<http://his.diva-portal.org/>

This is an author produced version of a paper published in Journal of Bioinformatics and Computational Biology. This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

Citation for the published paper:

Gamalielsson, J., Olsson, B.

Gene Ontology-based Semantic Alignment of Biological Pathways by Evolutionary Search

Journal of Bioinformatics and Computational Biology, 2008, 6, 4: 825-842.

URL: <http://dx.doi.org/10.1142/S0219720008003631>

Access to the published version may require subscription.

Published with permission from: World Scientific

Journal of Bioinformatics and Computational Biology
© Imperial College Press

GENE ONTOLOGY-BASED SEMANTIC ALIGNMENT OF BIOLOGICAL PATHWAYS BY EVOLUTIONARY SEARCH

JONAS GAMALIELSSON*

*Bioinformatics Research Group, University of Skövde, Box 408
Skövde, S-54128, Sweden
jonas.gamalielsson@his.se*

BJÖRN OLSSON

*Bioinformatics Research Group, University of Skövde, Box 408
Skövde, S-54128, Sweden
bjorn.olsson@his.se*

Received (Day Month Year)

Revised (Day Month Year)

Accepted (Day Month Year)

A large number of biological pathways have been elucidated recently, and there is a need for methods to analyse these pathways. One class of methods compares pathways semantically, in order to discover parts that are evolutionarily conserved between species or to discover intra-species similarities. Such methods usually require that the topologies of the pathways being compared are known, i.e. that a query pathway is being aligned to a model pathway. However, sometimes the query only consists of an unordered set of gene products. Previous methods for mapping sets of gene products onto known pathways have not been based on semantic comparison of gene products using ontologies or other abstraction hierarchies.

Therefore, we here propose an approach that uses a similarity function defined over Gene Ontology (GO) terms to find semantic alignments when comparing paths in biological pathways where the nodes are gene products. A known pathway graph is used as a model, and an evolutionary algorithm (EA) is used to evolve putative paths from a set of experimentally determined gene products. The method uses a measure of GO term similarity to calculate a match score between gene products, and the fitness value of each candidate path alignment is derived from these match scores. A statistical test is used to assess the significance of evolved alignments. The performance of the method has been tested using regulatory pathways for *S. cerevisiae* and *M. musculus*.

Keywords: gene ontology; pathways; alignment.

1. Introduction

The number of biological pathways that have been experimentally elucidated or computationally predicted is growing rapidly. Hence, there is a great need for meth-

*corresponding author

ods to compare pathways, so that similarities and differences can be analysed both within and between different species. Just like sequence alignments may help identifying evolutionary changes such as insertions, deletions and substitutions, a pathway alignment may help identifying evolutionary events at the pathway level, such as gene duplication and divergence of function. An alignment of two similar pathways from the same species may for example suggest that the aligned pathways have evolved from a common ancestor pathway by gene duplication followed by divergence¹.

Of particular interest is the class of methods that compare pathways semantically, i.e. using the annotation of the pathway components to discover homologies that are based on similarities regarding functional role, biological process or cellular location. Most previous work on such methods has focused on metabolic pathways, utilizing the EC enzyme hierarchy^a to calculate match scores^{1,2,3}. It has also been assumed previously that the comparison is done between pathways with known topologies. However, sometimes only a set of gene products is available on the query-side and the goal is to derive a putative pathway by finding the best possible matching of gene products onto the known model pathway. Some earlier methods^{4,5,6} for mapping gene products onto known pathways do this merely for presentation purposes and are not based on approximate matching using abstraction hierarchies or ontologies. We therefore propose EGOSAP, which is an Evolutionary Gene Ontology⁷-based method for finding Semantic Alignments between Paths in biological pathways where the nodes are gene products. EGOSAP uses a known pathway graph, from which a set of model paths are extracted. It then uses an evolutionary algorithm (EA) to derive putative paths, semantically similar to the model paths, from a set of experimentally derived gene products.

1.1. *Previous work*

Most previous work on pathway alignment has been focused on metabolic pathways and has used the EC enzyme hierarchy to calculate match scores for the enzymes and reactions^{1,2,3}. Dandekar et al.² proposed an approach involving analysis and comparison of biochemical data, pathway analysis using the elementary modes concept, and comparative analysis of a set of completely sequenced genomes. A method for deriving multiple alignments of paths in metabolic pathways was proposed by Tohsato and co-workers³, where the EC hierarchy is used for generalizing about enzymes. Pinter et al.¹ extended this idea in a method for pairwise alignment of metabolic pathway graphs using a technique known as approximate labeled sub-tree homeomorphism, where the EC hierarchy once again is used for generalization.

