BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 18 2011, pages 2478–2485
doi:10.1093/bioinformatics/btr428

Genome analysis

Advanced Access publication July 19, 2011

Discovering novel subsystems using comparative genomics

Luciana Ferrer*, Alexander G. Shearer and Peter D. Karp

Artificial Intelligence Center, SRI International, Menlo Park, CA, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Key problems for computational genomics include discovering novel pathways in genome data, and discovering functional interaction partners for genes to define new members of partially elucidated pathways.

Results: We propose a novel method for the discovery of subsystems from annotated genomes. For each gene pair, a score measuring the likelihood that the two genes belong to a same subsystem is computed using genome context methods. Genes are then grouped based on these scores, and the resulting groups are filtered to keep only high-confidence groups. Since the method is based on genome context analysis, it relies solely on structural annotation of the genomes. The method can be used to discover new pathways, find missing genes from a known pathway, find new protein complexes or other kinds of functional groups and assign function to genes. We tested the accuracy of our method in *Escherichia coli K-12*. In one configuration of the system, we find that 31.6% of the candidate groups generated by our method match a known pathway or protein complex closely, and that we rediscover 31.2% of all known pathways and protein complexes of at least 4 genes. We believe that a significant proportion of the candidates that do not match any known group in *E.coli K-12* corresponds to novel subsystems that may represent promising leads for future laboratory research. We discuss in-depth examples of these findings.

Availability: Predicted subsystems are available at http://brg.ai.sri.com/pwy-discovery/journal.html.

Contact: lferrer@ai.sri.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 12, 2011; revised on July 9, 2011; accepted on July 13, 2011

1 INTRODUCTION

In the last decade, the use of *in silico* approaches to aid in the discovery of novel pathways has been facilitated by the advent of high-throughput technology. Protein–protein interactions, genetic interaction, and metabolic networks have all been used individually and in combination to find gene groups that are likely to constitute novel pathways.

Kelley and Ideker (2005) propose a method that uses networks of physical interactions (protein–protein, protein–DNA and metabolic networks) to help interpret a set of genetic interactions. They define a pathway as a densely connected set of proteins in the network of physical interactions and use this definition to interpret genetic interactions as being between-pathways or within-pathway. They also use the resulting pathways to predict functions of proteins that belong in pathways in which most other proteins have a common functional annotation, and to predict new genetic interactions. Ma *et al.* (2008) proposed a method to find compensatory pathways using synthetic lethal interactions. Zhang and Ouellette (2010) combine protein–protein interactions, genetic interactions, domain–domain interactions and the similarity of Gene Ontology terms into a single network. Putative pathways are then generated by searching this network for groups of proteins that share similar neighborhoods.

A somewhat more modest goal than discovering completely novel pathways is to find the components of a known pathway in a new organism. Several methods have been proposed to this end. Cakmak and Ozsoyoglu (2007) present a method for finding pieces of a certain pathway in a new organism by abstracting the known instances of the pathway into functionality templates based on enzyme GO functions. Dale *et al.* (2010) compare different machine learning methods for finding known pathways in new organisms based on a large set of features. The features are functions of the pathway that is being searched for as well as of the organism in which the pathway's presence is being predicted. Some of these features are genome context features that reflect the relative location of the genes involved in each pathway in the target genome. Another method that uses genome context information for tasks related to pathway discovery is presented by Green and Karp (2007). They use various genome context methods (gene reaction co-adjacency, gene neighbors, gene clusters, gene fusions and phylogenetic profiles) to identify missing enzymes in predicted metabolic pathways and to predict functional relationships.

To our knowledge, all pathway discovery methods in the literature that aim to find novel pathways rely, at least partially, on interaction networks obtained from (mostly high throughput) laboratory experiments (Kelley and Ideker, 2005; Ma *et al.*, 2008; Zhang and Ouellette, 2010). On the other hand, as discussed above, some methods aimed at finding known pathways in new organisms have been proposed that use genome context information, although, in most cases, they also rely on experimental information. Furthermore, genome context methods have been used, along with other information, to infer functional associations between proteins (see for example, Jansen *et al.*, 2003; Lu *et al.*, 2005; Marcotte *et al.*, 1999b; Yamanishi *et al.*, 2005). Nevertheless, no further efforts toward finding pathways or functional groups from this data are made in those cases.

In this work, we present a method for discovery of subsystems based on genome context methods. We consider a subsystem to be a set of genes that work together within a biological pathway or a biological process, or that form a multimeric complex.

*To whom correspondence should be addressed.

Our system requires a collection of sequenced genomes with structural annotation (location of genes). Note that, for bacterial genomes, this kind of annotation is usually done automatically without the need for manual curation. Our method can be applied to any new bacterium, regardless of the amount of experimental data available. We say our system finds subsystems because we have found it to yield not just members of known metabolic pathways, but also genes that function within the same multimeric protein complex, and within other collections of genes that operate in a common biological process. The results obtained from this system can be used to guide future experimental research.

As an essential part of the method, we propose a measure of the likelihood of a certain group of genes being part of the same subsystem, which we refer to as their 'group score'. This measure is computed as the number of organisms in a large, carefully chosen, database of reference genomes that are enriched for the gene group. We show that the probability of finding more than one enriched organism for a random set of genes is 2.2%, while it is 67% for known functional gene groups. This measure can be used to filter out predicted groups, retaining only those for which at least one enriched organism is found.

