Method

Open Access

Inference of protein function and protein linkages in *Mycobacterium tuberculosis* based on prokaryotic genome organization: a combined computational approach

Michael Strong*, Parag Mallick*, Matteo Pellegrini†, Michael J Thompson† and David Eisenberg*

Addresses: *Howard Hughes Medical Institute, UCLA-DOE Institute for Genomics and Proteomics, Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90095-1570, USA. †Protein Pathways, 21111 Oxnard Street, Woodland Hills, CA 91367, USA.

Correspondence: David Eisenberg. E-mail: david@mbi.ucla.edu

Published: 29 August 2003

Genome ***Biology*** 2003, **4:**R59

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/9/R59

Received: 21 March 2003
Revised: 11 July 2003
Accepted: 28 July 2003

Abstract

The genome of *Mycobacterium tuberculosis* was analyzed using recently developed computational approaches to infer protein function and protein linkages. We evaluated and employed a method to infer genes likely to belong to the same operon, as judged by the nucleotide distance between genes in the same genomic orientation, and combined this method with those of the Rosetta Stone, Phylogenetic Profile and conserved Gene Neighbor computational methods for the inference of protein function.

Background

One difference between prokaryotic and eukaryotic genomes is the organization of the prokaryotic genome into multi-gene units, known as operons [1]. In its simplest form, as shown in Figure 1a, operon organization in prokaryotes results in a series of adjacent genes being transcribed onto a single polycistronic mRNA, containing the coding regions for the synthesis of multiple proteins. As opposed to eukaryotes, where the dominant transcription form is monogenic, prokaryotic operon organization enables the highly controlled co-expression of multiple genes, by transcribing them together onto a single transcript. More importantly, the encoded proteins of common operons often have related functions, form common complexes, or participate in shared biochemical pathways [2].

Although operon structure has been well studied at the biochemical level in microorganisms such as *Escherichia coli*, genome-wide operon organization in pathogenic organisms, such as *M. tuberculosis*, remains largely unknown. Even so, we can exploit the conservation of certain genetic elements present in many prokaryotic, eubacterial organisms, including *M. tuberculosis*, to learn about operon structure and gene function. Among these are the -10 and -35 bp promoter elements, the ribosome binding sites (RBS), and the 5' and 3' untranslated regions (UTR). Taking into account the orientation of genes on the chromosome, as well as the elements described above, we can build a model, as shown in Figure 2, which depicts the minimum requirements of adjacent genes that are either part of a common operon (Case 4) or not part of a common operon (Cases 1-3). In situations where there is a change in orientation (Cases 1 and 2), the operon boundaries (defined here as the first or last gene of an operon) are easily identifiable. It is rare for prokaryotic transcripts to traverse the length of a gene on a noncoding strand [3,4].

Identification of operon boundaries, however, becomes more challenging when dealing with adjacent genes in the same

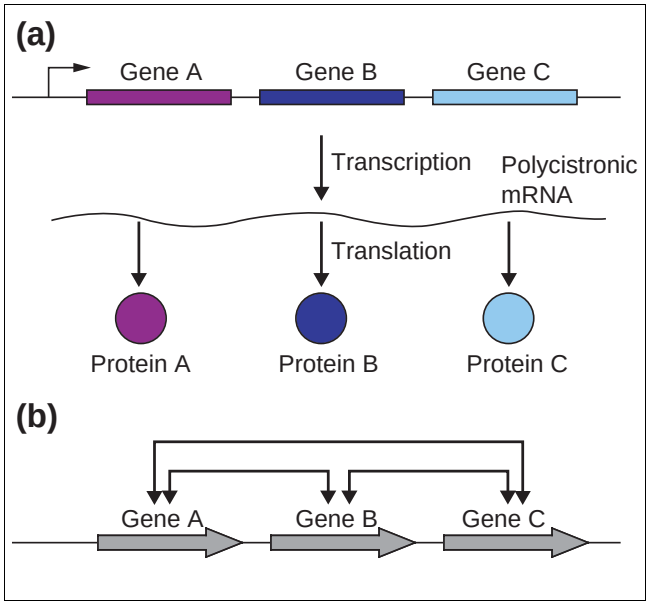


Figure 1
A simplified version of prokaryotic operon organization and functional linkages based on the Operon method. **(a)** Prokaryotic operon organization. Genes A, B, and C are transcribed together onto a single polycistronic transcript, which is then translated to produce three separate proteins. Proteins originating from genes of a common operon often have similar functions, interact physically through protein-protein interactions, or participate in shared biochemical pathways. **(b)** Functional Linkages based on the Operon method. Genes A, B and C are 'linked' if the intergenic nucleotide distance between pairs of adjacent genes is less than or equal to the specified threshold. In this case the distance between gene A and B, and the distance between gene B and C is less than the hypothetical distance threshold, thereby allowing links between all possible sets of genes.

orientation [3,4]. If two genes, A and B, are transcribed separately, as seen in Case 3, the minimum genetic elements required upstream from the start codon of gene B would be its corresponding -10 and -35 bp elements, transcription start site, 5' UTR, and any additional gene-specific promoter elements. As for the downstream elements of gene A, we would expect a 3' UTR and a transcription termination site. Although these genetic elements may overlap with both coding and non-coding elements of adjacent genes, the unique nucleic acid sequence requirements of each of these elements make substantial element overlap less likely if the genes are transcribed independently. If instead both genes are part of a multigene operon, the only upstream requirement of gene B is a single ribosome binding site (RBS) which may be in the intergenic region, or may overlap the coding region of gene A. Intuition suggests, as shown in Figure 2 Case 4, the intergenic spacing between genes in a common operon is shorter than the intergenic spacing of genes encoded by separate transcription units.

Previous studies have examined the nucleotide length distribution of the 5' UTRs, 3' UTRs, intergenic regions and space between RBSs and start sites of transcription in the genome of *E. coli* [5]. Salgado *et al.* utilized a set of experimentally determined *E. coli* operons to examine the distance distributions of intergenic regions between genes within operons and those found experimentally to be at transcription boundaries, and used these values to compute a log-likelihood score for predicting transcription units in *E. coli* [4]. Salgado *et al.* also demonstrated that short intergenic distances are common between adjacent genes of documented operons in *E. coli* [4], and subsequent analyses by Moreno-Hagelsieb *et al.* have suggested that short intergenic distances may be the case for operon members in most other prokaryotic genomes [3]. We have employed a similar dataset as Salgado *et al.* [4], obtained from RegulonDB [6], to evaluate the accuracy of operon inferences using various distance thresholds in *E. coli,* as well as to calculate a posterior probability of identifying *E. coli* genes of a common operon, given the intergenic distance separating two adjacent genes in the same orientation.

Intergenic distance thresholds have also been used both as single distance cutoffs [7,8] and for the construction of probabilistic models [9], to infer gene function based on predicted operon structure. Building on these studies, we evaluate a prokaryotic genome with minimal experimental evidence regarding operon organization. We provide a method for evaluating operon predictions, based on distance and orientation constraints, as well as a combined computational approach to infer protein function and protein linkages, based on the organization of the prokaryotic genome.

In addition to exploiting operon organization, we have also combined the Operon method (OP) with that of the Rosetta Stone (RS) [10], Phylogenetic Profile (PP) [11], and conserved Gene Neighbor (GN) [7,12] method. While the Operon method focuses on the analysis of a single genome, in this case *M. tuberculosis*, the Rosetta Stone, Phylogenetic Profiles, and conserved Gene Neighbor methods focus on the analysis of multiple genomes. Individual proteins that are functionally linked by the Rosetta Stone method occur as a single 'fusion' protein in another organism. The Phylogenetic Profile method functionally links proteins that occur in a correlated manner, for example: a group of genes which may have a shared pattern of presence or absence throughout various genomes. And finally, the conserved Gene Neighbor method links genes that occur as chromosomal neighbors in multiple genomes, often a characteristic of bacterial operons as well as clustered genes of related function.

Although the conserved Gene Neighbor method has been used previously to identify potential operon members [8], the method is distinct from that of the Operon method. Functional linkages established by the Operon method rely only on a single genome, where genes are functionally linked based on a specified intergenic distance threshold. The conserved Gene Neighbor method, in contrast, compares all available sequenced genomes in order to identify genes that are located

