# Detection of distant structural similarities in a set of proteins using a fast graph-based method

**Ina Koch**
Oranienburger Str. 12, 10178 Berlin, Germany
koch@pluto.chem.uni-potsdam.de

Corresponding author. Most of this work was performed while the first author was at the Institute for Algorithms and Scientific Computing, GMD-German National Research Center for Information Technology, St. Augustin, Germany

**Thomas Lengauer**
first address: Institute for Algorithms and Scientific Computing SCAI.ALG
GMD-German National Research Center for Information Technology,
D-53754 St. Augustin, Germany,
lengauer@gmd.de
second address: Department of Computer Science, University of Bonn
D-53117 Bonn, Germany

## Abstract

We introduce a method for finding weak structural similarities in a set of protein structures. Proteins are considered at their secondary structure level. The method uses a rigorous graph-theoretical algorithm which finds all structural similarities. Protein structures are modelled as undirected labelled graphs, the so-called *protein graphs*. We suggest that for detecting the similarities between two protein structures it is sufficient to find similarities in the protein core which consists of tightly packed secondary structure elements. Therefore, we can restrict ourselves to solving the maximal common *connected* subgraph problem instead of the maximal common subgraph problem. We have modified the algorithm by Bron and Kerbosch for solving that problem. The speed of the algorithm increases drastically. After calculating all maximal common connected substructures for all pairwise comparisons in a set of protein graphs the common substructure in all proteins can be calculated by intersecting them. In this paper we characterize the method briefly and explain the modelling of the protein structure in detail. For the pairwise alignment the similarity of *porin* (1OMF) with *bacteriochlorophyll a* (3BCL) and *BirA* protein (1BIB) with *DNA polymerase III* (2POL) will be discussed. In the case of the multiple structure alignment the similarity in variants of four *phosphatases* and in *subtilisin Carlsberg*, *carboxypeptidase*, *elongation factor Tu*, and *flavodoxin* will be represented. Our first experiments show that the method works correctly and fast. The method can be used for arbitrary graphs. Thus, different graph-theoretical models of protein structures can be examined.

Keywords : multiple structure alignment, graph algorithm, maximal common subgraph, structural similarity, phosphatases

## Introduction

The rapidly increasing number of known protein structures, 4787 proteins and 386 nucleic acids in the protein structure data bank PDB, release from January 1997 (Bernstein *et al.* 1977) makes it necessary to develop methods for fast and correct detection of weak and strong structural similarity in protein structures, analogous to the great number of different methods for finding sequence similarities in proteins, e.g. (Needleman & Wunsch 1970), (Gusfield, Balasubramanian, & Naor 1992), (Waterman, Eggert, & Lunder 1992). There are different methods for comparing protein structures, which use principles from fields of computer science, such as artificial intelligence and computer vision, for instance pattern matching methods (Lathrop, Webster, & Smith 1987), geometric hashing methods (Bachar *et al.* 1993), knowledge based approaches (Šali & Blundell 1990), Monte-Carlo methods (Holm & Sander 1993), clustering methods (Lessel & Schomburg 1994), graph-theoretical approaches (Grindley *et al.* 1993), and genetic algorithms (May & Johnson 1994).

The systematic comparison of protein families results in the identification of repeating folding motifs, which are the basic constituents of protein structures. At the primary structure level similar elements or patterns are detected mainly using automatic methods for multiple sequence alignment in protein families. However, for the tertiary and secondary structure level only

a few methods have been presented. Sutcliffe et al. (Sutcliffe *et al.* 1987) use methods for superpositioning of rigid bodies. This approach works well for very similar proteins, but cannot detect weak structural similarities. Methods, which are based on the comparison of structural environments (Taylor & Orengo 1996), (Šali & Blundell 1990), can solve this problem. These methods are used for the pairwise structure comparison and their results are assembled for multiple alignment (Johnson, Šali, & Blundell 1990), (Pickett, Saqi, & Sternberg 1992).

Taylor, Flores, and Orengo (Taylor, Flores, & Orengo 1994) were the first to develop an automatic approach for multiple structure alignment. The method is a combination of their programs for pairwise structure comparison *SSAP* and for multiple sequence alignment *MULTAL*. The programs are working at the atomic level. Structures resulting from the pairwise comparison are combined consecutively producing consensus structures, which are compared with each other.

Starting from the encouraging experiences with small molecules (Brint & Willett 1987) the graph-theoretic models have been used also for more complex chemical structures such as proteins. Mitchel et al. (Mitchell *et al.* 1993) and Artymiuk et al. (Artymiuk *et al.* 1990) search for substructures in proteins at the secondary structure level. Grindley et al. (Grindley *et al.* 1993) search for maximum common substructures in two proteins using the algorithm by Bron and Kerbosch (Bron & Kerbosch 1973), which enumerates all cliques in a single graph.

Despite of all these methods pairwise and multiple alignment of protein structures remains a combinatorially difficult problem, which leads to unacceptable runtimes as the proteins grow in size, especially if one wants to detect all structural similarities. For performing multiple structure alignment no algorithm has been developed so far, that is based on a discrete model.

## Methods

The method for multiple structure alignment works at the secondary structure level. There are two reasons for this choice. The first is that two proteins of the same fold exhibit the same secondary structure arrangement in their cores. In most cases the main differences in these proteins are restricted to the loop regions or their surfaces. The second reason is the resulting reduction of the complexity of comparing two protein structures. The largest protein chains contain about 70 secondary structure elements (SSEs) with about 800 residues and about 80 000 atoms. The arrangement of the atoms in helices and strands is repetitive, such that during structure comparison identical substructures would be compared many times. Therefore, several methods which work at the atomic description level use the results of the secondary structure arrangement in proteins as a filter before calculating at the atomic level (Orengo, Brown, & Taylor 1992), (Mizuguchi & Gō 1995), (Holm & Sander 1995).

