

Evolutionary Construction of Multiple Graph Alignments for Mining Structured Biomolecular Data

Thomas Fober, Eyke Hüllermeier, Marco Mernberger

FB Mathematik und Informatik

Philipps-Universität Marburg

thomas@mathematik.uni-marburg.de

Abstract

The concept of multiple graph alignment has recently been introduced as a novel method for the structural analysis of biomolecules. Using inexact, approximate graph-matching techniques, this method enables the robust identification of approximately conserved patterns in biologically related structures. In particular, using multiple graph alignments, it is possible to characterize functional protein families independent of sequence or fold homology. This paper first recalls the concept of multiple graph alignment and then addresses the problem of computing optimal alignments from an algorithmic point of view. In this regard, a method from the field of evolutionary algorithms is proposed and empirically compared to a hitherto existing greedy strategy.

1 Introduction

In the field of bioinformatics, multiple sequence alignment is an established approach for the identification of residues that are conserved across many members of a gene or protein family [12; 4; 11]. However, as this approach relies on evolutionary conserved sequences in DNA or protein chains to detect similarities between different molecules, it is only capable of detecting functional similarities based on heredity. Consequently, sequence analysis is not optimally suited for the identification of *functional* similarities between molecules, as functionality is more closely associated with structural than with sequential features. In fact, it is well-known that there is no strong correspondence between sequence similarity and structural similarity.

Focusing on the the identification of *structural* similarities of biomolecules, this paper presents

the concept of *multiple graph alignment* (MGA) as a structural counterpart to sequence alignment. As opposed to homology-based methods, this approach allows one to capture non-homologous molecules with similar functions as well as evolutionary conserved functional domains. Accordingly, MGA offers a much wider field of application, as the method is generally applicable to different types of biological or biochemical objects (and, in principle, even beyond the bioinformatics domain). In this regard, our special interest concerns the analysis of protein structures or, more specifically, protein binding sites. However, graph alignments can also be used for analyzing other types of biomolecules, e.g., to detect similarities between chemical molecules such as protein ligands (see Section 4.2).

Our previous approach to the MGA problem makes use of a greedy algorithm [13]. In this paper, we present an alternative method using evolutionary strategies. As will be shown experimentally, the latter significantly outperforms the former with regard to the quality of alignments, albeit at the cost of an increased runtime.

The paper is organized as follows: In Section 2, we introduce the concept of a multiple graph alignment. The problem of computing a MGA is then addressed in Section 3, where an evolutionary algorithm is proposed for this purpose. Section 4 is devoted to the experimental validation of the approach, and Section 5 concludes the paper.

2 Graph-Based Modeling and Multiple Graph Alignment

2.1 Graph-Based Modeling of Biomolecules

Graph models are often used to represent and analyze three-dimensional biomolecules, i.e., single biomolecules are represented in terms of a graph

G consisting of a set of (labeled) nodes V and (weighted) edges E . As mentioned previously, our special interest concerns protein binding pockets and chemical compounds.

Modeling Chemical Compounds

In the case of chemical compounds, atoms are represented as nodes labeled with their corresponding atom type, using the SYBYL atom type notation. The edges between nodes are labeled by the Euclidean distance between the respective atoms. This way, important geometric properties of the compound are captured. Alternatively, it is possible to use a representation in which edges correspond to molecular bonds and are weighted by the bond order.

Modeling Protein Binding Sites

Regarding the modeling of protein structures, the present work builds upon Cavbase [10; 8], a database system for the fully-automated detection and extraction of protein binding pockets from experimentally determined protein structures (available through the database PDB [6]). In Cavbase, graphs are used as a first approximation to describe binding pockets. The database currently contains 113,718 hypothetical binding pockets that have been extracted from 23,780 publicly available protein structures using the LIGSITE-algorithm [9].

To model a binding pocket as a graph, the geometrical arrangement of the pocket and its physicochemical properties are first represented by predefined *pseudocenters*—spatial points that represent the center of a particular property. The type and the spatial position of the centers depend on the amino acids that border the binding pocket and expose their functional groups. They are derived from the protein structure using a set of predefined rules [10]. As possible types for pseudocenters, hydrogen-bond donor, acceptor, mixed donor/acceptor, hydrophobic aliphatic and aromatic properties are considered. Pseudocenters can be regarded as a compressed representation of areas on the cavity surface where certain protein-ligand interactions are experienced.

