

Chapter 3

Protein Topologies

3.1 Introduction

Many proteins share similar geometric features in the conformation of the protein backbone [195]. The resulting compact folded form is called the tertiary structure of the protein. With only relatively few exceptions, the 3D structure of proteins is characterized by patterns of repeating SSEs [47]. One way to characterize the fold of a protein structure is by the arrangement of these SSEs as they pack together. Structural *protein topology* can then be defined as the relationship between the sequential ordering of SSEs and their spatial organization. It is one of the principal properties by which protein structures can be classified, categorized and compared. In fact, the same types of SSEs can come together in many different ways depending on their sequential ordering, for example, in the proteins *1tim*, a triose phosphate isomerase, and *1ai9*, a dihydrofolate reductase. Both folds consist of sequentially alternating helices and strands but different tertiary structures (see Figure 3.1). Supersecondary structure motifs, such as Greek-key or Zinc-finger motifs, describe the interaction and position of a few SSEs. The regular arrangements of such motifs open the possibility of a topological description of protein structures.

The first theoretical work on protein topologies considered only beta structure topologies [12, 44, 108, 109, 190] or only alpha topologies [91]. Koch *et al.* [128, 129] defined first a protein graph incorporating helices and strands. Protein topologies are also important for the study of the protein folding process, especially folding pathways [61, 189]. Zaki *et al.* [246] showed that using graph-based methods protein graphs can be applied to predict the sequence of unfolding events, and in reverse can be used to reconstruct the folding pathway.

The simplest representations of protein topology are schematic diagrams of protein folds illustrating the SSEs and their spatial neighborhood. Richardson [195] described first protein topology diagrams as a representation of a biological point of view. Several simple β -motifs of SSEs were described, such as the *Greek-key* or the *Jelly-roll* motifs. A detailed description is given in [30]. These cartoon representations are widely used, so, for example, in the protein classification databases CATH [176] and SCOP [169]. Another, but more complicated, representation is the hydrogen bond diagram which provides a graphical representation that is based on the hydrogen bonds between spatially

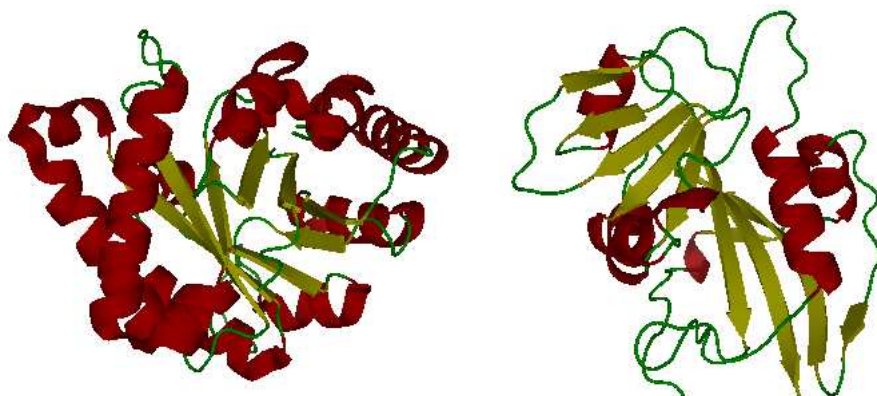


Figure 3.1: **Protein structure of *1timA* and *1ai9A*.** Left: *1timA*. Right: *1ai9A*. Helices are shown in red, strands in yellow. Protein structures are visualized with PyMol [58].

neighbored SSEs [108]. TOPS [81] generates topology diagrams automatically and provides the possibility to search for secondary structure motifs. TOPS cartoons represent a way of describing protein structures in two dimensions [71]. Figure 3.2 shows the TOPS diagram and the tertiary structure of the SCOP domain *2uaga1*. The domain consists of a parallel β -sheet of five strands and four helices above and beneath the sheet plane. In the PyMol [58] image in Figure 3.2 (right) only four strands are highlighted, because the last strand was not assigned by the SSE assignment method within PyMol. TOPS diagrams illustrate only the SSEs, with each helix as a circle and each beta strand as a triangle. Lines connecting two SSEs represent loops. The direction of the SSEs relative to the protein fold is shown through the connecting loops: if the connection is drawn to the entire of an SSE, it points downwards into the plane. If the connection is drawn to the boundary of the SSE, it points upwards out of the plane. For strands, the direction is also shown directly: an upper triangle points upwards and vice versa. The comparison of such diagrams is often difficult. For larger structures, TOPS diagrams are often hard to interpret, and sometimes impossible to draw, because of the large number of possible ways to order the SSEs according to their spatial neighborhood. Moreover, interactions between helices are not considered in TOPS diagrams and comparisons.

We developed PTGL [157]¹ (Protein Topology Graph Library), a web-based database application for the representation and retrieval of protein topologies and additional protein data combined with online search tools for data interrogation by sequence similarity and keyword queries. We provide a unique linear notation of four descriptions for protein structures on different abstraction levels based on graph-theory. The main idea is to mine known protein structures as protein graphs. The simplification of protein structure represented on secondary structure level will enable the implementation of graph-theoretic algorithms like the Bron-Kerbosch algorithm to detect structural similarities in proteins [126, 128] (see also Chapter 6 for a detailed description) and will make

¹PTGL is available on <http://ptgl.bcbio.de>

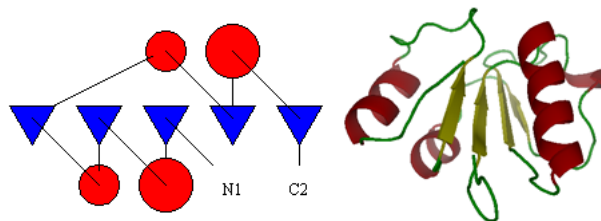


Figure 3.2: *2uagA1*: **T OPS [81] diagram and protein structure**. Left: T OPS diagram of the SCOP [169] domain *2uagA1* generated via the T OPS website (www.tops.leeds.ac.uk). Strands are colored blue, helices red. Right: protein structure visualized with PyMol [58]. Helices are shown in red, strands in yellow.

the representation of complex information such like protein structures more easy.

First, we explain the graph-theoretical description of protein topology. The linear notations are introduced on their different abstraction levels. Next, we describe the database application. As results, we first give statistical information of current holdings in PTGL. Then, we introduce the main classes of structural motifs together with their representations, and, finally, we compare the PTGL search tool based on the defined linear notations with the T OPS online search tool for structural motifs.