The novelty of the proposed method lies in a few key aspects: (i) as explained above, while most previous approaches to novel pathway discovery rely on large experimental interaction datasets, our method relies exclusively on sequence information that can be extracted from structurally annotated genomes alone; (ii) the use of a very fast clique-finding algorithm for the construction of gene groups that are likely to be functional groups; and (iii) our proposed filtering step based on the number of reference organisms enriched for the group.

We describe the method in detail and then discuss a selection of gene groups that were uncovered with this method, focusing on how they point to the potential applications of the method in guiding future biological research.

2 METHOD

In this section, we describe the three main components in the proposed system and briefly describe the final stage in which the results of the system are collected and enriched with all available information to generate html reports. Figure 1 shows an overview of the method. We call each group of genes generated by the method a *candidate* subsystem (or just candidate for short), since they are hypotheses that need to be confirmed by experimental methods.

2.1 Generation of pairwise scores

The inputs to the subsystem discovery method are a large list of bacterial reference genomes and the target genome for which gene subsystems are to be estimated. The first step in the method is the generation of a score for each pair of genes in the target genome that estimates the likelihood that two genes participate in a same subsystem, that is, that they are part of the same pathway, biological process or multimeric complex. These scores are computed using genome context methods. These methods rely exclusively on sequence similarity information between genes in the target genome and genes in the reference genomes. Four main genome context methods have been introduced in the literature: (i) the phylogenetic profile method (Pellegrini *et al.*, 1999), which uses the pattern of presence or absence of the genes in the pair across the reference genomes to generate a score; (ii) the gene neighbor method (Bowers *et al.*, 2004) that considers the distance between the homologs of the genes in the pair across the genomes in which both genes are found; (iii) the gene cluster method (Overbeek *et al.*, 1999), which simply computes the distance between genes in the target genome; and (iv) the Rosetta Stone or gene fusion method (Enright *et al.*, 1999; Marcotte *et al.*, 1999a), which looks for genes in the reference genomes that are likely to be a fusion of the two target genes.

In a recent paper (Ferrer *et al.*, 2010), we presented a thorough comparative study of these genome context methods and concluded that the gene neighbor method is significantly better than the phylogenetic profile method in most tested organisms, outperforming it by around 40% at most operating points of interest in *Escherichia coli K-12* gene pairs. The Rosetta Stone method gives significantly worse performance than any of the other three methods. The gene cluster method, while competitive to the gene neighbor method at very high specificity values, is not able to generate

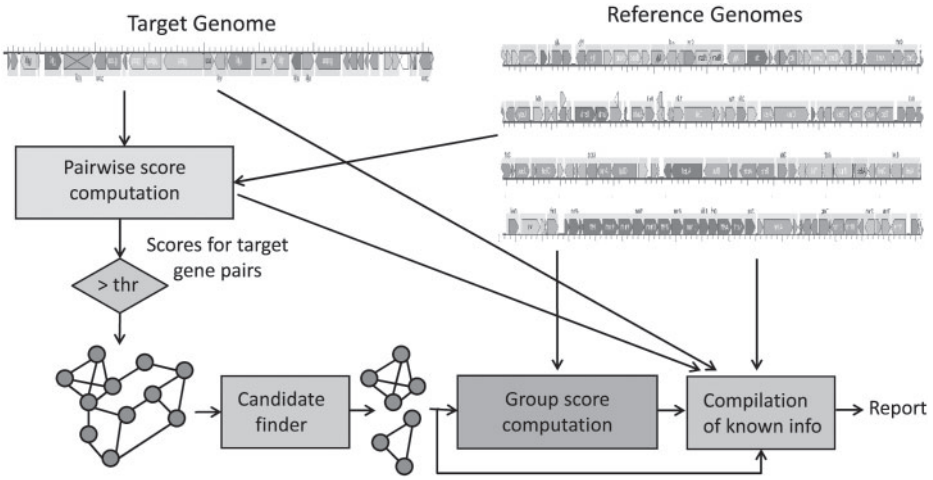

Fig. 1. Overview of the subsystem discovery method. The inputs to the system are a large list of bacterial reference genomes and the target genome for which candidate subsystems are to be created. Each genome should be structurally annotated, and higher quality annotations should yield better results from our method. For each gene pair in the target genome, a pairwise score is computed. A network is then created where the nodes are the genes and two nodes are connected by an edge if the pairwise score for the corresponding gene pair is larger than a certain threshold. A first set of candidate gene groups is then determined by finding sets of genes that are fully connected to each other. A group score is then computed and used to filter the resulting candidates. Finally, the information obtained from all previous steps and other information that can be gathered about the group is collected into html reports.

scores for all possible gene pairs since it is defined as the distance (in bases) between adjacent genes that are coded in the same strand. Gene pairs that are not adjacent and coded in the same strand lack gene cluster scores. Furthermore, we found that the combination of these methods leads to only small and unreliable gains. Hence, in this article, we use the gene neighbor method to generate the pairwise scores. Nevertheless, note that any other score could be used here, including other genome context scores or scores from experimental studies as long as scores are available for all (or most) gene pairs in the target genome.

Briefly, the gene neighbor method generates a summary measure of the observed distances between the best matching sequence for each gene in a pair of genes (G_i, G_j) from the target genome across the list of reference genomes. For each genome in the reference list, the best homolog of each gene G_i and G_j in the pair is obtained. The best homolog for gene G is determined by finding the gene in the reference genome with the smallest E-value (as returned by BLAST). If no E-value for G is below 10^{-4}, it is considered that no homolog exists in the reference genome for this gene. For genomes in which both target genes had a homolog, the distance (measured in number of intervening genes) between these homologs is recorded. Finally, a score is computed as a P-value for the observed distances. See Bowers *et al.* (2004) or Ferrer *et al.* (2010) for details on this method and experimental results.

