GOSAP: Gene Ontology Based Semantic Alignment of Biological Pathways

Jonas Gamalielsson and Björn Olsson

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

Abstract. A large number of biological pathways have been assembled in later years, and are being stored in databases. Hence, the need for methods to analyse these pathways has emerged. One class of methods compares pathways, in order to discover parts that are evolutionary conserved between species or to discover intra-species similarites. Most previous work has been focused on methods targeted at metabolic pathways utilising the EC enzyme hierarchy. Here, we propose a Gene Ontology (GO) based approach for finding semantic local alignments when comparing paths in biological pathways where the nodes are gene products. The method takes advantage of all three sub-ontologies, and uses a measure of semantic similarity to calculate a match score between gene products. 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. We demonstrate that the method is useful for studying protein regulatory pathways in *S. cerevisiae*, as well as metabolic pathways for the same organism.

1 Introduction

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[1] and EcoCyc[2]

There is a lack of and need for algorithms capable of searching for homologues to pathway queries in a collection of known pathways [3]. 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 [3, 4].

Previous work on comparative analysis of metabolic pathways has been addressed [5], 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 [6], where topological properties of metabolic pathways are considered. Another paper describes a method for topological motif search in biological pathways [7]. An approach for detecting frequent subgraphs in biological pathways has been reported [4], however this method does not directly address alignments. Furthermore, work on sequence similarity based alignments between protein interaction networks has been reported [8]. However, none of these papers address the concept of approximate matching and generalization using an abstraction hierarchy or ontology.

A method for deriving multiple alignments of paths in metabolic pathways has been proposed [9], where the EC hierarchy is used for generalizing about enzymes. A method for alignment of metabolic pathways using a technique known as approximate labeled sub-tree homeomorphism, has recently been proposed [3], where the EC hierarchy once again is used for generalization.

Here, we propose a Gene Ontology (GO) [10] 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.

2 Method

The method is summarized in figure 1, and involves the three procedures 1) GO term probability calculation, 2) path extraction and 3) path alignment. The first procedure calculates the probability of the GO terms using an annotation database for one or several organisms. This knowledge is used in the alignment procedure to calculate how semantically similar two gene products are. The second procedure systematically extracts paths (sequences) of gene products from the model- and query pathway graphs. This procedure enables the alignment algorithm to handle graph structures. The path alignment procedure is where each of the query paths is aligned with each of the model paths. This is done using a modified version of the Smith-Waterman [12] local alignment algorithm, tailored for the alphabet of gene product identifiers and using a GO semantic similarity match function. Furthermore, a test for statistical significance of alignments is performed. More details regarding the three procedures are given in the following.

GO term probability calculation

An annotation database D is used to calculate the probability of each GO term using the method proposed by Lord et al. [11].

- For each gene product G_i in D:

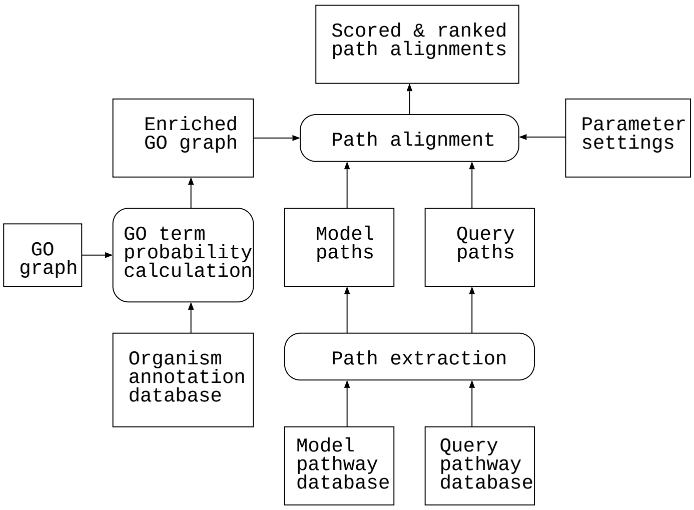

Fig. 1. The GOSAP method. Boxes with rounded corners represent procedures, and quadratic boxes represent information.

    - Increment a counter C_j for each annotated GO term T_j of G_i, and increment the counter of each ancestor term of T_j.
- For each GO term T_k:
    - Calculate the term probability $p(T_k) = \frac{C_k}{N}$, where N is total number of annotations in D.

Like in [11], both inheritance (is-a) and aggregation (part-of) relation types were considered in the term probability calculation. Terms of all evidence types were used.