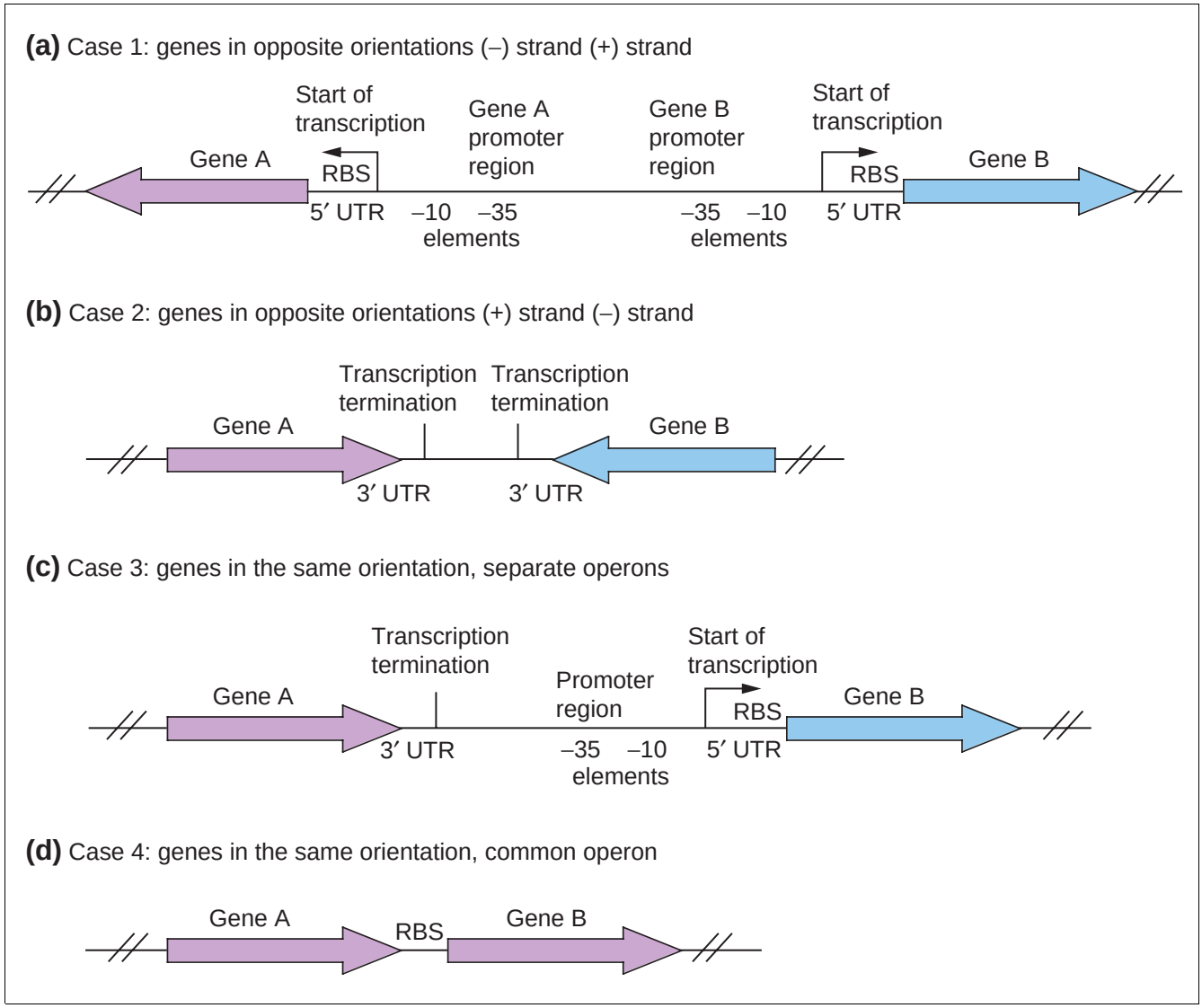


Figure 2
Schematic representation of the minimum genetic requirements for adjacent genes that are transcribed independently and those transcribed together as a single operon. Cases 1, 2 and 3 depict instances where gene A and gene B are transcribed independently as distinct transcriptional units, while Case 4 depicts genes organized into a common operon. The minimum requirement for genes of a common operon is only a RBS, while Case 3 emphasizes the numerous genetic elements required if gene A and gene B are organized into separate transcription units.

in close chromosomal proximity to each other in multiple genomes.

Results

Table 1 summarizes the number of *Mycobacterium tuberculosis* genes with links that we established by the Operon method at various distance thresholds. A pair of genes are considered functionally linked by this method if the intergenic nucleotide distance between adjacent genes in the same orientation is less than or equal to a specified distance threshold. Multiple genes are linked if a series of genes in the same orientation all have intergenic distances less than or equal to the defined distance threshold, as shown in Figure 1b. At a distance threshold of 0 bp we see that 1,279 genes are functionally linked by 2,034 functional linkages. We would expect a substantial percentage of these linkages to represent true operon linkages due to the minimal intergenic spacing and often overlapping open reading frames (ORFs). In *E. coli,* for example, we find that of the 654 links that overlap with experimentally documented *E. coli* transcription units at the 0 bp threshold, 89% correspond to true operon links, while only 11% link genes previously identified as independent transcription units.

Table 1

Total number of *Mycobacterium tuberculosis* genes with linkages based on the Operon method, employing orientation and a nucleotide distance threshold

Threshold (bp)	Predicted operon groups	Genes with links	Functional linkages
0	542	1279	2034
25	792	2071	4442
50	879	2420	5890
75	919	2665	7026
100	933	2870	8468

Column 1 (distance threshold) indicates the intergenic distance, in base pairs, of less than or equal to the indicated value. Notice that at the 0 bp threshold we have links connecting over 25% of the *M. tuberculosis* genes. As we increase the distance threshold from 0 bp to 100 bp we approach almost 75% of the genes having one or more links to other *M. tuberculosis* genes.

In *M. tuberculosis,* we find that six of the eight genes of the documented mammalian cell entry (mce1) operon are linked with a distance threshold of 0 bp, while the other two genes are included by slightly increasing this threshold to 5 bp. Also, we observe that all four members of the documented oligopeptide permease operon (oppA-D) are linked with a 0 bp distance threshold, and are flanked by genes in opposite orientations. As we increase the distance threshold from 0 bp to 100 bp we see the inclusion of genes accounting for a substantial percentage of the tuberculosis genome into putative operons. The expanded coverage enables inclusion of operon members such as the groEL1 groES operon pair, which is separated by 95 bp, and enables linkage of the eight members of the likely arginine biosynthesis operon (Rv1652-Rv1659), which includes intergenic separations ranging from -3 to 80 bp.

Evaluation of functional linkages (Operon method)

To evaluate the accuracy and coverage of functional linkages established at the various distance thresholds, we employed a keyword recovery scheme [13] to compare links between SWISS-PROT annotated proteins. Keyword recovery means that identical keywords are found in the annotations for both proteins connected by the link; the results are shown in Table 2. At a distance threshold of 0 bp, we see a 50% keyword recovery, indicating that half of the total keywords were shared between the linked pairs of genes. As the distance threshold increases from 0 to 100 bp we see the keyword recovery drop from 51% to 45%, and the maximum false positive fraction increase from 0.25 to 0.35 - both indicators of inclusion of links between genes that may not be true operon links. The maximum false positive fraction is calculated by dividing the number of pairwise linkages that do not have any SWISS-PROT keywords in common by the total number of pairwise linkages.

The method of keyword recovery allows us to evaluate a set of linkages based on known functional annotations. By comparing the SWISS-PROT keywords we can quantitatively evaluate different thresholds or different methods for inferring protein function. The maximum false positive fraction is the fraction of functionally-linked proteins that, based on their current annotation, do not share any function in common. We use the term 'maximum' because there are many reasons why two genes may not have any keywords in common, ranging from incomplete biochemical or genetic characterization to the use of different vocabularies to describe similar functions. The quantity 1-maximum false positive fraction indicates the fraction of pairwise links that have one or more keywords in common, and therefore some function in common.

Next we evaluated the keyword recovery at specified distance intervals using the combined intergenic distances between genes of a common run (defined as a series of adjacent genes in the same direction, bordered by genes in opposite orientations). Here we linked all gene pairs in the same run, and gave each linked pair a numerical value equal to the sum of the total intergenic distances between the two genes. The statistics for evaluated intervals are shown in Figure 3. While genes separated by combined intergenic distances of 150 bp or less

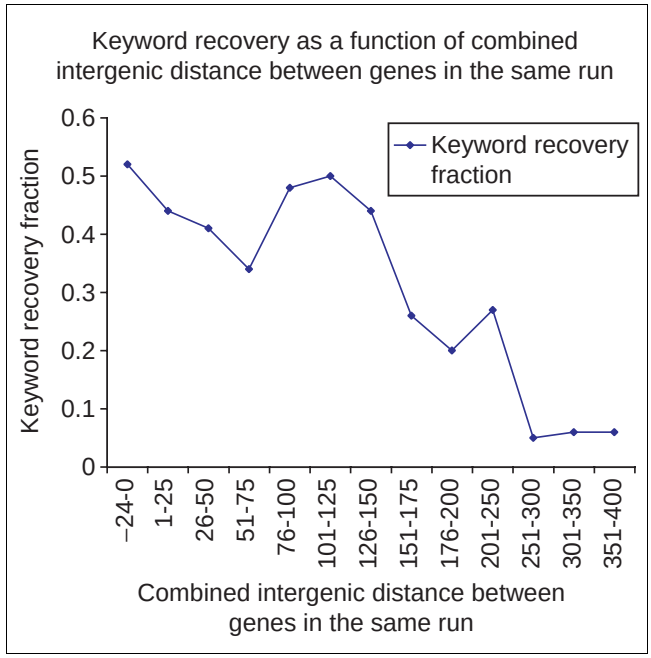


Figure 3
Keyword recovery scores as a function of combined intergenic distances between pairs of genes in a run. All gene members of a run (bordered on each side by genes in opposite orientations) were linked and given a value equal to the combined intergenic distances between them. While the keyword recovery of genes linked by a combined intergenic distance less than 150 bp is fairly high (34-52%), it is apparent that as the total intergenic distance increases above 150 bp, there is a decrease in keyword recovery. At combined intergenic distances above 250 bp the keyword recovery is comparable to that of randomly linked genes.

Table 2

Assessment, by keyword recovery, of the functional linkages established by the Operon method at various distance thresholds

Threshold (bp)	Functional links between SwissProt Annotated Proteins	Functional links with no keywords in common	Correct keywords recovered	Total keywords	Maximum false positive fraction*	Keyword recovery†
0	308	78	446	883	0.25	0.51
25	642	180	856	1766	0.28	0.48
50	818	254	1044	2226	0.31	0.47
75	912	326	1080	2453	0.36	0.44
100	1044	362	1224	2726	0.35	0.45

*The maximum false positive fractions were calculated as the fraction of pairwise links that do not have any SWISS-PROT keywords in common (ignoring the keywords 'hypothetical protein', 'three-dimensional structure', 'transmembrane' and 'complete proteome'). †Keyword recovery was calculated by comparing the SWISS-PROT keyword annotation between each pair of linked *M. tuberculosis* genes. The keyword recovery of all linkages was calculated as:

$$\langle \text{keyword recovery} \rangle = \frac{1}{X}\sum_{i=1}^{y}\sum_{j=1}^{x} n_j$$

where *X* is the total number of query protein keywords, *Y* is the total number of linked gene pairs, *x* is the number of query protein SWISS-PROT keywords, and *nj* is the number of times the query protein keyword *j* occurs in the linked protein. Notice that at 0 bp the keyword recovery is quite high, about 50%, while the maximum false positive rate is about 25%. As the distance threshold increases from 0 bp to 100 bp the keyword recovery decreases, while the maximum false positive fraction increases.

have keyword recovery scores ranging from 34% to 52%, as the distance increases past 125 bp the keyword recovery decreases steadily until the keyword recovery is no better than that of random links above 250 bp. This result suggests that, as expected, genes that are separated by greater distances are less likely to belong to common operons than those in close proximity, and are less likely to have similar functions.