An obvious drawback of using the EC hierarchy is that the method becomes limited to metabolic pathways, since all pathway elements must be enzymes. In our earlier work, we therefore proposed GOSAP⁸ (Gene Ontology based Semantic Align-

^awww.expasy.org/enzyme

ment of biological Pathways), a local alignment method for semantic comparison of biological pathways. A semantic similarity function defined over Gene Ontology (GO) was used to derive semantic alignments of paths in biological pathways, with the advantage that pathways can be analysed where nodes are not only enzymes but any kind of gene product. Another novelty was the use of combined alignment scores involving all three sub-ontologies of GO, so that the resulting alignments are based on more detailed information.

The approaches to pathway comparison described so far assume that the topology of both the model- and the query graph is known. However, sometimes only a query set of gene products is available, without any knowledge about how the gene products interact. In such cases it is of interest to apply a method capable of deriving putative paths using this query set. Such interactions may be discovered by EGOSAP, by evolving paths connecting the selected gene products to each other. A query set of gene products may for example be the products of genes that are differentially expressed between conditions in a microarray experiment.

There are tools 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. An example is GenMAPP⁴, where the genes and their colour-coded expression values are mapped onto known pathways. There is also the GenMAPP accessory software MAPPFinder⁹ where GO visualisation has been added. The Pathway tools software⁵ is another example where functionality is available for including gene expression data in pathway diagrams in a manner similar to GenMAPP. ArrayXPath⁶ is a similar tool where gene expression clusters can be mapped onto the best matching pathways in a database.

Since the existing methods for mapping gene products onto known pathways are restricted by relying on exact matching, we here propose a method that uses an EA to search for the best matching by semantic similarity. The new method, EGOSAP, is similar to our earlier proposed GOSAP algorithm, but features the novelty of using an EA to derive putative path alignments based on semantic similarity between gene products. The competence of the EA is verified using benchmark experiments, and example alignments are derived in a cross-species experiment. EGOSAP can for example be used to semantically map differentially expressed genes identified in a microarray experiment onto known regulatory pathways. This is particularly useful if the experiments have been conducted on a species where little is known about its pathways. It would also be possible to use gene products elucidated using other experimental methods, or to manually add gene products that are known to be important but which are not measured.

2. Method

EGOSAP, which is summarized in figure 1, is similar to the GOSAP method which compares a user-specified query pathway graph with a model pathway graph, using three procedures. The first two are preparatory procedures used to set up the GO

Fig. 1. The EGOSAP method. Boxes with rounded corners represent procedures, and rectangular boxes represent information.

annotation data and the model paths in such a way that semantic alignments can be derived. These two initial procedures are identical for GOSAP and EGOSAP, and are therefore only summarized here for convenience (see Gamalielsson and Olsson⁸ for a more detailed description).

1) GO term probability calculation

For every GO annotation term, a probability is calculated using an annotation database for one or several organisms. These probabilities reflect the frequencies with which the annotation terms are used, and are used in the alignment procedure to calculate the semantic similarity of each pair of gene products. This is based on the observation that more specific terms tend to have lower GO term probabilities, when these are calculated using the method proposed by Lord et al.¹⁰:

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

In the example in figure 3 (discussed in more detail later), the term probabilities appear as a p value for each GO term.

2) Path extraction

An algorithm involving depth-first search is used to derive all model paths originating from each node in the model 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 the *super-paths* are used in the subsequent path alignment, i.e. the set of paths where no path is included in its entirety as a sub-path of another path. The purpose is to obtain a minimal set of paths, while still covering the entire pathway graph. Putative paths are evolved for each of the extracted paths in the path alignment optimization procedure.

3) Path alignment optimization

The original GOSAP method aligns pairs of paths using the Smith-Waterman¹¹ algorithm with match scores calculated by a semantic similarity function over the "alphabet" of Gene Ontology annotation terms. However, this relies on the assumption that the topology of both the query- and model pathway graph is known. As mentioned earlier, sometimes only a query set of gene products is available, and there is no knowledge available about how the gene products interact. Therefore, the purpose of EGOSAP is to suggest putative paths using the query set of gene products. The EA evolves paths that are semantically similar to paths in the model pathway, and it is assumed that paths are permutations of gene products from the query set, i.e. a gene product can only appear once. This is the case in e.g. genetic networks.