Path extraction

A depth-first based algorithm is used to derive all paths originating from each node in the pathway graph. Extension of a path ends whenever a leaf node or a previously visited node is encountered (so that cycles are handled). Furthermore, only "super-paths" are used in the subsequent path alignment, i.e. the set of paths where no path is completely overlapping with another path. The objective is to obtain a small set of paths, while still covering the entire pathway graph.

Path alignment

The well-known Smith-Waterman algorithm, which was originally developed for identification of common molecular subsequences [12], was here adapted for the task of producing local alignments of paths of gene products. The scoring function s_f used for a match is defined in equation 1, which is similar to [11] , and in equation 2 [13].

$$s_f(G_i, G_j) = max(\{SS(T_k, T_l) : T_k \in t(G_i), T_l \in t(G_j)\}) \tag{1}$$

$$SS(T_k, T_l) = -log_2(p_{ms}(T_k, T_l)) \tag{2}$$

G_x refers to a gene product, $t(G_x)$ is the set of GO annotations for G_x, $SS(T_k, T_l)$ is the semantic similarity between GO molecular function terms T_k and T_l, $p_{ms}(T_k, T_l)$ is the probability of the minimum subsumer for T_k and T_l. The minimum subsumer refers to the ancestor term with lowest probability that is common to both terms. There is a reason for only using the molecular function sub-ontology for driving the alignment procedure. If the process sub-ontology was used instead, the alignment process would simply promote alignments where single gene product pairs belong to similar processes, without enforcing this process similarity throughout the complete alignment. The process- and cellular component sub-ontologies were instead used (in combination with the molecular function ontology) to post-evaluate an alignment by calculating overall alignment scores:

$$s_{fp}(G_i, G_j) = s_f(G_i, G_j) + s_p(G_i, G_j) \tag{3}$$

$$s_{fpc}(G_i, G_j) = s_f(G_i, G_j) + s_p(G_i, G_j) + s_c(G_i, G_j) \tag{4}$$

The total alignment scores, denoted as S_F, S_{FP} and S_{FPC}, are calculated by summing up gene product comparison scores s_f, s_{fp}, s_{fpc} from equations 1, 3 and 4, and possible gap penalties, over all positions in the locally aligned segment. Hence, the scores s_{fp} and s_{fpc} enhance the "resolution" of similarity between sequences (paths) of gene products, providing differentiation between alignments sharing similarity with respect to several sub-ontologies in combination.

When deriving an alignment using nucleotides or aminoacids, it is better to start a new alignment when the best dynamic programming matrix option is 0, motivated by the assumption that scores of random matches are negative. In GOSAP, a random match is expected to have the average term-to-term semantic similarity 0.7 or less, based on an empirical term-to-term comparison for the GO molecular function sub-graph. Furthermore, a linear gap penalty is used in the local alignment algorithm.

Statistical significance of alignments

The alignment score itself may not be sufficient for judging the quality of an alignment. Therefore, an assessment of the statistical significance of alignments was performed according to the procedure described in [14]. A query path is aligned against a large number of randomized versions of the model pathway, where nodes keep their connectivity but where edges are randomly switched. The p-value is defined as the share of randomized pathways that produce an alignment with equal or higher score than the original alignment.

3 Results

We have tested our algorithm on protein regulatory pathways as well as metabolic enzyme-to-enzyme pathways. Due to the space limitation we only report results for a few experiments in order to illustrate how GOSAP can be used.

Protein regulatory pathways

Protein regulatory pathways from KEGG [1] were used. When a gene product complex regulates another complex, a link is created from each gene product in the regulating complex to each gene product in the regulated complex. In an experiment, 195 super-paths extracted from the cell cycle regulatory pathway for *S. cerevisiae* were used as model paths and 37 super-paths of the MAPK regulatory pathway were used as queries. As these two pathways are quite different, few similarities are expected to be detected. Using a gap penalty of 1, 100 randomized models, and statistical significance threshold $p \leq 0.05$, six alignments in total were found for 3 of the 37 queries. For example, the query path "MSG5[i]>FUS3[p]>FAR1", resulted in the following alignment with $S_F = 25.8$ and a significance value $p_f = 0.04$:

```
Q: MSG5[i]>FUS3       (GAP)   FAR1
M: MIH1[d]>CDC28[ip]>SWI5[e]>SIC1
F: GO:0004725>GO:0004674 (GAP) GO:0019210
P: GO:0050875>GO:0006468 (GAP) GO:0000074
C: GO:0005737>GO:0005634 (GAP) GO:0005634
```