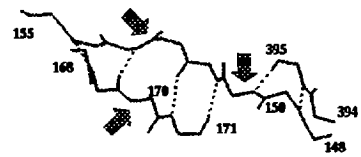


Figure 1: An example for the connection of two strands in *enolase* (4ENL). Only the backbone is represented. The arrows indicate the interruptions of the strands introduced by the DSSP-algorithm. The numbers denote residue numbers according to their order in the sequence from the N- to the C-terminus.

In our approach protein structures are modelled as undirected labelled graphs. The vertices represent SSEs according to the assignment of the DSSP-algorithm (Kabsch & Sander 1983) with the following modifications:

- If there are two strands which are disconnected by a single residue they are united to form a single strand (see figure 1). Dashed lines indicate hydrogen bonds. We can see clearly, that the DSSP-algorithm disconnects the strands, because the hydrogen bond pattern cannot be continued, although the backbone exhibits an extended structure.
- If an $H$-helix is followed by a $G$-helix or vice versa only the $H$-helix is considered (see figure 2). In figure 2 we can see the change of hydrogen bond patterns during the transition from one helix type to the other. The elimination of the $G$-helix is justified because of the change in the direction of the backbone often indicating the beginning of a loop region and because of the very short length of most $G$-helices.

The edges of the protein graph represent spatial adjacencies of SSEs. These adjacencies are defined through contacts between residues of SSEs. For a SSE $u$ is $T(u)$ the unification of all van-der-Waals-volumes of atoms, which belong to the corresponding reidues $u$. Two SSEs have a contact if $T(u) \cap T(v) \neq \emptyset$ (Kaden, Koch, & Selbig 1990), (Koch, Kaden, & Selbig 1992). We choose the van-der-Waals-radius of 2 Å for all atoms. According to the type of participating atoms there are backbone-backbone-contacts,


Figure 2: An example for consecutive $H$- and $G$-helices in *enolase* (4ENL). Only the backbone is represented.The arrow indicates the begin of the $G$-helix which is eliminated by the algorithm.

sidechain-sidechain-contacts, and sidechain-backbone-contacts. An edge will be drawn between two vertices in the protein graph if there are at least two backbone-backbone-contacts or two sidechain-backbone-contacts or three sidechain-sidechain-contacts. This definition also involves weak adjacencies between SSEs or spatial adjacencies between SSEs whose lengths differ by a large amount. According to the direction of two adjacent SSEs we distinguish between antiparallel, parallel, and mixed edges (see figure 3).


Figure 3: An antiparallel, parallel, and mixed spatial neighbourhood between two strands in *rubisco* (*ribulose-1,5-biphosphate-carboxylase*; 5RUB).

Domains of protein structures are regions, which are responsible for a certain function of the protein. Usually, these regions are characterized by tightly packed SSEs. Proteins can exhibit several domains and functions. In most cases proteins consist of one domain. Therefore, we can suggest that for detecting the similarities between two protein structures it is sufficient to find similarities in the protein core which consists of tightly packed secondary structure elements and corresponds to a domain. These domains can be described by *connected components* in the protein graph. Thus, we can restrict ourselves to solving the maximal common *connected* subgraph problem instead of the maximal common subgraph problem.

We can formulate the problem of protein structure comparison as the problem of finding all maximal common subgraphs in a set of protein graphs. First, we have to consider the pairwise comparison of two protein structures. We decide which edge pairs of one protein graph are compatible to which edge pairs of the other protein graph. In order to exclude the non-compatible edge pairs we transform the problem into the clique problem in a single graph, the so-called *product graph*. Because the problem belongs to the class of NP-hard problems, all known algorithms exhibit exponential runtimes. There are many protein graphs, for which the runtime for the fastest and most widely used algorithm by Bron and Kerbosch is unacceptably long. This algorithm works recursively and enumerates all cliques in a graph exact once. We have modified the algorithm in such a way that only cliques, which represent connected substructures in the protein graphs are considered during the search. Disregarding all cliques that represent disconnected subgraphs from the calculation procedure substantially reduces the size of the recursion tree. For example, in a random graph with 300 vertices and 4938 edges the recursion tree contains 3656 vertices when using the modified algorithm in contrast to 11 964 vertices when using the original algorithm. The number of cliques enumerated decreases significantly, as well. For the same example the modified algorithm finds 164 cliques of size 4 in 0.44 seconds, whereas the original algorithm enumerates 557 cliques of the same size in 1.26 seconds. The modified algorithm shows drastically decreased runtimes. The differences in the runtimes increase, when the graphs become larger and denser.

After calculating all maximal common connected substructures for all pairwise comparisons in a set of protein graphs the common substructures in all proteins are calculated by intersecting the edge sets of all maximal common connected substructures, represented in one protein graph. These edge sets are stored in ordered binary trees. For the details of the algorithm see (Koch, Lengauer, & Wanke 1996).

## Results

In this section we are going to discuss two examples for pairwise and two for multiple structure alignment. We distinguish between structural and topological similarity. The topology of the secondary stucture of a protein is defined by the order of SSEs in the sequence and by the type, antiparallel, parallel, or mixed, of the arrangement of two SSEs.

The calculations were performed on a SUN-Enterprise 4000 with six Ultra-Sparc-1-processors and 756 MB main memory. The runtimes for every pairwise comparison are smaller than one second in all examples. The total runtimes of all examples are between two and three seconds.

In the spatial representations of maximal common substructures the SSEs of the maximum common subgraphs are coloured gray and black. The remaining part of the protein is white.

## Pairwise structure alignment of porin and bacteriochlorophyll a

This similarity, first reported by Alexandrov and Fischer (Alexandrov & Fischer 1996), exhibits structural similarity of 13 secondary structure elements in each of the proteins. However, the proteins do not exhibit topological similarity.