Of particular interest are functional links that connect non-annotated proteins to annotated proteins. These links may represent likely operon pairs, and thus may be used to infer possible function for previously uncharacterized proteins. For this evaluation, our original database was expanded to include *M. tuberculosis* gene annotations obtained from the Sanger Institute web server. Of the 1,548 non-annotated genes, a substantial percentage (between 14% and 46%) can be linked to one or more of the 2,403 annotated genes at distance thresholds ranging from 0 bp to 100 bp, as summarized in Table 3. We expect that links established in this manner may suggest possible functional roles for hundreds of previously uncharacterized proteins, and used in combination with other techniques may aid, not only in the inference of protein function, but also in the identification of possible interacting partners.

Combined methods: Operon, Rosetta Stone, Phylogenetic Profiles and conserved Gene Neighbors

In order to further refine our functional inferences, we set out to evaluate the Operon method in combination with the Rosetta Stone [10], Phylogenetic Profile [11] and conserved Gene Neighbor [7,12] computational methods. The total number of functional linkages established by each of the methods alone, as well as in combination, are listed in Table 4. We see that there is substantial overlap among the four methods, and many gene pairs are linked by more than one method. For example, there are 414 links inferred by both the Operon and Rosetta Stone methods, 632 links inferred by both the Operon and Phylogenetic Profile methods, and 1,516 links inferred by both the Operon and conserved Gene Neighbor methods.

The most substantial overlap results from that of the Operon method and the conserved Gene Neighbor method. We see that 18% of the links identified by the Operon method at the 100 bp threshold are also identified by the conserved Gene Neighbor method. Both of these methods are used to identify potential operons in microbial genomes, but the conserved Gene Neighbor method only identifies operons that are conserved in multiple genomes. The Operon method, in contrast, is based solely on the intergenic distance between genes in the same orientation so it is able to identify potential operons even in the absence of homologous genes in other organisms. Ermolaeva *et al.* noted that many bacterial operons are organized in a similar manner in diverse genomes [8], and

Table 3

Total number of non-annotated genes that have one or more links to an annotated gene at the various distance thresholds

Threshold (bp)	Non-annotated genes with links	Non-annotated genes linked to one or more annotated genes
0	474	217 (14%)
25	786	412 (27%)
50	913	521 (34%)
75	1015	615 (40%)
100	1088	703 (46%)

The links from non-annotated genes to annotated genes may suggest biological functions for previously uncharacterized proteins.

employed a combination of the conserved Gene Neighbor method along with a distance threshold to predict operons in microbial genomes [8].

We also see that a number of functional linkages inferred by the Operon method overlap with linkages inferred by the Phylogenetic Profile and Rosetta Stone methods. Moreno-Hagelsieb *et al.* previously compared genes within operons to those at transcription boundaries, and demonstrated that genes in known operons are more likely to have similar Phylogenetic Profiles, are more likely to occur as conserved Gene Neighbors and are more likely to occur as fusion proteins than genes at transcription boundaries [14]. Yanai *et al.* also examined fusion genes, and noted many instances where individual components that constituted a 'fusion' gene were found as separate genes organized into common operons [15].

Table 4 also summarizes functional linkages inferred by the overlap of three methods and functional linkages inferred by all four computational methods. Each of the overlapped combinations was evaluated using the keyword recovery method described in the methods. Table 5 summarizes the keyword recovery for each of the combinations. Used independently, the Rosetta Stone, Phylogenetic Profile and conserved Gene Neighbor methods have a signal to noise ratio of 9.5, 2.8 and 5.6, respectively. The signal to noise ratio is calculated by dividing the keyword recovery score of functionally-linked proteins by the keyword recovery score of randomly-paired proteins. The combined overlap of the Rosetta Stone, Phylogenetic Profile and conserved Gene Neighbor linkages with the 100 bp Operon linkages increases the 100 bp Operon signal to noise from 7.9 to between 10 and 13.

The best keyword recovery score is achieved by a combination of the 100 bp Operon, Rosetta Stone and Phylogenetic Profile methods (Figure 4). This combination results in a 74% keyword recovery and a 0.03 maximum false positive fraction. The high keyword recovery and low maximum false positive fraction for the combined methods gives us additional confidence in the correctness of these functional linkages, and we indicate these as high confidence links.

Evaluation of coverage and establishing a distance threshold

To investigate further the coverage of the Operon method, we compiled a list of all adjacent genes in the same orientation that are functionally linked by either the Rosetta Stone, Phylogenetic Profile or conserved Gene Neighbor methods. We may hypothesize that many of these functional links overlap with true operon pairs, and therefore may yield a distance profile that would be indicative of the intergenic distance profile between genes that are in a common operon. Figure 5a summarizes the distance profiles of the intergenic distances between the 564 functionally-linked gene pairs, and Figure 5b indicates pairs not linked by any of the three methods. The linked pairs represent a distinct population, verified by the chi-square test (probability less than 0.005% that these distributions are drawn from the same population).

The linked population is more heavily weighted at the short distances, with a mean of 27 base pairs, while those that were not identified as linked by the three methods, have a mean of 94 base pairs. In fact, this result corresponds well with the

Table 4

Summary of the number of functional linkages between *M. tuberculosis* genes established by each of the four prediction methods of this paper, alone and in combination

Method(s)	Functional links	Proteins
One method		
Rosetta Stone	20,714	1,279
Gene Neighbor	15,550	1,658
Phylogenetic Profiles	67,684	1,815
Operon (#100 bp)	8,468	2,870
Two methods		
100 bp Operon and RS	414	
100 bp Operon and PP	632	
100 bp Operon and GN	1516	
Three methods		
100 bp Operon, RS and PP	186	
100 bp Operon, RS and GN	246	
100 bp Operon, PP and GN	472	
All four methods		
100 bp Operon, PP, GN and RS	138	

PP, Phylogenetic Profile method; GN, conserved Gene Neighbor method; RS, Rosetta Stone method.

Table 5

Keyword recovery scores for the Operon method alone and in combination with the Rosetta Stone (RS), Phylogenetic Profile (PP), and conserved Gene Neighbor (GN) methods

Method	Number of links between SWISS-PROT annotated proteins	Keyword recovery %	Maximum false positive fraction	Keyword recovery of random links (100 trials) %	Random links SD %	Keyword recovery signal to noise
0 bp Operon	308	51	0.25	5.6	1	9.1
25 bp Operon	642	49	0.28	5.6	0.7	8.8
50 bp Operon	818	47	0.31	5.7	0.6	8.2
75 bp Operon	912	44	0.36	5.5	0.5	8
100 bp Operon	1044	45	0.35	5.7	0.5	7.9
100 bp Operon and RS	88	67	0.05	5.5	1.8	12
100 bp Operon and GN	638	60	0.16	5.5	0.7	11
100 bp Operon and PP	268	62	0.11	5.8	0.9	11
100 bp Operon, PP and GN	242	61	0.11	5.6	1.2	11
100 bp Operon, PP and RS	62	74	0.03	5.6	2	13
100 bp Operon, RS and GN	78	65	0.05	6.3	1.9	10
All four methods	54	70	0.04	5.7	2.3	12

Keyword recovery and maximum false positive fraction calculated as described in Table 2. Random links were established between the same number of random pairwise Swiss-Prot annotated genes as exist real links between Swiss-Prot annotated genes (mean and standard deviation of 100 random trials indicated).
Signal to Noise calculated as:

$$\text{Signal to Noise} = \frac{\text{Keyword recovery}}{\text{Random Keyword recovery}}$$

study of Moreno-Hagelsieb *et al.*, where they observed short distances between adjacent genes predicted to be functionally linked by an extension of the conserved Gene Neighbor method [3]. If we assume the profile depicted in Figure 5a represents the profile for true operon members in *M. tuberculosis*, we infer that functional linkages based on the Operon method at a distance of 50 bp may have a coverage of more than 80%, while functional linkages established at the 100 bp cutoff may allow inclusion of more than 90% of true operon pairs. This method may also be useful for confirming short intergenic distances between genes in common operons in various microbial genomes.

A closer examination of Figure 5a and 5b reveals the similarities and differences among the two populations. The large occurrence at 0 distance in Figure 5b shows probable linkages that are detected by the Operon method but are undetected by the Rosetta Stone, Phylogenetic Profile or conserved Gene Neighbor methods. Furthermore, the large occurrence of potential linkages in Figure 5b above 200 bp is excluded by the Operon method, with a threshold of less than or equal to 100 bp.

We have also computed the keyword recovery and maximum false positive fraction for Operon distance thresholds ranging from 0 bp to 300 bp. As shown in Figure 6, as we increase the distance threshold from 0 bp to 300 bp, there is a steady decrease of the keyword recovery (from 0.51 to 0.33) and a steady increase of the maximum false positive fraction (from 0.25 to 0.50). Keyword recovery analysis allows evaluation of each of the distance thresholds and provides a method of evaluating operon inferences in organisms without extensive prior experimental data on operon structure. At a distance threshold of 0 bp we may expect approximately 75% ([1-maximum false positive] × 100%) of the linked genes to have one or more keywords in common, and therefore some function in common. If we increase the distance threshold to 50 bp, as was used previously by Blattner *et al.* to predict operons in *E. coli* [16], we may expect approximately 70% of the linked *M. tuberculosis* genes to have some function in common.