Q and M are the aligned paths for query and model, respectively. F shows the GO molecular function alignment sequence, where a GO identifier represents the minimum subsumer GO term for the two gene products under comparison. The corresponding information for biological process and cellular component is shown at P and C. Symbols within square brackets denote relation types as specified in KEGG; "i"=inhibition, "d"=dephosphorylation, "ip"=phosphorylation and inhibition, and "e"=expression. For example, in the molecular function alignment, gene products MSG5 and MIH1 have the minimum subsumer "protein tyrosine phosphatase activity" (GO:0004725). Even if the alignment is significant with respect to the chosen threshold for p_f, the score is considerably lower than the identity score, which represents a perfect match. The identity score is defined as the semantic score obtained when comparing the ungapped query alignment segment with itself, and is 36.4 for the query "MSG5[i]>FUS3[p]>FAR1". Hence, S_F is 71% of the identity score. However, another significant alignment for the same query path was obtained using a combined score where "FUS3[p]>FAR1" is aligned against exactly the same sub-sequence ("FUS3[p]>FAR1") in the model with $S_{FPC} = 53.3$ and a significance value $p_{fpc} = 0.04$, demonstrating the utility of using all three ontologies for scoring alignments. The path is present in both the MAPK and cell cycle pathways i KEGG, because it is both an exit point from the MAPK pathway and an entry point to the cell cycle pathway. All GO identifiers in alignments are not explained due to limited space, please refer to the Gene Ontology website (www.geneontology.org) for more details.

Assessing reverse engineered regulatory pathways

Apart from studying similarities among documented pathways, our method can be used to assess hypothetical regulatory pathways derived using reverse engineering techniques. An example is a pathway reported in [15] containing 12 gene products and 14 edges, which was derived using a dynamic Bayesian network technique and microarray gene expression data for the *S. cerevisiae* cell cycle.

It can be observed that only 3 of 14 edges in this pathway are identical to those in the known KEGG regulatory pathway of the *S. cerevisiae* cell cycle. Hence, it is expected that few similarities are detected by GOSAP.

The 11 super-paths that can be derived from this pathway were aligned against 195 super-paths found in a KEGG regulatory pathway of the *S. cerevisiae* cell cycle. Using a gap penalty of 1, 100 randomized models, and statistical significance threshold $p \leq 0.05$, one significant alignment was found for the query path "FAR1[?]>SIC1[?]>CLN2[?]>SIC1" with $S_{FP} = 76.6$ and a significance value $p_{fp} = 0.03$:

```
Q: FAR1[?]>SIC1    (GAP)   CLN2[?]>SIC1
M: FAR1[i]>CLN1[p]>SWI6[e]>CLN2[p]>SIC1
F: GO:0004861>GO:0019207 (GAP) GO:0016538>GO:0019210
P: GO:0007050|GO:0045786>GO:0000079 (GAP) GO:0000320|GO:0000321>GO:0000079
C: GO:0005634>GO:0005634 (GAP) GO:0005634>GO:0005634
```

A pipe sign "|" in the process alignment means that the terms it separates are equally good alternatives with respect to score. This alignment shows for example that SIC1 and CLN1 have the molecular function "kinase regulator activity" (GO:0019207) and biological process "regulation of cyclin dependent protein kinase activity" (GO:0000079), in common. The correct sub-path "CLN2[?]>SIC1" in the model was captured, and the gap in the query sequence is aligned against the transcription coactivator SWI6 in the model, suggesting that SWI6 is a step that is missing in the corresponding pathway in the hypothetical network. Additionally, clues to potential relation types for the query path are provided by the model path, e.g. a phosphorylation relation ("p") is feasible between CLN2 and SIC1. Once again, the utility of combined scores is demonstrated, since no significant alignment could be found using the molecular function score alone ($p_f \geq 0.36$). An identical significant alignment was found using all three sub-ontologies.