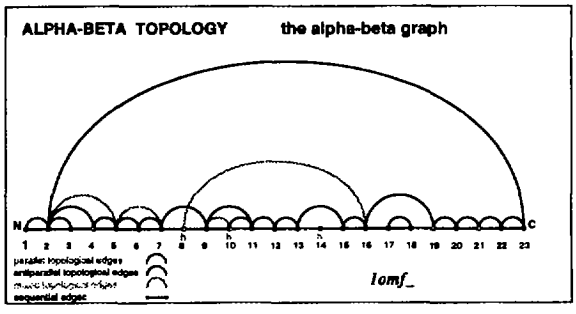


Figure 4: The protein graph of *porin*.

*Porin* (1OMF) is a transmembrane protein, which is located in the outer membrane of mitochondria. It has a large pore, through which a variety of ions and small molecules can penetrate. *Bacteriochlorophyll a* (3BCL) is a protein, which is very similar to the *chlorophyll protein*. Because of small structural modifications its absorption maximum is shifted to the near infrared, up to wavelengths of 1000 nm. The protein absorbs photons during photosynthesis.

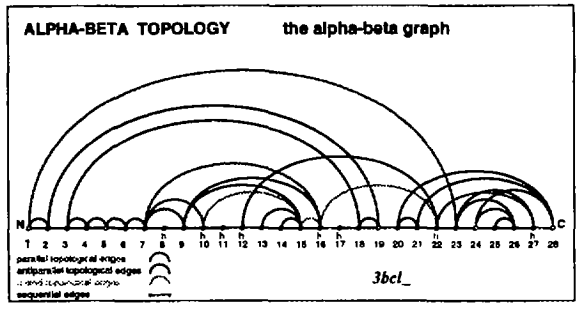


Figure 5: The protein graph of *bacteriochlorophyll a* .

The protein graphs of *porin* and *bacteriochlorophyll a* are represented in figures 4 and 5. The vertices are arranged in a row according to their order in the sequence from the N- to the C-terminus. The edges of the graphs are drawn as bows. SSEs adjacent in the sequence are connected by a line. Note that these lines do not belong to the protein graph. The protein graphs consist of three (*porin*) and four (*bacteriochlorophyll a*) connected components. In *porin* there are also two small components, a helix and two neighbouring strands. In *bacteriochlorophyll a* all three small components are helices. All small connected components in both proteins are located at the surfaces of both proteins. Thus, it is sufficient to compare the large connected components, because the smaller components do not belong to the protein core.

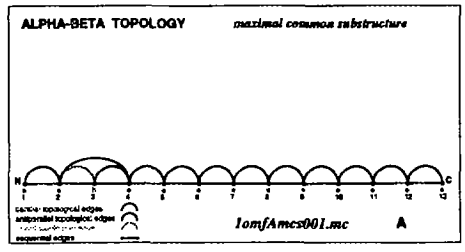


Figure 6: The maximum common subgraph of *porin*.

The protein graphs of *porin* and *bacteriochlorophyll a* contain many antiparallel pairs of strands. While in *porin* all strands form the classical barrel structure in which a helix is located between two strands in the sequence only three times, in the protein graph of *bacteriochlorophyll a* we can identify two barrel-like structures, one open barrel between the strands 3 to 18 and one small barrel between the strands 20 to 28. Although the topologies of both proteins are different our algorithm finds a common substructure, which consists of 13 SSEs. Because there are methodical differences, especially in the modelling and also in the algorithm, the residue ranges are not identical in both methods.

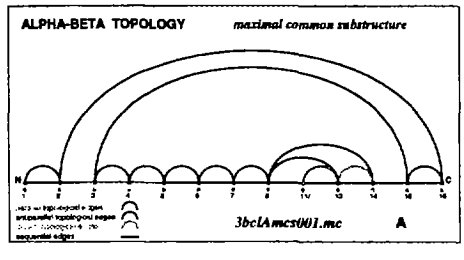


Figure 7: The maximum common subgraph of *bacteriochlorophyll a* .

Figures 6 and 7 show the maximum common subgraphs

in both proteins. Both graphs are identical. Note that we consider only the spatial adjacencies (the bows) between SSEs.

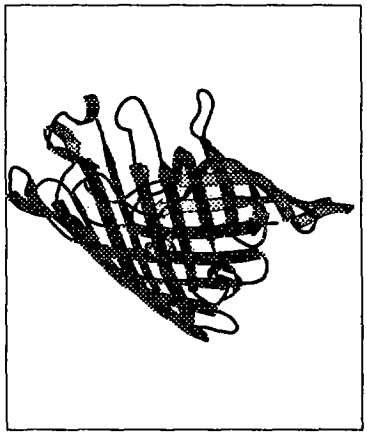

Figure 8: The maximum common substructure in *porin*.

In the maximum common substructure in *porin* the spatial order of SSEs is equal to the sequential order, whereas in *bacteriochlorophyll a* the sequential order is not equal to the spatial order of SSEs. In *bacteriochlorophyll a* we can detect two spatial adjacencies of SSEs which are located far away in the sequence, from SSE 2 to SSE 16 and from SSE 3 to SSE 15.

The spatial representations of the common substructures are shown in figures 8 and 9. Both strands marked dark grey in *bacteriochlorophyll a* are interpreted as a single strand by the program RASMOL (Sayle 1996).

In this case merging of both strands would be useful, because the direction of the backbone is not changing. In our model strands that are adjacent in the sequence are connected if a single residue is located between them. If more residues separate the strands usually a change in the direction of the backbone results such that no regular structure would be obtained by merging of both strands. It is also interesting that the two occurrences of the maximum common substructure in both proteins can only be superposed by a mirroring operation.