The assigned pseudocenters form the nodes $v \in V$ of the graph representation, and their properties are modeled in terms of node labels $l(v) \in \{P1, P2 \dots P5\}$, where P1 stands for donor, P2 for acceptor, etc. Two centers are connected by an edge in the graph representation if their Euclidean distance is below 11.0 Å and each edge $e \in E$ is la-

beled with the respective distance $w(e) \in \mathbb{R}$.¹ The edges of the graph thus represent geometrical constraints among points on the protein surface.

In Cavbase, a graph representation of a binding pocket has around 85 nodes on average; however, also graphs with several hundred nodes are frequently detected and extremes with thousands of nodes exist. The graphs are rather dense as approximately 20 percent of all pairs of nodes are connected by an edge.

2.2 Multiple Graph Alignment

In the following, we assume a set $\mathcal{G} = \{G_1(V_1, E_1) \dots G_m(V_m, E_m)\}$ of connected, node-labeled and edge-weighted graphs to be given, each of which represents a biomolecule. To make the discussion more concrete, we subsequently consider the case of protein binding sites.

When comparing homologs from different species in protein cavity space, one has to deal with the same mutations that are also given in sequence space. Corresponding mutations, in conjunction with conformational variability, strongly affect the spatial structure of a binding site as well as its physicochemical properties and, therefore, its graph descriptor. Thus, one cannot expect that the graph descriptors for two related binding pockets match exactly. In our approach, the following types of edit operations are allowed to account for differences between a graph $G_1(V_1, E_1)$ and another graph $G_2(V_2, E_2)$:

1. Insertion or deletion of a node $v_1 \in V_1$ (“In-Del”). A pseudocenter can be deleted or introduced due to a mutation in sequence space. Alternatively, a conformational difference can affect the exposure of a functional group toward the binding pocket, accordingly an insertion or deletion in the graph descriptor could result.
2. Change of the label $l(v_1)$ of a node $v_1 \in V_1$ (“Node Mismatch”). The assigned physicochemical property (“type”) of a pseudocenter can change if a mutation replaces a certain functional group by another type of group at the same position.
3. Change of the weight $w(e_1)$ of an edge $e_1 \in E_1$ (“Edge Mismatch”). The distance between two pseudocenters can change due to conformational differences.

¹An interaction distance of 11.0 Å is typically enough to capture the geometry of a binding site; ignoring larger distances strongly simplifies the graph representation.

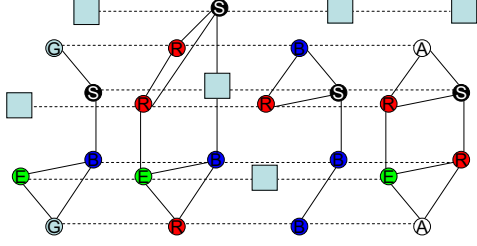


Figure 1: An alignment of four similar, but not identical graphs. The node labels are indicated by the letters assigned to the nodes (shown as circles). The edge labels are omitted for simplification. The assignments among the different graph nodes are indicated by the dashed lines. Large boxes in gray represent dummy nodes that have been introduced in the alignment to represent “missing” nodes.

By assigning a cost value to each of these edit operations, it becomes possible to define an edit distance for a pair of graph descriptors. The edit distance of two graphs G_1, G_2 is defined as the cost of a cost-minimal sequence of edit operations that transforms graph G_1 into G_2 . As in sequence analysis, this allows for defining the concept of an alignment of two (or more) graphs. The latter, however, also requires the possibility to use dummy nodes \perp that serve as placeholders for deleted nodes. They correspond to the gaps in sequence alignment (cf. Figure 1).

