Chapter 4

Structure Comparison and Classification

4.1 The Structure Alignment Problem

The three-dimensional (3D) structure of a protein provides much more information than the amino acid sequence alone to understand the function and evolutionary history of the protein. Thus, the comparison of structures becomes an even more important task than the comparison of sequences. While the comparison of sequences is almost on a residue level, the comparison of structures can be either on a *fine* level (mostly atoms, dihedral angles, or residues) or on a *coarse level* (mostly on SSE or supersecondary structure element level). The two levels have different purposes. Comparing protein structure on a fine level is important to deduce common local similarities, such as active sites or binding sites. Fine level structure descriptions can be used to make hypothesis about the possible function of a protein, e.g., the finding of catalytic triads in serine proteinases and lipases. Most of the existing methods use protein backbone descriptions, where one residue is represented by one or more atoms, most often Cα-atoms. The coarse level is mostly used for comparison of proteins on a global level, and is also used to classify proteins into structural classes of hierarchies [79, 98, 121, 134, 149]. As for sequence alignments, structural alignment methods can either be global or local: global means superimposing most of the atoms in the corresponding structures, whereas local alignments are searching for all matching maximum substructures in both proteins.

Following the tradition from sequence analysis, the results of comparing protein structures are expressed in form of an alignment: a set of one to one equivalences between positions in both proteins. But in structure comparison it is not always necessary that compared elements, i.e., residues or SSEs, occur in the same sequential ordering in every position, i.e., in structural alignments there exist correspondences that does not maintain the sequential order from the N- to C-terminus. This strictly contradicts the definition of alignments in literature (see, e.g., [65, 214]) that the sequential ordering has to be preserved. Due to this fact, it would be better to denote the correspondences in structure comparisons as mappings instead of using the term alignment, but, as also been done in literature, we will still use this somewhat misleading term throughout

this thesis. A general definition of an alignment, including the non-sequential case, can be given as follows:

Definition 15 (Alignment)**.** *Let Σ be some alphabet excluding the gap character '-', and let $\hat{\Sigma} = \Sigma \cup \{\text{-}\}$. Given a pair of proteins s^1 and s^2 represented each a string over Σ, we call $A = (\hat{s}^1, \hat{s}^2)$ an* alignment *with gaps if and only if the following conditions are satisfied: (a) The sequences $\hat{s}^1$, $\hat{s}^2$ are over the alphabet $\hat{\Sigma}$, (b) $\hat{s}^1$ and $\hat{s}^2$ have the same length $|A|$, (c) sequence $\hat{s}^i$ without '-' corresponds to s^i with $i = (1, 2)$, and (d) there is no index j such that $\hat{s}^1_j = \hat{s}^2_j =$ '-'. For $i = (1, 2)$ we define $M_i(j)$ as the mapping of s^i_j to its position in the alignment, and by $M_i^{-1}(j)$ the mapping from the position in the alignment to the actual position in the sequence. If $\hat{s}^1_j \neq$ '-' and $\hat{s}^2_j \neq$ '-', $1 \leq j \leq |A|$, then we say that $s^1_{M_i^{-1}(j)}$ is aligned to $s^2_{M_l^{-1}(j)}$, and to a gap otherwise.*

Alphabets commonly used for protein structure alignment are the 20 letter amino-acid alphabet (see also Table G.1 in the Appendix) for contact map alignments, and the two-letter alphabet $\Sigma = \{H, E\}$, denoting helices and strands, for protein graph alignments, respectively.