In table 1 the assignments of SSEs of the common substructures are given. The algorithm finds two matchings of SSEs. The strand (94 to 102) in *porin* is duplicated twice in *bacteriochlorophyll a*, once in strand (from residue 198 to 204) and once in strand (from residue 192 to 194).

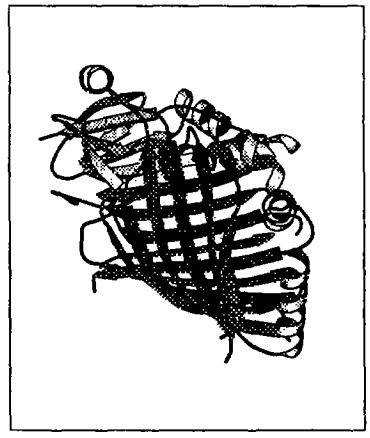

Figure 9: The maximum common substructure in *bacteriochlorophyll a*.

| row | *porin* | *bacteriochlorophyll a* | |
|---|---|---|---|
| 1. | 331 - 339 E | 4 - 13 E | 4 - 13 E |
| 2. | 307 - 316 E | 21 - 29 E | 21 - 29 E |
| 3. | 253 - 263 E | 40 - 50 E | 40 - 50 E |
| 4. | 225 - 235 E | 58 - 68 E | 58 - 68 E |
| 5. | 210 - 222 E | 71 - 84 E | 71 - 84 E |
| 6. | 185 - 195 E | 89 - 100 E | 89 - 100 E |
| 7. | 173 - 182 E | 103 - 117 E | 103 - 117 E |
| 8. | 151 - 158 E | 136 - 147 E | 136 - 147 E |
| 9. | 94 - 102 E | 198 - 204 E | 192 - 194 E |
| 10. | 132 - 141 E | 210 - 221 E | 210 - 221 E |
| 11. | 143 - 147 H | 226 - 229 H | 226 - 229 H |
| 12. | 269 - 283 E | 244 - 252 E | 244 - 252 E |
| 13. | 287 - 302 E | 257 - 265 E | 257 - 265 E |

Table 1: The assignments of SSEs of the common substructures in *porin* and *bacteriochlorophyll a*. The sequence range of a SSE is given. All assignments are the same except that represented in row 9.

## Pairwise structure alignment of BirA protein and DNA polymerase III

*BirA protein* (1BIB) is a bifunctional protein, which acts as biotin synthetase and as biotin operon repressor. The *DNA polymerase III* (2POL) catalyses a matrix dependent DNA synthesis. It consists of two chains.

The protein graph of *BirA protein* (see figure 10) exhibits one large and seven small components. The protein graph of *DNA polymerase III* has one large and four small components. We are comparing the large components, because the small components do not belong to the protein cores. The protein graphs are quite dissimilar. While in *BirA protein* the many parallel adjacencies are conspicuous, in *DNA polymerase III* par-

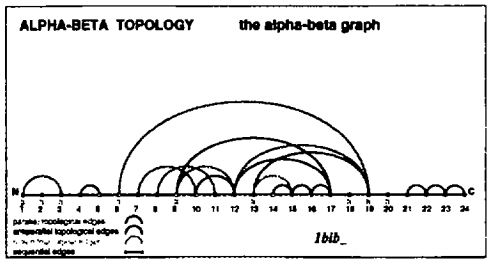


Figure 10: The protein graph of *BirA protein*.

allel adjacencies are rare, in most cases between SSEs which are adjacent in the sequence.

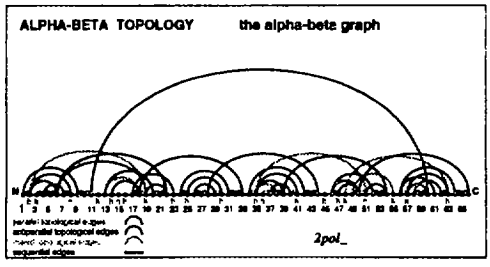


Figure 11: The protein graph of *DNA polymerase III*.

In these $\alpha + \beta$ proteins the algorithm calculates a maximum common substructure which consists of six antiparallel strands and two helices, which are arranged as an $\alpha + \beta$ *sandwich*, see figures 12 and 13. The substructure occurs twice in *DNA polymerase III*, once in chain A and once in chain B. Interestingly, the topology is also different in both common substructures.

| *BirA protein* | *DNA polymerase III* | |
|---|---|---|
| 90 - 97 H | B132 - B140 H | A197 - A206 H |
| 106 - 109 E | B176 - B196 E | A32 - A38 E |
| 130 - 139 E | B126 - B131 E | A66 - A71 E |
| 147 - 163 H | B7 - B18 H | A7 - A17 H |
| 170 - 172 E | B51 - B58 E | A51 - A58 E |
| 176 - 179 E | B230 - B235 E | A230 - A235 E |
| 182 - 192 E | B222 - B227 E | A222 - A227 E |
| 199 - 208 E | B213 - B218 E | A214 - A219 E |
| 235 - 254 H | B72 - B81 H | A72 - A81 H |

Table 2: The assignments of SSEs of the common substructures in *BirA protein* and *DNA polymerase III*. The sequence range of a SSE is given.

Alexandrov and Fischer, who first reported this similarity (Alexandrov & Fischer 1996), assign a biological meaning to this structural similarity. They assume that the part in *BirA protein* can exhibit a similar nonspecific binding to DNA as the helices of *DNA polymerase III*, although the first domain of the *BirA protein* has a specific DNA binding function.

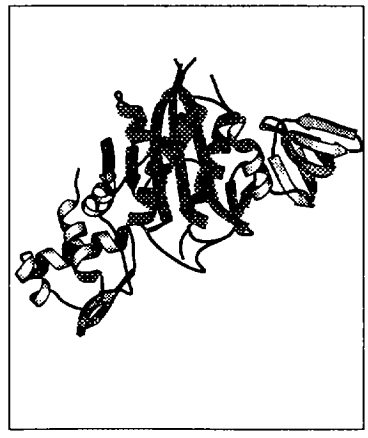

Figure 12: The maximum common substructure in *BirA protein*.