The path alignment optimization in EGOSAP is illustrated in figure 2. As input, EGOSAP takes a query set of gene products derived from experimental data (upper left in figure 2) and a model pathway graph obtained from a pathway database, e.g. KEGG¹² (upper right). The extracted paths are submitted one at a time to be aligned (indicated by filled circles). In order to search for an optimal alignment, EGOSAP samples the query set of gene products to generate an initial random population of candidate paths. Each individual is represented as an initially random permutation of gene products from the query set with the same length as the model path. The maximum length of the candidate paths is the same as the length of the model path, and in order to allow gaps, a special symbol is used to signify "no gene product" (indicated by circles with dashed borders). A user-defined number of "no gene product" symbols can be added to the query set. The match score between a specific model gene product (MGP) and a "no gene product" symbol in the evolved path is set to the average semantic similarity between the MGP and all gene products in the query alphabet.

During a number of iterations, the population of candidate paths is replaced by a new population by fitness-based tournament selection of "parent" paths, from which "offspring" paths are generated using recombination and mutation. Binary tournament selection is used, which means that two random individuals are drawn from the population and the best of these is selected¹³. Selection is done without re-

Fig. 2. Path alignment optimization procedure in EGOSAP. For explanations, see text.

placement, i.e. each individual can be selected several times for the next generation. Variation operators used are partially mapped crossover (PMX) and a mutation operator, and are performed with probabilities p_c (crossover) and p_m (mutation). PMX operates on two parents where a one-to-one mapping between the gene products of the two parents is created at a randomly chosen middle segment¹⁴. This mapping is used during crossover to ensure that feasible offspring are generated, i.e. an ordered set of elements with no duplicates. Mutation is done by a combined operator, which either (with 50% probability) switches the gene products at two random positions in a parent, or replaces a random gene product with a random one from the query set if possible. An elitist strategy is used where the worst individual of the old population is replaced by the best individual in the current population. This enforces a monotonous fitness growth of the population's best fitness.

The fitness of each alignment is calculated from the semantic similarities of the aligned pairs of gene products, according to the following equation:

$$F = \frac{\sum_{i=1}^L w_f \cdot s_f(M_i, Q_i) + w_p \cdot s_p(M_i, Q_i) + w_c \cdot s_c(M_i, Q_i)}{\sum_{i=1}^L w_f \cdot s_f(M_i, M_i) + w_p \cdot s_p(M_i, M_i) + w_c \cdot s_c(M_i, M_i)} \quad (1)$$

where L is the alignment length, and the weights w_f , w_p and w_c are adjustable

with the restriction that $\sum w_f + w_p + w_c = 1$. s_f is the molecular function semantic similarity between gene products at position i for the model path M and query path Q . s_p and s_c are the respective measures for the biological process and cellular component ontologies. The denominator part of equation 1 enforces the fitness interval $[0,1]$. s_f is calculated according to equation 2, which is similar to the one defined by Lord et al.¹⁰. s_p and s_c are calculated in the same manner, but using the biological process and cellular component annotations, respectively.

$$s_f(M_i, Q_i) = \max(\{SS(T_k, T_l) : T_k \in t(M_i), T_l \in t(Q_i)\}) \quad (2)$$

$$SS(T_k, T_l) = -\log_2(p_{ms}(T_k, T_l)) \quad (3)$$

In equation 2, $t(M_i)$ and $t(Q_i)$ are the sets of GO annotations for M_i and Q_i . The fitness function promotes individuals which are semantically similar to the model sequence with respect to all three sub-ontologies. In equation 3 (defined by Resnik¹⁵), $p_{ms}(T_k, T_l)$ is the probability of the minimum subsumer for GO terms T_k and T_l . The minimum subsumer ms is the ancestor term with lowest probability that terms T_k and T_l have in common.

The semantic similarity calculation for two example gene products, FUS3 and CDC28, is illustrated using figure 3. FUS3 is annotated with GO term GO:0004707 and CDC28 is annotated with GO:0004693. The minimum subsumer of these two terms is GO:0004674, and the probability of this GO term (reflecting the frequency with which it is used in the annotation database) is 0.011. Hence, the semantic similarity between FUS3 and CDC28 is $-\log_2(0.011) \approx 6.5$.

In the box to the middle right in figure 2 is shown the alignment between a query path, Q , and a model path, M . Below the alignment is shown a meta-alignment, indicating which GO terms have been used in the calculation of fitness for this alignment, and below the box is shown the meaning of the annotation terms from the three GO sub-ontologies for molecular function (F), biological process (P) and cellular component (C).

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 by Maslov and Sneppen¹⁶. In this procedure two edges $A \rightarrow B$ and $C \rightarrow D$ are randomly selected in a graph and rewired into $A \rightarrow D$ and $C \rightarrow B$. If the resulting edges are already present in the graph, a new pair of edges is selected. Hence, a randomization takes place while preserving the cardinality of each node (gene product). A series of random edge switches results in a randomized graph, with the restriction that the randomized graph must be different from the original graph. In EGOSAP, a query path can be aligned with a large number of randomized versions of the model pathway using the Maslov and Sneppen procedure. The p -value of an alignment is defined as the fraction of randomized graphs that produce an alignment with equal or higher score than the original alignment. Low p -values are desirable.