As with many other comparison problems, complications arise, as there is neither one best way to make the comparison nor to evaluate the answer. The situation also exists in sequence comparison where, although there exist optimal alignment algorithms for pairs of sequences [170, 214], the results depend on a model for the relatedness between sequences. In structure comparison, with the added complexity of structure relative to sequence, the models describing similarity varying much more. Additionally, it is not clear, if the one and only true alignment exists at all. The implicit assumption about the existence of such a unique structure alignment is only borrowed from the optimal sequence alignment problem, but there is a fundamental difference between these two problems. When comparing sequences of two related proteins, we are certain that there is a unique, correct solution, even if we are not able to find it. The assumption is that both proteins have evolved by a series of evolutionary events from a common ancestor, and there exists an actual molecular process that is linking one sequence to another. Therefore, there is a unique, one to one correspondence between positions in each protein and positions in the common ancestor. This correspondence could be used to create the true alignment. Of course, we do not know this correspondence, so the sequence alignments are approximate, subject to our approximate knowledge of the underlying evolutionary process, but optimal to the assumed model of evolution. On the other hand, the process of transforming one protein structure into another, or deriving both from a common ancestor, does not exist. The two proteins may fold due to a different balance of first principle forces and there is not necessarily a one to one correspondence between positions in both proteins. Therefore, it is entirely possible that the structural alignment between two distantly related proteins could be different from the correct evolutionary alignment, if, for instance, some fragments of the sequence changed their function during evolution [83].

It is important to differ between structure alignment and structure superposition. These terms are sometimes used interchangeably, but there is one major difference: structure superposition assumes that you already have an alignment of at least some residues between two protein structures. Typically, these residues are represented by one type of atom, e.g., the Cα-atoms, and

the task becomes to find an optimal transformation that minimizes the distance between aligned atoms. The most commonly used measure of the difference between two structures is the *root-mean-square deviation* (RMSD) in atomic position after optimal superposition. The RMSD between the two structures is given as follows:

Definition 16 (RMSD)**.** *Let* $(\vec{x_1}, \vec{y_1}), \ldots, (\vec{x_N}, \vec{y_N})$ *be the 3D coordinate vectors of aligned elements of the alignment between two protein structures A and B* ($\vec{x_i}$ *from A and* $\vec{y_i}$ *from B), then the RMSD is defined as*

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\vec{x_i} - \vec{y_i})^2} \tag{4.1}$$

where N is the number of aligned residues.

Superposition is a much easier problem than the structure alignment problem. Structure superposition methods have been around for a long time, see, for example, [60, 118, 122]. In many structure alignment algorithms superposition plays an important role once the alignment is complete. Then, for a defined RMSD threshold l, the task is to find the maximum number of elements that can be superimposed such that the distance between two aligned elements is less or equal to l.

There is a large amount of literature on methods for pairwise protein structure alignment. Orengo *et al.* [175] provided an overview of the field until 1994. Gibrat *et al.* [79] reviewed results from structure alignment using different methods and Lemmen and Lengauer [140] summarized protein structure alignment in the context of the general problem of structure alignment and superposition in drug design. Lancia and Istrail surveyed on various alignment models [137] and Ferrari and Guerra on geometric methods for protein structure comparison [67]. Additionally, there are recent studies [131,172] that performed large-scale evaluations on the most recommended protein structure alignment methods and similarity measures. In summery, the general protein structure alignment process can be described by three or four steps:

1. Represent two structures (polypeptide chains, domains, or other amino acid fragments) in some element-based representation form (atoms, angles, residues, SSEs, protein graphs).

2. Compare both structures by comparing the elements using an appropriate similarity measure together with a suitable optimization technique resulting in an element-based mapping.

3. Convert the element-based mapping into an atom- or residue-based alignment and optimize the superposition of the alignment between both structures.

4. Optional measure the statistical significance of the alignment against some random set of structure alignments.

Given this general approach there are three classes of problems that the defined algorithms try to solve. The first is to search for an optimal alignment between any given pair of proteins. The second, the optimal superposition for a given

alignment, can easily be found using standard algorithms as described above. The third is, given a target structure that have to be compared against a set of known template structures in some rank order, which structures in the set of known structures are most like the target structure (see Section 4.3).

With respect to the first problem, the apparently simple question about similarity between two protein structures is, in fact, quite difficult. On one hand, the considerable success of threading methods [29, 84, 114, 158, 243] seems to suggest that there is a significant similarity in interactions stabilizing structurally similar but sequentially unrelated proteins. On the other hand, other groups claimed that interactions stabilizing different proteins from the same structural family are, in fact, quite different [203]. There are different measures to quantify the similarity. As stated above, the RMSD is the most popular similarity measure [154]. Other groups use the difference between distance maps of two proteins [103], contact map overlap [86], or more complicated scoring systems [69, 143, 202, 204], including, among others, such additional structural features as local secondary structure, hydrogen bonding pattern, burial status, or interaction environment. The most important and commonly used similarity measures will be introduced in the next Section.