Let $\mathcal{G} = \{G_1(V_1, E_1) \dots G_m(V_m, E_m)\}$ be a set of graphs. Then $\mathcal{A} \subseteq (V_1 \cup \{\perp\}) \times \dots \times (V_m \cup \{\perp\})$ is an alignment of the graphs in \mathcal{G} if and only if

1. for all $i = 1 \dots m$ and for each $v \in V_i$ there exists exactly one $a = (a_1 \dots a_m) \in \mathcal{A}$ such that $v = a_i$ (i.e., each node of each graph occurs exactly once in the alignment).
2. for each $a = (a_1 \dots a_m) \in \mathcal{A}$ there exists at least one $1 \leq i \leq m$ such that $a_i \neq \perp$ (i.e., each tuple of the alignment contains at least one non-dummy node).

Each tuple in the alignment contains a certain number of nodes from the different graphs that are matched onto each other. If a node has no matching partner in a certain graph, it has to be mapped onto a dummy node \perp .

To assess the quality of a given alignment, a scoring function is needed. This scoring function corresponds to the above-mentioned edit distance, as each graph alignment defines a set of edit operations that have to be performed to transform one of the aligned graphs into another entry of the alignment. Here, a scoring function that follows a sum-of-pairs scheme is proposed, i.e., the score s of a

multiple alignment $\mathcal{A} = (a^1 \dots a^m)$ is defined by the sum of scores of all induced pairwise alignments:

$$s(\mathcal{A}) = \sum_{i=1}^n \text{ns}(a^i) + \sum_{1 \leq i < j \leq n} \text{es}(a^i, a^j), \quad (1)$$

where the *node score* (ns) is given by

$$\text{ns} \begin{pmatrix} a_1^i \\ \vdots \\ a_m^i \end{pmatrix} = \sum_{1 \leq j < k \leq m} \begin{cases} \text{ns}_m & l(a_j^i) = l(a_k^i) \\ \text{ns}_{mm} & l(a_j^i) \neq l(a_k^i) \\ \text{ns}_d & a_j^i = \perp, a_k^i \neq \perp \\ \text{ns}_d & a_j^i \neq \perp, a_k^i = \perp \end{cases}$$

Comparing two edges is somewhat more difficult than comparing two nodes, as one cannot expect to observe edges of exactly the same lengths. We consider two edges as a match if their respective lengths, a and b , differ by at most a given threshold t_{max} , and as a mismatch otherwise. The *edge score* (es) is then given by

$$\text{es} \left(\begin{pmatrix} a_1^i \\ \vdots \\ a_m^i \end{pmatrix}, \begin{pmatrix} a_1^j \\ \vdots \\ a_m^j \end{pmatrix} \right) = \sum_{1 \leq k < l \leq m} \begin{cases} \text{es}_{mm} & (a_k^i, a_k^j) \in E_k, (a_l^i, a_l^j) \notin E_l \\ \text{es}_{mm} & (a_k^i, a_k^j) \notin E_k, (a_l^i, a_l^j) \in E_l \\ \text{es}_m & d_{kl}^{ij} \leq t_{max} \\ \text{es}_{mm} & d_{kl}^{ij} > t_{max} \end{cases}$$

where $d_{kl}^{ij} = \|w(a_k^i, a_k^j) - w(a_l^i, a_l^j)\|$. The scores for the different cases have to be defined by the user. For the experiments, we used the following parameter setting: $\text{ns}_m = 1, \text{ns}_{mm} = -5.0, \text{ns}_d = -2.5, \text{es}_m = 0.2$ and $\text{es}_{mm} = -0.1$

2.3 Computing Multiple Graph Alignments

The problem of calculating an optimal MGA, that is, an alignment with maximal score for a given set of graphs is computationally very complex. The subgraph isomorphism problem (which is known to be NP-complete [1]) can be seen as a special case of the graph alignment problem where the cost for mismatches is set prohibitively high. Thus, one cannot expect to find an efficient algorithm that is guaranteed to find an optimal alignment for a given set of graphs. In [13], simple and effective heuristics for the MGA problem have been described that were found to be useful for the problem instances that were examined. The main idea of these methods is to reduce the multiple alignment problem to the problem of pairwise alignment (i.e., calculating an optimal graph alignment for only two graphs)

in a first step; these pairwise alignments are subsequently merged into a multiple alignment in a second step.

