NABEECO: Biological Network Alignment with Bee Colony Optimization Algorithm

Rashid Ibragimov
Max-Planck-Institut für Informatik
Campus E2 1, 66123
Saarbrücken, Germany
ribragim@mpi-inf.mpg.de

Jan Martens
Max-Planck-Institut für Informatik
Campus E2 1, 66123
Saarbrücken, Germany
jmartens@mmci.uni-saarland.de

Jiong Guo
Universität des Saarlandes
Campus E1 7, Saarbrücken
66123, Germany
jguo@mmci.uni-saarland.de

Jan Baumbach
University of Southern Denmark
Campusvej 55, DK-5230
Odense M, Denmark
jan.baumbach@imada.sdu.dk

ABSTRACT

Motivation: A growing number of biological networks of ever increasing sizes are becoming available nowadays, making the ability to solve Network Alignment of primer importance. However, computationally the problem is hard for data sets of real-world sizes.

Results: we developed NABEECO, a novel and robust Network Alignment heuristic based on Bee Colony Optimization. We use the so-called Graph Edit Distance (GED) as optimization criterion, which is defined as the minimal amount of edge and node modifications necessary to transform one graph into another. We compare NABEECO on a set of protein-protein interaction networks to the current state of the art tool for biological networks, MI-GRAAL.

Conclusion: We present the first Bee Colony Optimization algorithm for biological Network Alignment. NABEECO, in contrast to many other tools, can be applied to all kinds of networks and allows incorporating prior knowledge about node/edge similarity, though this is not required a priori. NABEECO together with a more detailed description and all data sets used are publicly available at `http://nabeeco.mpi-inf.mpg.de`.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Biology and genetics

Keywords

Bee Colony Optimization; Network Alignment; Graph Edit Distance; Protein-Protein Interaction Networks

1. INTRODUCTION

Many methods for aligning gene and protein sequences provide the basis for nowadays functional annotations. The alignment of networks is more challenging, but able to complement the tools, opening horizons for further developments in knowledge transfer between species [6]. More specifically, given two graphs, Network Alignment (NA) seeks for a mapping in which every node from the first graph is mapped to at most one node in the the second graph and vice versa, and which optimizes a quality creterion. This criterion may include topological properties of the graph as well as problem specific models. While topological quality is directly computed from a mapping, problem-specific biological information can be incorporated as (exclusive or fuzzy) node pre-matching, making NA computationally easier due to the restricted search space. However, the information increases ambiguity of inferences, since for many biological networks (and mainly the nodes therein) functional information is often incomplete and noisy. Here, we concentrate on protein-protein interaction (PPI) networks, where evolutionary preserved topology infers preservation of biological function [5]. Thus, we focus on graph topological aspects, rather than network specific biological information, such as BLASTscores, though these non-topological measures might be useful for pre-mapping and a network aligner should be able to incorporate the information, but not rely on them.

To tackle NA a set of techniques exploiting various approaches has been developed (see for example [3, 13, 1, 8, 11]). With highest values of Edge Correctnesses (taken from [9]) achieved by IsoRank [13], GRAAL [8], H-GRAAL [11], MI-GRAAL [9], and C-GRAAL [10] aligning *yeast2* vs. *human1* are 3.89, 11.72, 10.92, *23.26*, 22.55 respectively, and for *Meso*, *Syne* networks 5.33, 11.25, 4.59, *41.79*, 26.02, respectively, MI-GRAAL significantly outperforms the existing, recent tool.

In the following, we present NABEECO, a novel tool for PPI network alignment, which is based on bee colony optimization, minimizing the so-called Graph Edit Distance [2] (GED) as optimization criterion. We describe the strategy behind NABEECO and compare it with MI-GRAAL on a set of real PPI networks. An implementation of NABEECO, as well as its more detailed description and all data sets used are publicly available at `http://nabeeco.mpi-inf.mpg.de`.

GECCO'13 Companion, July 6–10, 2013, Amsterdam, The Netherlands.
ACM 978-1-4503-1964-5/13/07.

2. METHODS

Bee Colony Optimization (BCO) is a population based nature-inspired metaheuristic which mimics the behavior of a real honey bee colony in a hive, has been applied for solving hard optimization problems [7]. A food source in the abstraction points to a solution, i.e. a mapping between two graphs in case of NA. The quality of the solution that we aim to optimize is the Graph Edit Distance (GED) defined as the number of node/edge insertions/deletions induced by the given mapping between two graphs [2, 4]. Given a mapping, for a node and its image, NABEECO combines the precomputed graphlet degree signature distance [9] (local measure) with the GED (global measure) into a 'pair score', which is the sum of the relative number of edge deletions/insertions and GDSD between the two nodes. Intuitively, this 'pair score' reflects the contribution of the nodes to the total quality of the GED of the mapping: the lower the pair score, the better the node and its image fit.

The 'gathering' step performed by worker bees is the most crucial step in BCO, and has to guarantee a certain degree of diversity of candidates, at the same time, being restrictive enough. Exploring the local neighborhood of a solution NABEECO uses multiple techniques relying on the pair scores: *greedy swaps*, *local permutation*, *random swaps*, *random greedy swaps* (see NABEECO's webpage for details).