Fig. 3. Gene products (ovals) mapped to GO terms (rectangles) according to their molecular function annotation. Higher located GO terms are more specific (lower probability p) than lower located terms. GO:0004674 is the minimum subsumer of GO:0004707 and GO:0004693.

3. Results

3.1. On the problem complexity

EGOSAP aims to evolve an optimal alignment to a path of length L , by selecting an ordered subset (i.e. a permutation) of gene products from a gene product "alphabet" of size N . For a permutation, i.e. a candidate solution, the fitness is calculated according to equation 1. The number of possible permutations of length L that can be generated from a query alphabet of size N is given by:¹⁷

$$P(N, L) = N(N-1)(N-2)\dots(N-L+1) = \frac{N!}{(N-L)!} \quad (4)$$

Figure 4 shows the number of possible permutations for different combinations of N and L . It can be observed that $P(N, L)$ increases exponentially relative to L , and that the impact of N increases for larger values of L . Clearly, exhaustive enumeration of all possible candidate solutions is not feasible for paths of biologically realistic lengths.

3.2. Datasets

We used three different model graphs for *S. cerevisiae*. The first is a graph created using the transcriptional regulatory chain motifs described by Lee et al.¹⁸ This graph contains experimentally determined interactions between transcriptional elements. In our study we only use gene products which have annotations for all three sub-ontologies of GO. With this restriction the graph contains 64 gene products, 77 edges and 105 super-paths. Path lengths vary from two gene products to five, with an average path length of 3.3. The second model graph is the cell cycle regulatory pathway from KEGG containing 61 gene products, 81 edges and 151 super-paths. Path lengths vary from two to ten gene products, with an average length of 6.7. The third model is the MAPK signalling pathway from the same database containing

Fig. 4. Number of possible permutations for different sizes of alphabet (N) and path length (L).

48 gene products, 49 edges and 33 super-paths. Here, path lengths vary from 3 to 8 gene products, with an average length of 6.1.

Two query sets were used containing the products of genes for *M. musculus* that were found to be differentially expressed in an experiment comparing transgenic and knock-out mice with wild-type mice (for details regarding the experimental protocol, see Nilsson et al.¹⁹). The transgenic query set contained 460 gene products derived from 531 microarray probes, but since gene product annotation from all three GO sub-ontologies is desired, the number of gene products was reduced to 211. For the knock-out query set, the number of gene products was reduced from 256 (284 probes) to 119. In some of the cross-species experiments in section 3.4, the gene products from the corresponding *M. musculus* cell cycle and MAPK pathways are used in conjunction with the *M. musculus* transgenic and knockout datasets. The *M. musculus* cell cycle dataset contains 75 gene products, of which 59 have annotation from all three GO sub-ontologies. The corresponding figures for the MAPK dataset are 121 and 79.

It should be mentioned that EGOSAP works with only one GO sub-ontology, i.e. all three are not required. In our examples we chose to use all three sub-ontologies in order to be able to show as informative alignments as possible in our results. If only the molecular function sub-ontology, fewer gene products would typically be disqualified due to lacking GO annotation. The Lee et al.¹⁸ graph would contain 67 gene products, the cell cycle pathway 62 gene products, the MAPK pathway 49 gene products, and the *M. musculus* transgenic and knockout datasets would hold 331 and 192 gene products, respectively. Hence, requiring only one sub-ontology

typically leads to inclusion of more genes, and thus potentially more alignments, but also to less reliable results. Conversely, requiring additional sub-ontologies may reduce the quantity of results, while on the other hand increasing their reliability.

3.3. Benchmarking the EA

In order to assess under what conditions EGOSAP is able to evolve semantically optimal alignments, a number of benchmark simulations were performed. This was done by creating model paths from a query set, and evolving paths from the same query set. As an optimal solution is available in this set-up with a fitness $F = 1.0$, the competence of the EA is being tested. The ideal result would be that the algorithm finds this optimal solution within a small number of generations for problems of biologically realistic size. It is assumed that the EA is equally competent when the gene products in the model and in the query set are different, as in the cross-species experiments in section 3.4. No gaps were allowed in the benchmarking experiments.

