Developing Semantic Pathway Alignment Algorithms for Systems Biology

Jonas Gamalielsson

Systems Biology Group, Skövde University, Box 407, Skövde, 54128, Sweden,
jonas.gamalielsson@his.se,
WWW: http://www.ida.his.se/∼gam

1 Introduction

Systems biology is an emerging multi-disciplinary field in which the behaviour of complex biological systems is studied by considering the interaction of all cellular and molecular constituents rather than using a "traditional" reductionist approach [1, 2]. Systems are often studied over time [3], and the ultimate goal could be to develop models which predict various human diseases [4]. System biology also seeks to integrate information at different levels of organisation in order to understand the biology of cells (en.wikipedia.org). This multi-level structure can for example be illustrated by the pyramid of life [5], where there are four levels of organisation housed within a pyramid structure. The broad bottom level contains organism specific entities such as genes, mRNA, proteins and metabolites. The next level is where the bottom level entities are organised into regulatory motifs and metabolic pathways. At the third level, motifs and pathways are integrated into higher order networks where tightly connected proteins and metabolites form functional modules that perform certain cellular functions. Functional modules are organised hierarchically at the top level in the pyramid in order to describe the large-scale organisation of cells. Powerful computational tools are needed at each level in this pyramid and also for actually integrating information between levels. In this document we propose methods for semantic alignment of different kinds of biological pathways at the second level in the pyramid of life. We use the Gene Ontology (GO), documented pathways, hypothetical pathways or sets of gene products, GO annotation databases, various algorithms from computer science, and statistics. In a different track of research, not covered in this document, we have also developed different methods combining semantic and topologic analysis of protein interaction networks with the aim to identify functional modules [6–9], where GO was successfully used in the analysis of documented Y2H (yeast-two-hybrid) networks for *S. cerevisiae*. This shows that generalising about gene products using GO is also beneficial in tools for systems biology at higher levels in the pyramid of life.

Three methods are proposed as the core of the PhD dissertation and these are described in the following.

2 Method 1: GO-based regulatory template approach

2.1 Background

It is highly desirable to be able to derive causal gene regulatory networks using gene expression data. This is known as reverse engineering of genetic networks[10]. Time series expression data is often used, and the task of the reverse engineering algorithm is to find a set of activation rules that fits these data. Methods for reverse engineering of genetic networks that have been proposed include techniques such as boolean [11, 12], neural [13, 14] and Bayesian networks [15–17]. In most cases, however, many different reverse engineered networks are consistent with the observed data, but we can expect only a few of these networks to be biologically plausible. A drawback of reverse engineering methods which are based solely on fit to the data is that they do not provide any way of distinguishing between biologically plausible and implausible networks.

2.2 Method description

To be able to distinguish between plausible and implasible networks, we here propose a method for assessing the biological plausibility of regulatory hypotheses using prior biological knowledge in the form of Gene Ontology (GO) annotation of gene products and examples of regulatory interactions from pathway databases. Using GO, we derive templates encoding general knowledge by generalising from examples of known interactions to typical properties of interacting gene product pairs. By matching regulatory hypotheses to such templates, their plausibility can be assessed. The assumption is that if the properties of the gene products in a regulatory hypothesis are similar to those of gene products which are already known to interact, this increases the plausibility of the hypothesis.

Gene Ontology [18] is a structured vocabulary of molecular biology. It contains three different sub-ontologies covering the molecular functions, biological processes and cellular components of gene products, and is structured as a directed acyclic graph showing how terms are related to each other, using inheritance (IS-A) and aggregation (PART-OF). Using these relations, abstraction hierarchies of terms with different specificity are created. A gene product can be associated with several terms in each sub-ontology. One important use of an ontology is the calculation of semantic similarity between two terms. Different information theoretic measures have been proposed and applied for this purpose [19–21] and ideas from these measures are used in our method.