The output of this stage is a number for each gene pair (G_i, G_j) that is assumed to be monotonically related with the likelihood that the two genes participate in the same subsystem. For this, the negative of the logarithm of the P-value described above is used as a score (since smaller P-values indicate higher likelihood of two genes participating in the same subsystem). We call these numbers *pairwise scores*.

2.2 Finding candidate gene groups

Given the pairwise scores, a network can be created where the nodes are the genes in the target genome and where edges are added between each pair of genes for which the pairwise score is larger than a certain threshold. We define a candidate gene group to be any set of nodes (genes) that are fully connected to each other. That is, a candidate is composed of genes that are likely to be related to all other genes in the candidate. This is a strong constraint to impose, though a reasonable one if we are interested only in high-confidence groups. Furthermore, this constraint translates into a very common task in network analysis: finding all maximal cliques. Several algorithms have been designed to solve this task in a computationally efficient way. We use a python implementation of the Bron–Kerbosch algorithm, version 2 (Bron and Kerbosch, 1973).

The output of this stage is a set of gene groups, each group being a potential subsystem.

2.3 Selection using group scores

In this step, the groups obtained in the previous step are filtered to keep only high-confidence groups. A group-level score is computed that is used as a confidence measure for the group. Only groups with a group score larger than a certain threshold are kept. The group score is computed as the number of organisms in a reference list of bacterial genomes for which the group is significantly enriched.

Given a certain candidate group, we search for the genes within this group in the set of reference organisms and compute the enrichment P-values in each organism. The P-value for a candidate in a certain reference genome is computed using the formula

$$p = \sum_{i=n_r}^{n} \frac{\binom{n}{i}\binom{t-n}{r-i}}{\binom{t}{r}},$$

where t is the total number of genes in the target genome, r is the total number of genes in the target genome that have homologs in the reference genome, n is the number of genes in the candidate and n_r is the number of genes in the candidate that have homologs in the reference genome. This formula computes the probability of the number of genes in the candidate that have homologs in the reference genome being larger or equal than the observed value n_r. Each term in the sum is given by the hypergeometric distribution, where $\binom{x}{y}$ stands for the 'x choose y' operation. We use Bonferroni correction on this P-value to address the problem of multiple comparisons. That is, the final P-value is given by $p*N$, where N is the total number of organisms that were explored for enrichment.

Finally, the group score is the number of organisms enriched for the group. The P-value threshold used to determine enrichment is a parameter of the method. The performance of these scores as a measure of the likelihood that the group is, in fact, a subsystem is presented in Section 3.2.

2.4 Final selection and generation of reports

The candidates generated by the steps above might include groups that are not of interest and, hence, can be discarded from the output shown to the user. For example, they might include groups that are known subsystems for the organism. A scientist who is searching through the list for novel subsystems might wish to ignore these already-known groups. In the case of *E.coli K-12*, a large proportion of the candidates have an already-known function. This proportion is, of course, much smaller in organisms for which less curation is available. Another possible criterion for filtering candidates is to keep only groups that contain a certain set of genes of interest. For example, if the goal was to find metabolic pathways, one might decide to keep only groups that contain some percent of known enzymes. For the results presented in this paper, no such selection procedure is applied.

The last step in the process takes the filtered set of candidates, gathers all information about the genes in the groups that is present in our database (products and their functions, and reactions in which they are involved), infers new information using computational methods (including enrichment and depletion P-values for the group in the reference organisms, correlation between the group and known pathways based on their pattern of appearance in the reference organisms, and so on) and formats all of this information into a set of web pages. The goal is for these pages to record as much information as possible to allow the scientist looking through the lists to make informed decisions about which candidates are worth looking at in more detail. A manual describing the information available in the web pages is available in the website indicated in the Abstract. This website also contains a set of precomputed lists of candidates for *E.coli K-12*.

3 RESULTS

This section presents a set of experimental results obtained by applying our method to identify novel subsystems in *E.coli K-12*. We used EcoCyc version 14.5 (Keseler *et al.*, 2009) which contains 944 protein complexes and 235 metabolic pathways curated from the *E.coli K-12* literature, of which we only consider groups of size >3 for a total of 103 and 112 complexes and pathways, respectively. The list of reference organisms is the set of 623 bacterial genomes available in BioCyc 14.5 (Caspi *et al.*, 2008; Karp *et al.*, 2005). We show three sets of results. The first two sets evaluate the performance of the pairwise and group scores, and the last result set evaluates the performance of the system as a whole.

3.1 Performance of pairwise scores

We present the results for the gene neighbor method used to generate the pairwise scores for the current implementation of the system. The test set for these experiments is created by listing all pairs of genes in EcoCyc and labeling each pair as a positive sample if the products of the two genes are listed in the same pathway or protein complex in this database, and as a negative sample otherwise. A thorough study of genome context score methods, including the gene

neighbor method, can be found in Ferrer *et al.* (2010). In that work, we found that pruning the reference list of organisms to discard highly related ones resulted in an improvement in performance for most genome context methods. Hence, the gene neighbor scores used in this article are generated using a pruned list of reference organisms of size 468. This list is obtained by hierarchical clustering of organism profiles (vectors where each component corresponds to a gene in *E.coli K-12* and the entries are 1 if the gene has a homolog in the organism and 0 otherwise) from the original list of 623 bacterial genomes (see Ferrer *et al.*, 2010, for details). Note that this clustering method is probably not optimal, since organisms that are distant from *E.coli K-12* might look closer to each other than they really are. In future work, we plan to explore more general clustering methods that avoid this bias. Both reference genome lists are given in the website indicated in the Abstract.