3.2 Protein Topologies as Protein Graphs

Protein topologies, i.e., the relationship between the ordering and connectedness of SSEs, can generally be expressed in terms of graph-theory. A protein graph is defined as labeled undirected graph (see Definition 2). Here, the vertices correspond to SSEs (helices and strands). The edges of the protein graph represent spatial adjacencies of SSEs. Using PDB [22] structure data the SSEs are defined according to the assignment of the DSSP algorithm [119] (see Section 2.4.3) with some modifications [128]. The eight DSSP states are converted into three classes using the following convention: $[GHI] \rightarrow h$, $[EB] \rightarrow e$, $[TS] \rightarrow c$. For helices at least three consecutive residues are required for the 'h' SSE state, and for strands at least two neighbored residues need the 'e' SSE assignment. Then, the spatial contacts between the SSEs are generated as described in [127]. The edges are defined through contacts between SSEs [120] using van-der-Waals contacts (see Section 2.7.3 and Appendix Table G.2):

Definition 5 (SSE Contact). *For a SSE u $T(u)$ is the unification of all van-der-Waals-volumes of atoms, which belong to the corresponding residue u . Two SSEs have a contact if $T(u) \cap T(v) \neq \emptyset$. According to the type of participating atoms there are backbone-backbone contacts, sidechain-sidechain contacts, and sidechain-backbone contacts.*

This information build the basis for the description of protein structures as graphs as defined in the PTGL database:

Definition 6 (Protein Graph). *If R is the set of all strands e and H the set of all helices h in the secondary structure of a protein, the protein graph*

$G = (V, E)$ is undirected with the vertex set $V = R \cup H$ and the edge set $E \subseteq \mathcal{P}_2(V)$ and defined as: two vertices u, v are connected, i.e., $u, v \in E$, if there are at least two backbone-backbone contacts or two sidechain-backbone contacts or three sidechain-sidechain contacts.

The vertices of the protein graph G are enumerated according to the sequence of amino acids from the N - to the C -terminus. According to this direction two spatially neighbored SSEs, which are in contact, could have a *parallel*, *antiparallel*, or *mixed* orientation. We defined the edge labels p (parallel), a (antiparallel), and m (mixed neighborhood):

Definition 7 (Edge Labeling). u and v are two connected vertices in the protein graph G . A is the set of numbers of the amino acids of the SSE u and B the set of numbers of the amino acids of the SSE v using a residue numbering from the N - to the C -terminus of the polypeptide chain. Then, S is the sum of all pairs (A_i, B_j) of $A_i \in A$ and $B_j \in B$, which are connected, and D the set of differences of all pairs (A_i, B_j) . From the maximum and minimum values of S and D you can calculate the doubled difference DD :

$$DD = (S_{max} - S_{min}) - (D_{max} - D_{min}).$$

If $DD > 0$ the SSEs have parallel orientation, if $DD < 0$ antiparallel, and if $DD=0$ mixed.

If there are formulations of questions concerning only with one type of SSE, i.e., only helices or only strands, our modeling allows to exclude the SSE type that is not of interest. According to the SSE type of interest the protein graph can be defined as *Alpha*, *Beta*, or *Alpha-Beta* graph. The general protein graph G corresponds to the Alpha-Beta graph $G_{\alpha\beta}$.

Definition 8 (Alpha Graph). H is the set of helices in the secondary structure of a protein. The Alpha graph $G_\alpha = (V_\alpha, E_\alpha)$ is the undirected graph with the vertex set $V_\alpha = H$ and the edge set $E_\alpha \subseteq \mathcal{P}_2(V_\alpha)$ which are defined analogously to the protein graph G above.

Definition 9 (Beta Graph). R is the set of strands in the secondary structure of a protein. The Beta graph $G_\beta = (V_\beta, E_\beta)$ is the undirected graph with the vertex set $V_\beta = R$ and the edge set $E_\beta \subseteq \mathcal{P}_2(V_\beta)$ analogously to the protein graph G .

A protein graph has not necessarily to be connected; it can consist of one or more connected components:

Definition 10 (Connected Component). A graph is called connected, if there are all vertices connected through a path. The connected components of a graph are the sets of vertices or subgraphs, which are connected.

Every connected component within a protein graph is called a *folding graph* according to the protein *fold* it defines:

Definition 11 (Folding Graph). A connected component F (or F_α or F_β) of the protein graph G (or G_α or G_β) is called the folding graph. The connected components of a protein graph are enumerated according to their sequential topology, beginning at the N -terminus, alphabetically with capital letters.

Larger folding graphs, i.e., folding graphs consisting of many SSEs, often represent supersecondary structure motifs or protein domains. Protein domains are characterized by tightly packed SSEs that have no or less contacts to the rest of the protein structure. Therefore, single domains can be described as larger connected components within protein graphs of single polypeptide chains.

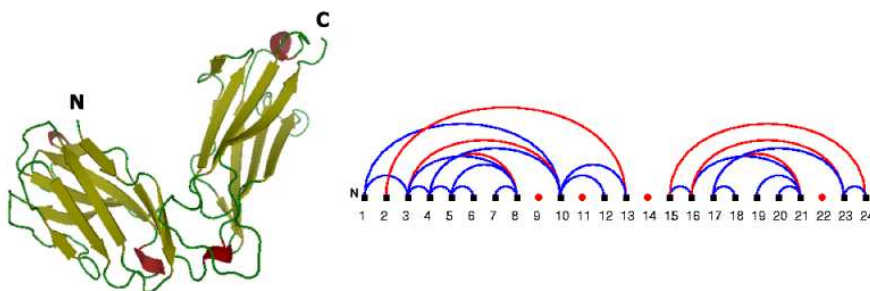


Figure 3.3: *1bec*: **structure and protein graph**. Left: the schematic representation of the topology of the *T-cell-antigen-receptor* (*1bec*) is drawn with PyMol [58]. Helices are colored red, strands yellow, the *N*- and *C*-termini are denoted. Right: the protein graph of *1bec*. Strands are shown as black squares, helices as red circles. Parallel edges are colored red, antiparallel edges blue.

Protein graphs consist often of several connected components. Many of these connected components represent only single SSEs. These SSEs are mostly located on the protein surface and not in the protein core and have, therefore, no contact to any other SSE. Figure 3.3 (left) shows the schematic representation of the antigen receptor protein *1bec*. The protein graph of *1bec* (Figure 3.3 right) consists of six connected components, two larger folding graphs consisting only of strands, and four single helices 9, 11, 14, and 22 shown in red. *1bec* is a transport membrane protein that detects foreign molecules on the cell surface. It has two domains that are represented by the two larger folding graphs, which are built each by two mainly antiparallel β -sheets (yellow arrows in Figure 3.3 (left) and blue edges in Figure 3.3 (right)). The other connected components consist in each case of one *G*-Helix on the protein surface.