Moreno-Hagelsieb *et al.* suggested a method for examining all adjacent gene pairs within the same direction (WD), and proposed a formula to calculate the fraction that exist in common operons [3]. Moreno-Hagelsieb estimated that 50% of *E. coli* WD pairs may be in shared operons [3,4], and using their formula we estimate that 57% of *M. tuberculosis* WD pairs are in common operons. According to Skovgaard *et al., E. coli* and *M. tuberculosis* appear to have a similar quality of assigned open reading frames [17], therefore it is likely that the higher percentage of *M. tuberculosis* WD pairs in operons is not a result of over assignment of open reading frames. Since *M. tuberculosis* has a higher percentage of WD pairs in operons, a higher distance threshold may be tolerated in *M. tuberculosis* than the 50 bp threshold used previously in *E. coli* [16].

Although we have selected the 100 bp threshold for evaluation of the combined computational approach, a lower threshold, such as 50 bp or 25 bp, results in a higher keyword recovery and a lower maximum false positive fraction (Table 2). The best scores for the Operon method, as we might expect, are achieved at a distance threshold of 0 bp. Although the keyword recovery and maximum false positive scores are

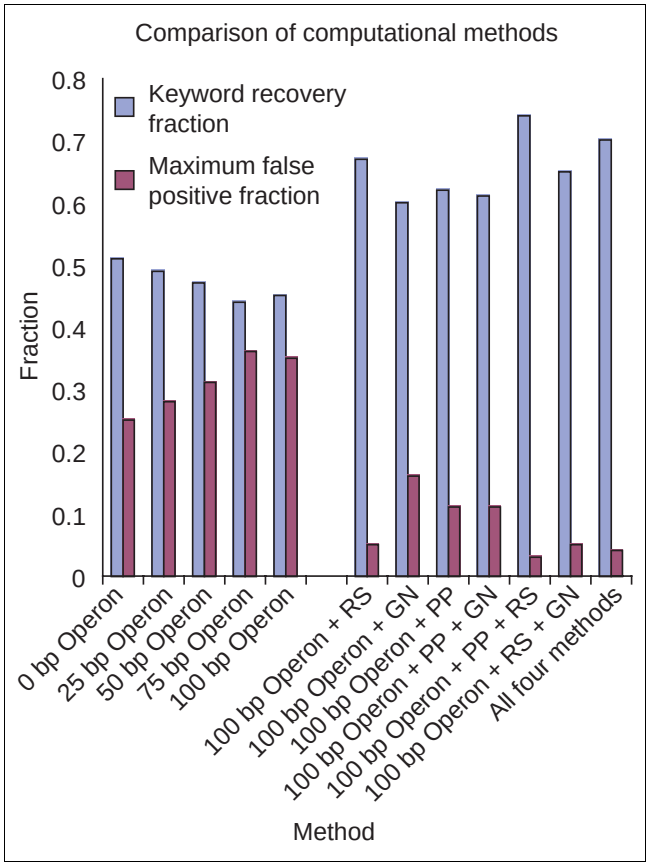


Figure 4
Keyword recovery scores for the Operon method alone and in combination with the Rosetta Stone (RS), Phylogenetic Profile (PP), and conserved Gene Neighbor (GN) methods. Notice that the combination of either the Rosetta Stone, Phylogenetic Profiles or conserved Gene Neighbor method has a dramatic effect on the keyword recovery, with the best score resulting from a combination of the 100 bp Operon, Rosetta Stone and Phylogenetic Profile methods.

improved at shorter distance thresholds, the coverage is decreased. Combining the Operon method with those of the Rosetta Stone, Phylogenetic Profile or conserved Gene Neighbor methods yield a dramatic improvement in both the keyword recovery and the maximum false positive fraction (Figure 4), and enables the use of a larger distance threshold, such as 100 bp. Gene pairs that are linked by two or more methods are very likely to share some function in common, even with a distance threshold of 100 bp (Figure 4).

For comparison, we have also included the distance profile of adjacent genes in experimentally documented *E. coli* operons (Figure 5c, data obtained from RegulonDB [6]). The distributions of both the *E. coli* operon data (Figure 5c) and the *M. tuberculosis* linked data (Figure 5a) tend to have shorter intergenic distances than the set of *M. tuberculosis* genes that are not functionally linked by the Rosetta Stone, Phylogenetic Profile or conserved Gene Neighbor methods (Figure 5b). Most prokaryotic organisms do not have extensive data regarding experimentally documented operons, so we suggest that this method may serve as an alternate method for identifying intergenic distance distributions of potential operon members in less characterized microbial genomes.

Discussion

Example of combined approach

Figure 7 depicts two *M. tuberculosis* genes, *leuD* and *leuC*, that are functionally linked by all four methods: Operon, Rosetta Stone, Phylogenetic Profiles and conserved Gene Neighbors. These two genes, like a number of other genes linked by all four methods, encode proteins that have a very close physical and functional association within the cell. The *leuD* and *leuC* genes encode individual protein subunits that dimerize to form a functional isopropylmalate isomerase heterodimer, involved in leucine biosynthesis [18]. These two genes have been identified as members of common operons in numerous prokaryotic organisms [19] often with additional members of the leucine biosynthesis pathway. This example exemplifies the use of a combined protocol to identify not only genes that are likely to belong to common operons, but also genes that encode proteins that have a strong functional link, possibly encoding proteins that may physically interact. In addition, *leuD* and *leuC* do not share any sequence similarity, emphasizing the ability of these methods to identify functional links between non-homologous proteins. In *S. pombe* the *leuD* and *leuC* genes occur as a single fusion protein, as shown in Figure 7b.

Inference of protein function and operon organization

Next we demonstrate the use of this combined approach to identify genes that are likely to belong to common operons, as well as to assign possible function to genes of previously unknown function. Figure 8a shows two genes, Rv0415 and Rv0417(*thiG*), that are linked by all four methods. Rv0415 has the Sanger annotation 'conserved hypothetical protein', while Rv0417 is annotated as 'thiamine synthesis'. These two genes flank and slightly overlap Rv0416, which also has the annotation 'conserved hypothetical protein'. Taken together, this evidence suggests that all three of these genes may be members of a common operon, and therefore we may assign a putative function to Rv0415 and Rv0416 as possibly involved with thiamine synthesis. Since Rv0415 and Rv0417(*thiG*) are linked by both the Operon method and the Rosetta Stone method, we may speculate that these two products may be both co-expressed as well as encode proteins that are functionally linked, either as physically interacting partners or as participants in a common biochemical pathway. These inferences are further supported by the recent change in the Pasteur Institute annotation of Rv0415 and Rv0416 from 'conserved hypothetical proteins' to 'Possible oxidoreductase thiO' and 'Possible Protein thiS' respectively, both believed to be involved in thiamine biosynthesis. The recent Pasteur Institute classification of Rv0416 is based on weak sequence similarity to other thiS proteins.