The concept of templates is appealing for identifying hypothetical regulatory relations that are similar to known regulatory relations. Templates have previously been used in the context of association rule discovery and expression data in [22], where a template-language was designed that allows users to group, filter and inspect a large number of rules produced by association rule discovery algorithms. Of particular interest for our work are rule templates $F1 \rightarrow F2$ which detect rules where the antecedent part contains any of the genes in a predefined functional group $F1$ and the consequent part contains a gene from another predefined group $F2$.

The basic idea behind the method proposed here is to use what we know about regulation in documented pathways, generalise this knowledge, and apply it for assessment of the plausibility of regulatory hypotheses. The GO molecular function classification of the gene products participating in regulatory relations in pathways is used to derive templates. The templates encode the types of gene products known to be involved in a particular type of regulatory relation. By searching for matches in the set of templates, the plausibility of regulatory hypotheses can be assessed. We evaluate to what degree the collection of templates can separate true from false positive interactions, and we illustrate the practical use of the method by applying it to example network reconstruction problems.

The method for assessing the biological plausibility of regulatory hypotheses is described thoroughly in [23, 24].

3 Method 2: GOSAP; GO-based semantic alignment of biological pathways

3.1 Background

A large number of biological pathways are being derived for many different organisms such as *S. cerevisiae* and *E. coli*, and these are stored in various databases such as KEGG[25] and EcoCyc[26].

There is a lack of and need for algorithms capable of searching for homologues to pathway queries in a collection of known pathways [27]. These algorithms should also return alignments between matching pathway fragments. Furthermore, these pathway alignment methods should rely on approximate, rather than exact, matching in biological pathways [27, 28].

Previous work on comparative analysis of metabolic pathways has been addressed [29], where a combined approach was used which involves analysis and comparison of biochemical data, pathway analysis using the elementary modes concept, and comparative analysis of a set of completely sequenced genomes where the EC hierarchy was used. A method for detection of functionally related enzyme clusters has been proposed [30], where topological properties of metabolic pathways are considered. Another paper describes a method for topological motif search in biological pathways [31]. An approach for detecting frequent subgraphs in biological pathways has been reported [28], however this method does not directly address alignments. Furthermore, work on sequence similarity based alignments between protein interaction networks has been reported [32]. However, none of these papers address the concept of approximate matching and generalisation using an abstraction hierarchy or ontology.

A method for deriving multiple alignments of paths in metabolic pathways has been proposed [33], where the EC hierarchy is used for generalising about enzymes. A method for alignment of metabolic pathways using a technique known as approximate labeled sub-tree homeomorphism, has recently been proposed [27], where the EC hierarchy once again is used for generalisation.

3.2 Method description

Here, we propose a Gene Ontology (GO) [18] based local alignment method for comparing biological pathways. To our knowledge, GO has not been used for deriving semantic alignments of paths in biological pathways earlier. GO enables the analysis of pathways where nodes are not only enzymes, but any kind of gene product. Another novelty is the use of combined alignment scores involving all three sub-ontologies of GO. Our proposed method is applicable to all types of biological pathways, where nodes are gene products, e.g. regulatory pathways, signalling pathways and metabolic enzyme-to-enzyme pathways. It would also be possible to extend the method to work with other types of nodes, as long as there is an ontology or abstraction hierarchy available for categorising the nodes.

In our method, one known pathway graph is used as a model graph, and another pathway graph serves as query graph. Simple paths are extracted from both model- and query graphs, and every path from the set of query paths are aligned against each of the model paths using a modified Smith-Waterman [34] local alignment algorithm. Special cost matrices featuring different measures of semantic similarity between GO terms are used. The specificities of different GO terms used for semantic similarity comparisons are derived using GO annotation databases for the organisms under study. In order to detect statistically significant path alignments, a p-test is performed using an ensemble of randomised model graphs. Alignments below a certain p-value threshold are regarded significant. It is demonstrated that the method is useful for studying protein regulatory pathways in *S. cerevisiae*, as well as metabolic pathways for the same organism.

The method for GO-based semantic alignment of biological pathways is described in detail in [35].