Both steps, finding optimal pairwise alignments and merging them into a multiple alignment, are only heuristics that do not guarantee to find a good solution on every problem instance. Even worse, as there is no information about the distance to the optimum, it is hardly possible to get an idea about the quality of a solution produced by the above strategy.

Motivated by these problems, we investigated the use of evolutionary algorithms as an alternative approach. As will be detailed in Section 3, we resorted to an evolution strategy with a non-standard representation and examined two different approaches: While the first approach seeks to optimize a complete multiple alignment directly, the second approach sticks to the decomposition strategy that is also used by the greedy algorithm and, hence, uses evolutionary optimization only for the pairwise case.

On the one hand, evolutionary optimization is of course more expensive from a computational point of view. On the other hand, the hope is that this approach will be able to improve the solution quality, i.e., to produce alignments that are better than those obtained by a simple greedy strategy. To see if the increased cost of an evolutionary algorithm is justified, we have also compared it with a simple hill-climbing strategy.

3 An Evolutionary Algorithm for Multiple Graph Alignment

Evolutionary Algorithms seeks to optimize a fitness function, which in our case is given by the sum-of-pairs score (1). To this end, it simulates the evolution process by repeatedly executing the following loop [3]: (i) Initially, a population consisting of μ individuals, each representing a candidate solution, is generated at random. (ii) In each generation, $\mu \cdot \nu$ offspring individuals are created; the parameter ν is called selective pressure. To generate a single offspring, the mating-selection operator chooses ρ parent individuals at random and submits them to the recombination operator. This operator generates an offspring by exchanging the genetic information of these individuals. The offspring is further modified by the mutation operator. (iii) The offsprings are evaluated and added to the parent population. Among the individuals in this temporary population T , the selection operator chooses the best μ candidates, which then form the population of the next

A:	1	2	3	4	.	5
B:	.	3	.	4	5	.	1	6	.	.	2	.	.	.
C:	5	2	.	.	4	.	.	.	1	.	.	3	.	.
D:	4	.	.	1	5	2	.	.	3	.

Figure 2: Matrix representation of a MGA. Dummies are represented by a dot.

generation. (iv) The whole procedure is repeated until a stopping criterion is met.

Regarding the representation of individuals, note that in our case candidate solutions correspond to MGAs. Given a fixed numbering of the nodes of graph G_i from 1 to $|V_i|$ (not to be confused with the labeling), an MGA can be represented in a unique way by a two-dimensional matrix, where the rows correspond to the graphs and the columns to the aligned nodes of these graphs. Fig. 2 shows an example of such a matrix. The first column indicates a mutual assignment of the first node of graph A, the fifth node of graph C, and the fourth node of graph D, while there is no matching partner in graph B.

In the course of optimizing an MGA, the graphs can become larger due to the insertion of dummy nodes. For the matrix representation, this means that the number of columns is in principle not known and can only be upper-bounded by $n_1 + \dots + n_m$, where $n_i = |V_i|$. This, however, will usually be too large a number and may come along with an excessive increase of the search space. From an optimization point of view, a small number of columns is hence preferable. On the other hand, by fixing a too small length of the alignment, flexibility is lost and the optimal solution is possibly excluded.

To avoid these problems, we make use of an *adaptive* representation: Starting with a single extra column filled with dummies, more such columns can be added if required or, when becoming obsolete, again be removed (see below). Thus, our matrix scheme is initialized with m rows and $n + 1$ columns, where $n = \max\{n_1, n_2, \dots, n_m\}$. For each graph G_i , a permutation of its nodes is then inserted, with dummies replacing the index positions $j > |V_i|$. As an aside, we note that dummy columns are of course excluded from scoring, i.e., the insertion or deletion of dummy columns has no influence on the fitness.

Among the proper selection operators for evolution strategies, the deterministic plus-selection, which selects the μ best individuals from the union of the μ parents and the λ offsprings, is most convenient for our purpose. In fact, since the search space of an MGA problem is extremely large, it would be very

unfortunate to lose a current best solution. This excludes other selection techniques such as fitness-proportional or simulated annealing selection.