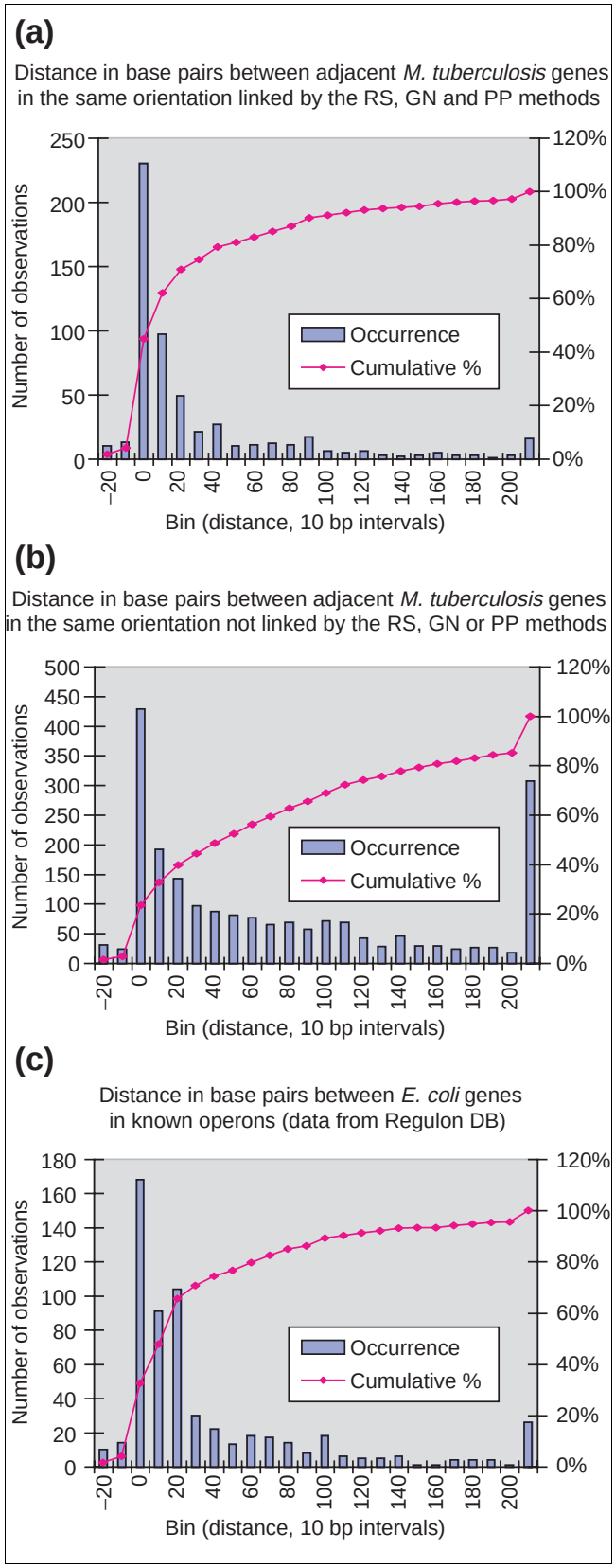


Figure 5

Figure 5
Distance profile of adjacent *M. tuberculosis* genes in the same orientation that are functionally linked by the Rosetta Stone, Phylogenetic Profiles or conserved Gene Neighbor methods, compared to adjacent genes in the same orientation that are not linked by these methods. **(a)** Distance profile of adjacent *M. tuberculosis* genes in the same orientation linked by either the Rosetta Stone, Phylogenetic Profile or conserved Gene Neighbor method in *M. tuberculosis*. **(b)** Distance profile of all other adjacent *M. tuberculosis* genes in the same orientation, excluding those linked by the Rosetta Stone, Phylogenetic Profiles or conserved Gene Neighbor methods in *M. tuberculosis*. **(c)** Distance profile of adjacent genes in the same orientation in experimentally documented operons in *E. coli*. *E. coli* operon data obtained from RegulonDB [6]. The linked profile (a) yielded a mean intergenic distance of 27 base pairs, as compared with (b) 94 base pairs for the mean intergenic distance for genes not linked by any of the three methods. This demonstrates that adjacent genes in the same orientation that have small intergenic spacing are more likely to be functionally linked that those that are separated farther apart.

A possible operon involved in RNA degradation may be encoded by the three genes Rv2925c(*rnc*), Rv2926c and Rv2927c, as shown in Figure 8b. Functional links between Rv2926c and Rv2925c(*rnc*) were identified by the Operon, Rosetta Stone and conserved Gene Neighbor methods, while Rv2926c was linked to Rv2927c by the Operon and conserved Gene Neighbor methods. Rv2926c and Rv2927c have the Sanger annotation 'hypothetical protein' and 'conserved hypothetical protein', respectively. Since Rv2925c(*rnc*) has the annotation 'RNAse III', we may assign a putative function to both Rv2926c and Rv2927c as possibly involved in RNA degradation. Notice that the functional link between Rv2926c and Rv2925c(*rnc*) is supported by three separate forms of evidence (OP, RS and GN), while the only direct link between Rv2927c and Rv2925c(*rnc*) is by the Operon method. Further support for this link is established indirectly due to the two functional links to Rv2926c.

While the functional linkages established by the Rosetta Stone method in combination with the Operon method seem to be useful at identifying functionally-linked partners, the overlap between the Operon method and the conserved Gene Neighbor methods is more prevalent. Figure 8c shows a region with numerous functional linkages to the penicillin-binding protein, Rv2163c(*pbpB*). Here we see a string of four genes, three of which have the Sanger annotation of 'hypothetical protein' or 'conserved hypothetical protein'. Based on the functional linkages established by the Operon method, Phylogenetic Profiles and conserved Gene Neighbors methods, we may assign a putative function to Rv2164c, Rv2165c and Rv2166c, similar to that of the *pbpB*, which is involved in cell wall biosynthesis.

Applications to the identification of possible drug targets

In some situations, the function of uncharacterized genes cannot be inferred directly from the Operon links. In these cases, we can examine the Rosetta Stone, Phylogenetic Profiles and conserved Gene Neighbors functional linkages to

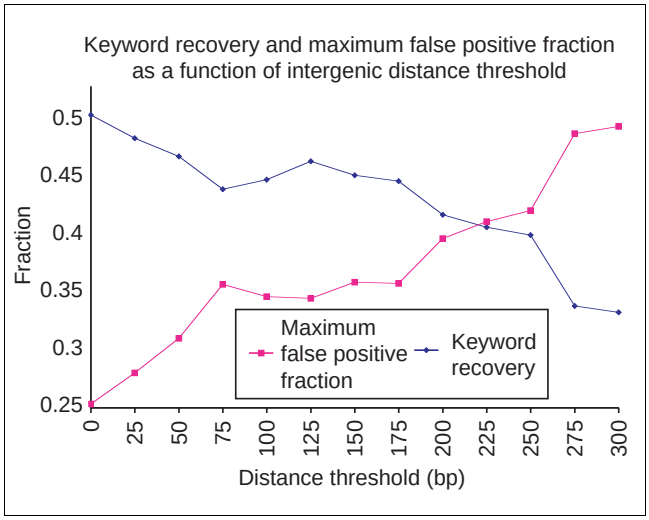


Figure 6
Keyword recovery and maximum false positive fraction scores as the Operon distance threshold increases from 0 bp to 300 bp. Notice the decrease in the keyword recovery and the increase in maximum false positive fraction as the distance threshold increases.

other genes throughout the genome. In Figure 9 we see two genes, Rv1503c and Rv1504c, linked by both the Operon and the Rosetta Stone method. Although neither of these genes have functions assigned to them, we may hypothesize that they may be co-expressed as a single transcript and may function together.

In order to link Rv1503c and Rv1504c to a possible function or pathway we can examine all of the other functional linkages. Although both Rv1503c and Rv1504c have a number of functional linkages, there are some common linkages between them. Both Rv1503c and Rv1504c have functional linkages to the genes Rv1302(*rfe*) and Rv3464(*rmlB*). Interestingly, both of these genes, *rfe* and *rmlB*, are important elements of the arabinogalactan biosynthesis pathway [20]. Arabinogalactan is an essential component of the *M. tuberculosis* cell wall, and the arabinogalactan biosynthetic pathway is of major medical relevance, since two of its downstream members, EmbA and EmbB, are primary targets of the tuberculosis drug Ethambutol. Although there have been efforts to identify members of this pathway, only a fraction of the pathway members are known [20]. Here we propose that Rv1503c and Rv1504c may be important members of the arabinogalactan biosynthesis pathway, possibly organized into a shared operon, encoding functionally-linked proteins.

Another potential drug target is turned up by our methods, as a link to one of the *M. tuberculosis* glutamine synthetase homologues, as seen in Figure 10. This emerges from the genomic region containing the genes Rv1878(*glnA3*) and Rv1879. Rv1878 is annotated as enocoding a 'probable glutamine synthase' while Rv1879 is annotated 'conserved hypothetical protein'. There are four homologous proteins in the *M. tuberculosis* genome, annotated as either 'glutamine synthase' or 'probable glutamine synthase'. Glutamine synthetase plays a vital role in the incorporation of ammonia into biomolecules through a two-step mechanism involving glutamate, ammonia and ATP. Although it is not known why there are four homologous glutamine synthase-like genes in *M. tuberculosis*, at least one of these, *glnA1*, has been proposed as a drug target in *M. tuberculosis* [21].

The 'probable glutamine synthase' gene Rv1878 is linked to the 'conserved hypothetical protein' Rv1879 by both the Operon Method and the Rosetta Stone method, possibly indicating a functional relationship between these two genes. In *Arabidopsis thaliana*, both domains encoded by *M. tuberculosis* genes Rv1878(*glnA3*) and Rv1879 occur as a single fused protein, as seen in Figure 10. A closer look at the phylogenetic distribution of proteins homologous to Rv1879 reveals that homologs of this protein are present in only a handful of prokaryotic genomes, and as fusion proteins with glutamine synthase-like domains in some fungi and plants. Glutamine synthetase homologs, on the other hand, are found in all living organisms. We speculate that Rv1879 may have a functional relationship to the 'probable glutamine synthase' gene Rv1878, and may encode for a protein that links this glutamine synthase homolog to a previously undescribed pathway. The absence of an *M. tuberculosis* Rv1879 homolog in mammals, and the possible association with a glutamine synthetase homolog, may make this protein a promising drug target candidate.

In this study we have focused on the Rosetta Stone, Phylogenetic Profiles and conserved Gene Neighbor linkages that overlap with the Operon linkages. The Operon method links genes to other genes at a particular genetic locus. In contrast, the Rosetta Stone, Phylogenetic Profiles and conserved Gene Neighbor methods can link genes that lie near or far along the chromosome. Thus by combining these methods we are able to identify functional linkages involving genes that are not part of operons as well as genes that are.

Conclusions

The organization of the *M. tuberculosis* genome holds many clues to the possible function of hundreds of previously uncharacterized proteins. The use of the Operon method, based on distance between genes in the same orientation, appears to provide a useful tool for the prediction of protein function as well as the identification of possible operon members. The coverage of linkages at various distances may be inferred by the distance profile of genes known to be functionally linked (Figure 5a), and the accuracy of these functional links may be represented by the maximum false positive rate (Table 2).