3.3 Linear Notations and Graphical Representation

To explore topological aspects of SSEs it is useful to represent protein graphs as well as folding graphs as two-dimensional diagrams [30, 71, 81, 195]. There are two possibilities of representing protein or folding graphs: first, you can order the SSEs in one line after their position within the amino acid sequence, or second, you can order the SSEs after their spatial arrangement, a linear positioning, which often occurs in Beta graphs. In the first case, representing directly a protein or folding graph (or a simplified contact map), the SSEs are ordered as points on a straight line after their sequential ordering from the *N*- to the *C*-terminus. The spatial neighborhoods are described by semicircles between the points. Figure 3.3 (right) shows an example for a protein graph in such a

graph representation as defined in PTGL. Sequential edges between neighbored SSEs are not represented explicitly, because the sequential ordering is given by the numerical numbering of the SSEs from the N - to the C -terminus. For the second description type a unique arrangement of the SSEs is more difficult to draw. Here, we refer to the topology diagrams (see [30]) that are used often from biologists. Cylinders and arrows represent helices and strands, respectively. Semicircles and lines between the SSEs describe the sequential neighborhood. This description type is only useful for folding graphs with vertices having only two neighbors. With these somewhat idealistic representation folding motifs and similarities can be visually detect very easily. Figure 3.4 shows two beta graphs with a frequently occurring supersecondary structure motif built by four antiparallel strands, the *Greek-key motif*. The motif is named after an ornamental pattern in ancient Greece. The folding graph B of the protein chain C of the protein *lectin* (*1len*) represents the Greek-key motif, exactly as the SSEs 3,6,5, and 4 of the folding graph B in the protein *tonin* (*1ton*). The motif is also easily detectable in the first representation type. Figure 3.5 shows the Beta graph representations of folding graph B of lectin chain C and the folding graph B of the protein tonin, respectively.

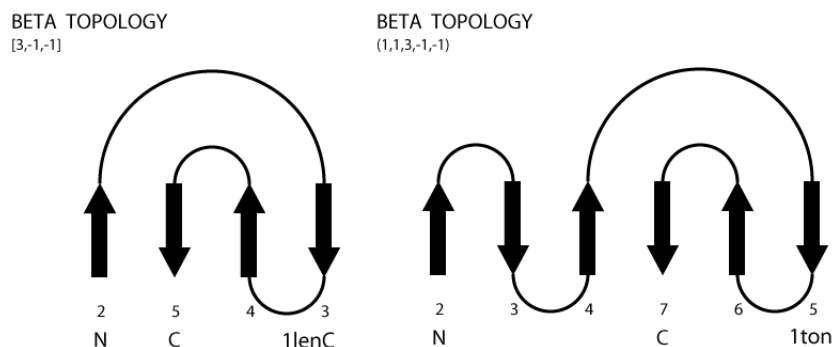


Figure 3.4: **Topology diagrams of *1lenC* and *1ton*.** Left: the Beta folding graph B in the protein *lectin* (*1lenC*), a vegetable protein with affinity to specific sugar residues. Right: the Beta graph B in the protein *tonin* (*1ton*)

In general a linear notation for a protein graph is a unique description for protein structures. Linear notations are well-known for representing the complex topologies of chemical structures in a linearized form [14, 238]. There exist also notations for protein topologies decomposing branching structures or structures containing cycles in a tree form that can be represented as nested parenthesis [73]. The problem using such nested notations is that regular expressions cannot be used anymore for direct pattern searching. Johannissen and Taylor [113] encoded protein topologies as layers of SSEs that can be represented as text strings using a dynamic programming algorithm to compare these strings. This representation is only useful for β -sandwich structures. Here, we present different, but more general, notations, that reveal to the features of protein structure like the linear sequence and the spatial arrangement of SSEs describing the secondary structure arrangement correctly and completely, the including of all types of SSEs, and the additional feature of supporting primitive regular

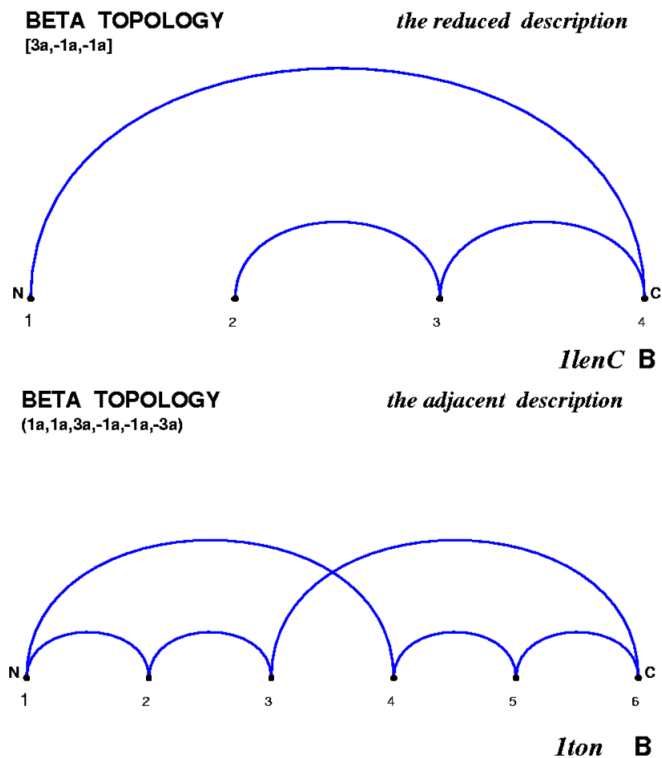


Figure 3.5: **Graph representations for *1lenC* and *1ton*.** Top: the reduced notation type of the Beta graph *B* in *lectin* (*1lenC*). Bottom: adjacent notation for the Beta graph *B* in *tonin* (*1ton*). Strands are marked with black squares. Antiparallel edges are shown in blue.

expression searches. A notation is here defined as a string, which represents a folding graph or a protein graph. We define four different linear notations: the *KEY*, the *adjacent*, the *reduced*, and the *sequence* notation.

3.3.1 The *KEY* Notation

The *KEY* notation represents the often-used topology diagrams [30] that were first introduced by Richardson [195] for β -topologies. Here, protein topologies are described with diagrams of arrows for strands and cylinders for helices. The SSEs are ordered and numbered in spatial order and are connected sequentially via lines. Figure 3.6 shows the folding graph of a β -sheet of the *aspartate amino transferase* (*1ars*), an enzyme of the amino acid metabolism. The sheet has parallel and antiparallely arranged strands. The *KEY* notation follows a path through the sequential order of SSEs noting their sequential distance within the described folding graph. Beginning at the SSE, which is nearest to the *N*-terminus, the SSEs are described by counting the spatial distance in numbers of SSEs in between. If the two SSEs are parallel, the letter 'x' is noted [195]. In this case the protein chain has to move on the other side of the sheet by crossing the sheet. This arrangement is called *cross over*. Antiparallel arrangements are

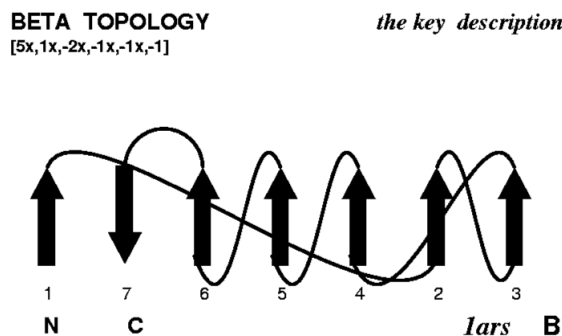


Figure 3.6: **KEY notation for *1ars***. The KEY notation of the Beta folding graph *B* of *1ars*. The strands are numbered in their sequential ordering from the *N*- to the *C*-terminus.