3. RESULTS AND DISCUSSIONS

We compare our approach with a state of the art tool for biological NA, MI-GRAAL [9]. MI-GRAAL's quality was measured by using the so-called *Edge Correctness* (EC), which can be understood as the proportion of aligned edges given a mapping between two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$. The original definition of EC used in [9] assumes that $|V_1| \leq |V_2|$ and $|E_1| \leq |E_2|$, which does not cover all pairs of networks. We generalize the EC to $\frac{\#alignedEdges}{\min(|E_1|,|E_2|)}$. With the costs used in the paper for edit operations (edge insertions and deletions are of cost 1, and the other operations are of cost 0), EC and GED are inversely proportional.

We assess the solution quality by applying NABEECO and MI-GRAAL to a set of real PPI networks from [12] (*hprd*), [14] (*ulitsky*), [15] (*HS, SC, DM*) and [9] (*cjejuni*, *Meso*, *Syne*, *ecoli_fi*, *yeast2*, *human1*). The highest EC value for the pairs of networks are summarized in Table 1. Quality-wise both, MI-GRAAL and NABEECO, perform almost equally good. Comparing larger networks, however, NABEECO converges to results of higher EC. Note that MI-GRAAL failed aligning the bigger networks.

Table 1: The highest values of EC (%) aligning pairs of PPI networks with NABEECO and MI-GRAAL.

Network 1	Network 2	NABEECO	MI-GRAAL
ecoli_fi	*cjejuni*	**28.24**	24.60
Meso	*Syne*	32.25	**39.88**
yeast2	*human1*	**36.78**	21.38
HS	*SC*	**28.26**	26.15
SC	*DM*	14.14	**17.73**
DM	*human1*	**19.09**	-
ulitsky	*hprd*	**23.51**	-
human1	*hprd*	**43.57**	-

4. CONCLUSION

We have presented NABEECO, a novel method for solving the Network Alignment problem. NABEECO is the first method exploiting artificial bee colony to optimize the graph edit distance. Our NABEECO software comes with flexibility as a major advantage: in contrast to many other methods, NABEECO can work on topology only, but it can incorporate non-topological node as pre-mappings to speed-up convergence towards an increased accuracy. It can further be executed on any kind of graph, directed graphs, for instance, thereby generally allowing to compare, for example, gene regulatory networks. We will further use NABEECO in follow-up studies and compare different types of networks, predicted as well as known networks.

5. REFERENCES

[1] G. Blin, et al. Querying Protein-Protein Interaction Networks. ISBRA '09, 52–62, Berlin, Heidelberg, 2009. Springer.

[2] H. Bunke, K. Riesen. *Graph Edit Distance – Optimal and Suboptimal Algorithms with Applications*, 113–143. Wiley-VCH Verlag GmbH & Co. KGaA, 2009.

[3] D. Conte, et al. Thirty Years Of Graph Matching In Pattern Recognition. *Int. J. Pattern Recog. and Artif. Intelligence*, 2004.

[4] X. Gao, et al. A survey of graph edit distance. *Pattern Analysis & Applications*, 13(1):113–129, 2010.

[5] L. Hakes, et al. Protein-protein interaction networks and biology - what's the connection? *Nat. Biotechnol.*, 26(1):69–72, 2008.

[6] A. P. Heath, L. E. Kavraki. Computational challenges in systems biology. *Comp. Sci. Rev.*, 3(1):1 – 17, 2009.

[7] D. Karaboga and B. Akay. A comparative study of artificial bee colony algorithm. *Appl. Math. and Comp.*, 214(1):108–132, 2009.

[8] O. Kuchaiev, T. Milenković, V. Memišević, W. Hayes, and N. Pržulj. Topological network alignment uncovers biological function and phylogeny. *J. R. Soc. Inter.*, 7(50):1341–1354, 2010.

[9] O. Kuchaiev and N. Pržulj. Integrative Network Alignment Reveals Large Regions of Global Network Similarity in Yeast and Human. *Bioinformatics*, 27(10):1390–1396, 2011.

[10] V. Memišević and N. Pržulj. C-GRAAL: Common-neighbors-based global GRAph ALignment of biological networks. *Integr. Biol.*, 4:734–743, 2012.

[11] T. Milenković, W. Ng, W. Hayes, and N. Pržulj. Optimal network alignment with graphlet degree vectors. *Cancer inf.*, 9:121, 2010.

[12] K. Prasad, et al. Human Protein Reference Database – 2009 update. *Nucl. acids res.*, 37, D767–72, 2009.

[13] R. Singh, et al. Pairwise global alignment of protein interaction networks by matching neighborhood topology. RECOMB'07, 16 – 31, Berlin, Heidelberg, 2007. Springer.

[14] I. Ulitsky, et al. Detecting Disease-Specific Dysregulated Pathways Via Analysis of Clinical Expression Profiles Research in Computational Molecular Biology. LNCS 4955, 347–359. Springer Berlin, Heidelberg, 2008.

[15] I. Xenarios et al. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucl. acids res.*, 30(1):303–305, 2002.