As we use a non-standard representation of individuals, namely a matrix scheme, the commonly used recombination and mutation operators are not applicable and have to be adapted correspondingly. Our recombination operator randomly selects ρ parent individuals from the current population (according to a uniform distribution). Then, $\rho - 1$ random numbers r_i , $i = 1 \dots \rho - 1$ are generated, where $1 \leq r_1 < r_2 < \dots < r_{\rho-1} < m$, and an offspring individual is constructed by combining the sub-matrices consisting, respectively, of the rows $\{r_{i-1} + 1 \dots r_i\}$ from the i -th parent individual (where $r_0 = 0$ and $r_\rho = m$ by definition). Simply stitching together complete sub-matrices is not possible, however, since the nodes are not ordered in a uniform way. Therefore, the ordering of the first sub-matrix is used as a reference, i.e., the elements of the first row serve as pivot elements. General experience has shown that recombination increases the speed of convergence, and this was also confirmed by our experiments (see Section 4).

The mutation operator selects one row and two columns at random and swaps the entries in the corresponding cells. To enable large mutation steps, we have tried to repeat this procedure multiple times for each individual. As the optimal number of repetitions was unknown in the design phase of the algorithm, it was specified as a strategy component adjusted by a self-adaptation mechanism [5].

To adapt the length of an MGA (number of columns in the matrix scheme), it is checked in randomly chosen intervals whether further dummy columns are needed or existing ones have become unnecessary. Three cases can occur: (i) There exists exactly one dummy column, which means that the current length is still optimal. (ii) There is more than one dummy column: Apparently, a number of dummy columns are obsolete and can be removed, retaining only a single one. (iii) There is no dummy column left: The dummy column has been “consumed” by mapping dummies to real nodes. Therefore, a new dummy column has to be inserted.

3.1 Combining Evolutionary Optimization and Pairwise Decomposition

As mentioned above, the search space of a MGA problem grows exponentially with the number of graphs. Moreover, the evaluation time of the fitness function grows quadratically with the number

of graphs. Therefore, the application of our EA algorithm becomes expensive for large problems.

One established strategy to reduce complexity is to decompose a multiple alignment problem into several pairwise problems and to merge the solutions of these presumably more simple problems into a complete solution. This strategy has already been exploited in our greedy approach, where the merging step has been realized by means of the star-alignment algorithm [13]. In star-alignment, a center structure is first determined, and this structure is aligned with each of the other $m - 1$ structures. The $m - 1$ pairwise alignments thus obtained are then merged by using the nodes of the center as pivot elements. As the quality of a MGA derived in this way critically depends on the choice of a suitable center structure, one often tries every structure as a center. In this case, all possible pairwise alignments are needed, which means that our evolutionary algorithm must be called $\frac{1}{2}(m^2 - m)$ times.

As star-alignment is again a purely heuristic aggregation procedure, the gain in efficiency is likely to come along with a decrease in solution quality, compared with the original EA algorithm. This is not necessarily the case, however. In fact, a decomposition essentially produces two opposite effects, a positive one due to a simplification of the problem and, thereby, a reduction of the search space, and a negative one due to a potentially suboptimal aggregation of the partial solutions. For a concrete problem, it is not clear in advance which among these two effects will prevail.

4 Experimental Results

This section presents two experimental studies. The purpose of the first study was to optimize our evolutionary algorithms by means of a proper adjustment of its exogenous parameters, which have a strong influence on performance and runtime. In our case, the following parameters are concerned:

- the population size μ and the selective pressure ν ;
- selfadaption and recombination, which can assume values $\{\text{on}, \text{off}\}$, and allow the automatic step size control and the recombination operator to be enabled or disabled, respectively;
- initial step size, which defines the initial step size for the mutation; if the automatic step size control is disabled, this parameter is ignored and a constant step size of 1 is used for the mutation.

In the second study, we used our algorithms to analyze two data sets with different characteristics, the first consisting of a set of small molecules (benzamidine) and the second of a set of protein binding sites (thermolisin). For comparison purpose, we also included a simple hill-climbing strategy, namely a $(1 + 1)$ -EA.