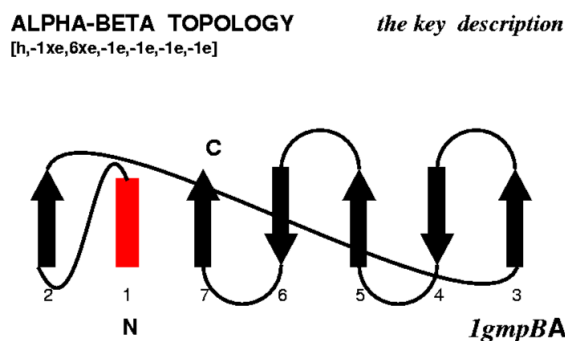


Figure 3.7: **The KEY notation of the Alpha-Beta graph *A* in *ribonuclease S* (*1gmpB*)**. Numbering from *N*- to *C*-terminus. Helices in red, strands in black.

called *same end*, which are more stable [47]. Mixed arrangements are defined in the KEY notation as *same end*. In Alpha-Beta graphs helices and strands are labeled by 'h' and 'e', respectively. Then, the notation starts with the type of the first SSE. Figure 3.7 shows the KEY-Notation of the Alpha-Beta graph of the *ribonuclease S* (*1gmp*), an enzyme that hydrolyses RNA. The folding graph consists of one helix and six strands. The helix is colored red, and the strands black. The notation '[h,-1xe,6xe,-1e,-1e,-1e,-1e]' starts at the *N*-terminus with a helix, denoted with 'h'. The next strand (strand 2) is in parallel orientation, i.e., the loop has to crossover to the beginning of the sheet, denoted by '-1xe'. The following strand 3 is six positions away ('6xe'). The strands 7, 6, 5, and 4 have antiparallel orientations, denoted each with '-1e'.

3.3.2 Adjacent and Reduced Notation

This kind of notation is based on following a path through the connectedness of neighboring SSEs and denoting their sequence distance. Rather than following the sequential ordering of the SSEs a depth-first path is traced through the fold-

ing graph and the distance in terms of SSEs between spatial neighbored SSEs is denoted, whether this goes backwards or forwards in the sequence and the orientation of the SSEs. If the SSEs of a folding graph are ordered sequentially, the spatial neighborhood between two SSEs is called *adjacent* (see, e.g., Figure 3.5 (right)). If you note the sequential distance between adjacent vertices, you can differ between the *adjacent* (ADJ) or *reduced* (RED) notation (Figure 3.5). In the ADJ notation all SSEs of the considered protein chain are counted to calculate their sequential distance, whereas in the RED notation only the SSEs of the considered folding graph are taken into account. You start with the first vertex, which has vertex degree 1, i.e., that has only one edge to another vertex, nearest to the N -terminus. Barrel structures have folding graphs without vertices of vertex degree 1. Then, you start with the first SSE in the sequence beginning at the N -terminus. When two SSEs are in parallel orientation the letter 'p' is noted, for an antiparallel neighborhood an 'a' and for mixed arrangements a 'm'. Figure 3.8 shows the ADJ and the RED notation of the folding graph B in the protein *aspartate amino transferase* (*1ars*). *1ars* consists only of one chain with overall 10 SSEs, but only 9 strands that are included into the Beta ADJ folding graph on the left. The Beta folding graph B consists of 7 SSEs. In the RED notation on the right, the SSEs that are not part of the folding graph are excluded. The resulting notations are:

- ADJ notation: $[8a, -1a, -1p, -1p, -4a, 1p]$, and
- RED notation: $[6a, -1a, -1p, -1p, -2a, 1p]$.

Both notations start at the first vertex near the N -terminus with vertex degree 1, here strand 1. Strand 1 has antiparallel neighborhood to strand 10 denoted with '8a' in the ADJ notation and '6a' in the RED notation. Note, both folding graphs are Beta graphs, i.e., only strands are included that have not to be labeled with an 'e' anymore, like in Alpha-Beta graphs. The path is ending in strand 4. At each vertex the distance and orientation to the predecessor vertex are noted. The RED notation is implicitly included in the ADJ notation, because the only

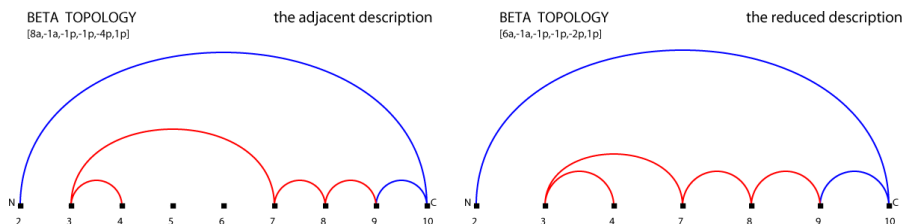


Figure 3.8: **ADJ and RED notations for *1ars***. The ADJ (left) and RED (right) notation of the Beta graph B in *Aspartat-Aminotransferase* (*1ars*). Strands are shown as black squares. Parallel edges are colored red, antiparallel edges blue.

difference in maintaining both notations is the additional counting of SSEs of non-considered folding graphs. Therefore, different ADJ notation can be traced back to one RED notation, but not vice versa.

3.3.3 The Sequence Notation

The sequence (SEQ) notation describes the sequential ordering of protein and folding graphs. As in the ADJ notation all SSEs of a protein chain are included. For a certain graph the consisting SSEs are considered in sequential order and the sequential differences between the SSEs are counted. Figure 3.9 shows the Alpha-Beta graph of folding graph D in the protein *1ars* in ADJ and SEQ notation. The SEQ notation is kind of trivial, but it can be useful if sequential distances and spatial information are both important.

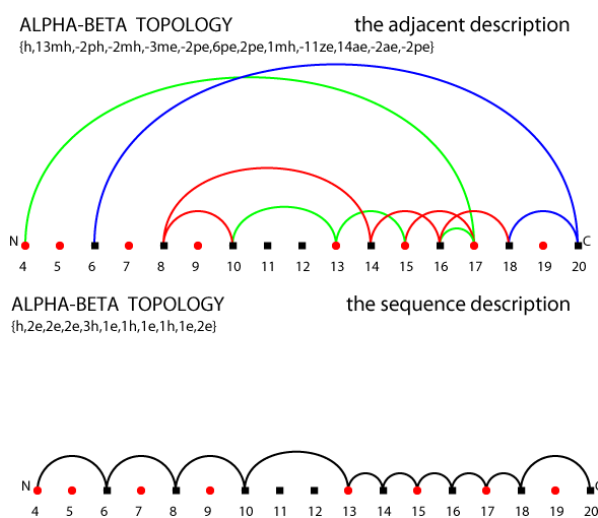


Figure 3.9: **The ADJ and the SEQ notation of *1ars*.** Left: Alpha-Beta folding graph D of *1ars* in ADJ notation. Right: Alpha-Beta folding graph D of *1ars* in SEQ notation. Helices are shown as red circles, strands as black squares. Parallel edges are colored red, antiparallel edges blue, and mixed edges green. For the SEQ notation, the sequential edges are shown in black.