In table 2 the assignments of SSEs of the common substructures *BirA protein* and *DNA polymerase III* are given.

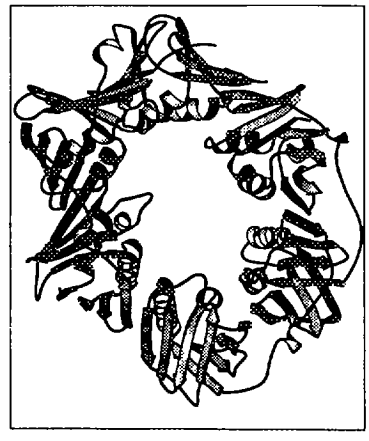

Figure 13: The maximum common substructures in *DNA polymerase III*.

## Multiple structure alignment of four phosphatases

The phosphorylation of tyrosine residues of intracellular proteins is an important cellular regulatory mechanism involved in processes such as cell growth, pro-

liferation, and differentiation. In recent years attention has turned to the enzymes, that dephosporylate the tyrosine residues, the so-called *protein tyrosine phosphatases*. Many enzymes which are capable of hydrolysing phosphorylated tyrosine residues are now known.

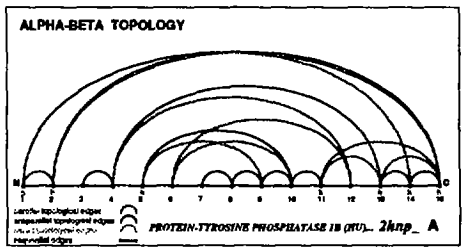


Figure 14: The connected component A in *2HNP*.

Despite of the very low similarity in their sequences they have a common sequence pattern in their active centers, which consists of a cysteine residue and an arginine residue, which are separated by five arbitrary residues. This pattern is called $Cx_5R$ motif. Fauman and Saper (Fauman & Saper 1996) distinguish the following four main families on the basis of function, structure, and sequence:

1. the tyrosine-specific phosphatases,
2. the VH1 dual-specific phosphatases,
3. the cdc25 phosphatases, and
4. the *low molecular weight* (LMW) phosphatases.

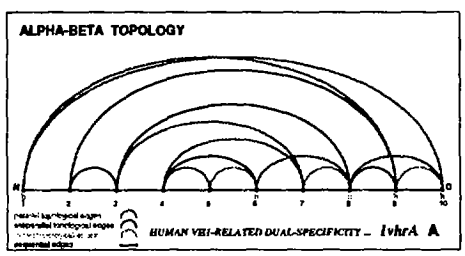


Figure 15: The connected component A in *1VHR*.

While the first three families exhibit similarity in their sequences the family of LMW phosphatases exhibits no sequential similarity to the other three families except for the very small sequence pattern, the $Cx_5R$ motif, in the active center. All the more surprising was the detection of a common arrangement of SSEs in the protein core (Fauman & Saper 1996). We have detected the same similarity with our algorithm. We consider four proteins, each being a representative of the four classes,

1. *protein tyrosine phophatase 1B* (2HNP) from human,
2. *VHR phosphatase* (1VHR) from human,
3. *tyrosine phosphatase* (1YPT) from yersinia, and
4. *tyrosine phosphatase* (1PNT) from ox.

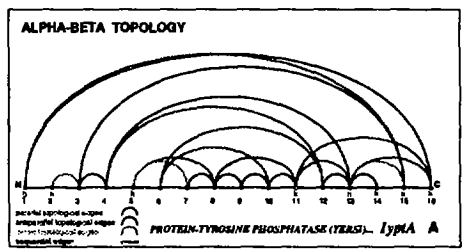


Figure 16: The connected component in *1YPT*.

The graphs to be compared are small (see figures 14, 15, 16, and 17). They exhibit up to 17 vertices and 25 edges. For the pairwise comparisons the resulting product graphs in which we search for cliques are also small, with up to 61 vertices and 1264 edges, such that the runtimes are below one second.

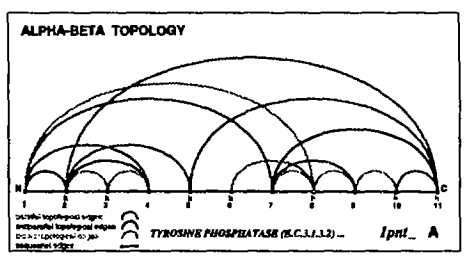


Figure 17: The connected component in *1PNT*.

The graphs of *1YPT*, *2HNP*, and *1VHR* are characterized by antiparallel spatial adjacencies, for which the participating SSEs are distant in the sequence. On the other hand the graph of *1PNT* is smaller and different. The antiparallel adjacencies that are characteristic for the other three phosphatases are missing, here. Moreover, there are seven mixed edges in *1PNT*, which do not occur in *2HNP* and occur only three times in *1VHR*.

Despite of the differences in their graphs the four phosphatases share a common structural pattern,

| SSE | *1PNT* | *1VHRA* | *1YPTA* | *2HNP* |
|---|---|---|---|---|
| E | 6- 12 | A120-A123 | A400-A402 | 211-214 |
| H | 18- 32 | A102-A113 | A409-A420 | 221-237 |
| E | 39- 45 | A58- A61 | A254-A259 | 79- 84 |
| H | 57- 65 | A164-A179 | A448-A459 | 264-281 |
| E | 87- 90 | A37- A40 | A282-A284 | 106-109 |
| H | 135-156 | A130-A142 | A362-A385 | 189-200 |

Table 3: The assignments of SSEs of the common substructures in the four phosphatases. The sequence ranges of the SSEs are given.

which consists of three helices and three parallel strands. Table 3 lists the assignments of SSEs. The spatial representations are shown in figures 18, 19, 20, and 21.