In an initial experiment all 64 gene products in the Lee et al. model graph were arranged (in random order) into a linear path consisting of 64 nodes, and the same set of gene products was used as a query set. The following EA parameters were used: population size=100, crossover probability=0.7, mutation probability=0.03. Equal weight was given to the three sub-ontologies in the fitness function, i.e. $w_f = w_p = w_c = \frac{1}{3}$. Since the EA is stochastic, the performance may differ between runs. However, a typical run resulted in perfect fitness, $F = 1$ (see equation 1), after approximately 1000 generations, meaning that 10^5 fitness calculations were done in total. An exhaustive enumeration of permutations, instead of using an EA, would require $\sim 10^{89}$ fitness calculations (see equation 4). This shows that the EA is performing very well on this particular dataset, even when using as long paths as 64 gene products, which is likely to include most biologically realistic situations.

An equivalent experiment was performed by using the *M. musculus* transgenic data set containing 211 gene products. The same EA parameters were used as in the initial experiment using the model graph by Lee et al., except for the mutation probability which was set to 0.0095. It turns out that this dataset requires considerably more iterations to optimize than the Lee et al. dataset. The EA reached $F = 1.0$ after approximately 15000 generations (1.5 million fitness calculations). An enumeration would require $e^{918} \approx 10^{399}$ fitness calculations, i.e. the search space is extremely large. This number is too large to handle for most computers, so the Stirling²⁰ factorial approximation $\ln(n!) \approx n \cdot \ln(n) - n$ was used instead of equation 4. As a contrast, optimizing a path containing a randomly selected subset of 10 gene products would only require approximately 200 generations. It should also be emphasized that the time complexity of the fitness function in equation 1 is $O(L)$, i.e. the execution time increases linearly as a function of alignment length L .

One reason for the quicker EA convergence on the first dataset is probably that it is more semantically homogeneous than the second dataset, i.e. the gene products are more semantically similar to each other because they are all related to

Table 1. Example alignment obtained when using the transgenic data set as query set and the Lee et al. graph as model. For explanations, see text.

Alignment				
Q:	NFIX	→	AKAP8	→ STAT5B
M:	MBP1	→	ABF1	→ STP1
Meta-alignment				
F:	GO:0003700	→	GO:0003682	→ GO:0003700
P:	GO:0006260	→	GO:0051276	→ GO:0045944
C:	GO:0005634	→	GO:0000790	→ GO:0005634

transcription. In the second dataset, gene products differ more with respect to their GO annotations, making the EA's optimization task harder. Furthermore, the query set is more than three times larger in the second case. To sum up, the benchmark results indicate that the GO annotations of gene products in both model graph and query set clearly affect the convergence speed of the EA, but that good performance can be expected for paths of biologically realistic length.

3.4. Cross-species experiments

For the cross-species experiments the same EA parameter settings were used as in the benchmark experiments, and 100 randomized model graphs were used for calculation of statistical significance. In the first experiment, the Lee et al. model and both *M. musculus* query sets were used. Figure 5 shows the percentage of paths for which alignments with significant p values were found, as a function of p -value threshold. This test was performed both for the 105 super-paths of the model graph, and for the complete set of 204 possible paths of all lengths. For both query sets it can be observed that significant alignments were found for only a few of the paths with threshold $p \leq 0.02$. Furthermore, significant alignments are found for a larger proportion of super-paths compared to the case with all possible paths. One reason for this effect is that many of the later mentioned paths are short and therefore less likely to be significant. Given that this is a cross-species scenario, comparing two distantly related species, and considering the rather small numbers of gene products in the two datasets, we do not find it surprising that significant alignments are found only for a small proportion of the paths.

To illustrate the output of EGOSAP, we here show examples of putative path alignments evolved using the knock-out query set of differentially expressed genes for *M. musculus* and the super-paths extracted from the Lee et al. model graph. An example is the query path "NFIX → AKAP8 → STAT5B", which resulted in the alignment shown in table 1. This alignment has $F = 0.72$ and $p = 0$. Q is the evolved path, and M is the model path, respectively. F shows the GO molecular function meta-alignment, where each identifier represents the minimum subsumer GO term for the two gene products under comparison. The corresponding infor-

Fig. 5. Percentage of paths for which significant alignments are found as a function of p -value threshold for the transgenic dataset (left) and knock-out dataset (right). Solid lines represent the case when all possible paths were used and dashed lines represent the case with super-paths.

mation for biological process and cellular component is shown in the rows labelled by P and C . For example, in the molecular function meta-alignment, gene products NFIX1 and MBP1 have the minimum subsumer "transcription factor activity" (GO:0003700), and are both involved in DNA replication (GO:0006260) and expressed in the nucleus (GO:0005634). In the second position, both gene products have the function "chromatin binding" (GO:0003682) and are involved in chromosome organization and biogenesis (GO:0051276) and are expressed in the nuclear chromatin. In the third position, the two gene products STAT5B and STP1 share function "transcription factor activity", the biological process "positive regulation of transcription from RNA polymerase" and the cellular localization "nucleus". Thus, the alignment clearly indicates that these are homologous paths of transcriptional regulation, which have been preserved between yeast and mouse.