Table 1 shows the recall and precision of the gene neighbor methods at three different operating points. An operating point is determined by the score threshold used to declare two genes as belonging to the same subsystem. Each threshold, in turn, results in a network where edges are given between two genes if the pairwise score for the gene pair (given by the negative of the logarithm of the gene neighbor *P*-value) is larger than the threshold. We specify the operating point by the percent of edges in the resulting network with respect to the total number of edges if the network was fully connected. For example, in a network with four nodes, the fully connected network would have six (four choose two) edges. If the network instead has three edges, the percent of edges would be 50%.

Given a certain threshold, the number of true positives (*tp*) and true negatives (*tn*) can be obtained by counting the number of gene pairs with score higher than the threshold (existing edges) that participate in a common subsystem in our test set, and the number of gene pairs with score lower than the threshold (missing edges) that do not take part in any common subsystem, respectively. We measure the performance of the scores using recall and precision, which are defined as $tp/dp*100$ and $tp/pp*100$, respectively, where *dp* is the total number of positive samples in the database and *pp* is the total number of predicted positive pairs.

We present results for three networks: one with 0.07% edges (only one gene pair out of 1400 is connected by an edge), one with 0.15% edges (one pair out of 666 is connected by an edge) and one with 2% edges (one pair out of 50 is connected by an edge). The corresponding threshold on the gene neighbor *P*-value is also shown. The first setting coincides with the percent of edges in the actual *E.coli K-12* network obtained by putting edges between any two genes that are found in the same pathway or protein complex in EcoCyc. The ROC for the gene neighbor method on different subsets of the *Escherichia coli K-12* genes can be found in Ferrer *et al.* (2010).[1]

We can see that, for the first setting corresponding to a sparser network, 26.94% of the predicted pairs are correct with respect to both pathways and protein complexes. This percent is almost halved for the second setting and 16 times smaller for the third setting. At these levels of precision, the recall of these scores is around 26, 35 and 47%, respectively. We can see that the increase in recall comes at the expense of a drastic increase in the percent of incorrect predictions. Given that the number of negative samples in the database is around 8 million, these percents correspond to around 4000, 12 000 and 150 000 mislabeled negative samples. Meanwhile, the numbers of correctly labeled positive samples are 1600, 2146 and 2874, since the total number of positive samples in the database is 6115. The gain in recall for the higher setting does not seem to justify the sharp increase in the number of mislabeled negative samples which would result in a large proportion of false subsystems. For this reason, in Section 3.3 we show performance for only the first two, sparser, settings.

Finally, note that the gene neighbor method is able to find protein complexes much better than pathways. That is, genes involved in protein complexes are more likely to be highly conserved across the reference organisms than genes involved in pathways.

Table 1. Recall (percent of true positive samples found) and precision (percent of predicted positives that are correct) for the gene neighbor scores at three different operating points determined by three different thresholds each of which results in a certain percent of edges in the graph

Edges (%)	Threshold	Recall (%)			Precision (%)		
		All	pwy	cplx	All	pwy	cplx
0.07%	1.7e-61	26.41	15.13	42.35	26.94	8.14	21.93
0.15%	3.9e-28	35.10	21.36	53.43	16.70	5.36	12.91
2.00%	1.4e-05	47.08	34.51	63.54	1.68	0.64	1.15

Recall and precision are computed with respect to three set of labels: considering pairs coming from pathways only, protein complexes only or both types of groups as true positives. Results are computed over all pairs of genes from *E.coli K-12*. The thresholds are shown in terms of gene neighbor *P*-values before transforming them to scores with the negative logarithm operation.

3.2 Performance of group scores

We evaluate the performance of the group enrichment score described in Section 2.3 using the known pathways and protein complexes in EcoCyc version 14.5. The test database is generated using the following procedure. For each known group (pathway or protein complex) in the EcoCyc database, positive samples are created as random subsets of the genes in the group. For each subset size, up to 10 samples are created (or less, if fewer than 10 combinations of genes of the corresponding size are possible). Negative samples are created as random sets of different sizes of all the genes in EcoCyc. The distribution of negative group sizes follows the distribution of positive groups sizes: for each created positive sample, 10 negative samples of the same size are randomly created. The resulting database contains 130 730 negative samples and 13 073 positive samples. For more details on the generation of this database, see the Supplementary Material.

For each gene group in this database, we compute the number of organisms in the reference list that are enriched for the gene group. We explored the use of different reference lists and different *P*-value thresholds for the determination of which organisms are enriched. As shown in the Supplementary Material, the best settings corresponds to using a pruned set of 468 reference organisms (obtained as described in the previous section) with a *P*-value threshold of 0.10. Using these parameters, we find that no negative sample has a number of enriched organisms >25. The percent of positive samples for which there are at least that many enriched organism is 34%. That

[1]Results in Ferrer *et al.* (2010) are presented on a subset of genes for which we have higher confidence on the labels. Table 1 shows the results on all genes. This is why results in the article and the table are not identical.

is, if we set the third step of our method to filter out all candidates with a number of enriched organisms <25, we can expect to discard most (if not all) of the negative samples and keep around 34% of the positive ones. In the other extreme, we can choose to keep all candidates that have at least one enriched organism in the reference list of size 468. This would result in 2.2% of the negative samples and 67% of the positive samples being accepted as final candidates.

3.3 Performance of subsystem candidates

We measure the performance of the candidates resulting from the subsystem prediction method using two metrics: the percent of correct candidates defined as the percent of candidates that match at least one known pathway or protein complex in EcoCyc, and the percent of groups found defined as the percent of pathways and protein complexes in EcoCyc that match at least one candidate.