3.3.4 Bifurcated and Non-bifurcated Graphs

Definition 12 (Vertex Degree). *The vertex degree of vertex u of the Graph $G = (V, E)$ is denoted as $grad(u)$. The neighborhood of $u \in V$ is given as $N(u) = \{v \in V | (u, v) \in E\}$. The vertex degree of vertex u is then defined as $grad(u) = |N(u)|$.*

Protein graphs may contain cycles, i.e., all vertices have at least vertex degree 2, or multiple branches, indicating single vertices with vertex degree greater than 2. For example, topological cycles in β -sheets are well-known geometrical features of protein structures, so-called β -barrels [72]. Such complex topological characteristics cannot readily expressed by topology notations [127] such as that of Richardson [195] or the KEY notation. Instead, linear notations based on a graph-theoretical description of protein topologies should deal with branching and cycle closings. Generally, the definition of linear notations for graphs with

vertex degree greater than 2 is a problem, because the usage of simple string matching algorithms is not possible anymore.

Definition 13 (Bifurcated Structure). *A protein graph G or a folding graph F is called bifurcated, if there is one vertex $v \in V$ with a vertex degree $\text{grad}(v) > 2$. If not, the graph is called non-bifurcated.*

The '{}' brackets at the beginning and the end of a notation are marking a bifurcated folding graph, e.g., in Figure 3.9 (left). Notations of folding graphs representing barrels are marked with '()' brackets. '[']'-brackets denote the notations of all other non-bifurcated folding graphs.

Definition 14 (Hamiltonian Path, and Cycle). *A Hamiltonian path is a path through a graph that meets each vertex exactly once. A cycle is a path such that the start vertex and end vertex are the same. Notice however that unlike with paths, any vertex of a cycle can be chosen as the start, so the start is often not specified.*

The vertices of a non-bifurcated folding graph represent a *Hamiltonian path*. If such a folding graph is forming a cycle, a barrel structure is defined. Bifurcated graphs are building no Hamiltonian paths at all, but trees. The RED and ADJ notations for bifurcated structures follow the following rules:

1. The SSEs are ordered sequentially from N - to C -terminus.
2. Search the first SSE with vertex degree 1.
3. Beginning with this vertex, you have to follow the spatial edges and note the sequential distances. If a SSE has more than 2 spatial neighbors, you choose that SSE nearest the N -terminus. Passing through the vertices every vertex degree is reduced by 1. That is done as long, as finding a vertex with vertex degree 1, and the path is ended.
4. Now, search the SSE nearest the N -terminus with vertex degree 1. The sequential difference to this SSE is noted together with the letter 'z'. The sequential differences are calculated in the same way as the sequential differences in ADJ and RED by non-bifurcated folding graphs.
5. Step 3 and 4 are repeated until no vertex of vertex degree > 0 is left over.

Figure 3.10 shows the folding graph D in the protein *adenylosuccinate-synthase* (*1adeA*). The Beta folding graph is bifurcated indicated by the brackets and the 'z'. It has the RED notation: $\{1p, 1a, -1z, 2a\}$.

Bifurcated folding graphs make a unique linear notation of the spatial order of the SSEs impossible, which is needed for the KEY notation. It will always depend on the way the path through the graph is defined. Therefore, there are up to now only diagrams of bifurcated folding graphs in the ADJ, RED, and SEQ notation. Additionally, the usage of regular expressions for fast string searches is not possible for bifurcated graphs.

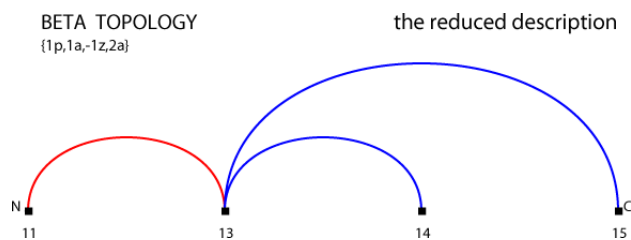


Figure 3.10: **The RED notation of *1adeA* Beta folding graph *D*.** RED notation of a Beta folding graph *D* in the protein *adenylosuccinat-synthase* (*1adnA*). Strands are shown as black squares. Parallel edges are colored red, antiparallel edges blue.

3.4 The PTGL Database

3.4.1 Database Design

PTGL [157] (Protein Topology Graph Library) is stored locally in an object-relational PostgreSQL database running on a Linux server. Programs for data export and import have been implemented in Perl and C. All diagrams of protein and folding graphs in PTGL are generated automatically. PTGL input data are acquired from PDB [22] and DSSP [119]. The online search tool consists of four different search interfaces:

- a simple keyword search, where keywords can be connected by Boolean operators,
- a more specific query, which searches in selected fields and tables, e.g., search for subtopologies in all Beta graphs,
- a query form for sets of PDB identifier, and
- a sequence search based on a local BLAST [4] search against the PDB database.

We have incorporated the ability to select a subset of homologous protein structures. The clustering of the polypeptide chain sequences is done according to re-generated PDB cluster tables based on the cd-hit algorithm [147]. The 'query result browser' provides possibilities to select single polypeptide chains or all found proteins, the graph type, as well as the notation type of folding graphs. In the 'topology browser' for every selected PDB chain the topology diagrams for the protein graph and all folding graphs consisting of more than one SSE are presented together with a table showing all SSEs (Figure 3.11). The graph descriptions and SSE information is available as ASCII files. Furthermore, the topology can be represented as a 3D Rasmol [205] picture. Additionally, users can upload their own PDB files to generate the different linear notations and diagrams.

3.4.2 Statistics

PTGL currently holds topology information for 29,247 proteins and 1,300,685 SSEs. Only proteins with at least one defined SSE according to DSSP [119], a

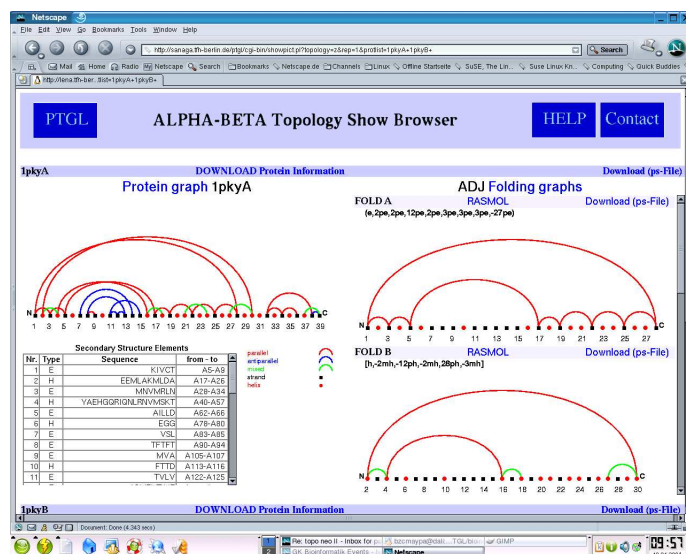


Figure 3.11: Screenshot of the PTGL 'topology browser'. Screenshot of the Alpha-Beta 'topology show browser' for the *1pkyA*.