4 Method 3: Deriving hypothetical pathways by an evolutionary approach

4.1 Background

Method 2 (GOSAP) assumes that the topologies of both model- and query graphs are known. However, sometimes only a query set of gene products is available, and there is no knowledge available about how the gene products interact. This is for example the case when microarray experiments are performed and a list of genes is derived where genes are differentially expressed between two conditions. Such analysis is particularly common when time series data is not available. Experiments could be performed using animal models where there is a number of wildtype phenotypes and a number of treated phenotypes used for statistical analyses. If time series data is available, it is common to perform cluster analysis. Co-expressed genes usually end up in the same cluster, and it is of interest to find out how genes interact within such a cluster.

There are existing tools available that are capable of mapping groups of gene products onto known pathways, but this is only done using the identity of gene products and for visualisation purposes only. An example is GenMAPP

[36], where the genes and their colour-coded expression values are mapped onto known pathways. There is also the GenMAPP accessory software MAPPFinder [37] where GO visualisation has been added. Pathway tools [38] is another example where functionality is available for including gene expression data in pathway diagrams in a manner similar to GenMAPP. ArrayXPath [39] is a similar tool where gene expression clusters can be mapped onto the best matching pathways in a database. Additionally, there are numerous algorithms for reverse engineering regulatory networks using gene expression data (see e.g. section 2), but this is a different problem as no model networks are used to compare derived networks with.

Hence, no generalisation with respect to gene products is performed in earlier efforts. As stated in section 3, it is important to be able to reason about pathways and gene products using similarities rather than identity. Additionally, related methods do not attempt to create hypothetical paths using a query set of gene products.

4.2 Method description

In our method, hypothetical paths are evolved by an evolutionary algorithm using a query set of gene products and a known pathway graph as model. All paths are extracted from the model and for each path a semantically similar path is assembled using the query set of gene products. Measures of semantic similarity are used calculate a match score between each pair of gene products in the model path and in the evolved query path. For example, if the model path contains five gene products, there will be five semantic similarity comparisons, each adding to a total score describing the similarity of the two paths. All three GO subontologies can be used in the calculations. The evolutionary algorithm tested so far is a steady state (elitism) algorithm using a fixed number of individuals and generations. Operators such as tournament selection, partially mapped crossover and mutation, are used. A test for statistical significance of hypothetical paths is also performed using randomised versions of the model graph. It is also possible to include path alignment features (gap insertion) in the evolutionary algorithm itself or by aligning the hypothetical paths using GOSAP. Regarding representation of solutions, one possibility is that paths are represented as permutations of gene products from the query set, i.e. a gene product can only appear once. This is, for example, the case regarding the transcriptional regulator chain motifs for *S. cerevisiae* in [41]. It is also possible to allow each gene product to appear at several positions in a path, but it would require different operators in the evolutionary algorithm. Pathways with multiple instances of the same gene product can be found in e.g. KEGG (www.genome.jp/kegg) and MetaCyc (metacyc.org). We demonstrate that the method is able to derive hypothetical pathways that are significantly similar to documented pathways of regulation in several different biological scenarios. Biological experiments are required to validate any derived hypothetical pathway.

The method for deriving hypothetical pathways by an evolutionary approach is described more thoroughly in [40].

5 Conclusion and contribution

A thesis is proposed where three related methods are developed for semantic analysis of biological pathways. The methods have several traits in common; Biological pathways are being analysed. Gene Ontology, gene product annotation databases, and information theory are used in all methods. Furthermore, the concept of generalisation using GO categories of gene products is used throughout. Methods also aim to derive, in some sense, biologically plausible results. There are also some major differences between the methods; Method 1 uses binary interactions between gene products in pathways, whereas methods 2 and 3 use paths of gene products. Additionally, Method 1 derives structures of knowledge (templates) from a model pathway prior to comparison with a query pathway, whereas methods 2 and 3 derive structures of knowledge (alignments) during the comparison itself. Furthermore, Methods 2 and 3 use statistical tests to assess the significance of alignments whereas method 1 uses a more arbitrary score threshold approach to template match assessment. Finally, Methods 1 and 2 assume a query graph, whereas method 3 only assumes a set of query gene products.