Figure 18: The maximum common substructure in *2HNP*.

Interestingly, the four phosphatases exhibit a maximal common substructure with an additional common strand. This common substructure can be superimposed in all four structures (Fauman and Saper 1996). This strand cannot be detected by the algorithm, because it occurs in an antiparallel neighbourhood in *2HNP* , *1VHR*, and *1YPT*, whereas it occurs in a parallel adjacency to the three strands in the common substructure in *1PNT*. Thus, the labels of the edges in the protein graphs matched cannot be considered as compatible edges.

LWM phosphatases have been found in mammals, yeast and various bacteria. While many enzymes of the other families possess regulatory or targeting domains, the LWM phosphatases seem to be composed of a catalytic domain alone. No specific biological function is yet known for this family. But in vitro they are specific for aryl phosphates, such as phosphotyrosine.

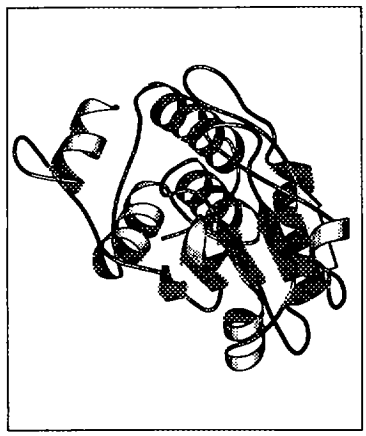

Figure 19: The maximum common substructure in *1VHR*.

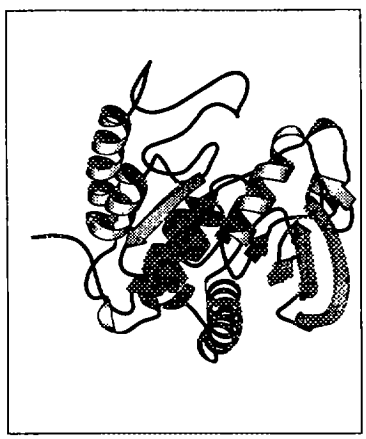

Figure 20: The maximum common substructure in *1YPT*.

Besides the $Cx_5R$ motif no sequence similarity exists between the LWM phosphatases and the other phosphatases. Therefore, the similar arrangement of SSEs in the four enzymes is quite surprising. The common substructure of four $\beta$-strands and three $\alpha$-helices can be superimposed with an RMSD of 3.3 Å.

## Multiple structure alignment of subtilisin Carlsberg, elongation factor Tu, flavodoxin, and carboxypeptidase

*Subtilisin Carlsberg* (1CSE) is a bacterial proteolytic enzyme and belongs to the class of *serine proteases*. It cleaves peptide bonds by hydrolysis. *Elongation factor Tu* (1ETU) is a bacterial transport protein, that catalyses the elongation in the protein synthesis. *Flavodoxin*

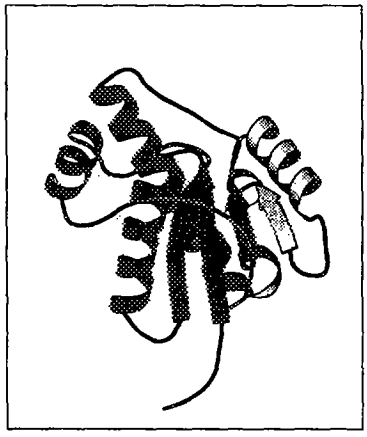

Figure 21: The maximum common substructure in *1PNT*.

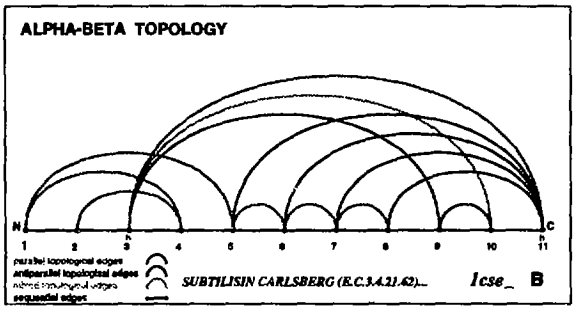


Figure 22: The connected component B of *subtilisin Carlsberg*.

(4FXN) is a flavinmononucleotide(FMN)-binding protein. *Carboxypeptidase A* (5CPA) is another proteolytic enzyme. It is a degestive enzyme, that hydrolyses carboxyl endstanding peptide bonds in polypeptide chains.

The similarity of the four proteins is discussed by Rufino and Blundell (Rufino & Blundell 1994).

The graphs to be compared (see figures 22, 23, 24, and 25) consist of up to 17 vertices and 25 edges. A similarity is not recognizable at first sight.

For the pairwise comparisons the resulting product graphs are also small, with up to 75 vertices and 1544 edges. The runtime for all pairwise comparisons is below one second.

The algorithm calculates several maximum common substructures, which consist of four parallel strands and two helices. The spatial arrangement of one maximum common substructure is shown in figures 26, 27, 28, and 29.

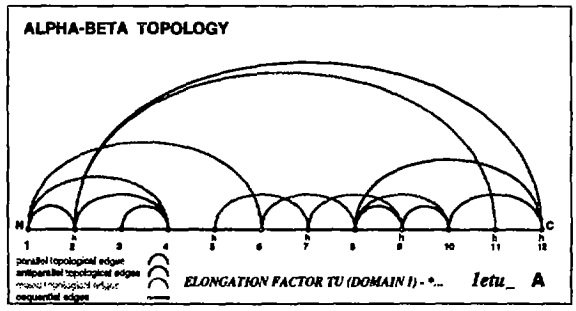


Figure 23: The connected component A of *elongation factor Tu*.