4.1 Speeding up the Evolutionary Algorithm

The main goal of this study was to reduce the runtime of the EA without sacrificing solution quality. To this end, we have experimented with the *sequential parameter optimization toolbox* (SPOT) [2] that enables a quasi-automatic adjustment of exogenous parameters. The standard parametrization by Bäck [3] gives a first idea of a useful range. We have chosen a range between 1 and 50 for the population size μ and a range between 1 and 20 for the selective pressure ν . For `initial stepsize`, we allowed values between 1 and $\frac{1}{2} \times n$.

Test problems were produced by generating a random graph first, and replicating this graph (with renumbered nodes) $m - 1$ times afterward. Obviously, the optimal solution of a problem of that kind is a perfect match of these m graphs; we modified the scoring system (by a linear transformation) such that the score of this solution is 0. SPOT generates a sequence of design points in the parameter space with the goal to minimize the number of required function evaluations for reaching a fitness of zero. The experiments were conducted for alignment problems of size $m \in \{2, 4, 8, 16\}$ which, however, all yielded similar results. According to these results, the parameter configuration $\mu = 4, \nu = 20, \text{selfadaption} = \text{off}, \text{recombination} = \text{on}$ and `initial stepsize` = 21 seems to be well-suited for the problem. As can be seen, a small value for the population size (only large enough to enable recombination) is enough, probably due to the fact that local optima do not cause a severe problem. On the other hand, as the search space is extremely large, a high selective pressure is necessary to create offsprings with improved fitness. The self-adaptation mechanism is disabled and, hence, the mutation rate is set to one (only two cells are swapped by mutation). This appears reasonable, as most swaps do not yield an improvement and instead may even produce a deterioration, especially during the final phase of the optimization. Thus, an improvement obtained by swapping two cells is likely to be annulled by a second swap in the same individual. Finally, our experiments suggest that a recombination is very

useful and should therefore be enabled.

4.2 Mining Molecular Fragment Data

As a first proof-of-concept for the algorithms presented in the previous section, we analyzed a data set consisting of 87 compounds that belong to a series of selective thrombin inhibitors and were taken from a 3D-QSAR study [7]. The data set is suitable for conducting experiments in a systematic way, as it is quite homogeneous and relatively small (the graph descriptors contain 47–100 nodes, where each node corresponds to an atom). Moreover, as the 87 compounds all share a common core fragment (which is distributed over two different regions with a variety of substituents), the data set contains a clear and unambiguous target pattern. From this data set, 100 subsets of 2, 4, 8 and 16 compounds have been selected at random, and for each subset, a MGA has been calculated using the greedy heuristic (Greedy), our evolutionary algorithm with optimized parametrization (EA), and in combination with a star-alignment procedure (EA*).

Before performing these experiments, we have compared the optimized EA with a simple hill-climbing strategy, namely a $(1 + 1)$ -EA. To ensure a fair comparison, each search strategy was allowed a fixed number of fitness function evaluations. The results indicate, that the $(1 + 1)$ -EA performs very poorly in comparison with EA.

The results of the main experiment, comparing the two EA-variants with the greedy strategy, are shown in Fig. 3. As a measure of comparison, we derived the relative improvement of the fitness value, defined as

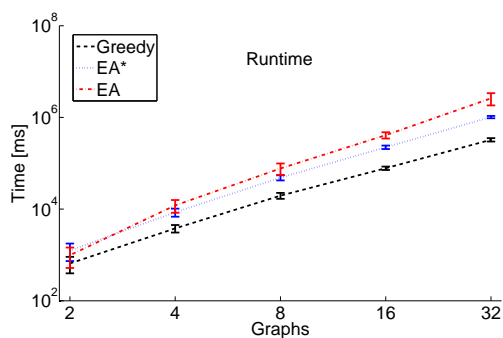
$$\frac{\text{fitness}(\text{EA}) - \text{fitness}(\text{Greedy})}{\min\{|\text{fitness}(\text{EA})|, |\text{fitness}(\text{Greedy})|\}} \quad (2)$$

We obtain a value of zero if both methods construct an alignment with the same score. The relative improvement is positive, if the EA solution yields a higher score than the Greedy solution (e.g. a relative improvement of one would mean an increase in score by a factor of two), else it is negative.