Interestingly, it is unlikely that this similarity would have been found by using a traditional sequence homology-based approach. We found that the average sequence identity between the three pairs of amino acid sequences was only 14.4%. Further, for each one of the three query sequences another match with slightly higher sequence identity (16.1% on average) could be found by BLAST-searching the *M. musculus* subset of REFSEQ²¹. When studying the annotations of these "closer homologs", we found that their function was much less similar and that cellular location included cytoplasm and membrane, which clearly indicates that a sequence-based approach would have produced spurious hits for this query path.

Table 2. Example alignment obtained when using the transgenic data set as query set and the Lee et al. graph as model. GO term codes (shown without initial zeros) have the following interpretations: 3713: transcription coactivator activity, 3700: transcription factor activity, 16563: transcriptional activator activity, 8237: metalloproteinase activity, 16564: transcriptional repressor activity, 6366: transcription from RNA polymerase II promoter, 7049: cell cycle, 6357: regulation of transcription from RNA polymerase II promoter, 6508: proteolysis, 122: negative regulation of transcription from RNA polymerase II promoter, 5634: nucleus, 16021: integral to membrane, 5694: chromosome.

Alignment						
Q:	MEF2C	→	NR2F6	→	NRBF2	→ AFG3L2 → TRIM28
M:	SWI6	→	SWI4	→	NDD1	→ ACE2 → SFL1
Meta-alignment						
F:	0003713	→	0003700	→	0016563	→ 0008237 → 0016564
P:	0006366	→	0007049	→	0006357	→ 0006508 → 0000122
C:	0005634	→	0005634	→	0005634	→ 0016021 → 0005694

In the example alignment in table 1, it can also be observed that despite the obviously high similarity between the two paths, they do not have exactly the same annotations throughout the whole alignment. At position two, the shared biological process annotation term is "chromosome organization and biogenesis", since this is the minimum subsumer of the terms found in the annotation of AKAP8 and ABF1. However, looking at the concrete annotation of the gene products, we find that AKAP8 is annotated with "mitotic chromosome condensation" and that ABF1 is annotated with "nucleotide-excision repair / DNA damage recognition". This difference accounts for the fitness score being $F = 0.72$, rather than $F = 1.0$, but also demonstrates that relatively modest fitness scores may correspond to high quality alignments between closely homologous paths. Another example of a significant alignment ($F = 0.73$, $p = 0.01$), obtained when using the transgenic query set, is shown in table 2.

In the second experiment we used the cell cycle pathway and the MAPK pathway for *S. cerevisiae* according to KEGG as models. In each case, the *M. musculus* transgenic dataset was used as query set, but with the addition of all gene products from the corresponding pathway (cell cycle or MAPK, respectively) for *M. musculus*. The percentage of significant paths as a function of p -value is shown in figure 6. This test was performed for both pathways, and also both for the super-paths and the complete set of possible paths. For both model pathways it can be observed that significant alignments were derived for a large proportion of the paths, even for such a conservative significance threshold as $p = 0$. One reason for this is probably that paths on average are considerably longer for the cell cycle and MAPK pathways than for those from the Lee et al. model. Optimized alignments containing long paths are less likely to appear by chance when using the Maslov and Sneppen graph randomization procedure. Furthermore, the cell cycle and MAPK pathways

Fig. 6. Percentage of paths for which significant alignments are found as a function of p -value threshold for the cell cycle pathway (left) and MAPK pathway (right). Solid lines represent the case when all possible paths were used and dashed lines represent the case with super-paths.

are less semantically homogeneous compared to the Lee et al. model, which only contains gene products related to transcription. Lower semantic homogeneity would probably yield a larger number of significant paths. Another reason may be that the scores of alignments are generally higher since the transgenic dataset has been "injected" with all the gene products from the corresponding *M. musculus* pathway (cell cycle or MAPK). In fact, all optimized alignments contain query paths where all gene products are from the corresponding *M. musculus* pathway, i.e. no gene products were selected from the original transgenic dataset. Gene products from the corresponding pathway are exclusively selected also if the transgenic dataset is replaced with the knockout dataset.