From a mathematical point of view protein structure alignment is either a *NP-hard* problem [138], arising from the combination of the non-locality of the used scoring functions and similarity measures and the existence of insertions/deletions in the alignment or using contact maps [88], or even *NP-complete* when searching for maximal common subgraphs [144] due to structure representation and modeling. More generally, protein structure alignment is computationally very expensive, and, furthermore, up to now there is no fast structural alignment algorithm that guarantees to be optimal in reasonable time. Therefore, all of the existing structural alignment methods use some simplifications, either of the scoring function or of the search procedure, to arrive at a reasonable, if not the optimal, alignment. Although different heuristics employed by different methods tend to recognize similar folds, they will not provide exactly the same structure alignments. In fact, two structure alignment methods may produce alignments that differ in every position [83]. Breaking down the dimensionality of the problem by performing at least part of the search at the level of SSEs is a commonly used approach [103, 129, 162, 204]. Different groups used different optimization methods, such as dynamic programming, two level dynamic programming [115], or Monte Carlo minimization [86, 103, 204]. Algorithms are described in the formalism of sequence homology analysis, graph-theory [129, 162], or computer vision technique [69]. As a result, protein structure comparison is a vibrant, very active field where different techniques are used to answer the same question of how different/similar protein structures are, and where is still need for new fast and accurate alignment strategies. In the Appendix E we provide of an alphabetically ordered list of state-of-the-art methods for protein structure alignment that were used in recent evaluation tests for structural alignment methods [37, 172] or that were used throughout this thesis.

4.2 Similarity Measures

When comparing protein structures it is widely known that, beyond close sequence similarity, there is no uniquely correct structural alignment of two proteins [83]. Different alignments are achieved depending on which properties and representations are compared. This adds another difficulty into the assessment of the result of a comparison. However, it is reasonable to assume that unique alignments or better mappings exist for essential 'core' regions of homologous protein structures.

Several measures for protein structure comparison have been proposed. Most of the methods for protein structure alignment quantify the quality of the alignment on the basis of geometric properties of the set of points representing the structures. The RMSD [118] (Equation 4.1) is an essential part of most scoring systems for structure alignment methods, but implicates some serious drawbacks [3, 24, 154]:

- The best structural alignment does not always yield the minimal RMSD.
- The significance of RMSD depends on the size of the structures.
- The significance of RMSD varies with the type of proteins.
- RMSD it is not a good measure when all equivalent parts of the proteins cannot be simultaneously superposed.
- Using RMSD all atoms are usually treated equally, but, for example, residues on the surface have a higher degree of freedom than those in the core.
- RMSD depends more on the worst fitting atoms than on the best-fitting atom.
- RMSD does not penalize gaps.

As a result, small local structural deviations can result in high RMSD values, even when global topologies of the compared structures are similar. Therefore, RMSD is a useful measure of structural similarity only for closely related proteins [165].

Godzik *et al.* [83] could show that different similarity measures used for structural comparisons may lead to different alignments. In principle, for the same pairs of proteins there exist different alignments using Cα-RMSD, Cα distance difference, contact map overlap, or some other structural based scoring. Even within any given measure, alignments are often degenerate and whole families of alignments can be generated with almost the same alignment score. Thus, it should be clear that there is not such a thing as the one structural alignment that could be used as a 'gold standard' to judge and validate other alignment methods, such as threading or sequence alignments. Instead, there are different structural alignments, emphasizing structural similarity as seen by a certain similarity measure. For closely homologous proteins, the differences between various structural alignments are minor and confined to residues outside the hydrophobic core. For more distance homologs and unrelated proteins with similar structure alignments start to differ more increasingly from each other, not only in loop regions or irregular fragments of the structure, but also for well-defined

arrangements within the protein core. Here, different similarity measures, such as RMSD or distance difference, contradict other similarity measures such as hydrogen bond networks or contact maps. This problem greatly influences the analysis of structural similarities, leading to contradictory results reported by groups using different types of alignments [131]. Every measure of structural similarity is, in fact, based on the actual representation form.