A pathway or protein complex and a candidate are considered to match if their adjusted rand index (ARI) is larger than 0.5. This is the same definition used by Zhang and Ouellette (2010), which facilitates comparison. ARI scores are commonly used to determine agreement between partitions in a graph. They lie between 0 and 1 for any two groups that overlap in at least one gene, being 1 when the two gene groups are identical. See Zhang and Ouellette (2010) for the formula used to compute the ARI score.

We focus on pathways, protein complexes and candidates of size 4 or larger. The method reduces to the genome context method alone for groups of size 2. Furthermore, for these smaller groups of size 2 and 3 the enrichment *P*-values are not useful, since they rarely reach values small enough to consider an organism enriched.

Table 2 shows the two performance measures and the number of candidates generated for different settings of the gene neighbor score threshold and the minimum number of enriched organisms considering performance with respect to both pathways and protein complexes, to pathways only and to complexes only. We can see that the parameters of the system have significant effects on the performance. Nevertheless, there is no clear winner. The selection of parameters should be made based on the way the candidates will be used. For example, if the goal of running the system is to find a promising group of genes that will be further evaluated in the laboratory, we might wish to set the parameters to obtain only high-confidence candidates (for example, % edges = 0.07% and # enr = 30). This results in around half of the candidates being correct, but it finds very few of the known pathways and protein complexes. On the other hand, if the purpose for running the system is to complement knowledge obtained from a separate source, such as large-scale gene expression or proteomics results, then a large number of candidates might be desirable, since additional filtering criteria can be applied based on the original source of information. Finally, note that the threshold on the number of enriched organisms does not necessarily need to be selected *a priori* by the user of the system. By setting the threshold to 0, all candidates for a certain % edges can be generated and the number of enriched organisms can then be used as a confidence measure to prioritize experiments or as a score to be combined with measures coming from other sources.

An interesting observation from this table is that protein complexes are more easily found by our proposed method than pathways. This is probably mostly due to the fact that the gene neighbor method is able to detect gene pairs involved in the same protein complex better than those involved in the same pathway (Table 1). Relaxing the constraint that groups have to be fully connected by the gene neighbor scores (enforced by the clique finding algorithm) could improve our performance on pathways and is part of our planned work for the near future.

Table 2. Performance of the subsystem discovery method

Edges (%)	# enr.	# cand.	Correct (%)			Found (%)		
			All	pwy	cplx	All	pwy	cplx
0.15%	0	1130	17.9	8.1	10.6	55.8	50.9	61.2
	1	472	26.9	12.3	16.1	40.9	38.4	43.7
	10	206	33.0	14.6	20.4	20.5	17.9	23.3
	30	111	39.6	16.2	26.1	12.6	13.4	11.6
0.07%	0	413	24.9	8.0	18.9	41.9	25.9	59.2
	1	215	31.6	10.7	23.7	31.2	19.6	43.7
	10	110	32.7	12.7	24.5	15.8	12.5	19.4
	30	56	42.9	17.9	30.4	11.2	9.8	12.6

Columns correspond to the percent of edges in the network [edges (%)], the minimum number of enriched organisms (# enr.), the number of candidates output by the system (# cand.), the percentage of correct candidates [correct (%)] and the percentage of gene groups (pathways or protein complexes) in EcoCyc that are found by the system [found (%)]. For correct (%), we show three numbers: considering matches to both pathways and protein complexes, to pathways only and to complexes only (these last two numbers do not add up to the first because some candidate groups match both complexes and pathways). For found (%), we also show three numbers: the percentage of both pathways and complexes from EcoCyc that are found, the percentage of pathways only and the percentage of complexes only.

The results in Table 2 may seem to contradict those presented in the previous section. As we discussed, the enrichment scores are highly reliable: no random gene group has >25 enriched organisms. Nevertheless, Table 2 shows that when we set the minimum number of enriched organisms to 30, we still get a percent of incorrect candidates >50%. This contradiction is partially explained by the fact that the negative samples in Section 3.2 were random groups of genes, while the ones in Table 2 are chosen for having high gene neighbor scores for all pairs in the group. Hence, the negative samples in this table are a small subset of all possible negative samples and they are, by the very definition of the gene neighbor method, more likely to result in higher enrichment scores. On the other hand, based on the study of a few high-confidence examples, we believe that this apparent contradiction is also partially explained by the fact that many of the candidates with a large number of enriched organisms are, indeed, subsystems—but not the kind that EcoCyc formally represents, and therefore not detectable as true positives. Other candidates may fail to match any protein complexes or pathways in EcoCyc for the very simple reason that they represent correctly identified novel gene groups for which no experimental work has been done yet. Both possibilities are discussed further in the following section.