An example of a significant alignment ($F = 0.86$, $p = 0$) is shown in table 3. This alignment was derived using the cell cycle pathway as model. In the model path, SIC1 inhibits the cyclin-dependent protein kinase CDC28. CDC28 in turn phosphorylates and inhibits SWI5, which is a transcriptional activator. The corresponding *M. musculus* gene products in the alignment are not connected in the KEGG cell cycle for *M. musculus*. It should be emphasized that the cell cycle pathways for *S. cerevisiae* and *M. musculus* have similar regulatory mechanisms, but the pathway topologies are quite different. This may explain why the evolved query path is not part of the *M. musculus* cell cycle pathway, even though close gene product homologs were found. An interesting observation is that SWI5 expresses SIC1 according to KEGG, and the same regulation is present between SMAD4

Table 3. Example alignment obtained when using the *M. musculus* transgenic data set as query set, with the cell cycle gene products for the same organism added. The cell cycle pathway for *S. cerevisiae* was used as model. A pipe sign "|" separates GO terms resulting in equal score. GO term codes have the following interpretations: 19210: kinase inhibitor activity, 4674: protein serine/threonine kinase activity, 16563: transcriptional activator activity, 79: regulation of cyclin dependent protein kinase activity, 51320: S phase, 84: S phase of mitotic cell cycle, 6357: regulation of transcription from RNA polymerase II promoter, 5634: nucleus.

Alignment				
Q:	CDKN1B	→	DBF4	→ SMAD4
M:	SIC1	→	CDC28	→ SWI51
Meta-alignment				
F:	0019210	→	0004674	→ 0016563
P:	0000079	→	0051320 0000084	→ 0006357
C:	0005634	→	0005634	→ 0005634

and CDKN1B in the *M. musculus* cell cycle pathway. Thus, the same relationship between the last and first position in alignment is present in both paths.

Another example of a significant alignment ($F = 0.80$, $p = 0$) is shown in table 4. This alignment was derived using the MAPK pathway as model. The model path represents a part of the high osmolarity induced sub-pathway where STE20 is a signal-inducing kinase which phosphorylates STE11, which is a signal-transducing MEK-kinase active during the MAPKKK phase of the MAPK pathway. STE11 in turn phosphorylates PBS2, a kinase active in the MAPKK phase. PBS2 finally phosphorylates HOG1, which is a another kinase operating during the MAPK phase. The *M. musculus* gene products in the query path are not connected in the *M. musculus* version of the MAPK pathway, but gene products in positions 2 through 4 appear in the correct phase in the pathway (MAPKKK, MAPKK and MAPK). As for the cell cycle pathway, the MAPK pathways for *S. cerevisiae* and *M. musculus* have similar mechanisms, but are rather different in their topologies.

4. Discussion

We have developed EGOSAP, a method which uses a known pathway graph as a model from which model paths are extracted, and an EA to evolve putative paths that are semantically similar to these model paths using a set of experimentally determined, or by other means derived, gene products. We have tested the performance of the method on example datasets, and shown by examples the output produced by the method.

The execution time of the path optimization procedure in EGOSAP depends on many parameters, e.g. number of model paths, the model path length and the number of randomized models used in the significance calculation. To optimize the

Table 4. Example alignment obtained when using the *M. musculus* transgenic data set as query set, with the gene products of the MAPK pathway for the same organism added. The MAPK pathway for *S. cerevisiae* was used as model. GO term codes have the following interpretations: 4674: protein serine/threonine kinase activity, 4709: MAP kinase kinase activity, 5076: receptor signaling protein serine/threonine kinase signaling protein activity, 5078: MAP-kinase scaffold activity, 4707: MAP kinase activity, 19953: sexual reproduction, 6468: protein amino acid phosphorylation, 187: activation of MAPK activity, 43406: positive regulation of MAPK activity, 42995: cell projection, 5737: cytoplasm, 5634: nucleus.

Alignment					
Q:	AKT1	→	B230120H23RIK	→	MAPK8IP3 → MAPK1
M:	STE20	→	STE11	→	PBS2 → HOG1
Meta-alignment					
F:	0004674	→	0004709	→	0005076 0005078 → 0004707
P:	0019953	→	0006468	→	0000187 0043406 → 0006468
C:	0042995	→	0005737	→	0005737 → 0005634

alignment of one path containing a reasonable number of gene products (< 10) is done in a matter of seconds or at most one minute on a modern PC with a processor speed of 3 GHz and 1 GB of RAM using the current implementation.

Currently, the optimization procedure has the restriction that a gene product can only appear once in a path. This restriction is reasonable for regulatory pathways, but may not be desirable when metabolic pathways are studied. The method could easily be adapted for this scenario by allowing each gene product to appear at several positions in a path, i.e. paths are no longer limited to permutations of gene products. This modification would also require that some of the EA operators are replaced or modified.