A quantitative measure of the similarities of protein structure is also essential for a critical assessment of the quality of protein structure predictions, such as generated for CASP (a community-wide experiment on the Critical Assessment of techniques for Structure Prediction) [167]. In the special case of comparing a predicted structure with the corresponding experimental structure, the equivalence list of residues is known because the two sequences are identical, which reduces the complexity of the problem. On the other hand, each prediction may omit different residues and different parts of the structure may have different accuracies. In CASP, the GDT_TS score is used. GDT_TS measures the number of equivalent residues for a given RMSD threshold [247]:

$$GDT\_TS = \frac{1}{4}\left[N1 + N2 + N4 + N8\right] \quad , \tag{4.2}$$

where Nn is the number of residues superimposed under the distance threshold nÅ.

4.2.1 Geometry-based Measures

Many methods compare the respective distance matrices of each structure, trying to match the corresponding intramolecular distances for selected aligned substructures [103, 164, 210, 227]. Other methods compare the structures directly after superposition of aligned substructures, trying to match the positions of corresponding atoms [3,152,216,234,242]. Interestingly, there is no consensus on the definition of a match of distances or of atomic positions needed for either of these two schemes. When comparing two pairs of atoms between two structures, Taylor and Orengo [227] defined straight forward a distance or similarity score based on the RMSD in the form

$$\frac{a}{D+b} \quad ,$$

where D is the difference between two intramolecular distances, and a and b are arbitrarily defined constant values. Holm and Sander [103] defined for DALI a similarity score as

$$\left(a - \frac{D}{< D >}\right) exp\left(-\left(\frac{< D >}{b}\right)^2\right) \quad ,$$

where $< D >$ is the average of the two intramolecular distances. In [198] and [202] a score is defined as follows:

$$exp\left(-\left(\frac{D}{a}\right)^2\right) exp\left(-\left(\frac{S}{a}\right)^2\right) \quad ,$$

where S takes into account local neighbors for each pairs of atoms. As another example of a scoring scheme for minimizing intermolecular distances, Levitt and

co-workers [78, 216] defined a score

$$LG = \frac{a}{1 + \left(\frac{R}{b}\right)^2} \ ,$$

where R is the distance between a pair of corresponding atoms in the two strutures. This score was defined as a more reliable indicator of structure similarity than RMSD, because it depends especially on the best-fitting pairs of atoms, whereas RMSD gives equal weight to all pairs of atoms. Zhang and Skolnick [249] varied the LG score defining the TM-score:

$$TM - score = \max \left[\frac{1}{L_N} \sum_{i=1}^{L_T} \frac{1}{1 + \left(\frac{d_i}{d_0}\right)^2} \right] \tag{4.3}$$

where L_N is the aligned length of the first protein structure (native structure), L_T the sequence length of the second structure (template), d_i is the difference between the i-th pair of aligned residues and d_0 is a scaling factor. Given the alignment the TM-score is used to search for the best superposition by maximizing the LG-score instead of minimizing RMSD [118].

Interestingly, Lesk [141] proposed replacing the Euclidean norm in the RMSD definition by the L_∞ norm, also called the *Chebyshev norm*, yielding a new score:

$$S_\infty = max_{i \in [1,..,N]} \left\{ \|\vec{x_i} - \vec{y_i}\| \right\} \ .$$

S_∞ reports the worst fitting pair of atoms (after optimal superposition of the two structures), and, as such, is even more sensitive to outliers than the RMSD.

4.2.2 Contact Map Overlap

The *contact map overlap* (CMO) [85, 136] is based on the basic notion of contacts between two residues. A contact map is an abstraction of a 3D structure, where the 3D conformation is represented by a graph or a symmetrical matrix of contacts (see Definition 3). The CMO is a measure for contacts that are conserved in the alignment. The CMO for the alignment of two protein structures is defined as follows [21]:

Definition 17 (Contact map overlap)**.** *The contact map overlap q between two proteins A and B represented by their residue contact maps C_A and C_B and their respective sequence lengths N_A and N_B is defined as*

$$q(C_A, C_B) = \frac{\sum_{\substack{i,j \\ i<j}}^{M} C^A_{i,j} C^B_{map(i),map(j)}}{min \left[\sum_{\substack{i,j \\ i<j}}^{N_A} C^A_{i,j}, \sum_{\substack{i,j \\ i<j}}^{N_B} C^B_{map(i),map(j)} \right]} \ , \tag{4.4}$$

with M the number of aligned residues, $C^S_{i,j}$ representing a contact between residues i and j in structure S, and $map(x)$ as a mapping of residue x from A onto residue $map(x)$ in B.

