HUMBOLDT-UNIVERSITÄT ZU BERLIN
INSTITUT FÜR INFORMATIK

Exposé for Bachelor's Thesis

# Survey on the Graph Alignment Problem and a Benchmark of Suitable Algorithms

Christoph Döpmann

3 May 2013

Supervisors: Prof. Dr. Ulf Leser, André Koschmieder

# 1 Introduction

Graph alignment is a challenging but very important problem within the field of graph theory that has a wide range of applications.

In general, graph alignment is the problem of mapping two or more graphs to each other such that a given cost function is optimized [9]. That is, it aims at aligning graphs in a way that they become as "similar" as possible. There are several different definitions of what similarity between graphs might mean and thus, there are also different definitions of the graph alignment problem [9, 12, 8]. Graph alignment can therefore be regarded as a superset of problems from which each one incorporates its own subtle adaptation as required by the specific domain.

During the last years, graph alignment has become more and more important for systems biology [5]. It is used to compare biological networks, for example protein-protein interaction networks (PPI), metabolic networks or gene regulatory networks and is consequently mostly referred to as *network alignment*. Due to the enormous growth of such data over the last decade [7], there has been much research on how to gain biological insight from network comparisons, leading to a variety of new and enhanced algorithms for network alignment. These often aim at finding similar parts in networks of different species or at calculating an overall similarity score that can, for instance, be used to recover phylogenetic trees. Techniques for the alignment of biological networks are expected to prove at least as valuable for biological research as did sequence alignment [5].

Even though the definitions of graph alignment differ slightly, the main principle of mapping nodes of different graphs to each other is always the same. With that in mind, one general definition that holds for many applications and expresses the idea of graph alignment is the following.

Let $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$ be two simple graphs (undirected, unweighted and without loops and multiple edges). The *graph alignment problem* consists of finding an *alignment function* $f : V_1 \to V_2$ such that the *quality* of the alignment, denoted by $Q(G_1, G_2, f)$, is maximized. $Q$ expresses the similarity between the two graphs with respect to that specific alignment – the higher it is, the better the graphs could be aligned. We require $f$ to be a bijection between $V_1$ and $V_2$ in the general case, but sometimes allow graphs of slightly different sizes, in which case we assume $|V_1| \leq |V_2|$ and only require $f$ to be total and injective. By requiring an injective function, we ensure that the nodes of graph $G_1$ are all mapped to different nodes of $G_2$.

Clearly, the concrete problem strongly depends on the similarity function $Q$. The following types of measuring alignment quality are commonly used:

**topological similarity** In this case, the graphs are aligned solely based on their topology. Thus, a perfect alignment would imply that the two graphs are isomorphic to each other. One standard measurement for this is the so-called *edge-correctness (EC)*. It is defined as the ratio of edges in $G_1$ that are aligned to edges in $G_2$: [16]

$$EC(G_1, G_2, f) = \frac{|\{(f(v_1), f(v_2)) \mid (v_1, v_2) \in E_1\} \cap E_2|}{|E_1|}$$

However, the edge-correctness does not reflect, whether the correctly aligned edges are near each other and form a connected graph, which might be a desirable property. [14] Therefore, often the size of the largest common connected subgraph that is preserved under the alignment, is also used as an indicator of the alignment's quality. Formally, it may be defined as the number of nodes or edges in the largest connected subgraph of the following *alignment graph* $G_A(V_A, E_A)$:

$$V_A = V_2, \qquad E_A = \{(f(v_1), f(v_2)) \mid (v_1, v_2) \in E_1\} \cap E_2$$

**node similarity** In some applications, it does not make sense to align graphs only with respect to their topology, since the nodes themselves can be more or less similar to each other. So the aim could be to align similar nodes to each other. With a given node similarity function $s : V_1 \times V_2 \to \mathbb{R}$, one measurement of the alignment's quality might be the overall similarity between all nodes that are aligned to each other:

$$\sum_{v \in V_1} s(v, f(v))$$

Note, however, that this kind of topology-independent node similarity measure is not always meaningful. While, for example, when aligning protein-protein interaction networks, sequence similarity of proteins can be used, there is no general node similarity in arbitrary graphs.