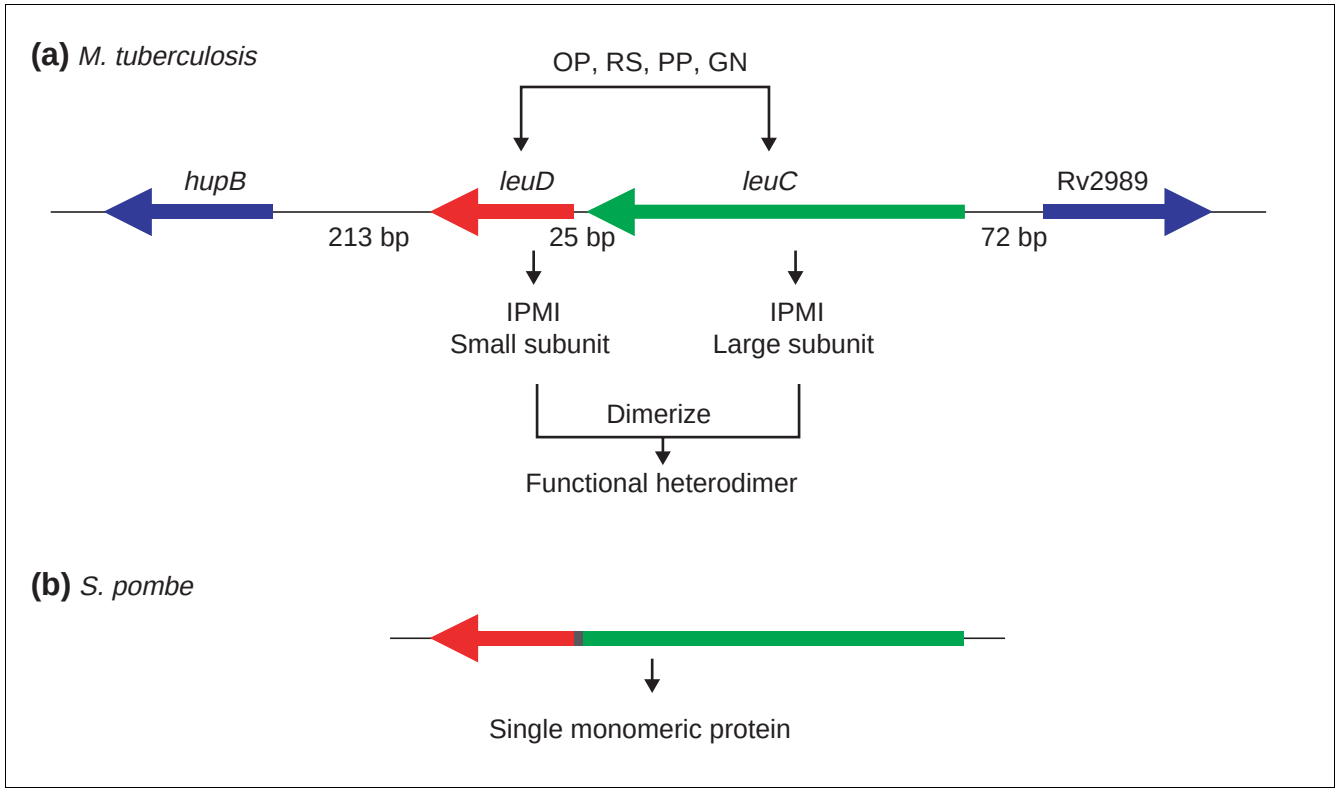


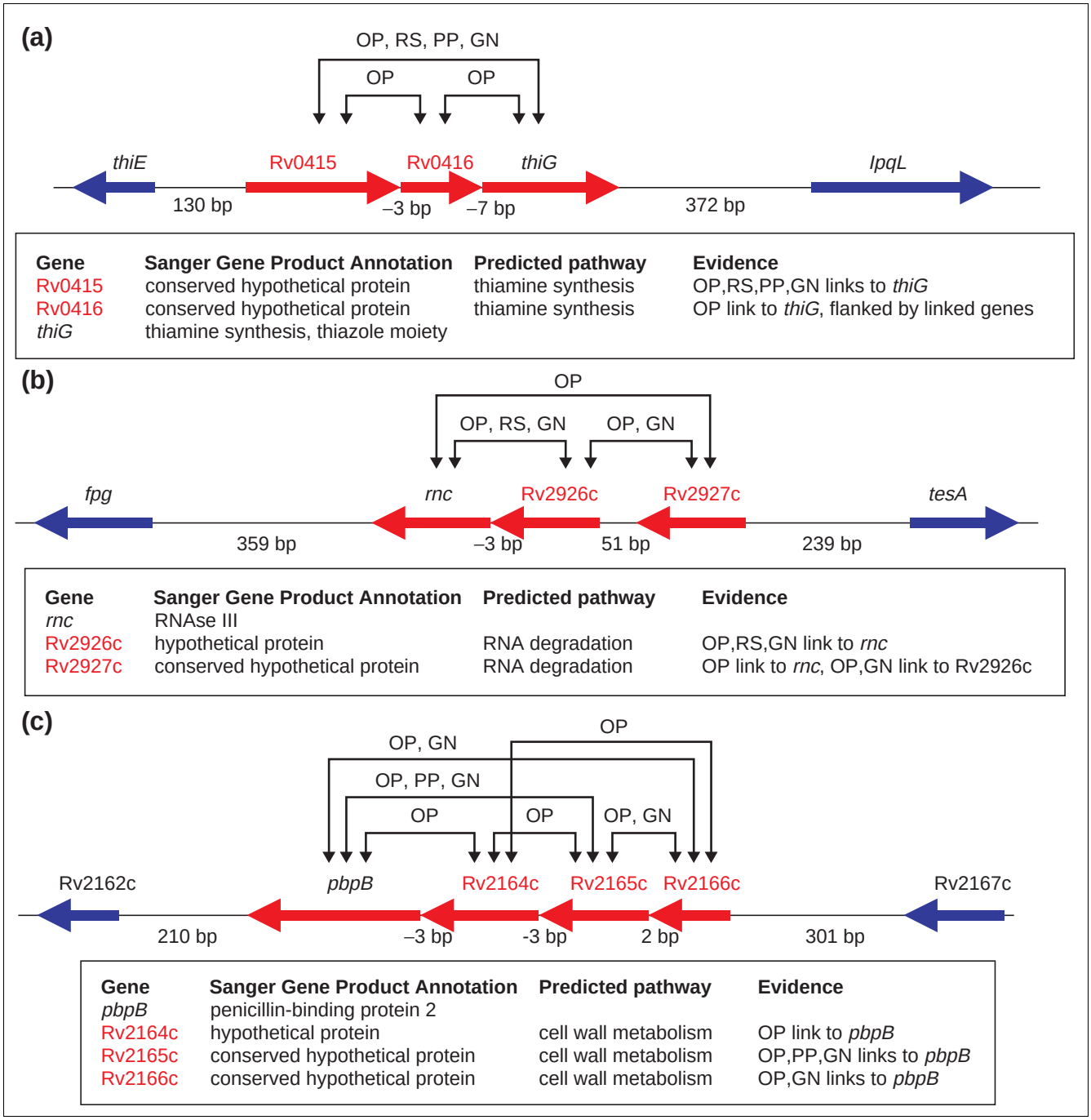
Figure 7
Comparison of the genomic organization of the leucine biosynthesis genes in *M. tuberculosis* and *Schizosaccharomyces pombe*. **(a)** Genomic organization of the *leuC* and *leuD* genes of *M. tuberculosis*. **(b)** *S. pombe* alpha-isopropylmalate isomerase, containing both the *leuC* and *leuD* coding regions in a single fusion gene. This example illustrates the power of the Rosetta Stone, Phylogenetic Profile, Gene Neighbor and Operon methods to infer a functional linkage, in this case one that is already established [18].

Taken together, using the Operon method alone with a threshold distance of 100 bp, we may expect to link over 90% of the true operon members in the *M. tuberculosis* genome. The accuracy of the 8,468 functional links established at that distance, may be inferred from the maximum false positive fractions in Table 2, roughly 60% ([1-max. false positive fraction] × 100%) of these links may represent links between genes that have at least some functional similarity, many probably representing true operon members. We expect that the maximum false positive fractions represent the upper limit of true false positive pairwise linkages since the annotations of many genes may be incomplete.

In further support of the idea that the functional similarities between genes in the same orientation is due primarily to operon structure, is the observation that common function is related to the distance between genes in the same orientation, as depicted in Figure 3. Genes separated by a combined intergenic distance of more than 250 bp are no more likely to share a common function than randomly selected pairs. In contrast to operon prediction methods based on conserved gene strings [8,22], methods based on intergenic distance thresholds allow identification of operon members without the dependency on identifiable homologues in other sequenced genomes [3,23].

From our examination of experimentally documented operons in *E. coli* we expect that the Operon method would be able to identify functional relationships among proteins involved in a wide variety of functional categories. For example, in *E. coli* we observe operons containing genes involved in common metabolic pathways, multi-protein complexes, membrane-bound transport complexes, as well as genes involved in cell structure, cell adaptation, DNA replication, transcription, translation, regulatory functions and a number of other cellular activities.

Although the coverage of the Operon method alone allows us to identify thousands of potentially functionally-linked genes, a combined approach with the Rosetta Stone, Phylogenetic Profiles and conserved Gene Neighbor methods allows us to establish higher confidence links, as demonstrated in Figure 4 and Table 5. The Operon method in combination with any of the other methods results in an increase in the keyword recovery and a decrease in the maximum false positive fractions. The combination of the 100 bp threshold Operon



Gene	Sanger Gene Product Annotation	Predicted pathway	Evidence
Rv0415	conserved hypothetical protein	thiamine synthesis	OP,RS,PP,GN links to *thiG*
Rv0416	conserved hypothetical protein	thiamine synthesis	OP link to *thiG*, flanked by linked genes
thiG	thiamine synthesis, thiazole moiety		

Gene	Sanger Gene Product Annotation	Predicted pathway	Evidence
rnc	RNAse III		
Rv2926c	hypothetical protein	RNA degradation	OP,RS,GN link to *rnc*
Rv2927c	conserved hypothetical protein	RNA degradation	OP link to *rnc*, OP,GN link to Rv2926c

Gene	Sanger Gene Product Annotation	Predicted pathway	Evidence
pbpB	penicillin-binding protein 2		
Rv2164c	hypothetical protein	cell wall metabolism	OP link to *pbpB*
Rv2165c	conserved hypothetical protein	cell wall metabolism	OP,PP,GN links to *pbpB*
Rv2166c	conserved hypothetical protein	cell wall metabolism	OP,GN links to *pbpB*