The CMO between two contact maps is equal to the number of contacts the contact maps have in common. The CMO ranges from $q = 1$ if the two contact maps are identical to $q = 0$ for structures having no contact in common.

4.2.3 Other Measures

Suyama *et al.* [217] proposed another approach in which they ignored the 3D geometry at all and compared structures on the basis of sequence profiles [29] using dynamic programming. These profiles include information on solvent accessibility, hydrogen bonds, local secondary structure states, and sidechain packing. Although this method is able to align two-domain proteins with different relative orientations of the two domains, it often generates inaccurate alignments. Jung and Lee [116] improved this method by iteratively refining the initial profile alignment using dynamic programming and 3D superposition. Their method, referred as SHEBA, was found to be fast and as reliable as other alignment techniques. Kawabata and Nishikawa [121] proposed a novel scoring scheme for generating structure alignments based on the Markov transition model of evolution.

Yang and Honig [244] described a new measure for protein structure similarity, the protein structural distance (PSD). The PSD includes both secondary structure alignment score and RMSD. It thus incorporated the resolution power of RMSD for closely related structures and the secondary structure score for proteins that can be very different. They showed that there is continuous aspect of protein conformation space, what is in apparent disagreement with structural classification schemes like SCOP [169] and CATH [176] (see Section 4.4).

4.3 Statistical Significance

The straightforward approach for assessing the significance of the result of comparing two structures is to do the same as for sequence alignment: to calculate a distribution over what scores can be expected only by chance. The idea is to relate the alignment score to the density function of the scores between random sequences. For structure alignment, this can be done by constructing a set of random structures and comparing those with one of the native structures. However, as for sequences, some basic non-random properties should remain. For sequence these include the length and amino acid distribution, for structures it would also be length and amino acid distribution, but also overall shape, secondary structure content and packing density. As an extreme example, imagine that the aligned proteins contain only helices, and the unrelated set of structures contains only proteins consisting of sheets. Relative to the background, the two structures containing helices would appear more related than they should do. This demonstrates that the set of 'random' structures should have not only the 'non-random' properties intact but also include representatives of all typical protein structure classes (see Section 4.4 for details). Ideally, random structure models should be calculated for each comparison to match the non-random properties of the query structures. There are essentially two methods proposed to generate random background distributions for a given alignment [65]: constructing random structures from scratch, or using known structures from structural databases.

4.3.1 Geometry-based Random Models

When generating random structure models one has to ensure that the models can represent native structures, i.e., in steric terms that the model has a compact fold built of non-overlapping atoms. The simplest procedure to build random protein structures is a self-avoiding walk [18, 223]. A polypeptide chain is built successively by addition of random residues to the C-terminal end of the last added residue such that steric clashing is avoided and residue packing is favored resulting in compact structures and preserving the total sequence length. Another geometry-based method is to randomly change the values in the distance matrix. Depending on the degree of noise introduced, a range of structural variants (also called *decoy* structures) can be generated [15] by maintaining most of the main properties of the native structure such as secondary structure and the hydrophobic core.

4.3.2 Use of Databases

The generation of random models by geometry cannot be made for each comparison without extensive computation. Instead, it is easier to compare one of the target structures with a structure database and try to remove those scores coming from structures that are homologous to the target structure. When using a database to generate random background, one should work with a non-redundant subset of structures, which have limited similarity to each other. This is to avoid the results being biased, e.g., overrepresenting many very similar structures.

To test structural or sequential comparison and scoring methods, one must define sets of homologous and non-homologous structures. To construct such sets classification databases like SCOP [169] or CATH [176] have become the 'gold standard'. Each structure in the database is being compared to all other structures in the database providing a similarity score for each pairwise alignment. The scores can be arranged in a list sorted after decreasing (or increasing) score and the analysis can be used to calculate *P-values* or *Z-scores*.

