

Similarity Analysis of Protein Binding Sites: A Generalization of the Maximum Common Subgraph Measure Based on Quasi-Clique Detection

Imen Boukhris¹, Zied Elouedi¹, Thomas Fober²,
Marco Mernberger², and Eyke Hüllermeier²

¹ LARODEC Laboratory, Higher Institute of Business, Tunis, Tunisia

boukhris.imen@gmail.com, zied.elouedi@gmx.fr

² Department of Mathematics & Computer Science, Philipps-Universität, Marburg, Germany

{thomas, mernberger, eyke}@mathematik.uni-marburg.de

Protein binding sites are often represented by means of graphs capturing their most important geometrical and physicochemical properties. Searching for structural similarities and identifying functional relationships between them can thus be reduced to matching their corresponding graph descriptors. In this paper, we propose a method for the structural analysis of protein binding sites that makes use of such matching techniques to assess the similarity between proteins independently of sequence or fold homology. More specifically, we propose a similarity measure that generalizes the commonly used maximum common subgraph measure in two ways. First, using algorithms for so-called quasi-clique detection, our measure is based on maximum ‘approximately’ common subgraphs, a relaxation of maximum common subgraphs which is tolerant toward edge mismatches. Second, instead of focusing on equivalence, our measure is a compromise between a generalized equivalence and an inclusion measure. An experimental study is presented to illustrate the effectiveness of the method and to show that both types of relaxation are useful in the context of protein structure analysis.

1 Introduction

The progress in medicine and drug design largely hinges on discoveries in bioinformatics. Indeed, with the exponential growth of molecular data, computational techniques are needed to extract, store and process this data. The structural comparison of proteins is one of the main tasks in bioinformatics, since it is well-known that functional similarity does not necessarily come along with sequence similarity [7].

Our focus in this paper will be on the special case of protein binding sites derived from crystal structures. To model such structures in a formal way, we resort to a graph representation which is able to capture the most important geometrical and physicochemical properties of a binding site. For a long time, graphs have been used in cheminformatics for the modeling of chemical compounds [4]. In bioinformatics, they are becoming more and more important, too, due to their general versatility in modeling complex structures such as proteins or interaction networks [2]. It is hence not surprising that a number of methods has been developed for comparing graphs representing protein structures (e.g. [6, 10, 23]), and for computing related similarity measures, for example based the concepts of maximum (minimum) common subgraph (supergraph) [18, 19] or graph edit distance [15].

Considering the definition of the maximum common subgraph, a drawback of this measure is its sensitivity toward errors and small deviations. This becomes especially obvious in the case of graphs with real-valued edge weights. Due to mutations, molecular flexibility, and noise in the data, one cannot expect to find exact matches in the context of comparing protein binding sites. This may result in very small common subgraphs that fail to capture the structural similarities of two or more protein binding sites in a proper way. To overcome this problem, we relax the condition of exact matches and propose a method for detecting “approximately” common subgraphs, which is arguably more appropriate to search for common substructures in biological data. To this end, we employ the concept of a so-called *quasi-clique* of a graph that has recently been studied in the literature [1, 14, 16].

The remainder of the paper is organized as follow: In Section 2, we introduce protein binding sites and their graph representation. Section 3 discusses the problem of finding a maximum common subgraph using clique detection techniques. The concept of a quasi-clique and our novel similarity measure are introduced in Section 4. An experimental validation is presented in Section 5. Section 6 concludes the paper.

2 Graph-Based Representation of Protein Binding Sites

To model protein binding sites as graphs, we build upon CavBase [21, 22], a database developed for the purpose of identifying and extracting putative protein binding sites from structural data deposited in the protein database (PDB) [3]. CavBase detects putative binding sites as cavities on the surface of proteins by using the LIGSITE algorithm [9]. The geometry of a protein binding site is internally represented by a set of pseudocenters, spatial points that represent the physico-chemical properties of a surface patch within the binding site. Pseudocenters can be seen as a compressed spatial representation of areas on the cavity surface where certain protein-ligand interactions are experienced. Currently, CavBase uses seven types of pseudocenters (donor, acceptor, donor-acceptor, pi, aromatic, aliphatic and metal) that account for different types of possible interactions between residues of the binding site and the substrate of the protein. These pseudocenters are derived from the amino acid composition of the binding site.