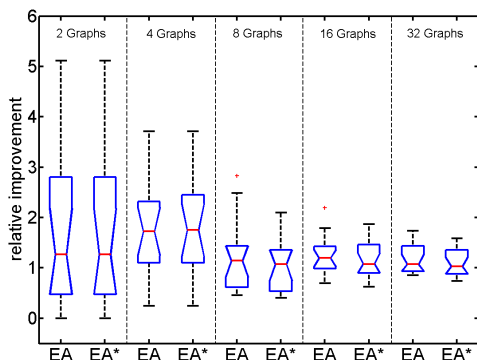
The results confirm our expectations: Both EA variants outperform the Greedy significantly with respect to alignment quality. In general, the full EA-variants also perform better than EA*.

The runtimes, shown in Fig. 3, confirm that EA produces better results. The smallest runtime among all alternatives is still produced by the greedy strategy, albeit at the cost of a lower quality. A good compromise between solution quality and efficiency is achieved by EA*, as the runtime is much

better than the EA runtime, especially for a higher number of graphs.



(a) Runtimes in milliseconds



(b) Relative improvements as defined in (2)

Figure 3: Mean and standard deviation of runtime and relative improvement as a function of the number of graphs.

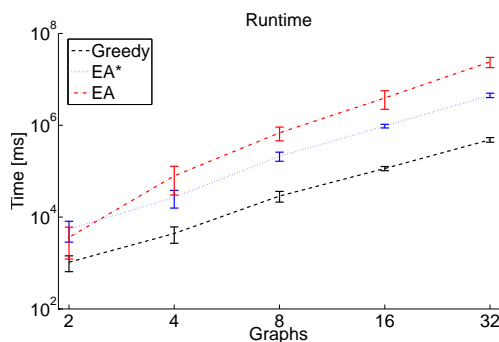
Regarding the aforementioned benzamidine core fragment, which consists of 25 atoms (11 hydrogens), it turned out that it was fully conserved, throughout all experiments, in all the solutions produced by our EA. In fact the EA was able to retrieve much more structure especially for closer related derivatives. For the greedy strategy, the corresponding degree of conservation was significantly smaller.

4.3 Mining Protein Binding Pockets

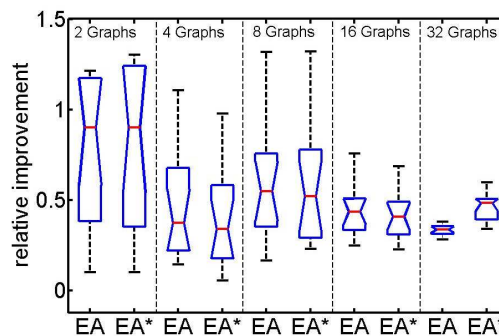
As our main interest concerns the characterization of protein binding pockets, we subsequently examined the performance of our algorithms on a second data set consisting of 74 structures derived from the Cavbase database. Each structure represents a protein binding pocket belonging to the protein family of thermolysine, bacterial proteases frequently used in structural protein analysis and annotated with the E.C. number 3.4.24.27 in the ENZYME classification database. The data set is suited for our purpose, as all binding pockets belong to the same

enzyme family and, therefore, should share evolutionary related, highly conserved substructures. On the other hand, with binding pockets ranging from about 30 to 90 pseudocenters, the data set is also diverse enough to present a real challenge for graph matching techniques.

Again, 100 graph alignments of size 2, 4, 8, 16 and 32 were produced, respectively, for randomly chosen structures, and the results are summarized in Fig. 4. The improvements achieved by the EA are not as high as in the previous experiment. This is possibly due to the high diversity of the protein cavities. For the benzamidine derivatives a common core fragment was present in all molecules. Given that the greedy solution is sub-optimal one can be sure that the EA can always improve the score by identifying more common structure. For the thermolysin cavities, it is not known a priori if a common structure exists, as not all cavities are binding pockets for the same type of ligand. If we compare cavities that are completely dissimilar, the EA solution might not yield a much higher score since there is simply not much to optimize. That being said, the EA solutions are still significantly better than the greedy solutions, as shown in Fig. 4.