Levitt and Gerstein [146] have proposed a unified statistical framework for sequence and structure comparison using SCOP domains to define related and non-related structures. They performed an all-against-all comparison of a non-redundant SCOP subset, and then fitted an extreme-value distribution [94] to the observed scores of the non-related structure alignments. Doing this, they could assign statistical significance to each comparison score in the form of a *P-value*. Similar methods have been proposed by [1, 179, 232]. A detailed description is given in Section 5.2.6.

4.4 Protein Structure Classification

Early work on protein structures showed that there are striking regularities in the way in which SSEs are arranged [48, 145] in 3D and that there are recurring SSE topologies [195]. These regularities in secondary and tertiary structure arise from the intrinsic physical and chemical properties of proteins [45, 68] (see Chapter 2) and provide the basis for the classification of protein folds. Within the hierarchy of classification, only the relationships among classes of proteins within the same family reflect evolutionary divergence. At higher level of the

Table 4.1: **Protein classification according to structure [142, 145].**

Class	Characteristics	Examples
α-helical	Secondary structure exclusively or almost exclusively α-helical	Myoglobin, cytochrome c, citrate synthase
β-sheet	Secondary structure exclusively or almost exclusively β-sheet	Chymotrypsin immunoglobulin domain
α/β	Helices and sheets assembled from β-α-β units	
α/β linear	Line through centres of strands roughly linear	Alcohol dehydrogenase, flavodoxin
α/β barrel	Line through centres of strands roughly circular	Triose phosphate isomerase, glycolate oxidase
$\alpha + \beta$	α-helices and β-sheets separated in different parts of molecule. Absence of β-α-β supersecondary structure	Papain, staphylococcal nuclease
Few SSEs		Wheat germ agglutinin ferredoxin

hierarchy, the classification of sets of unrelated proteins is based purely on architectural similarity, independent of provable evolutionary history and relationship. Although many proteins are composed of single structural domains, most proteins are built up in a modular fashion from two or more domains fused together within one or more polypeptide chains. In some cases, each domain has a characteristic biochemical function and the function of the entire protein is determined by the sum of the properties of the individual domains. Therefore, most classification schemes for protein structures are based on domains as the discrete units of evolution and 3D structure.

Levitt and Chothia [145] proposed first a general structural classification approach for protein domains. The general classes for protein domains together with some example proteins are given in Table 4.1. Next, we will introduce the most commonly used databases for the classification of protein structures: SCOP [169], CATH [176], and DALI/FSSP/DDD [106]. Comparisons between these three classification approaches revealed a reasonable degree of correspondence (more than 80%) between individually classified protein families [96].

4.4.1 SCOP

The SCOP (Structural classification of proteins) database [169] organizes proteins hierarchically according to their structures and evolutionary origin. The method used to construct the protein classification is visual inspection and comparison of structures using automated procedures. Within the hierarchy, the unit of categorization is the protein domain, since domains are typically the units of protein evolution, protein structure, and its function. Small- and medium-sized polypeptide chains usually have a single domain and are treated as such. The domains in larger proteins are classified separately. A protein do-

main is defined as a region of a protein that has its own hydrophobic core and has relatively little interaction with the rest of the protein, making it structurally independent. Typically, domains are collinear in sequence, but can occasionally consist of several non-collinear sequence regions. In SCOP, *families* contain protein domains that share a clear common ancestor, indicated by high sequence identity or very high structure and function similarity. *Superfamilies* consist of families whose proteins share common structure and function, and therefore there is reason to believe that the different families are evolutionary related. *Folds* consist of one or more superfamilies that share a common core structure, i.e., the same SSEs in the same spatial arrangement with the same topological connections. Finally, depending on the type and organization of the secondary structure elements, folds are grouped into four major *classes*: all-α, all-β, α/β, and $\alpha+\beta$ (see Table 4.1). In addition, there are several other classes of proteins that are atypical and therefore difficult to classify, like membrane proteins or very small proteins.