Table 3.1: Statistics of PTGL. The first column indicates the type of the folding graph, 'total' gives the total amount of folding graphs of the particular graph type, 'different' the amount of different graphs among the total number, 'non-bifurcated' the number of different non-bifurcated graphs including barrels, and 'barrels' the number of different barrel structures.

Graph type	total	different	non-bifurcated	barrels
Alpha	439,328	14,874	587	42
Beta	228,396	12,241	1,732	80
Alpha-Beta	613,369	46,068	3,297	156

protein structure resolution less than 3.5\AA and a sequence length of minimal 20 amino acids have been considered. Table 3.1 shows the total number of folding graphs for every graph type. The number of Alpha-Beta graphs is less than the sum of Alpha and Beta graphs, because in Alpha-Beta graphs additional contacts between helices and strands are considered.

3.5 Supersecondary Structure Motifs

Protein domains are built from simple combinations of few SSEs with specific geometric arrangements that have been found to occur frequently in protein structures. These supersecondary structures or motifs are organized in a limited number of combinations within protein structures. Sequentially adjacent motifs are usually close together in the tertiary structure [30]. Simple motifs combine to form more complex motifs or whole domains. Thus, similar motifs frequently occur in different protein domains with different functions. Domains

can be classified according to their dominating SSE or SSE motifs: α -domains are built entirely of helical motifs, β -domains contain mainly antiparallel β -sheets, α/β -domains contain β -sheets with connecting helical segments, and $\alpha+\beta$ -domains contain separated sheet and helical regions. We have used our linear notations for protein topologies to define search patterns for the most common supersecondary structure motifs. Table G.3 in the Appendix summarizes the linear notations for these motifs according to the definitions given, for example, in Brandon and Tooze [30]. Next, we will introduce typical motifs for the most important structural classes. At the end we show the results how the simple string search performs in comparison to the TOPS [81] website search tool that uses algorithms from graph-theory to search for topology patterns.

3.5.1 α -helical Motifs

The most common α -helical motifs are the *four-helix-bundle* and the *globin fold*. The globin fold consists of eight helices forming a hydrophobic pocket in the core in which large hydrophobic organic and organometallic groups can bind. The fold has got its name from myoglobin, a single-chain oxygen-storage molecule in which the eight helices are arranged around a central heme group. Globins have to be found monomeric, dimeric, tetrameric, or in higher aggregates. The cytochromes *C* are another important family of heme proteins. Even large proteins can consist of purely helical secondary structures, like the ras-GTPase activating domain of P120gap (*1wer*). Transcription factors like the 'leucine zipper' (*2dgc*) or 'zinc finger' proteins are also purely helical. Many transmembrane segments consist only of helices, e.g., the bacteriorhodopsin (*1brd*). The simplest packing arrangement of a domain of two helices is an antiparallel topology, connected by a short loop. This constitutes, for example, the structure of the small (63 residue) RNA-binding protein Rop (*1rop*), which is found in certain plasmids (small circular molecules of double-stranded DNA occurring in bacteria and yeast) and involved in their replication.

Four-Helix-Bundle

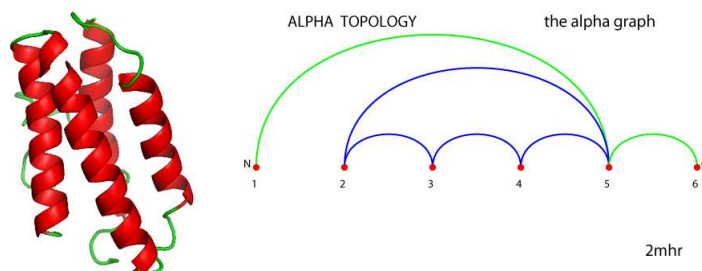


Figure 3.12: **Four-helix-bundle.** Left: structure of *2mhr* (PyMol [58]). Right: topology diagram for *2mhr*. Helices are illustrated as red circles. Antiparallel edges are colored blue, mixed edges green.

The most common α -helical motif is the *four-helix-bundle* consisting of 4 helices packed into a compact bundle. It has been found in many α -domains with a

range of diverse functions such as oxygen transport, nucleic acid binding, and electron transport. In four-helix-bundle proteins, the interfaces between the helices consist mostly of hydrophobic residues while polar side chains on the exposed surfaces interact with the aqueous environment. The four helices can have different topologies defined by the helix ordering:

- All helices have purely antiparallel organization (up-and-down motif).
- All helices have purely parallel organization.
- The helices have parallel and antiparallel organization.

Figure 3.12 shows the four-helix-bundle motif as well as the topology diagram of myohemerythrin, a protein composed of a single four-helix-bundle that serves as oxygen-storage protein in marine worms. The four helices have all antiparallel orientations. The two small helices at the *N*- and the *C*-terminus (helices 1 and 6 in the topology diagram) are not shown in the PyMol [58] representation on the left.

3.5.2 β -Sheet Motifs

β -domains contain only sheets, tight turns and irregular loop structures. Most β -sheet proteins consist of antiparallel strands dictated by the directionality of the polypeptide chain from the *N*- to the *C*-terminus. The most common contacts between strands are immediately adjacent strands, β -hairpins, forming up-and-down motifs, or to one three strands away in the sequence leading to the Greek-key motif. Antiparallel sheets can form barrel or sandwich structures. In sandwiches two separate sheets are packed against each other, e.g., the β -sheet sandwich structures in concanavalin *A*. In β -barrel motifs, a single sheet forms a closed cylindrical structure in which all strands are forming hydrogen bonds to one another, i.e., in the sheet the last strand is hydrogen bonded to the first strand. The distinction into sandwich and barrel is often useful but not absolute, since β -sheets have a natural twist, and two twisted sheets can either form a closed barrel or an open sandwich structure. Both types of antiparallel connectivity, up-and-down as well as Greek-key, can be found in both structure types.

Greek-key Motif

A very simple motif is the *Greek-key* motif, formed out of four antiparallel strands. This structural motif is not associated with any function but often occurs in proteins of almost all kinds of function. Figure 3.13 shows the four possible arrangements forming a Greek-key motif. In the KEY notation of Beta graphs the strings '-1,-1,3', '1,1,-3', '3,-1,-1', and '-3,1,1' represent these four possibilities. A string search can be done using the 'SearchFields' query browser of the PTGL web application, but it is important to note that if you search for the string '1,1,-3' all possible strings '1,1,-3*' are found including the string '1,1,-3x', which does not represent a Greek-key motif at all. In Table 3.2 we count the numbers of motifs we found searching for a certain Greek-key string and subtracted the number of non-Greek-key strings including an 'x'. The total as well as the non-redundant numbers, i.e., only proteins chains are reported that have sequence identity lower 90%, for every search string are shown. All