To model such structures, we make use of node-labeled and edge-weighted graphs

$G = (V, E, l_V, l_E)$ where V is the set of nodes, $E = V \times V$ is the set of edges, $l_V : V \rightarrow \{1, \dots, 7\}$ assigns labels to nodes (each number represents one physicochemical property), and $l_E : E \rightarrow \mathbb{R}$ assigns weights to edges that represent the distance between the adjacent nodes.¹ To reduce the complexity of the representation and increase algorithmic efficiency, we use an approximate representation in which edges exceeding a certain length are ignored; in this regard, a threshold of 11 Ångström has proved to be a reasonable choice [6]. Despite this approximation, our representation will produce graphs that are rather dense, as approximately 20 percent of all pairs of nodes are connected by an edge.

3 Similarity Based on the Maximum Common Subgraph

A simple though intuitively appealing and frequently used approach to graph comparison is to define the similarity between two graphs in terms of the size (number of nodes) of their maximum common subgraph (MCS). To obtain a normalized variant of this measure, the size of the MCS is often divided by the number of nodes of the larger of the two graphs [8, 17]. This leads to a similarity measure $s : \mathcal{G} \times \mathcal{G} \rightarrow [0, 1]$, where 1 indicates that both graphs are isomorphic and 0 that both graphs have nothing in common. Before turning our attention to the problem of finding (maximum) common subgraphs, we recall some terms that will be needed in the further discussion.

Graph isomorphism: Given two graphs

$$G = (V, E, l_V, l_E), \quad G' = (V', E', l'_V, l'_E),$$

a graph isomorphism is a bijection $f : V \rightarrow V'$ satisfying the following properties: For all $u, v \in V$,

$$(u, v) \in E \Leftrightarrow (f(u), f(v)) \in E'.$$

Moreover, for node-labeled and edge-weighted graphs, $l_V(u) = l'_V(f(u))$, $l_V(v) = l'_V(f(v))$ and $l_E(u, v) = l'_E(f(u), f(v))$ must hold for all $u, v \in V$ and $(u, v) \in E$. G and G' are called isomorphic, $G \approx G'$, if there exists a graph isomorphism between them. Obviously, isomorphism is an equivalence relation on graphs.

Subgraph: A graph $G_S = (V_S, E_S)$ is a subgraph of a graph $G = (V, E)$ if $V_S \subseteq V$ and $E_S \subseteq E \cap (V_S \times V_S)$. It is an induced subgraph if $V_S \subseteq V$ and $E_S = E \cap (V_S \times V_S)$.

Maximum common subgraph: Given two graphs G and G' , G_{cs} is called a common subgraph of G and G' if there is an induced subgraph G_S of G and an induced subgraph G'_S of G' such that $G_{cs} \approx G_S$ and $G_{cs} \approx G'_S$. A common subgraph is called a maximum common subgraph (MCS) if there is no other common subgraph of G and G' with more nodes than G_{cs} .

¹Since our edges are undirected, it would be more correct to use a subset instead of a tuple representation. For convenience, however, we stick to the simpler tuple notation, with the implicit understanding that $(u, v) \in E$ implies $(v, u) \in E$ and $l_E((u, v)) = l_E((v, u))$.

Clique: A clique in a graph $G = (V, E)$ is an induced subgraph $G_C = (V_C, E_C)$ which is fully connected, i.e., such that $(u, v) \in E_C$ for all $u, v \in V_C$.

Product graph: The product graph $G_\bullet = (V_\bullet, E_\bullet)$ of two graphs $G = (V, E)$ and $G' = (V', E')$ is defined by its node set $V_\bullet \subseteq V \times V'$ and its edge set $E_\bullet \subseteq V_\bullet \times V_\bullet$ as follows:

$$\begin{aligned} V_\bullet &= \{ (v_i, v'_j) \mid l_V(v_i) = l_V(v'_j) \} \\ E_\bullet &= \{ ((v_i, v'_j), (v_k, v'_l)) \mid l_E(v_i, v_k) = l_E(v'_j, v'_l) \} \end{aligned}$$

The product graph has a number of interesting properties, one of them being especially important for our purpose: A clique in the product graph of two graphs G and G' corresponds to a common subgraph of G and G' [12]. Thus, to detect common subgraphs, one can simply search for cliques in the product graph $G_\bullet = G \times G'$, and finding a maximum common subgraph amounts to finding a maximal clique in G_\bullet . In other words, the problem of finding a maximum common subgraph can be reduced to the problem of clique detection, and any algorithm for the latter can be used to solve the former. In this regard, it is worth mentioning that clique detection is an NP-complete problem [11]. Therefore, exact algorithms are feasible only for very small graphs, while practically relevant problems are usually solved in an approximate way by means of heuristic algorithms.