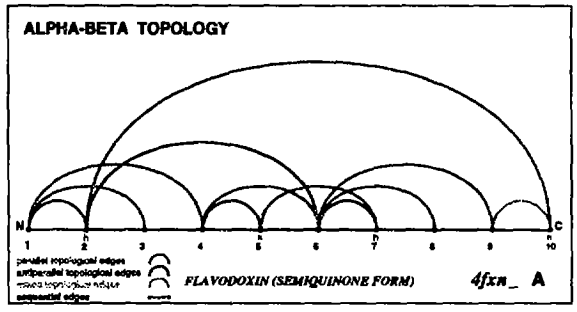


Figure 24: The connected component A of *flavodoxine*.

The appropriate SSEs assignments are listet in table 4.

| SSE | 1CSE | 1ETU | 4FXN | 5CPA |
|---|---|---|---|---|
| E | E27 - E32 | 11 - 17 | 48 - 53 | 104 - 108 |
| H | E64 - E73 | 113 - 125 | 125 - 136 | 286 - 306 |
| E | E89 - E94 | 100 - 105 | 81 - 88 | 61 - 66 |
| E | E121 - E121 | 129 - 135 | 2 - 6 | 189 - 197 |
| E | E148 - E152 | 169 - 172 | 31 - 34 | 265 - 271 |
| H | E220 - E237 | 143 - 160 | 11 - 25 | 73 - 89 |

Table 4: The assignments of SSEs of the common substructures in *subtilisin Carlsberg*, *elongation factor Tu*, *flavodoxin* und *carboxypeptidase a*. The sequence ranges of the SSEs are given.

## Concluding remarks

We have presented results obtained using a fast graph-theoretic algorithm for pairwise and multiple structure alignment. We have considered examples with low topological similarity or sequence similarity, which nevertheless exhibit structural similarity. The algorithm finds the reported similarities quickly and cor-

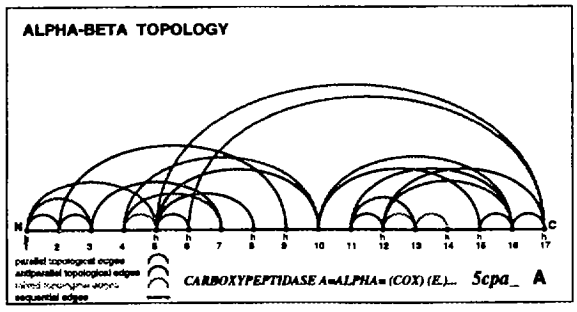


Figure 25: The connected component A of *carboxypeptidase a*.

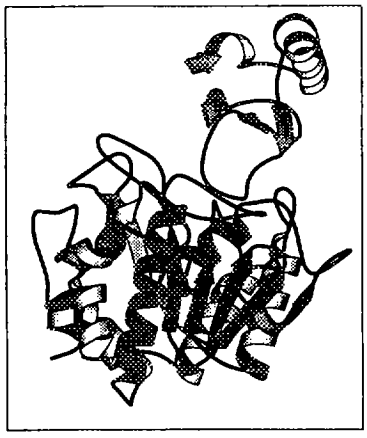

Figure 26: One maximum common substructure in *1CSE*.

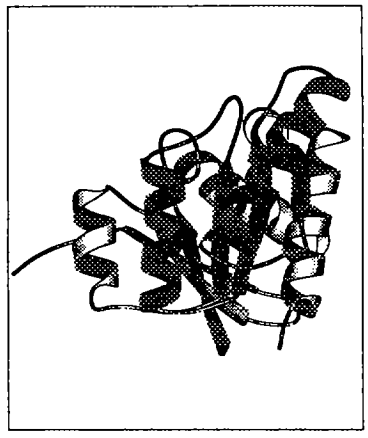

Figure 27: One maximum common substructure in *1CSE*.

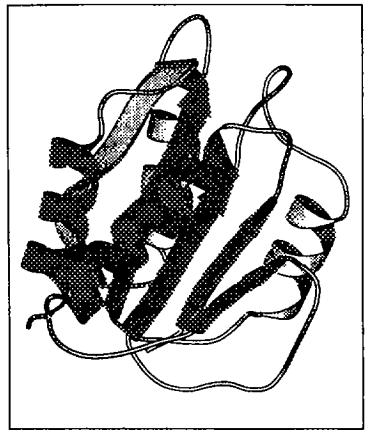

Figure 28: One maximum common substructure in *1ETU*.

rectly. Other examples, that we have calculated, but cannot report here for lack of space, confirm these results. The method is very fast, also for large proteins with a lot of possible matches, such as proteins that contain the jelly-roll motif. If the common substructure is highly symmetric, the algorithm reports all possible matches of the same substructure. In such cases the number of matches can be reduced by using a subgraph isomorphism algorithm.

By modelling the protein structure at the secondary structure level a lot of substantial abstractions have been done. The resulting similar substructures, found by the algorithm cannot always been superimposed with a small RMSD. Thus, our method finds *weak* structural similarity by a *weak* modelling of the protein structure, although we use an *exact* graph algorithm.

The results show that modelling proteins at the secondary structure level is useful for detecting distant similarities in the core of protein structures. At the present time we have not performed a comparison of the whole structural database against itself. A systematic application of the method to the whole database would be very interesting.

## References

Alexandrov, N., and Fischer, D. 1996. Analysis of topological and nontopological structural similarities in the PDB: new examples with old structures. *PROTEINS: Structure, Function, and Genetics* 25:354–365.