(a) Runtimes in milliseconds



(b) Relative improvements as defined in (2)

Figure 4: Mean and standard deviation of runtime and relative improvement as a function of the number of graphs.

5 Conclusions

Multiple graph alignment (MGA) has recently been introduced as a novel method for analyzing biomolecules on a structural level. Using robust, noise-tolerant graph matching techniques, MGA is able to discover approximately conserved patterns in a set of graph-descriptors representing a family of evolutionary related biological structures. As the computation of optimal alignments is a computationally complex problem, this paper has proposed an evolutionary algorithm (EA) as an alternative to a hitherto existing greedy strategy.

Our experiments, carried out on several data sets with different characteristics, have shown the high potential of this approach and give rise to the following conclusions: The EA significantly outperforms the greedy strategy and is able to produce alignments of much higher quality, albeit at the cost of a considerable increase in runtime. In fact, as runtime seems to increase exponentially with the number of graphs to be aligned, we also considered a combination of evolutionary optimization with decomposition techniques, namely the star-alignment method. This approach appears to be a reasonable compromise as it achieves a good trade-off between solution quality and runtime.

As part of ongoing work, we currently elaborate on decomposition techniques in more detail. In particular, we investigate alternative decomposition schemes and aggregation procedures. An especially interesting idea in this regard is to replace a heuristic aggregation strategy again by an evolutionary optimization step. This line of research is complemented by more standard means to improve evolutionary optimization, such as alternative representations and more sophisticated genetic operators.

References

- [1] Atallah, M. J. (editor): *Algorithms and Theory of Computation Handbook*. CRC Press LLC. 1999.
- [2] Bartz-Beielstein, T., Lasarczyk, C. and Preuß, M.: Sequential Parameter Optimization Toolbox. *Technical Report, Universität Dortmund, Germany*. 2006.
- [3] Bäck, T.: Evolutionary Algorithms in Theory and Practise. *Dissertation, Universität Dortmund, Germany*. 1994.
- [4] Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C. and Eddy, S. R.: The Pfam protein families database. *Nucl. Acids. Res.*, 32(90001):D138–141. 2004.
- [5] Beyer, H.-G. and Schwefel, H.-P.: Evolution strategies – A comprehensive introduction. *Natural Computing*, 1(1):3–52. 2002.
- [6] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P.E.: The protein data bank. *Nucleic Acids Research*, 28:235–242. 2000.
- [7] Böhm, M., Stürzebecher, J. and Klebe, G.: Three-dimensional quantitative structure-activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor xa. *Journal of Medicinal Chemistry*, 42(3):458–477. 1999.
- [8] Hendlich, M., Bergner, A., Günther, J. and Klebe, G.: Relibase: Design and Development of a Database for Comprehensive Analysis of Protein-Ligand Interactions. *Journal of Molecular Biology*, 326:607–620. 2003.
- [9] Hendlich, M., Rippmann, F. and Barnickel, G.: LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal of Molecular Graphics and Modelling*, 15:359–363. 1997.
- [10] Schmitt, S., Kuhn, D. and Klebe, G.: A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology. *J. Mol. Biol.*, 323(2):387–406. 2002.
- [11] Servant, F., Bru, C., Carrère, S., Courcelle, E., Gouzy, J., Peyruc, D. and Kahn, D.: Prodom: Automated clustering of homologous domains. *Briefings in Bioinformatics*, 3(3):246–251. 2002.
- [12] Thompson, J. D., Higgins, D. G. and Gibson T. J.: Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673–4680. 1994.
- [13] Weskamp, N., Hüllermeier, E., Kuhn, D. and Klebe, G.: Graph Alignments: A New Concept to Detect Conserved Regions in Protein Active Sites. *German Conference on Bioinformatics 2004*, 131–140. 2004.