Suppose that V_C is the set of nodes in a largest clique found in G_\bullet and, hence, its cardinality the size of the maximum common subgraph of G and G' . The similarity between G and G' can then be defined as follows:

$$sim(G, G') = \frac{|V_C|}{\max\{|V|, |V'|\}} . \quad (1)$$

4 Graph Similarity Based on Quasi-Cliques

Cliques are the densest form of subgraphs, since each pair of nodes must be connected by an edge. Considering the retrieval of the maximum common subgraph by searching for cliques in the product graph G_\bullet , this means that all node and edge labels must be equal. As mentioned previously, this requirement is overly restrictive in the context of biological data analysis, especially in the case of structure analysis where edges are labeled with real-valued distances. An obvious approach to introduce some tolerance is to define the set of edges $E_\bullet \subseteq V_\bullet \times V_\bullet$ in the product graph G_\bullet as

$$\{ ((v_i, v'_j), (v_k, v'_l)) \mid \|l_E(v_i, v_k) - l_E(v'_j, v'_l)\| \leq \epsilon \},$$

which means that, in the maximum common subgraph, the length of isomorphic edges is allowed to differ by at most a constant ϵ . We consider complete graphs and weight "missing" edges with infinity. Furthermore we assume that the distance between two such edges is always smaller ϵ . Yet, looking for cliques in G_\bullet still means that this condition must hold for *all* pairs of edges in the MCS. Roughly speaking, this approach is

tolerant toward possibly numerous though small (measurement) errors but not toward single though exceptionally large deviations. To become flexible in this regard, too, our idea is to replace the detection of cliques in G_\bullet by the detection of *quasi-cliques*.

4.1 Quasi-Cliques

Roughly speaking, quasi-cliques are “almost complete” graphs $G = (V, E)$. In the literature, different definitions of quasi-cliques have been proposed. Some of them are based on the degree of the nodes [14, 16], calling G a quasi-clique if every node in V is adjacent to at least $\gamma \cdot (|V| - 1)$ other nodes, where $\deg(v)$ is the number of nodes adjacent to v . This is the definition that we shall adopt in this paper. Yet, other definitions do exist, for example referring to the edge density: A graph G is a quasi-clique if $|E| \geq \gamma \cdot \binom{|V|}{2}$ [1].

In both cases, $\gamma \in]0, 1]$ is a relaxation parameter. Note that the concept of a γ -quasi-clique is a proper generalization of the concept of a cliques, since each clique is a 1-quasi-clique.

4.2 Quasi-Clique Detection

As mentioned earlier, the problem to find a maximum clique in a graph is NP-complete [11]. Since quasi-cliques are a generalization of cliques, it immediately follows that finding a maximum γ -quasi-clique is an NP-complete problem, too. Therefore, to solve the problem, one has to resort to heuristic algorithms.

Heuristic methods for clique detection typically exploit a downward-closure property, namely that a supergraph of a non-clique cannot be a clique either. Unfortunately, this property does not hold for quasi-cliques, as one can easily show by counter-examples. Instead, any subset of the set of nodes V in a graph $G = (V, E)$ may form a γ -quasi-clique.

Nevertheless, alternative heuristic methods for quasi-clique detection have been developed. In our approach, we make use of the method proposed in [13], which represents all potentially maximal γ -quasi-cliques by its nodes in a set-enumeration tree [20]. Thus, the search space is given by the powerset of the set of nodes V . Searching for maximal quasi-cliques is performed by means of a depth-first search on the set-enumeration tree. Once a quasi-clique has been discovered, it is stored in a prefix-tree, so that a maximal γ -quasi-cliques is provably found in a leaf of the prefix-tree. For technical details, we refer to [13].