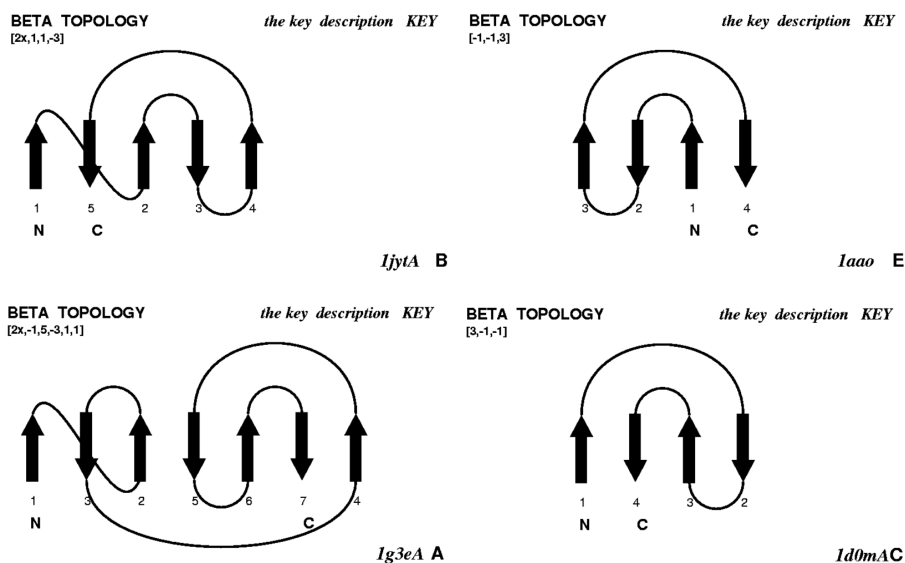


Figure 3.13: **The Greek-key motifs.** The four types of Greek-key motifs in Beta graphs (KEY notation) .

four possible topologies are equally distributed in the non-redundant case, but it is very significant that the Greek-key string '3,-1,-1' is more often to be found in barrel structures; meanwhile the symmetrical string '-3,1,1' is almost never a component of a barrel.

Jelly-roll Motif

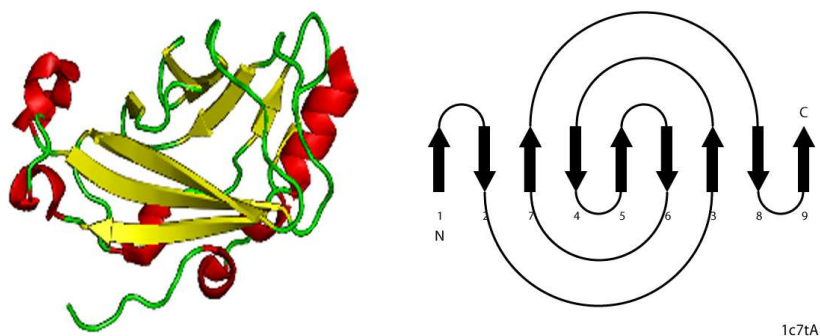


Figure 3.14: **Jelly-roll motif.** Left: structure of 1c7tA (residues 1 – 180 (drawn with PyMol [58])). Right: the Beta topology diagram for 1c7tA.

The *jelly-roll* motif is formed of overall eight strands from two sheets of each four strands. This motif has been found in different structures like coat proteins of many spherical viruses, the plant lectin concanavalin A and the hemagglutinin

Table 3.2: **The Greek-key motifs.** Number of proteins with Greek-key folds in Beta graphs. The first column shows the KEY linear notations for the different Greek-key arrangements, the second column the total number of proteins found in the PTGL database, and the third column shows the number of those proteins containing barrel structures including the Greek-key motif. For each search the total number of non-redundant proteins is denoted in brackets.

Greek-key	total	barrel
-1,-1,3	1,207(434)	81(42)
1,1,-3	655(311)	36(7)
3,-1,-1	1,835(536)	966(168)
-3,1,1	1,308(332)	6(5)

protein from the influenza virus. It is called jelly-roll, because the two sheets are wrapped around a barrel core like a jelly-roll. Figure 3.14 shows the structure and the topology diagram of a typical jelly-roll motif. Here, the first domain from *1c7t*, a beta-N-acetylhexosaminidase, is shown.

Immunoglobulin Fold

An immunoglobulin molecule consists of a total of 12 immunoglobulin domains. These domains have many sequence features in common and adopt a common structure, the immunoglobulin fold. Remarkable is that this structural motif is found in many other proteins that play key roles in the immune system. An immunoglobulin fold is a common β -sheet motif that consists of a 2-layer β -sandwich built of antiparallel strands arranged in two β -sheets that surround a central hydrophobic core. Although the motif consists of two separate sheets, one three and one four stranded sheet, the topology often includes a Greek-key motif where one strand can switch between sheets. The sheets are approximately aligned, in contrast, for example, to orthogonal beta sandwich structures.

Antiparallel Beta Barrel

Antiparallel beta barrels consist of antiparallel β -strands. The first and the last strand are connected via hydrogen bonds. A well-known example is the plasma borne *retinol binding protein 1rbp* that contains eight antiparallel strands forming a barrel structure. This protein is responsible for transporting vitamin A from its storage site in the liver to the various vitamin A dependent tissues. A retinol molecule is bound inside the barrel. Figure 3.15 shows the structure of *1rbp* and the corresponding protein graph. The topology consists of three folding graphs, the two single helices plus the big barrel structure with an additional strand at the C terminus. The closing of the barrel is only in the ADJ description in evidence but not in the standard topological diagrams or in KEY notation diagrams.

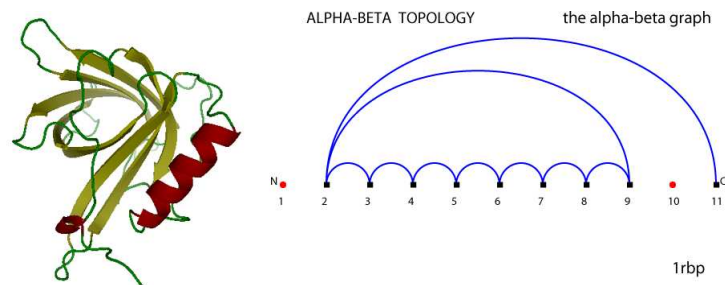


Figure 3.15: **Beta barrel protein.** Left: the structure of the retinol binding protein *1rbp* (PyMol [58]). Right: the Alpha-Beta protein graph, and the Alpha-Beta folding graph B. Strands are shown as black squares, helices as red circles. Antiparallel edges are colored blue.

3.5.3 α/β Motifs

α - β motifs are composed of β - α - β -units containing two hydrogen bonded strands in parallel orientation. The parallel strands must be joined by long connections, because the linking segment has to crossover onto the other side of the sheet. These outer connections are usually made by helices packing against a central core of parallel strands. The crossover can either be right- or left-handed, but the right-handed topology is seen in 95% of α/β motifs. There are two main classes of α/β motifs: in some cases there is a linear arrangement of the β - α - β -units forming twisted sheets, like the *Rossmann fold* (see Figure 3.2), in other cases the β -sheet closes on itself forming parallel barrel structures, the TIM barrel folds (see Figure 3.1 left). Many enzymes, including all those involved in glycolysis, are α/β structures. Most of these proteins are cytosolic.