**combination of both** As both of the aforementioned types of quality measure might fail to express what is meant by a good alignment of graphs, it is common practice to combine both of them [16]. One possibility to do so is as simple as adding them up using a user-defined weight $\alpha$, assuming two quality measures $q_1$ and $q_2$ that map to $[0, 1]$:

$$Q(G_1, G_2, f) = \alpha \cdot q_1(G_1, G_2, f) + (1 - \alpha) \cdot q_2(G_1, G_2, f)$$

This definition of an alignment's quality is rather versatile and is thus supported by many algorithms [8, 11, 12].

The specific quality measure used, only depends on the actual data modelled by the graph. For example, when aligning PPI networks, one might want to avoid proteins being aligned to each other that are completely different, as far as sequence or structural similarity is concerned. Here, it would make sense to use node similarity. Since it is, nevertheless, possible to draw surprisingly much benefit from only topological analysis [12], one might want to use a combination of both. In other cases, one might, for instance, want to align networks whose nodes represent different and non-comparable things. There, purely topological alignment would be appropriate.

Of course, these different definitions also require different algorithms. Aligning two graphs only based on their node similarity is the easiest of the mentioned problems. It can be regarded as an instance of the classical assignment problem and can, for example, be solved efficiently by the Hungarian algorithm in $O(n^3)$. In contrast, maximizing edge-correctness and the maximum common connected subgraph are both NP-hard problems [14] [8]. Consequently, the exact computation of an optimal alignment is infeasible if the graphs under consideration are large, which is the case for biological networks, which consist of thousands of nodes and edges. Heuristics are necessary for approximating the optimal solution.

## 2 Purpose of the Thesis

In my Bachelor's thesis, I am going to give an overview of existing graph alignment algorithms. Besides introducing the graph alignment problem itself, I am going to outline the main principles, these algorithms are built on. I am going to compare the different approaches and identify similarities and differences concerning their internal functioning as well as their applicability to different domains. Due to the fact that most progress on this topic was made concerning the comparison of biological networks, I will present algorithms that take into account the special characteristics of these networks, such as the fact that they are mostly very sparse – which is true for many graphs that model aspects of the real world [7].

Moreover, I am going to compare these algorithms on a quantitative basis in form of a benchmark. I am going to focus on runtime as well as on the topological quality of the alignments produced. Currently, I consider benchmarking some of the following algorithms: IsoRank [8], Græmlin 2.0 [10], GRAAL [12], MI-GRAAL [14], C-GRAAL [15], Natalie 2.0 [13], GHOST [16] and SPINAL [17]. I am going to test them on publicly available real PPI data as well as on artificially created random graphs.

I am also going to give a short overview on applications or extensions of the pure pairwise graph alignment problem. For instance, the alignment of multiple graphs at a time might prove useful and it would also offer new perspectives to systems biology research to be able to query a network databases for the network that produces the best alignment with a given query network, always bearing in mind that even pairwise alignment is a very hard problem, so algorithms must be scalable in order to be of practical relevance.

## 3 Related Work

The graph alignment problem as defined above is just one of several approaches to comparing graphs. A lot of research has been done in this field that will not directly be part of my thesis.

The definition of graph alignment presented above is widely regarded as the definition of *global network alignment* [9, 8]. The nodes of two graphs are globally mapped to each other in a one-to-one relation. In contrast, in *local network alignment*, the nodes do not need to be mapped to each other injectively, but one-to-many or many-to-many relations are allowed. This way, similar regions can be matched to each other independently of the rest of the graphs. The mapping becomes ambigous, but the local alignments might be of higher quality than global alignment would have achieved. This is because the global aligner might have found a different global alignment that reaches a higher overall quality score but does destroy the mapping of single similar regions. Therefore, local alignment is, for example, used to find conserved subnetworks across species instead of comparing the graphs as a whole [1]. In fact, local network alignment was used before the first global aligner IsoRank was developed [8]. Examples include Græmlin [2], MaWISh [3] and NetAlign [4].