4.3 Similarity Based on Quasi-Cliques

Finding a maximum γ -quasi-clique in the product graph of two graphs G and G' means finding a maximum *approximately* common subgraph (MACS) of these two graphs. Fig. 1 illustrates this correspondence through a simple example: In the upper part of the figure, two node labeled and edge weighted graphs are shown. Note that both graph share a roughly similar subgraph consisting of the five nodes labeled A to E. From these graphs a product graph is calculated ($\epsilon = 0.5$) and the MCS and MACS are derived by

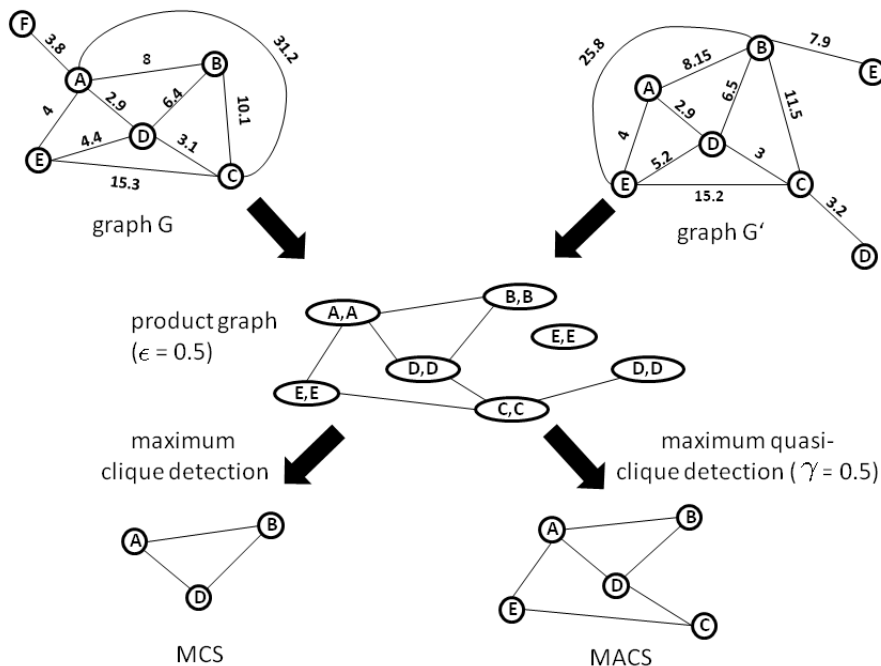


Figure 1: Illustration of the correspondence between quasi-cliques and MACS.

clique detection and γ -quasi-clique detection ($\gamma = 0.5$), respectively. Obviously, the γ -quasi-clique detection is able to capture all five nodes of the approximately common subgraph.

The MACS computed by quasi-clique detection can in turn be used to define a similarity degree via (1). Obviously, the smaller γ is, i.e., the more tolerant the comparison, the larger the MACS and, hence, the larger the degree of similarity becomes.

Despite this obvious possibility, we opt for another type of similarity measure. In fact, (1) may become problematic for the comparison of structures of different size. For example, since protein binding sites do not have a clear-cut boundary, it often happens that a structure is larger than the actual binding site. In such cases, where, for instance, one structure is a subpocket of the other one containing the most important catalytic residues (while the rest of the binding site is functionally less important), it might be desirable to consider $G \subseteq G'$ instead of $G = G'$ as a sufficient condition for a high similarity degree, a property which is not supported by (1).

Our idea, therefore, is to express similarity in terms of subset relations, proceeding from the following equivalence known from set theory:

$$A = B \Leftrightarrow A \subseteq B \wedge B \subseteq A. \quad (2)$$

Let $G = (V, E)$ and $G' = (V', E')$ be two graphs and let $QC = (V_{QC}, E_{QC})$ be the

maximum γ -quasi-clique of their product graph $G_{\bullet} = G \times G'$. Then, the fraction

$$\alpha = \frac{|V_{QC}|}{|V|} \in [0, 1]$$

can be considered as a degree to which G is a subset of G' . Likewise,

$$\beta = \frac{|V_{QC}|}{|V'|} \in [0, 1]$$

corresponds to the degree to which G' is a subset of G . Obviously, (1) is then given by $\min\{\alpha, \beta\}$. An interesting generalization to using the minimum operator for combining the two degrees of inclusion has been proposed in [5], namely the use of an Ordered Weighted Averaging (OWA) operator [24]. In our special case, this leads to the measure

$$\text{sim}(G, G') = \varphi \min(\alpha, \beta) + (1 - \varphi) \max(\alpha, \beta), \quad (3)$$

where $\varphi \in [0, 1]$ is a compromise parameter. Note that $\varphi = 1$ recovers the original measure (1), which yields $\text{sim}(G, G') = 1$ only if $G = G'$, while $\varphi = 0$ corresponds to a set inclusion measure for which $G \subseteq G'$ or $G' \subseteq G$ is sufficient to obtain a similarity degree of 1. Parameter values $0 < \varphi < 1$ produce measures in-between these extreme cases.