Figure 8

Inference of *M. tuberculosis* protein function and operon organization based on multiple method overlap. **(a)** Inference of an operon encoding members involved in thiamine biosynthesis. **(b)** Operon inference for a region possibly involved in RNA degradation. **(c)** Functional links and operon inference for a region likely to be involved in cell wall metabolism. In these cases, inferences are made for the functions of uncharacterized genes by their functional linkages to genes of known function.

inferences with either the Rosetta Stone, Phylogenetic Profile or conserved Gene Neighbor method exemplifies this. Although the 100 bp threshold Operon inferences alone have a keyword recovery of 45% and a maximum false positive fraction of 0.35, used in combination with either the Rosetta Stone, Phylogenetic Profile or conserved Gene Neighbor method yields a keyword recovery increase to 60-67% and a maximum false positive fraction decrease to 0.05-0.16, depending on the combination. Especially notable are the Operon inferences that overlap with Rosetta Stone

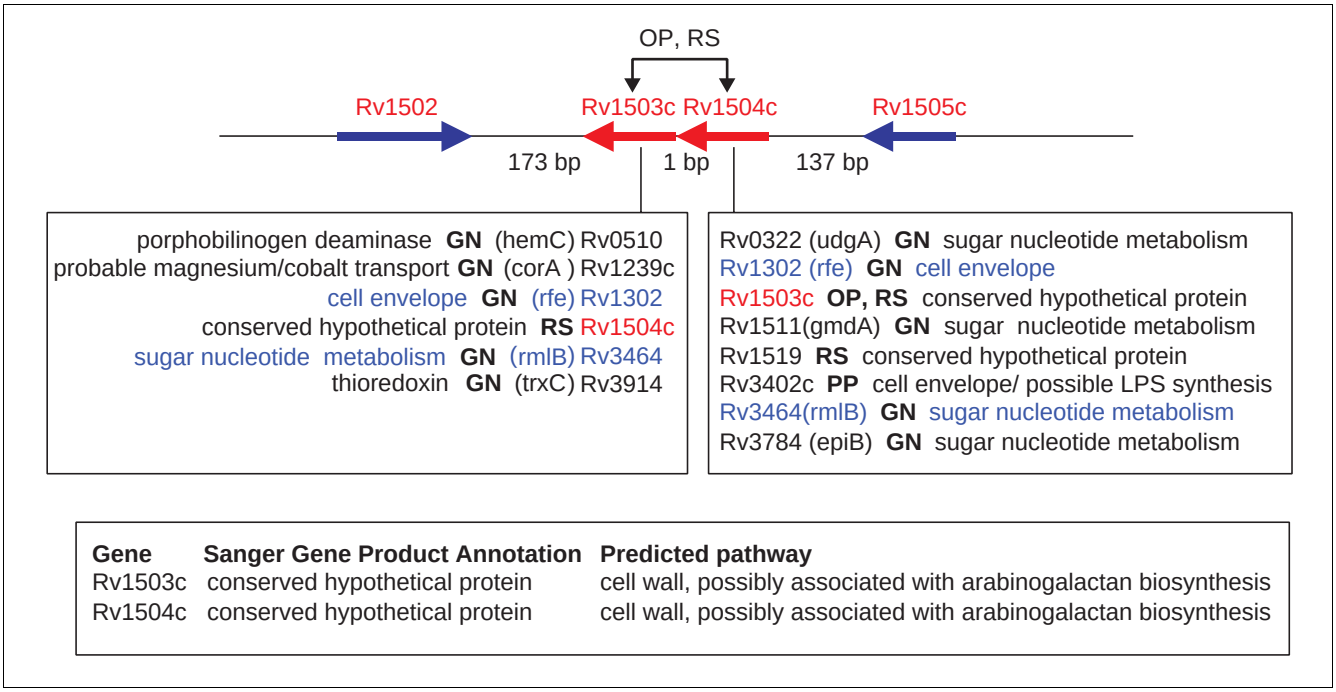


Figure 9
Identification of two novel genes linked to the arabinogalactan biosynthesis pathway, an important target of *M. tuberculosis* specific drugs. Based on the close proximity of adjacent genes (Operon method) and the functional linkage established by the Rosetta Stone method, we infer that Rv1503c and Rv1504c may be organized into a common operon. Both genes also have functional links to the genes rfe and rmlB, important components in the arabinogalactan biosynthesis pathway.

inferences. The high keyword recovery and low maximum false positive fractions for this combination may be an indication that these links represent not only genes with similar functions that are organized into a common operon, but also may suggest proteins that may physically interact.

Many linkages inferred by the Rosetta Stone, Phylogenetic Profile and conserved Gene Neighbor methods overlap with those of the Operon method. The highest overlap results from that of the conserved Gene Neighbors method. In many cases this would be expected since genes organized in an operon would have a tendency to coevolve as a single unit, rather than as separate units, therefore these genes would be observed as 'Neighbors' in multiple prokaryotic genomes. The Phylogenetic Profiles overlap would result from the observation that operon members are often involved in shared pathways or complexes, and therefore would be expected to evolve in a correlated fashion. Finally, the overlap with the Rosetta Stone links may reflect the observation that the Rosetta Stone method, like the Operon method, often links proteins that either physically interact or are in the same pathway. The Phylogenetic Profile and conserved Gene Neighbor methods have also been used previously to confirm operon predictions based on intergenic distances [3,23].

There are a number of potential applications for this combined method, ranging from the prediction of protein function based on functional linkages to annotated proteins, to the reconstruction of biochemical pathways. Zheng *et al.* have employed a combination of gene proximity and phylogenetic profiles to examine the co-evolution of gene clusters in *E. coli* [24], while Pellegrini *et al.* have used similar methods to those described here to construct a network of interconnected proteins within the *Mycoplasma genitalium* genome [9].

Here we have applied a combined method to investigate the genome of the pathogenic bacterium *M. tuberculosis*, and have demonstrated functional links from a number of previously uncharacterized proteins to specific biochemical pathways. Included in these, we have identified five novel proteins that may be functionally involved with pathways involved in the biosynthesis of components of the mycobacterium cell wall. By applying these methods to the entire genome of the pathogenic *M. tuberculosis*, we have identified many other novel genes that are linked to numerous biochemical pathways, some that may eventually serve as potential drug targets. Combined, these methods will also enable the genome-wide analysis of other prokaryotic genomes, and will aid in the identification of novel partners in both characterized and

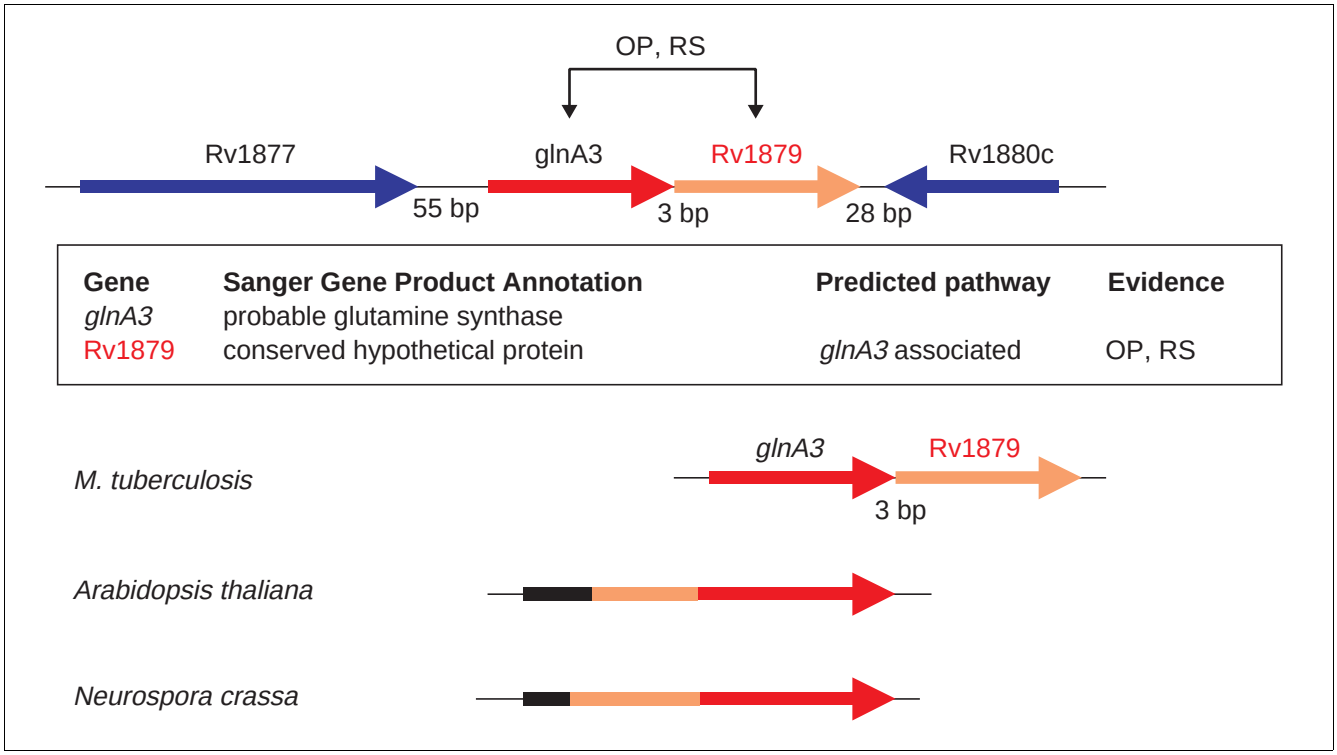


Figure 10
A unique *M. tuberculosis* gene linked to a glutamine synthetase paralog. Few homologs of Rv1879 exist in prokaryotes, but some plants and certain fungi contain a fusion protein containing domains homologous to both Rv1879 and to glutamine synthetase. The Operon and Rosetta Stone linkages suggest a possible role for Rv1879, and a possible functional association with the *glnA3* gene product.

uncharacterized biochemical pathways. Newly inferred functional linkages are given at [25].