Comparing the results of subsystem prediction methods presented in the literature is a complicated matter due to the lack of a standard test set and standard performance measures. The closest comparison we have found is with Zhang and Ouellette (2010). They obtain a 5.6% of correct prediction on the BioCyc database out of 195 predicted pathways. This number should be compared with the *p*wy column under % correct in Table 2 for a similar number of candidates. For two different setting of the algorithm, we get 12.56 and 16.99% correct, an improvement larger than two times over

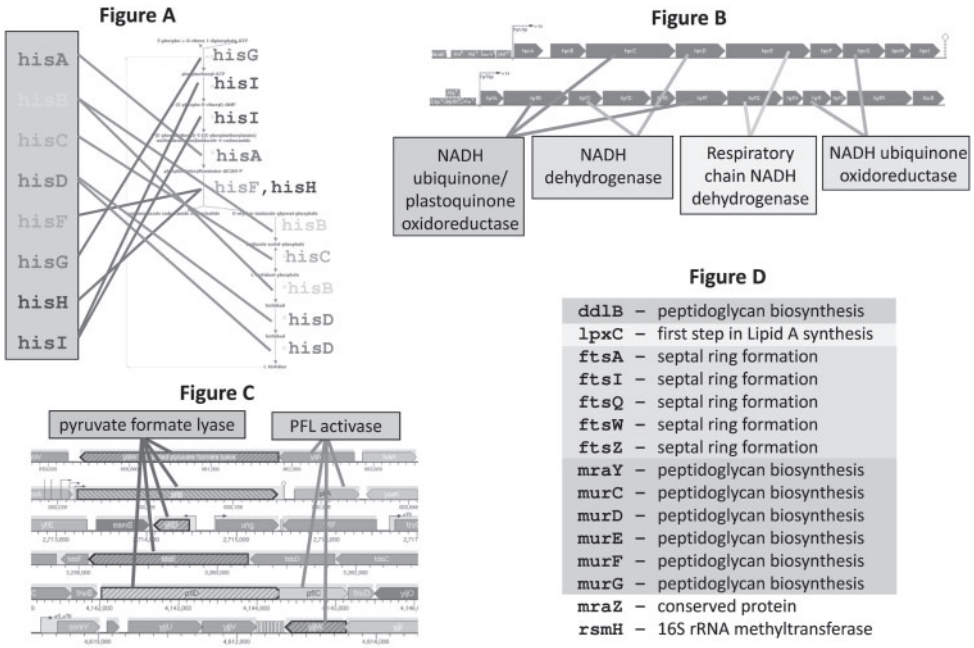

Fig. 2. Genome context analysis successfully identifies known gene groups and likely novel subsystems. (**A**) Many known pathways are completely identified via genome context analysis. Here, a candidate group is revealed to be the complete histidine biosynthesis pathway. The histidine pathway genes are shown in operon order on the left, and as they map onto pathway steps on the right. (**B**) A 'hydrogenase core' is identified, containing the key genes in both predicted hydrogenase operons in *E.coli K-12*. Pictured are the hyc operon (top) and hyf operon (bottom). Each operon contains representative genes from each of four matched gene superfamilies. The lines from each colored box show where a given gene superfamily member appears at least once in each operon. (**C**) Novel groups such as this highly conserved set of pyruvate formate lyase (orange) and PFL activase (blue) orthologs can assist in assigning more accurate gene functions and directing future research. Genes are shown in their local genome context, as captured using the EcoCyc Genome Browser. (**D**) This combined group of septal genes (green), peptidoglycan synthesis genes (blue) and lipid A synthesis genes (yellow) are a clear subsystems, but are not represented formally in EcoCyc. The appearance of mraZ in this group also suggests a role for this largely uncharacterized gene in cell division.

their algorithm (which was, in turn, shown to be better than other methods in the literature).

4 DISCUSSION

In applying our method to the biology of *E.coli K-12* as it is represented in EcoCyc, we have the opportunity to 'discover' known groups, some that are not currently formally represented, and some that are likely to be entirely novel, in addition to rediscovering groups already represented in EcoCyc.

Groups mentioned in the following discussion were obtained with the network with % edges = 0.07%. The candidate numbers mentioned in this section correspond to those used in the website indicated in the Abstract for the corresponding % edges. They can all be found in the link corresponding to # enr = 0, which is a superset of all the other sets of candidates with larger # enr.

4.1 Rediscovered known subsystems

As described above, we identify from 11% to 55% of the protein complexes and pathways that are formally described within the EcoCyc database depending on the settings of the system. Examples of this kind of 'perfect match' include the histidine biosynthesis pathway, tryptophan biosynthesis and the ATP synthase complex. The candidate that mapped to histidine biosynthesis (number 0073 in the web page corresponding to % edges = 0.07% and # enr = 0) included all eight genes coding for the enzymes in the histidine biosynthesis pathway (Fig. 2A). This is representative of a typical 'perfect fit' match between a candidate and the EcoCyc data.

The best percent found value of 55% (Table 2) may seem low at first, but in the context of what is actually within our test set it actually makes a great deal of sense. Pathways and protein complexes are both conceptual units, identified by human researchers and entered into databases such as EcoCyc by human curators. Although the reasoning behind the selection of this kind of subsystem is typically quite sound—for example, the histidine biosynthesis pathway in EcoCyc starts with a reaction that is subject to feedback regulation by histidine and ends with a reaction that produces histidine—it is not a given that this kind of human-defined conceptual unit is also evolutionarily conserved (Green and Karp, 2006). In other words, the group that is predicted based on genome context information may not precisely match the subsystem that exists within our databases, leading to a 'failure to find' those human-defined groups.

One fascinating example of this kind of mismatch between our concept of a pathway and the results of this method is a candidate we call the 'core hydrogenase pair' (Fig. 2B, candidate number 0056 in the web page corresponding to % edges = 0.07% and # enr = 0). This candidate counterintuitively collects half the genes from each of two putative hydrogenase complexes in *E.coli K-12* (Andrews *et al.*, 1997; Bagramyan *et al.*, 2001; Bohm *et al.*, 1990; Self *et al.*, 2004). A close, protein domain-centered analysis of the results, however,

reveals that this group appears to comprise a pair of redundant 'hydrogenase cores', rigorously conserved across the group of enriched organisms. The experimental support for the specific composition of the hyc (hydrogenase 3) and hyf (hydrogenase 4) complexes as represented in EcoCyc 14.5 is sparse. This candidate group may point to a conserved hydrogenase core and an array of more organism-specific adapters, indicating a possible direction for future evaluation in the laboratory.