5 Experimental Results

We conducted a performance study using a data set from [6], namely a set of binding sites belonging to the two classes of ATP- and NADH-binding proteins. For complexity reasons, however, we removed from this data set all structures whose size exceeded 200 nodes. In the construction of the product graph, we have considered two edges as a match if their lengths differ by at most a threshold $\epsilon = 0.2$ which has proved to be a reasonable choice [6].

To assess the performance of our new similarity measure based on quasi-cliques, we compare it to the standard clique measure (1). In comparison with this measure, our approach has two degrees of freedom, namely the parameter γ which controls the relaxation of the clique concept for pattern matching, and the parameter φ that determines the type of comparison and interpolates between a (generalized) equivalence measure and a measure of inclusion. Note that (1) corresponds to the most stringent type of measure obtained for $\gamma = \varphi = 1$. Our conjecture is that less stringent variants, obtained for $0 < \gamma, \varphi < 1$ will be more appropriate in the context of protein structure analysis.

To assess the usefulness of a similarity measure, we used it in the context of k -nearest-neighbor classification. The idea is that, the more suitable a similarity measure is, the better is the performance of a k -NN classifier using this measure for determining the nearest neighbors of a query. We measured performance in terms of classification accuracy (percent of correct classifications, PCC), which in turn was estimated by means of a leave-one-out-cross-validation.

Table 1: Accuracy of the clique-measure depending on φ .

φ	PCC (k=1)	PCC (k=3)	PCC (k=5)
0	65.68	75.49	75.49
0.1	67.64	74.50	74.50
0.2	67.64	75.49	74.50
0.3	68.62	74.50	74.50
0.4	69.60	74.50	75.49
0.5	78.43	76.47	79.41
0.6	74.50	76.47	75.49
0.7	74.50	79.41	71.56
0.8	71.56	73.52	74.50
0.9	65.68	71.56	74.50
1	64.70	64.70	67.64
mean PCC	69.86	74.23	74.32

Table 1 summarizes the results obtained for the standard clique-measure where $\gamma = 1$ for different values of φ and different sizes k of the neighborhood in k -NN classification. As can be seen, the best results are indeed achieved for values $\varphi \approx 0.5$, suggesting that neither a pure equivalence nor a pure inclusion measure is optimal. Instead, a mixture of the two yields a good compromise and seems to produce improved similarity degrees.

This result was confirmed by experiments with the quasi-clique measure. Therefore, we fixed the value $\varphi = 0.5$ for this measure which corresponds to the best results found in the case of the clique measure and analyzed the effect of the clique-parameter γ . The results are shown in Table 2. As can be seen, medium-sized values of $\gamma \approx 0.6$ yield the best results, which means that a relaxation of the clique concept does indeed pay off. Compared to the strongest result of the clique-measure (79.41%), the best performance of the quasi-clique measure (89.21%) is significantly higher.

Table 2: Accuracy of the quasi-clique-measure depending on $\gamma(\varphi = 0.5)$

γ	PCC (k=1)	PCC (k=3)	PCC(k=5)
0.1	55.88	55.88	57.84
0.2	69.60	72.54	74.50
0.3	74.50	75.49	78.43
0.4	78.43	79.41	81.37
0.5	84.31	85.29	85.29
0.6	84.31	85.29	89.21
0.7	83.33	80.39	83.33
0.8	80.39	77.45	81.37
0.9	78.43	76.47	79.41
1	78.43	76.47	79.41

These results confirm our conjecture that an increased tolerance toward mismatches and minor differences due to conformational flexibility or measurement errors allows for the detection of a larger MACS of a pair of graphs and thus leads to a more useful similarity measure for binding pockets. Of course, it can also be seen that, as expected, decreasing γ beyond a certain level is not meaningful. In fact, for very small γ , geometrical constraints are not only relaxed but essentially ignored. For example, when we take $\gamma \rightarrow 0$, the similarity measure will only take the distribution of physicochemical properties into account, without paying any attention to the geometry of the binding site.

6 Conclusion

Maximum common subgraphs have been used successfully as similarity measures for graphs. In this paper, however, we have argued that this measure is overly stringent in the context of protein structure comparison, mainly since graph descriptors of such structures are only approximate models afflicted with noise and imprecision.

Therefore, we have proposed an alternative measure relaxing the MCS in two different ways. First, using algorithms for *quasi-clique* detection, our measure is based on maximum *approximately* common subgraphs, a relaxation of MCS which is tolerant toward edge mismatches. Second, instead of focusing on equivalence, our measure is a compromise between a generalized equivalence and an inclusion measure. First empirical studies, in which similarity measures are used for the purpose of classification, suggest that both types of relaxation are useful and lead to improved measures of similarity between protein binding sites.