4.4.2 CATH

CATH (Class, Architecture, Topology, Homologous superfamily) [176] presents a classification scheme similar to that of SCOP. In the CATH hierarchy, protein domains with very similar structures, sequences, and functions are grouped into *sequence families*. A *homologous superfamily* contains proteins for which there is evidence of a common ancestor, based on sequence and/or structure similarity. A *topology* comprises sets of homologous superfamilies that share the spatial arrangement and connectivity of SSEs. *Architectures* are groups of protein domains with similar arrangement of helices and sheets, but with different connectivity. For instance, different four-helix-bundles with different connectivity would share the same architecture but not the same topology. At the top, the overall protein class is determined by the secondary structure composition and packing. The main classes in CATH are: *mainly* α, *mainly* β, *mixed* α-β (subsuming α/β and $\alpha+\beta$ classes in SCOP, see Table 4.1), and domains with only few SSEs. The architecture and topology levels in CATH corresponding to the fold level in SCOP; the homologous superfamily in CATH corresponds to the superfamily level in SCOP, and the sequence family in CATH with the family level in SCOP. In both classification schemes the single domain denotes the lowest level. CATH uses many automated methods for sequence and structure alignments, but for difficult cases, the structures are manually classified.

4.4.3 FSSP/DALI/DDD

A third commonly used classification approach for protein domains—FSSP (Fold classification based on Structure-Structure alignments) and the DDD (DALI Domain Dictionary) [106]—is based on the DALI method [103]. DALI performs pairwise structure alignments of the entire PDB providing two classification schemes of protein structures:

1. FSSP presents the results from applying DALI to all polypeptide chains from known protein structures. Here, first all structures are clustered according to sequence identity on a 25% level and a single representative for each cluster is determined. Then, all the representatives are aligned

using the DALI method inducing a classification presented in the FSSP database.

2. DDD is the corresponding classification of recurrent protein domains automatically extracted from PDB [22] entries.

4.5 Protein structures with Non-trivial Relationships

The systematic analysis of protein structures have given important insights into protein evolution. Proteins that have descended from a common ancestor generally share a common fold but also retain structural and functional features. This empirical observation is used as a basis for protein classification (see above). The structures are projected on a hierarchical tree, which evolves with the increasing amount of structural data. This tree-like classification was based on several assumptions for the nature of sequence-structure relationships that were generally accepted at the time of its creation and that are still applied. Commonly, it was assumed that (see also Sections 2.6.1 and 4.4):

- Sequences of proteins performing the same molecular function diverged with speciation of the organisms.
- A protein can adopt only one native structure.
- Homologous proteins fold into similar structures.
- Protein structures are evolutionary more conserved than sequences.
- A protein fold could have evolved independently more than once.

Generally, it was thought that the protein fold is physically and biologically invariant, and that the number of folds in nature is very limited [46] (see Introduction). By the increasing structural data the classification of new protein structures has revealed numerous exceptions to these original assumptions, providing new insights into evolution of protein structure and shifting the paradigm of the protein fold [7]. The evolvability of protein folds was further supported by the results of several recent experimental studies applying multiple gene arrangement and non-homologous recombination approaches [56,183]. The possible mechanisms of fold changes include circular permutations, segment swapping, or presence of chameleon sequences that can adopt alternative conformations [148, 209, 232] resulting in non-trivial structural relationships at any evolutionary level of SCOP or CATH. Thus, for example in SCOP, homologous levels within the classification may contain proteins with different architectures. These cases add extra complexity and create practical difficulties for their presentation on the tree-like hierarchical classification scheme. In addition to the structural changes observed amongst related protein structures, the active sites of many functionally similar proteins were found to share structural common motifs embedded in otherwise totally different folds. These structural motifs can have substantial sequence similarity that often results in significant sequence hits between members of different superfamilies in SCOP [8]. The origin of these motifs is unclear and can be attributed to either divergent of

convergent evolution [201, 215]. Their non-trivial structural relationships are readily identified during expert analysis but their automatic identification still remains difficult or impossible, because such relationships can not be found by multiple sequence alignments, comparative modeling studies, or structural alignment methods preserving the sequential ordering of polypeptide chains. SISYPHUS [8] is a database containing a collection of manual curated structural alignments and their inter-relationships.