4.2 Known, but not represented, subsystems

Some of the candidates represent the successful identification of known subsystems that happen not to be represented formally within our test dataset. These cases are incorrectly counted as errors in our statistics in Table 2. The 'cell division' candidate (Fig. 2D, candidate number 0009 in the web page corresponding to % edges = 0.07% and # enr = 0) clearly contains the components of the septal ring, a critical structure that forms during cell division, but which is not represented formally within EcoCyc (Weiss, 2004). It also maps quite well onto the *division cell wall* (dcw) cluster of genes that has been identified as a set of conserved cell division genes (Mingorance *et al.*, 2004; Real and Henriques, 2006; Vicente *et al.*, 1998).

This candidate also represents another potential opportunity in this type of result—what could be called the 'N+1' result. In the case of the candidate shown in Figure 2D, the majority of the genes are clearly directly involved in the structural aspects of cell division, with the majority being split between peptidoglycan biosynthesis (blue) and septal ring formation (green). The gene *rsmH*, which is not in either of these groups, is active against membrane-located substrates and is at least indirectly required for proper cell division, with disruption leading to an alteration in doubling times (Carrion *et al.*, 1999; Kimura and Suzuki, 2010). However, the group also includes the gene *mraZ*, which has been the subject of crystallization and attempted enzyme analysis, but for which a specific cellular role is unknown (Adams *et al.*, 2005). The strong association between *mraZ* and this subsystem suggests narrowing future analysis of this gene to first examine a potential role in cell division, with a specific emphasis on peptidoglycan and the septal ring. This is generally representative of how such results have the potential to reduce the extensive time and effort spent trying to assign functions to the increasingly recalcitrant set of uncharacterized genes.

A related case is represented by the molybdopterin biosynthesis candidate, comprising *moaA*, *moaB*, *moaC* and *moaE* (candidate number 0306 in the web page corresponding to % edges = 0.07% and # enr = 0). These genes have been experimentally identified as forming a pathway for molybdopterin biosynthesis, but the steps in that pathway have not been fully elucidated (Gutzke *et al.*, 2001; Leimkuhler *et al.*, 2001). As a consequence, there is insufficient information to include a formal pathway in EcoCyc. Nonetheless, this is a clear subsystem. The prediction of this group actually offers some assistance to experimental elucidation of the pathway, as it can help constrain the set of genes that must be examined as biochemists attempt to pin down the enzymatic steps involved in making molybdopterin.

4.3 Likely novel subsystems

Of the candidates that do not match known EcoCyc pathways and complexes, some are likely to be novel subsystems. It would be curious if this were not the case, as ∼30% of *E.coli K-12* genes still lack experimental functional characterization.

For example, one predicted group combines apparently disparate genes that are not found in the same operon within *E.coli K-12* (candidate number 0127 in the web page corresponding to % edges = 0.07% and # enr = 0). These include two genes known to be required for anaerobic carnitine reduction (*fixA* and *fixB*), two that are predicted to code for acyl-CoA dehydrogenase subunits (*ydiQ* and *ydiR*) and two predicted flavoproteins (*ygcR* and *ygcQ*). Taken together, this may represent a previously uncharacterized electron transfer pathway. As these genes occur in three different places in the *E.coli K-12* genome, the group is not likely to be obvious by a simple examination of the genome.

The 'PFL' candidate (Fig. 2C, candidate number 0061 in the web page corresponding to % edges = 0.07% and # enr = 0) is another example of scattered genes that are clearly coordinated when viewed through the lens of our approach. They are functionally related in terms of individual sequence, most likely representing a history of ancestral gene duplications. Despite these duplications, this candidate displays a surprising tenacity, appearing as a conserved group across a large number of organisms even though its members are often very far from each other on a given genome. Many of the constituent genes are not well characterized in *E.coli K-12*, but researchers may benefit from approaching them as a system, in light of their identification as a conserved group.

5 CONCLUSIONS

We propose a method for the discovery of novel subsystems in a target genome. The method relies exclusively on sequence information and, hence, only requires structurally annotated genomes as input, not needing any other experimental or manually curated information. The system can be tuned to generate few high-confidence gene groups or a larger number of lower confidence groups. The output of the system is stored in a set of web pages that integrate a large amount of information about the gene groups.

We assessed the performance of the method on *E.coli K-12*, using the manually curated information about pathways and protein complexes available in EcoCyc version 14.5. We find that, in one particular setting of the parameters, the system finds 31.2% of all known pathways and protein complexes in EcoCyc that include at least four genes, and that 31.6% of all the candidate groups of size 4 or larger generated by the system match a known pathway or protein complex closely. Careful analysis of the highest confidence groups obtained for *E.coli K-12* leads us to conclude that a large percent of the remaining groups is likely to consist of a mix of known groups that were not formally represented in our test data and novel subsystems that have not yet been evaluated in the laboratory.

These novel and 'near match' predictions can serve as a valuable resource in guiding future laboratory research, such as the effort to characterize the remaining 30% of genes of unknown function in *E.coli K-12*. Predictions can help constrain the size of the experimental problem, as in the case of molydopterin synthesis, identify conserved subsystems that may be larger or smaller than we expected, as was seen with the mixed *hyc/hyf* group, or even identify scattered subsystems, such as the possible electron transport pathway, that are nearly impossible to identify through manual examination of a genome. Although changing the parameters of the method can lead to different results, the ability to focus

initial experiments on a handful of genes instead of all of them represents a significant opportunity for savings in time and effort at the bench. These predictions, most likely in combination with high-throughput methods such as gene expression, proteomics or phenotype microarrays, can also provide guidance in addressing the otherwise opaque problem of assigning function to uncharacterized genes.