It is believed that the methods will be useful to biologists in order to assess the biological plausibility of derived pathways, compare different pathways for semantic similarities, and to obtain hypothetical pathways using a query set of gene products and documented pathways as input. To our knowledge, all methods are novel, and will therefore extend the bioinformatics toolbox that biologists can use to make new biological discoveries.

References

1. Kirschner, M. W.: The Meaning of Systems Biology. Cell **121** (2005) 503–504
2. Liu, E. T.: Systems Biology, Integrative Biology, Predictive Biology. Cell **121** (2005) 505–506
3. Aderem, A.: Systems Biology: Its Practice and Challenges. Cell **121** (2005) 511–513
4. Butcher, E. C., Berg, E. L., Kunkel, E. J.: Systems biology in drug discovery. Nature biotechnology **22** (2004) 1253–1259
5. Oltvai, Z. N., Barabasi, A-L.: Life's Complexity Pyramid. Science **298** (2002) 763–764
6. Lubovac, Z., Gamalielsson, J., Olsson, B., Lindlf, A.: Exploring protein networks with a semantic similarity measure. In Proceedings of the 6:th International Symposium on Computational Biology and Genome Informatics, USA, (July 2005)
7. Lubovac, Z., Olsson, B. and Gamalielsson, J.: Combining topological properties and domain knowledge reveals functional modules in protein interaction networks. In Proceedings of the 2:nd International Conference on Algorithms and Computational Methods for Biochemical and Evolutionary Networks (CompBioNets), Lyon, France (December 2005)
8. Lubovac, Z., Gamalielsson, J., Olsson, B.: Combining functional and topological properties to identify core modules in protein interaction networks, PROTEINS: Structure, Function and Bioinformatics **XXX** (2006) YYY–ZZZ

9. Lubovac Z, Olsson B, and Gamalielsson J.: Weighted Clustering Coefficient for Identifying Modular Formations in Protein-Protein Interaction Networks (to appear). In proceedings of the Third International Conference on Bioinformatics and Computational and Systems Biology, Prague (August 25-27 2006)
10. D'haeseleer, P., Liang, S., Somogyi, R.: Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics **16** (2000) 707–726
11. Liang, S., Fuhrman, S., Somogyi, R.: REVEAL - a general reverse engineering algorithm for inference of genetic network architectures. Pac Symp Biocomput **3** (1998) 18–29
12. Akutsu, T., Miyano, S., Kuhara, S.: Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. Pac Symp Biocomput **4** (1999) 17–28
13. D'haeseleer, P., Wen, X., Fuhrman, S., Somogyi, R.: Linear modeling of mRNA expression levels during CNS development and injury. Pac Symp Biocomput **4** (1999) 41–52
14. Weaver, D. C., Workman, C. T., Stormo, G. D.: Modeling regulatory networks with weight matrices. Pac Symp Biocomput **4** (1999) 112–123
15. Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using Bayesian networks to analyze expression data. RECOMB '00: Proceedings of the fourth annual international conference on computational molecular biology (2000) 127–135
16. Kim, S. Y., Imoto, S., Miyano, S.: Inferring gene networks from time series microarray data using dynamic Bayesian networks. Briefings in Bioinformatics **4** (2003) 228–235
17. Husmeier, D.: Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. Bioinformatics **19** (2003) 2271–2282
18. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G.: Gene Ontology: tool for the unification of biology. Nature Genetics **25** (2000) 25–29
19. Lin, D.: An information-theoretic definition of similarity. Proceedings of the 15th International Conference on Machine Learning (1998) 296–304
20. Jiang, J. J., Conrath, D. W.: Semantic similarity based on corpus statistics and lexical taxonomy. Proceedings of the 10th International Conference on Research on Computational Linguistics, ROCLING X (1997) 19–33
21. Resnik, P.: Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research **11** (1999) 95–130
22. Tuzhilin, A., Adomavicius, G.: Handling very large numbers of association rules in the analysis of microarray data. Proceedings of the Eighth ACM SIGKDD International Conference on Data Mining and Knowledge Discovery (2002) 396–404
23. Gamalielsson, J., Olsson, B., Nilsson, P.: A Gene Ontology based Method for Assessing the Biological Plausibility of Regulatory Hypotheses. Technical report, HS-IKI-TR-05-004, University of Skövde, Sweden (2005)
24. Gamalielsson, J., Nilsson, P., Olsson, B.: A GO-based Method for Assessing the Biological Plausibility of Regulatory Hypotheses. In proceedings of the 2nd International Workshop on Bioinformatics Research and Applications (IWBRA 2006), Reading, Great Britain (May 2006)
25. Kanehisa, M., Goto, S.: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research **28** (2000) 27–30