Materials and methods

M. tuberculosis gene coordinates

Gene name, length, coordinates and orientation were downloaded from the Pasteur Institute TubercuList web server [26]. Gene coordinates were adjusted to include the stop codon of each gene.

Sanger Institute Functional Annotations

Sanger Institute *M. tuberculosis* H37Rv Functional Annotations were obtained from the Sanger *M. tuberculosis* web server [27].

SWISS-PROT Functional Annotations

M. tuberculosis SWISS-PROT Keywords were obtained from the Swiss Institute of Bioinformatics and European Bioinformatics Institute (EBI) SWISS-PROT web server [28].

Keyword recovery

Pairwise links between functionally-linked proteins were evaluated by a keyword recovery scheme [13] using the SWISS-PROT annotation for each of the tuberculosis proteins. For keyword recovery scores, pairs were evaluated only when both members of the pair had at least one SWISS-PROT keyword. The uninformative keywords: hypothetical protein, three-dimensional structure, transmembrane and complete proteome were discarded. Each pair of functionally-linked proteins received a preliminary score for the number of corresponding keywords which were shared between the two linked proteins. For example, consider the functional link at the 0 bp threshold between Rv0350 and Rv0351. Rv0350 has three SWISS-PROT keywords: ATP-binding, Chaperone and Heat shock. Rv0351 has two SWISS-PROT keywords: Chaperone and Heat shock. This link is assigned a preliminary score of two. This process was repeated for all linked genes, and a global measure of keyword recovery was derived by summing the individual link keyword scores and dividing by the total number of query keywords, as shown in Table 2.

The maximum false positive fraction was calculated by dividing the number of pairwise functional links that had no keywords in common by the total number of pairwise links. This estimate of false positives is presumably an upper limit because two linked genes might have related functions even in the absence of overlapping annotated functions.

The keyword recovery of random links is calculated by establishing the same number of random pairwise links between SWISS-PROT annotated genes as there is for real links, and calculating the keyword recovery as described above.

Operon method

A database consisting of the gene name, start coordinate, end coordinate and gene orientation was constructed and employed to determine functional links between genes using distance and orientation parameters. A series of genes is considered functionally linked if the nucleotide distance between genes in the same orientation was less than or equal to a specified distance threshold. Multiple genes were linked if a series of genes in the same orientation all had intergenic distances less than or equal to the defined distance threshold, as shown in Figure 1b.

Rosetta Stone method

Proteins were functionally linked by the Rosetta Stone method if individual proteins were found to be present as a single fused protein in another organism, as described by Marcotte *et al.* [10]. In this case, if individual *M. tuberculosis* proteins have significant homology to distinct regions of a single 'fusion' protein in another organism then they are indicated as functionally linked by this method. A probabilistic score is calculated by estimating the likelihood of observing Rosetta Stone proteins given the number of homologs each protein has.

Phylogenetic profile method

Phylogenetic profiles were used to identify proteins that evolved in a correlated fashion, as described by Pellegrini *et al.* [11]. A phylogenetic profile for each *M. tuberculosis* protein was created in the form of a bit vector, by searching for the presence or absence of homologs in each of the available fully-sequenced genomes. The presence of an identifiable homolog in a particular genome was indicated by the integer 1 in the bit vector at the position corresponding to that genome, while the absence of a homolog was indicated by the integer 0. Phylogenetic profiles were then clustered based on the similarity of profiles, resulting in clusters of genes with similar profiles and likely related functions.

Conserved Gene Neighbor method

Functional links were established by the conserved Gene Neighbor method where genes appear as chromosomal neighbors in multiple genomes, as described by Overbeek *et al.* [7] and Dandekar *et al.* [12]. For all possible pairs of *M. tuberculosis* genes, the nucleotide distance between homologs of these genes in all available sequenced genomes was calculated. Genes that were in close proximity in multiple genomes were indicated as functionally linked by this method. A probabilistic score reflects the likelihood of observing the intergenic distance between a pair of genes across all sequenced genomes.

Estimated fraction of adjacent gene pairs within the same direction (WD) that belong to operons

We employed the equation given by Moreno-Hagelsieb *et al.* [3] to estimate the fraction of *M. tuberculosis* WD pairs that are in common operons. The fraction of *M. tuberculosis* WD pairs with an intergenic distance between -20 bp and 30 bp were divided by the fraction of *E. coli* WD pairs with an intergenic distance between -20 bp and 30 bp. This number was then multiplied by 0.5, which was previously estimated to be the fraction of *E. coli* WD pairs that are in operons [3,4].

Acknowledgements

M.S. is supported by a USPHS National Research Service Award GM07185.

References

1. Madigan M, Martinko J, Parker J: *Brock Biology of Microorganisms* 9th edition. New Jersey: Prentice Hall; 2000.
2. Lodish H, Baltimore D, Berk A, Zipursky SL, Matsudaira P, Darnell J: *Molecular Cell Biology* 3rd edition. New York: Scientific American Books; 1995.
3. Moreno-Hagelsieb G, Collado-Vides J: **A powerful non-homology method for the prediction of operons in prokaryotes.** *Bioinformatics* 2002, **18:**329S-336S.
4. Salgado H, Moreno-Haelsieb G, Smith T, Collado-Vides J: **Operons in *Escherichia coli* : genomic analysis and predictions.** *Proc Natl Acad Sci USA* 2000, **97:**6652-6657.
5. Yada T, Nakao M, Totoki Y, Nakai K: **Modeling and predicting transcriptional units of *Escherichia coli* genes using hidden Markov models.** *Bioinformatics* 1999, **15:**987-993.
6. Salgado H, Santos-Zavaleta A, Gama-Castro S, Millan-Zarate D, Diaz-Peredo E, Sanchez-Solano F, Perez-Rueda E, Bonavides-Martinez C, Collado-Vides J: **RegulonDB (version 3.2): transcriptional regulation and operon organization in Escherichia coli K-12.** *Nucleic Acids Res* 2001, **29:**72-4.
7. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling.** *Proc Natl Acad Sci USA* 1999, **96:**2896-2901.
8. Ermolaeva MD, White O, Salzberg SL: **Prediction of operons in microbial genomes.** *Nucleic Acids Res* 2001, **29:**1216-1221.
9. Pellegrini M, Thompson M, Fierro J, Bowers P: **Computational method to assign microbial genes to pathways.** *J Cell Biochem Suppl* 2001, **Suppl 37:**106-109.
10. Marcotte EM, Pellegrini M, Ho-Leung N, Rice D, Yeates T, Eisenberg D: **Detecting protein function and protein-protein interactions from genome sequences.** *Science* 1999, **285:**751-753.
11. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc Natl Acad Sci USA* 1999, **96:**4285-4288.
12. Dandekar T, Snel B, Huynen M, Bork P: **Conservation of gene order: a fingerprint of proteins that physically interact.** *Trends Biochem Sci* 1998, **23:**324-328.
13. Marcotte EM, Pellegrini M, Thompson MJ, Yeates T, Eisenberg D: **A combined algorithm for genome-wide prediction of protein function.** *Nature* 1999, **402:**83-86.
14. Moreno-Hagelsieb G, Trevino V, Perez-Rueda E, Smith TF, Collado-Vides J: **Transcription unit conservation in the three domains of life: a perspective from *Escherichia coli*.** *Trends Genet* 2001, **17:**175-7.
15. Yanai I, Wolf YI, Koonin EV: **Evolution of gene fusions: horizontal transfer versus independent events.** *Genome Biol* 2002, **3:**research0024.1-0024.13.
16. Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, *et al.*: **The complete genome sequence of *Escherichia coli* K-12.** *Science* 1997, **277:**1453-1462.
17. Skovgaard M, Jensen LJ, Brunak S, Ussery D, Krogh A: **On the total number of genes and their length distribution in complete**

microbial genomes. *Trends Genet* 2001, **17:**425-8.
18. Fultz P, Kemper J: **Wild-type isopropylmalate isomerase in *Salmonella typhimurium* is composed of two different subunits.** *J Bacteriol* 1981, **148:**210-219.
19. Tamakoshi M, Yamagishi A, Oshima T: **The organization of the *leuC*, *leuD* and *leuB* genes of the extreme thermophile *Thermus thermophilus*.** *Gene* 1998, **222:**125-132.
20. Hatfull GF, Jacobs WR: *Molecular Genetics of Mycobacteria,* Washington, DC: ASM Press; 2000.
21. Harth G, Horwitz MA: **An inhibitor of exported Mycobacterium tuberculosis glutamine synthetase selectively blocks the growth of pathogenic mycobacteria in axenic culture and in human monocytes: extracellular proteins as potential novel drug targets.** *J Exp Med* 1999, **189:**1425-1436.
22. Wolf Y, Rogozin IB, Kondrashov AS, Koonin EV: **Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context.** *Genome Res* 2001, **11:**356-372.
23. Moreno-Hagelsieb G, Collado-Vides J: **Operon conservation from the point of view of *Escherichia coli*, and inference of functional inter-dependence of gene products from genome context.** In *Silico Biol* 2002, **2:**87-95.
24. Zheng Y, Roberts RJ, Kasif S: **Genomic functional annotation using co-evolution profiles of gene clusters.** *Genome Biol* 2002, **3:**research0060.1-0060.9.
25. **Computational Functional Linkages** [http://www.doe-mbi.ucla.edu/TB/computational/linkages]
26. **Pasteur Institute TubercuList** [http://genolist.pasteur.fr/TubercuList/index.html]
27. **The Sanger Institute: *M. tuberculosis*** [http://www.sanger.ac.uk/Projects/M_tuberculosis/Gene_list/]
28. **SWISS-PROT** [http://www.expasy.org/sprot/sprot_details.html]