References

- [1] J. Abello, M. G. C. Resende, and R. Sudarsky. Massive quasi-clique detection. In *Latin American Theoretical Informatics*, pages 598–612, 2002.
- [2] Johannes Berg and Michael Lässig. Local graph alignment and motif search in biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(41):14689–14694, 2004.
- [3] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.
- [4] Horst Bunke and Xiaoyi Jiang. Graph matching and similarity. *Intelligent systems and interfaces*, 15:281 – 304, 2000.
- [5] Thomas Fober and Eyke Hüllermeier. Fuzzy modeling of labeled point cloud superposition for the comparison of protein binding sites. In *IFSA World Congress, EUSFLAT World Conference*, 2009.
- [6] Thomas Fober, Marco Mernberger, Gerhard Klebe, and Eyke Hüllermeier. Evolutionary construction of multiple graph alignments for the structural analysis of biomolecules. *Bioinformatics*, 2009.

- [7] J. F. Gibrat, T. Madej, and S. H. Bryant. Surprising similarities in structure comparison. *Current Opinion in Structural Biology*, 6(3):377–385, 1996.
- [8] M. Hattori, Y. Okuno, S. Goto, and M. Kanehisa. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.*, 125(39), 2003.
- [9] M. Hendlich, F. Rippmann, and G. Barnickel. LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal of Molecular Graphics and Modelling*, 15:359–363, 1997.
- [10] M. Jambon, A. Imberty, G. Deleage, and C. Geourjon. A New Bioinformatic Approach to Detect Common 3 D Sites in Protein Structures. *Proteins Structure Function and Genetics*, 52(2):137–145, 2003.
- [11] R.M. Karp, R.E. Miller, and J.W. Thatcher. Reducibility among combinatorial problems. *The Journal of Symbolic Logic*, 40(4):618–619, 1975.
- [12] G. Levi. A note on the derivation of maximal common subgraphs of two directed or undirect graphs. *Calcolo*, 9:341–352, 1972.
- [13] G. Liu and L. Wong. Effective pruning techniques for mining quasi-cliques. In *ECML PKDD '08: Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases - Part II*, pages 33–49, 2008.
- [14] H. Matsuda, T. Ishihara, and A. Hashimoto. Classifying molecular sequences using a linkage graph with their pairwise similarities. *Theor. Comput. Sci.*, 210(2):305–325, 1999.
- [15] Michael Neuhaus and Horst Bunke. *Briding the Gap between Graph Edit Distance and Kernel Machines*. World Scientific, New Jersey, 2007.
- [16] J. Pei, D. Jiang, and A. Zhang. On mining cross-graph quasi-cliques. In *In KDD*, pages 228–238, 2005.
- [17] J. Raymond and P. Willet. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *Journal of Computer-Aided Molecular Design*, 16:521–533, 2002.
- [18] J. Raymond and P. Willett. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *Journal of Computer-Aided Molecular Design*, 16(7):521–533, 2002.
- [19] J.W. Raymond, E.J. Gardiner, and P. Willett. Heuristics for Similarity Searching of Chemical Graphs Using a Maximum Common Edge Subgraph Algorithm. *Jornal of Chemical Information and Computer Sciences*, 42(2):305–316, 2002.
- [20] R. Rymond. Search through systematic set enumeration. In *Proc. of the International Conference on Principles of Knowledge Representation and Reasoning*, pages 268–275, 1992.
- [21] S. Schmitt, M. Hendlich, and G. Klebe. From structure to function: A new approach to detect functional similarity among proteins independent from sequence and fold homology. *Angewandte Chemie International Edition*, 40(17):3141 – 3144, 2001.

- [22] S. Schmitt, D. Kuhn, and G. Klebe. A new method to detect related function among proteins independent of sequence and fold homology. *Journal of Molecular Biology*, 323(2):387–406, 2002.
- [23] N. Weskamp, E. Hüllermeier, D. Kuhn, and G. Klebe. Multiple graph alignment for the structural analysis of protein active sites. *IEEE Transactions on Computational Biology and Bioinformatics*, 4(2):310–320, 2007.
- [24] R. R. Yager. On ordered weighted averaging aggregation operators in multicriteria decision making. *IEEE Trans. Syst. Man Cybern.*, 18(1):183–190, 1988.