26. Karp, P., Arnaud, M., Collado-Vides, J., Ingraham, J., Paulsen, I. T., Saier, M. H. Jr.: The E. coli EcoCyc Database: No Longer Just a Metabolic Pathway Database. ASM News **70** (2004) 25–30
27. Pinter, R. Y., Rokhlenko, O., Yeger-Lotem, E., Ziv-Ukelson, M.: Alignment of Metabolic Pathways. Bioinformatics **21** (2005) 3401–3408
28. Koyutürk, M., Grama, A., Szpankowski, W.: An efficient algorithm for detecting frequent subgraphs in biological networks. Bioinformatics **20** (2004) i200–i207
29. Dandekar, T., Schuster, A., Snel, B., Huynen, M., Bork, P.: Pathway alignment: application to the comparative analysis of glycolytic enzymes. Biochemistry Journal **343** (1999) 115–124
30. Ogata, H., Fujibuchi, W., Goto, S., Kanehisa, M.: A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. Nucleic Acids Research **28** (2000) 4021–4028
31. Berg, J., Lssig, M.: Local graph alignment and motif search in biological networks. PNAS **101** (2004) 14689–14694
32. Kelley, B. P., Sharan, R., Karp, R. M., Sittler, T., Root, D. E., Stockwell, B. R.: Conserved pathways within bacteria and yeast as revealed by global protein network alignment. PNAS **100** (2003) 11394–11399
33. Tohsato, Y., Matsuda, H., Hashimoto, A.: A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. Proc. 8th International Conference on Intelligent Systems for Molecular Biology (ISMB 2000) 376–383
34. Smith, T. F., Waterman, M. S.: Identification of Common Molecular Subsequences. Journal of Molecular Biology **147** (1981) 195–197
35. Gamalielsson, J., Olsson, B.: GOSAP: Gene Ontology Based Semantic Alignment of Biological Pathways. Technical report, HS-IKI-TR-05-005, University of Skövde, Sweden (2005)
36. Dahlquist, K. D., Salomonis, N., Vranizan, K., Lawlor, S. C., Conklin, B. R.: GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nature Genetics **31** (2002) 19–20
37. Doniger, S. W., Salomonis, N.,Dahlquist, K. D., Vranizan, K., Lawlor, S. C., Conklin, B. R.: MAPPFinder: using Gene Ontologu and GenMAPP to create a global gene-expression profile from microarray data. Genome Biology **4** (2003) R7
38. Karp, P. D., Paley, S., Romero, P.: The pathway tools software. Bioinformatics **18** (2002) S1–S8
39. Chung, H-J., Kim, M., Park C. H., Kim, J., Kim, J-H.: ArrayXPath: mapping and visualizing microarray gene-expression data with integrated biological pathway resources using Scalable Vector Graphics. Nucleic Acids Research **32** (2004) W460–W464
40. Gamalielsson, J., Olsson, B.: EGOSAP: Evolutionary Gene Ontology Based Semantic Alignment of Biological Pathways. *Manuscript in preparation* (2006)
41. Lee, T. I., Rinaldi, K., med flera: Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*. Science **298** (2002) 799–804