Funding: National Library of Medicine (award number R01LM009651); SRI International. The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of SRI International, the National Library of Medicine, or the National Institutes of Health.

Conflict of Interest: none declared.

REFERENCES

Adams,M.A. *et al.* (2005) Mraz from escherichia coli: cloning, purification, crystallization and preliminary x-ray analysis. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.*, **61**(Pt 4), 378–380.

Andrews,S.C. *et al.* (1997) A 12-cistron escherichia coli operon (hyf) encoding a putative proton-translocating formate hydrogenlyase system. *Microbiology*, **143 (Pt 11)**, 3633–3647.

Bagramyan,K. *et al.* (2001) Participation of hyf-encoded hydrogenase 4 in molecular hydrogen release coupled with proton-potassium exchange in escherichia coli. *Membr. Cell. Biol.*, **14**, 749–763.

Bohm,R. *et al.* (1990) Nucleotide sequence and expression of an operon in escherichia coli coding for formate hydrogenlyase components. *Mol. Microbiol.*, **4**, 231–243.

Bowers,P. *et al.* (2004) Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol.*, **5**, R35.

Bron,C. and Kerbosch,J. (1973) Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM*, **9**, 575–âŁ“577.

Cakmak,A. and Ozsoyoglu,G. (2007) Mining biological networks for unknown pathways. *Bioinformatics*, **23**, 2775–2783.

Carrion,M. *et al.* (1999) mraw, an essential gene at the dcw cluster of escherichia coli codes for a cytoplasmic protein with methyltransferase activity. *Biochimie*, **81**, 879–888.

Caspi,R. *et al.* (2008) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **36**, D623–D631.

Dale,J. *et al.* (2010) Machine learning methods for metabolic pathway prediction. *BMC Bioinformatics*, **11**, 15.

Enright,A. *et al.* (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.

Ferrer,L. *et al.* (2010) A systematic study of genome context methods: calibration, normalization and combination. *BMC Bioinformatics*, **11**, 493.

Green,M. and Karp,P. (2006) The outcomes of pathway database computations depend on pathway ontology. *Nucleic Acids Res.*, **34**, 3687–3697.

Green,M.L. and Karp,P.D. (2007) Using genome-context data to identify specific types of functional associations in pathway/genome databases. *Bioinformatics*, **23**, i205–i211.

Gutzke,G. *et al.* (2001) Thiocarboxylation of molybdopterin synthase provides evidence for the mechanism of dithiolene formation in metal-binding pterins. *J. Biol. Chem.*, **276**, 36268–36274.

Jansen,R. *et al.* (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449–453.

Karp,P. *et al.* (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.*, **33**, 6083–6089.

Kelley,R. and Ideker,T. (2005) Systematic interpretation of genetic interactions using protein networks. *Nat. Biotech.*, **23**, 561–566.

Keseler,I. *et al.* (2009) EcoCyc: a comprehensive view of *E. coli* biology. *Nucleic Acids Res.*, **37**, D464–D470.

Kimura,S. and Suzuki,T. (2010) Fine-tuning of the ribosomal decoding center by conserved methyl-modifications in the escherichia coli 16s rrna. *Nucleic Acids Res.*, **38**, 1341–1352.

Leimkuhler,S. *et al.* (2001) Characterization of escherichia coli moeb and its involvement in the activation of molybdopterin synthase for the biosynthesis of the molybdenum cofactor. *J. Biol. Chem.*, **276**, 34695–34701.

Lu,L. *et al.* (2005) Assessing the limits of genomic data integration for predicting protein networks. *Genome Res.*, **15**, 945–953.

Ma,X. *et al.* (2008) Mapping genetically compensatory pathways from synthetic lethal interactions in yeast. *PLoS One*, **3**, e1922.

Marcotte,E. *et al.* (1999a) Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751–753.

Marcotte,E.M. *et al.* (1999b) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.

Mingorance,J. *et al.* (2004) Genomic channeling in bacterial cell division. *J. Mol. Recog.*, **17**, 481–487.

Overbeek,R. *et al.* (1999) Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol.*, **1**, 93–108.

Pellegrini,M. *et al.* (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci.*, **96**, 4285–4288.

Real,G. and Henriques,A.O. (2006) Localization of the bacillus subtilis murb gene within the dcw cluster is important for growth and sporulation. *J. Bacteriol.*, **188**, 1721–1732.

Self,W.T. *et al.* (2004) Expression and regulation of a silent operon, hyf, coding for hydrogenase 4 isoenzyme in escherichia coli. *J. Bacteriol.*, **186**, 580–587.

Vicente,M. *et al.* (1998) Regulation of transcription of cell division genes in the escherichia coli dcw cluster. *Cell. Mol. Life Sci.*, **54**, 317–324.

Weiss,D.S. (2004) Bacterial cell division and the septal ring. *Mol. Microbiol.*, **54**, 588–597.

Yamanishi,Y. *et al.* (2005) Supervised enzyme network inference from the integration of genomic data and chemical information. *Bioinformatics*, **21**, i468–i477.

Zhang,K.X. and Ouellette,B.F. (2010) Pandora, a pathway and network discovery approach based on common biological evidence. *Bioinformatics*, **26**, 529–535.