the given structures simultaneously, rather than initiating from pairwise alignments. They all try to detect structurally similar common pieces, which are then extended to compute global alignments. The majority of the methods for multiple structure alignment, with the exception of MASS, use dynamic programming [170]. As a result, they have the disadvantage of being dependent on the sequence order of the polypeptide chain.

As described in this thesis, GANGSTA is a sequence-order independent method. Additionally, GANGSTA is able to produce different non-sequential alignments for one pair of protein structures, either using the GA or Exact-GANGSTA for the first stage, the SSE alignment level. Therefore, GANG-STA has not the limitation that local similarities have to be necessarily optimal for all pairs of proteins. We propose the following schema for multiple structure alignment using GANGSTA, called the MultipleGANGSTA method:

1. Given n protein structures represented as protein graphs $PG_n = (V_n, E_n)$, perform all $\frac{n!}{(n-2)!2!2} = f$ pairwise GANGSTA alignments.

2. For every pairwise alignment A_i $(i = (1, \ldots, f))$, use the m best SSE alignments according to the GANGSTA *score* (Figure 7.1a).

3. Generate a *SSE-alignment graph* G_c consisting of the vertex set $V_c = V_1 \cup V_2 \cup \ldots V_n$. The edge set is defined by all pairwise SSE mappings of all $n \times m$ valid GANGSTA-SSE-alignments (Definition 23). An example is given in Figure 7.1b.

4. Search all cliques of minimal size k in the SSE-alignment graph. A clique in G_c represents then a column of a valid multiple alignment, where minimum k structures are not gapped. In Figure 7.1b only cliques of size 3 are marked.

5. To find the maximal global multiple structure alignment, the maximal number of columns has to be combined. Therefore, one has to define when two columns are consistent: no two columns are allowed two contain the same SSE from the same structure. Then, a column-consistency graph can

be build. The vertex set represents all valid multiple alignment columns, and an edge is defined, if two columns are consistent (Figure 7.1c).

6. A clique in the *column-consistency graph* represents a maximal global multiple structure alignment. In Figure 7.1a only cliques of size 3 are marked. The resulting multiple structure alignment is shown in Figure 7.1d.

The advantage of such an implementation would be that for a single pairwise alignment not only the optimal alignment is used, but in addition the $m-1$ best suboptimal alignments, and that all optimal and suboptimal alignments from all pairs of proteins could be merged into one multiple alignment. This would be a great advantage over the progressive multiple structure alignment methods that are using one alignment as a pivot element and align all other structures to this particular structure. Therefore, the resulting alignments are often strongly biased toward the pivot protein structure. Additionally, the MultipleGANGSTA method would combine the global view coming from the pairwise alignments with local search strategies: The global pairwise GANGSTA alignments provide a pool of pairwise SSE mappings. The SSE-alignment graph contains pairs of SSEs that can be combined into a greater alignment, a strategy, for example, MASS is also employing. This capability prevents the loss of good alignments due to local structural outliers, and will be highly useful in protein classification of heterogeneous ensembles of superfamilies or folds. Due the extensive use of clique searching the method is presumably only applicable to small numbers of structures, but again heuristic search strategies could be developed to make the search faster. Since we are searching in the SSE-alignment graph for all cliques with a minimum size and in the column-consistency graph only for the maximum clique representing the multiple alignment with the maximal number of aligned SSEs, the BK-algorithm (see Section 6.2.3) should be able to traverse the search tree quickly.

The MultipleGANGSTA method would be a new sequence-independent multiple structure alignment method combining the advantages of global and local search strategies by applying alternative pairwise alignments and local SSE similarities that could be used to search for non-trivial spatial arrangements in sets of proteins showing no sequential similarity at all.