The method allows the user to set different weights for the different GO sub-ontologies, although we have only used $w_f = w_p = w_c = \frac{1}{3}$ in the presented evaluations. Setting different weights for the different types of annotation can be useful since the certainty of different types of annotation can vary a lot for different organisms. If it is found that a large proportion of the annotation in a particular sub-ontology is, for example, inferred by homology, rather than based on experimental evidence, then it may be desirable to give this type of annotation a lower impact in the alignment optimization process.

5. Acknowledgement

This work was supported by project grant number 2003/0215 from the Knowledge Foundation, and by the Information Fusion Research Program (University of Skövde, Sweden) under grant 2003/0104 in partnership with the Knowledge Foundation (URL: <http://infofusion.se>).

References

1. Pinter RY, Rokhlenko O, Yeger-Lotem E, Ziv-Ukelson M, Alignment of Metabolic Pathways, *Bioinformatics* **21**:3401–3408, 2005.
2. Dandekar T, Schuster A, Snel B, Huynen M, Bork P, Pathway alignment: application to the comparative analysis of glycolytic enzymes, *Biochemistry Journal* **343**:115–124, 1999.
3. 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)*, pp. 376–383, 2000.
4. Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR, GenMAPP - a new tool for viewing and analyzing microarray data on biological pathways, *Nature Genetics* **31**:19–20, 2002.
5. Karp PD, Paley S, Romero P, The pathway tools software, *Bioinformatics* **18**:S1–S8, 2002.
6. Chung H-J, Kim M, Park CH, 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**:W460–W464, 2004.
7. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G, Gene Ontology: tool for the unification of biology, *Nature Genetics* **25**:25–29, 2000.
8. Gamalielsson J, Olsson B, GOSAP: Gene Ontology Based Semantic Alignment of Biological Pathways, to appear in *International Journal of Bioinformatics Research and Applications*.
9. Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, Conklin BR, MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data, *Genome Biology* **4**:R7, 2003.
10. Lord PW, Stevens RD, Brass A, Goble CA, Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation, *Bioinformatics* **19**:1275–1283, 2003.
11. Smith TF, Waterman MS, Identification of Common Molecular Subsequences, *Journal of Molecular Biology* **147**:195–197, 1981.
12. Kanehisa M, Goto S, KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucleic Acids Research* **28**:27–30, 2000.
13. Michalewicz Z, Fogel DB, *How to Solve It: Modern Heuristics*, Second edition, Springer, Berlin, 2004.
14. Goldberg DE, Lingle R, Alleles, Loci and the TSP, *Proc. 1st International Conference on Genetic Algorithms and their Applications*, pp. 154–159, 1985.
15. 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 (JAIR)* **11**:95–130, 1999.
16. Maslov S, Sneppen K, Specificity and Stability in Topology of Protein Networks, *Science* **296**:910–913, 2002.
17. Rosen KH, *Discrete mathematics and its applications*, McGraw-Hill, New York, 1995.
18. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne J-B, Volkert TL, Fraenkel E, Gifford DK, Young RA, Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*, *Science* **298**:799–804, 2002.
19. Nilsson EC, Long YC, Martinsson S, Glund S, Garcia-Roves P, Svensson TL, Anders-

- son L, Zierath JR, Mahlapuu M, Opposite transcriptional regulation in skeletal muscle of AMPK γ 3 R225Q transgenic versus knock-out mice, *Journal of Biological Chemistry* **281**:7244–7252, 2006.
20. Feller KH, *An Introduction to Probability Theory and Its Applications*, Vol. 1, 3rd ed., Wiley, New York, 1968.
21. Pruitt KD, Tatusova T, Maglott DR, NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic Acids Research* **35**:D61–D65, 2007.

Jonas Gamalielsson received his BS and MS degrees in computer science from University of Skövde in 2000 and 2001, respectively. Currently he is a senior PhD candidate at the School of Mathematical and Computer Sciences of Heriot-Watt University in Edinburgh, but he is geographically located at and sponsored by University of Skövde. Gamalielsson's research interests include development of semantic pathway alignment algorithms for systems biology, as well as evolutionary algorithms, data mining algorithms and information fusion applied within the field of bioinformatics. He has so far published 5 journal articles, 2 book chapters and 9 conference papers.

Björn Olsson received his BS and MS degrees in computer science from University of Skövde in 1992 and 1993, respectively, and his PhD degree in computer science from University of Exeter in 2000. Currently he is associate professor of bioinformatics at University of Skövde. Olsson's research interests include bio-sequence analysis, gene expression analysis, and other core topics in bioinformatics, as well as evolutionary algorithms and information fusion, with a focus on their application in systems biology and bioinformatics. Since his PhD graduation in 2000, he has published 18 journal articles and 22 conference papers on these topics