Artymiuk, P.; Rice, D.; Mitchell, E.; and P. Willett, P. 1990. Structural resemblance between the families of bacterial signal-transduction proteins and of

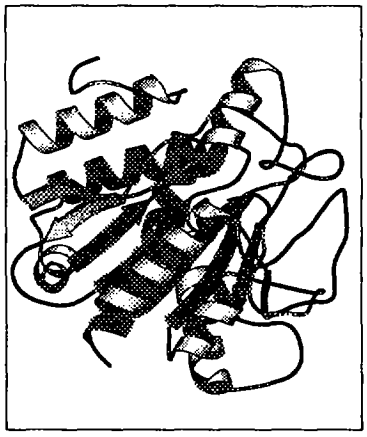

Figure 29: One maximum common substructure in *5CPA*.


G proteins revealed by graph theoretical techniques. *Protein Engineering* 4:39–43.

Bachar, O.; Fischer, D.; R. Nussinov, R.; and H. Wolfson, H. 1993. A computer vision based for 3-d sequence-independent structural comparison of proteins. *Protein Engineering* 6:279–288.

Bernstein, F.; Koetzle, T.; Williams, G.; Meyer, Jr., E.; Brice, M.; Rodgers, J.; Kennard, O.; Shimanouchi, T.; and Tasumi, M. 1977. The protein data bank: a computer based archival file for macromolecular structures. *J.Mol.Biol.* 112:535–542.

Brint, A., and Willett, P. 1987. Pharmacophoric pattern matching in files of 3D chemical structures: comparision of geometric searching algorithms. *J.Mol.Graphics* 5:49–56.

Bron, C., and Kerbosch, J. 1973. Algorithm 457 - finding all cliques of an undirected graph. *Commun.ACM* 16:575–577.

Fauman, E., and Saper, M. 1996. Structure and function of the protein tyrosine phosphatases. *Trends in Biochem.Sci.* 21:413–417.

Grindley, H.; Artymiuk, P.; Rice, D.; and Willett, P. 1993. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J.Mol.Biol.* 229:707–721.

Gusfield, D.; Balasubramanian, K.; and Naor, D. 1992. Parametric optimization of sequence alignment. In *Proceedings of the Third Annual ACM-SIAM Symposium on Discrete Algorithms*, 432–439.

Holm, L., and Sander, C. 1993. Protein structure comparison by alignment of distance matrices. *J.Mol.Biol.* 233:123–138.

Holm, L., and Sander, C. 1995. 3-D-lookup: Fast protein struckture database searches at 90% reliability. In Rawlings, C.; Clark, D.; Altman, R.; Hunter, L.; Lengauer, T.; and Wodak, S., eds., *Proceedings Third International Conference on Intelligent Systems for Molecular Biology.* Menlo Park, California: AAAI Press. 179–187.

Johnson, M.; Šali, A.; and Blundell, T. 1990. Phylogenetic relationships from three-dimensional protein structures. *Methods in Enzym.* 183:670–690.

Kabsch, W., and Sander, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637.

Kaden, F.; Koch, I.; and Selbig, J. 1990. Knowledge-based prediction of protein structures. *J.theor.Biol.* 147:85–100.

Koch, I.; Kaden, F.; and Selbig, J. 1992. Analysis of protein sheet topologies by graph-theoretical methods. *PROTEINS: Structure, Function, and Genetics* 12:314–323.

Koch, I.; Lengauer, T.; and Wanke, E. 1996. An algorithm for finding maximal common subtopologies in a set of protein structures. *J.Comp.Biol.* 3:289–306.

Lathrop, R.; Webster, T.; and Smith, T. 1987. ARIADNE: pattern-directed inference and hierarchical abstraction in protein structure recognition. *Commun.ACM* 30:909–921.

Lessel, U., and Schomburg, D. 1994. Similarities between protein 3-D structures. *Protein Engineering* 7:1175–1187.

May, A., and Johnson, M. 1994. Protein structure comparisons using a combination of a genetic algorithm, dynamic programming and least-squares minimization. *Protein Engineering* 7:475–485.

Mitchell, E.; Artymiuk, P.; Rice, D.; and Willett, P. 1993. Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J.Mol.Biol.* 212:151–166.

Mizuguchi, K., and Gō, N. 1995. Comparison of spatial arrangements of secondary structural elements in proteins. *Protein Engineering* 8:353–362.

Needleman, S., and Wunsch, C. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J.Mol.Biol.* 48:443–453.

Orengo, C.; Brown, N.; and Taylor, W. 1992. Fast structure alignment for protein databank searching. *PROTEINS: Structure, Function, and Genetics* 14:139–167.

Pickett, S.; Saqi, M.; and Sternberg, M. 1992. Evaluation of the sequence template method for protein structure prediction. *J.Mol.Biol.* 228:170–187.

Rufino, S., and Blundell, T. 1994. Structure-based identification and clustering of protein families and superfamilies. *J.Comp.-Aid.Mol.Design* 8:5–27.

Sayle, R. 1996. RasMol Molecular Graphics Visualisation Tool. *Glaxo Research & Development, Greenford, Middlesex* Version 2.5.

Sutcliffe, M.; Haneef, I.; Carney, D.; and Blundell, T. 1987. Knowledge based modelling of homologous proteins, part I: three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Engineering* 1:377–384.

Taylor, W., and Orengo, C. 1996. Protein structure alignment. *J.Mol.Biol.* 208:1–22.

Taylor, W.; Flores, T.; and Orengo, C. 1994. Multiple protein structure alignment. *Protein Science* 3:1858–1870.

Šali, A., and Blundell, T. 1990. Definition of general topological equivalence in protein structures: a procedure involving comparison of properties and relationships through simulated annealing and dynamic programming technique. *J.Mol.Biol.* 212:344–346.

Waterman, M.; Eggert, M.; and Lunder, E. 1992. Parametric sequence comparisons. *Proc.Nat.Acad.Sci.USA* 89:6090–6093.