TIM Barrel Fold

The very common TIM barrel motif is an α/β motif consisting of alternating β - α units, where the eighth strand is hydrogen bonded to the first strand. The sheet has cylindrical topology with the helices on the outside and strands mostly in parallel organization. The motif is named after the structure of the chicken triose phosphate isomerase (*tim*, see Figure 3.1 (left)). In SCOP [169] there are currently 32 different superfamilies classified containing the TIM barrel motif. Although their folds are very similar, TIM barrel enzymes catalyze a variety of different reactions, and the amino acid sequences are very dissimilar. A.D. McLachlan classified ideal β -barrels [160] by the number of strands n and the shear number S , i.e., the difference, along the sequence, of SSEs when closing the barrel. All TIM barrel structures have eight strands, i.e., $n = 8$ and $S = 8$, but the number of helices can range from 5 to 9.

Rossmann Fold

NAD-dependent dehydrogenases comprise one of the largest and best studied protein structure families. All the NAD-binding domains have similar structures despite of lacking amino acid sequence homology. They consist of an

open parallel sheet surrounded by helices on both sides and is divided into two similar motifs, the mononucleotide-binding motif or *Rossmann fold* containing a typical linear β - α - β - α - β unit (see Figure 3.2). The sheet normally consists of six strands, but can be extended by additional strands. Other proteins than dehydrogenases bind nucleotide-containing cofactors in a manner similar to the binding of NAD or NADPH to dehydrogenases, and create binding sites with similar supersecondary structures. The fold creates a natural cavity contributing to the binding of the adenine ring of NAD or other nucleotide-containing fragments. Because each Rossmann fold can bind only one nucleotide, binding domains for dinucleotides such as NAD consist of two paired Rossmann folds that each binds one nucleotide of the cofactor molecule. Single Rossmann folds can bind mononucleotides such as the cofactor FMN. The motif is named for Michael Rossmann who first pointed out that this was a frequently occurring motif in nucleotide binding proteins [192].

3.5.4 α + β Motifs

α + β domains contain both sheets and helices, but they are separated. The helical motifs are usually just clusters of interacting helices, while the sheets tend to be antiparallel or mixed. Sometimes the layer of helices is used to form recognition sites, such as the peptide-binding binding site in major histocompatibility proteins.

Ubiquitin Roll

Ubiquitin is a small protein that is involved in the degradation of proteins in eukaryotic cells. The fold has a rather common architecture of a small alpha and beta roll, with a five-strand β -sheet wrapped around a helix. The ubiquitin sequence is highly conserved from fungi to mammals but ubiquitin itself is found only in eukaryotic cells (Eukarya) and not in prokaryotic cells (Archaea and Bacteria). There are, however, proteins found in Archaea and Bacteria that have a structural fold similar to ubiquitin [25, 102].

3.5.5 Comparison PTGL versus TOPS

To evaluate how capable our simple linear notations (Table G.3 in the Appendix) are to search for common structural motifs we have used the SCOP [169] version 1.69 (see next chapter for a detailed introduction into protein classification) domain classification scheme. We compared the results from the PTGL database with those of the TOPS² [81] pattern matching online search tool, where you can search with pre-defined TOPS patterns for the most important supersecondary structure motifs. For PTGL, additionally, linear notations for four-helix-bundles and one type of beta barrels ($n = 8, S = 10$) were defined. First, we searched in SCOP for fold entries with domain descriptions containing the name of each defined motif. Column three in Table 3.3 gives the absolute number of domains we found for each motif in SCOP that were also found in the PTGL database. The numbers in brackets give the total number of domains that were classified in SCOP 1.69. In column four the number of SCOP domains is given that we found with at least one of the defined linear notations via the PTGL web interface.

²<http://www.tops.leeds.ac.uk/>

Table 3.3: **Comparison PTGL versus TOPS [81]**. See text for details.

Class	Motif	SCOP (total)	PTGL (total)	TOPS
α motifs	Four-helix-bundle	340 (476)	87 (399)	-
β motifs	Greek-key	8,514 (10,047)	1,538 (4,655)	118
	Jelly-roll	1,287 (1,747)	36 (65)	23
	Immunoglobulin-like	4,709 (5,532)	509 (1,674)	248
α/β motifs	Beta Barrel ($n = 8, S = 10$)	300 (302)	212 (355)	-
	Rossmann Fold	1,643 (1,820)	823 (1,905)	538
	TIM Barrel	2,642 (2,951)	832 (1,284)	224
$\alpha+\beta$ motifs	Ubiquitin Roll	398 (620)	151 (520)	95

The number in brackets is the total number of hits we found in PTGL denoting false positive hits or folds missing the attribute of the searched domain. Column five shows the number of SCOP domains reported by the TOPS search tool when searched with the pre-defined TOPS representations. The TOPS website does not give any information about which SCOP version they are using, how many false positives they find, and how often their database is updated. Thus, a fair comparison is not realizable. Totally, PTGL detects for every defined structural motif more true positives than TOPS. For β -motifs barrel structures are better defined than sandwich structures. Here, the SCOP motif definition is much less strict than the PTGL definitions, i.e., small variants are classified into the same fold category. Moreover, the PTGL edge definition includes no geometric information. Therefore, small changes can lead to changes from antiparallel or parallel orientation to mixed arrangements or vice versa. Thus, similarities are no more detectable with simple string searches. Very surprisingly, PTGL and TOPS have both difficulties to detect the well defined Greek-key motif. Often short strands do not contain enough backbone-backbone contacts, so the PTGL edge definition cannot capture these cases. Many motifs from different classes include up-and-down sheets. For the exact detection of, for example, Rossmann fold motifs more complex search patterns would be more adequate. Currently, the linear notations for the Rossmann fold motif are only defined for Beta graphs ignoring the fact that there have to be helices in between the strands.

3.6 Discussion

In order to define a unique description of protein topology, we described the secondary structure topology of proteins by methods of applied graph-theory [127]

as an undirected labeled graph on three description levels of its SSEs: the Alpha graph, the Beta graph, and the Alpha-Beta graph. The three graph types were defined for most of the proteins in the PDB [22], and DSSP [119] was used for SSE assignment. For each graph, four linear notations and corresponding graphical representations exist. We defined the adjacent, reduced, and sequence notation for the sequential order of SSEs and the KEY notation for a topological ordering of SSEs depending on the spatial arrangement of their SSEs. The database enables the user to search for specific proteins or for certain topologies or subtopologies using the linear notation as input for simple regular expressions. The PTGL linear notations detect more and additional structural motifs than the TOPS [81] database. For a more accurate and flexible search tool, protein topologies have to be represented within in rule-based languages like Prolog [229, 230] or the protein graphs have to be compared with computational expensive methods from graph-theory (see chapters 5 and 6). The PTGL notations include for the first time helical contacts, barrel structures, as well as bifurcated structures into unique representations that can be used for the fast detection of the most common structural motifs within protein domains. The linear notations could be combined with established pattern recognition or graph-theoretic methods to automate the search.