Metabolic pathways

Metabolic enzyme-to-enzyme pathways are derived from ordinary metabolic pathways by creating directed links between enzymes if the product of one enzymatic reaction is the substrate of another enzymatic reaction. GOSAP was used to rediscover one of the findings in [3], where a significant alignment was detected between the isoleucine biosynthesis- and valine biosynthesis metabolic pathways for *S. cerevisiae*. Using a gap penalty of 1 and 100 randomized models, an alignment was found where the query path "ILV2[E]>ILV5[E]>ILV3[?]>BAT1" from the isoleucine biosynthesis pathway is aligned with an identical path in the valine biosynthesis model pathway, with $S_F = 51.3$ and a significance value $p_f = 0.07$. Another almost identical alignment with identical S_F and p_f was also found, but where the last model gene product BAT1 was replaced with BAT2. This is expected, since the gene products have the same functional annotation and the corresponding reaction is annotated with both enzymes. These results are consistent with those in [3]. Other alignments in this case had $p_f \geq 0.17$ and $S_F \leq 49.5$. Using a combined score $S_{FPC} = 118.0$ and $pfpc = 0.02$ the alignment with BAT1 as last gene product in the model path was separated from the alignment where BAT2 is last gene product ($S_{FPC} = 113.4$, $pfpc = 0.11$). This is due to differing GO cellular component annotation for BAT1 and BAT2. BAT1 is active in the mitochondrial matrix, whereas BAT2 is annotated with the less specific "cytoplasm" term.

4 Discussion

We have developed a method for extracting and aligning paths from biological pathways containing gene products, where GO is used to annotate the gene products and GO based semantic similarity for all three sub-ontologies is used to score alignments. In the application of assessing hypothetical pathways, GOSAP is potentially useful for correction of query pathway segments that diverge from the model pathway, and also for predicting what gene products that may be missing in the query pathway.

Execution times for pathway comparisons on a Sun workstation varied from a few seconds for the metabolic network comparison (pathways have 4 and 5 nodes, and 4 super-paths each), to one hour when comparing the MAPK pathway against the cell cycle pathway (pathways have 49 and 65 nodes, 37 and 195 super-paths, respectively). This is when using 100 randomized models for statistical significance calculation. We plan to further develop the alignment algorithm by introducing heuristics, in order to reduce the computational complexity.

There are different parameter settings in GOSAP that must be considered, especially the gap penalty and the significance value. Currently, these settings must be set manually by the user. It was for example empirically observed that the significance values for alignments in general increase as a function of decreasing gap penalty, i.e. many alignments involving randomized graphs get higher scores, which in turn makes it harder to get significant alignments. As for the gap penalty, increasing gap penalties promote the matching of less similar gene products in order to avoid gaps, whereas very low penalties (or no penalty at all) results in fragmented alignments where only small sub-segments match. Different gap penalty strategies, such as affine gap costs, would also be of interest to investigate.

The performance of GOSAP depends on the quality of the GO annotations. Both regarding experimental evidence (e.g. "traceable author statement") and regarding specificity of GO terms for individual gene products. Currently, annotations of all evidence types are used, since we found it unfair to disqualify any specific type. But it is generally the opinion that "traceable author statement" is the most reliable type of annotation. As for specificity, some gene products are annotated with very specific terms and some are not. GOSAP would benefit from future, more fine grained versions of GO.

Furthermore, a multiple alignment extension of GOSAP similar to the work of [9] would enable the study of more than two species or paths at a time. It would also be possible to develop a measure of how semantically similar entire pathways are, based on the current method of comparing individual paths in pathways. GOSAP could in fact be generalised to cover any type of pathway or graph from any domain, as long as there are abstraction hierarchies or ontologies available for the different node types.

References

1. Kanehisa, M., Goto, S.: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research **28** (2000) 27–30
2. 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
3. Pinter, R. Y., Rokhlenko, O., Yeger-Lotem, E., Ziv-Ukelson, M.: Alignment of Metabolic Pathways. Bioinformatics **21** (2005) 3401–3408
4. Koyutürk, M., Grama, A., Szpankowski, W.: An efficient algorithm for detecting frequent subgraphs in biological networks. Bioinformatics **20** (2004) i200–i207
5. 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
6. 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
7. Berg, J., Lssig, M.: Local graph alignment and motif search in biological networks. PNAS **101** (2004) 14689–14694
8. 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
9. 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
10. 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
11. Lord, P. W., Stevens, R. D., Brass, A. and Goble, C. A.: Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics **19** (2003) 1275–1283
12. Smith, T. F., Waterman, M. S.: Identification of Common Molecular Subsequences. Journal of Molecular Biology **147** (1981) 195–197
13. 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
14. Maslov, S., Sneppen, K.: Specificity and Stability in Topology of Protein Networks. Science **296** (2002) 910–913
15. 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