As a special case of local alignment, one might regard the alignment with only one region that can be significantly smaller than the other network. This kind of search is often ambigously referred to as "network querying" [6], but is essentially the same as *approximate subgraph matching*.

All of the mentioned problems are closely related to the classical graph problems of graph isomorphism and subgraph isomorphism, but are more general because they also compute a solution if there is no exact match. If there is an exact match, however, they should preferably find it. As the classical problems are known to be NP-hard, it is generally also computationally demanding to find exact solutions for the mentioned related problems. Hence, they are usually approximated by heuristics, too.

## 4 References

[1] Roded Sharan et al. "Conserved patterns of protein interaction in multiple species". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.6 (2005), pp. 1974–1979.

[2] Jason Flannick et al. "Græmlin: general and robust alignment of multiple large interaction networks". In: *Genome research* 16.9 (2006), pp. 1169–1181.

[3] Mehmet Koyutürk et al. "Pairwise alignment of protein interaction networks". In: *Journal of Computational Biology* 13.2 (2006), pp. 182–199.

[4] Zhi Liang et al. "Comparison of protein interaction networks reveals species conservation and divergence". In: *BMC bioinformatics* 7.1 (2006), p. 457.

[5] Roded Sharan and Trey Ideker. "Modeling cellular machinery through biological network comparison". In: *Nature biotechnology* 24.4 (2006), pp. 427–433.

[6] Banu Dost et al. "QNet: a tool for querying protein interaction networks". In: *Research in Computational Molecular Biology*. Springer. 2007, pp. 1–15.

[7] Nataša Pržulj. "Biological network comparison using graphlet degree distribution". In: *Bioinformatics* 23.2 (2007), e177–e183.

[8] Rohit Singh, Jinbo Xu, and Bonnie Berger. "Pairwise global alignment of protein interaction networks by matching neighborhood topology". In: *Research in computational molecular biology*. Springer. 2007, pp. 16–31.

[9] Jason Flannick. "Algorithms for biological network alignment". PhD thesis. Stanford University, 2008.

[10] Jason Flannick et al. “Automatic parameter learning for multiple network alignment”. In: *Research in Computational Molecular Biology*. Springer. 2008, pp. 214–231.

[11] Mikhail Zaslavskiy, Francis Bach, and Jean-Philippe Vert. “Global alignment of protein–protein interaction networks by graph matching methods”. In: *Bioinformatics* 25.12 (2009), pp. i259–1267.

[12] Oleksii Kuchaiev et al. “Topological network alignment uncovers biological function and phylogeny”. In: *Journal of the Royal Society Interface* 7.50 (2010), pp. 1341–1354.

[13] Mohammed El-Kebir, Jaap Heringa, and Gunnar W Klau. “Lagrangian relaxation applied to sparse global network alignment”. In: *Pattern Recognition in Bioinformatics*. Springer, 2011, pp. 225–236.

[14] Oleksii Kuchaiev and Nataša Pržulj. “Integrative network alignment reveals large regions of global network similarity in yeast and human”. In: *Bioinformatics* 27.10 (2011), pp. 1390–1396.

[15] Vesna Memišević and Nataša Pržulj. “C-GRAAL: Common-neighbors-based global GRAph ALignment of biological networks”. In: *Integrative Biology* 4.7 (2012), pp. 734–743.

[16] Rob Patro and Carl Kingsford. “Global network alignment using multiscale spectral signatures”. In: *Bioinformatics* 28.23 (2012), pp. 3105–3114.

[17] Ahmet E Aladağ and Cesim Erten. “SPINAL: scalable protein interaction network alignment”. In: *Bioinformatics* (2013).