Nucleic Acids Research, 2005, Vol. 33, No. 8 ***2521–2530***
doi:10.1093/nar/gki545

Nebulon: a system for the inference of functional relationships of gene products from the rearrangement of predicted operons

Sarath Chandra Janga, Julio Collado-Vides and Gabriel Moreno-Hagelsieb[1,*]

Program of Computational Genomics, CCG-UNAM, Apdo Postal 565-A, Cuernavaca, Morelos, 62100 Mexico and [1]Department of Biology, Wilfrid Laurier University, 75 University Avenue West, Waterloo, ON, N2L 3C5 Canada

Received March 2, 2005; Revised March 23, 2005; Accepted April 13, 2005

ABSTRACT

Since operons are unstable across Prokaryotes, it has been suggested that perhaps they re-combine in a conservative manner. Thus, genes belonging to a given operon in one genome might re-associate in other genomes revealing functional relationships among gene products. We developed a system to build networks of functional relationships of gene products based on their organization into operons in any available genome. The operon predictions are based on inter-genic distances. Our system can use different kinds of thresholds to accept a functional relationship, either related to the prediction of operons, or to the number of non-redundant genomes that support the associations. We also work by shells, meaning that we decide on the number of linking iterations to allow for the complementation of related gene sets. The method shows high reliability benchmarked against knowledge-bases of functional interactions. We also illustrate the use of Nebulon in finding new members of regulons, and of other functional groups of genes. Operon rearrangements produce thousands of high-quality new interactions per prokaryotic genome, and thousands of confirmations per genome to other predictions, making it another important tool for the inference of functional interactions from genomic context.

INTRODUCTION

Several methods have been proposed to infer functional relationships among gene products from genomic context. The three main methods, which we call The Three Musketeers of genomic context analysis, might be intuitively classified by their levels of confidence. The first and lowest confidence level would be the modular gain and loss of genes across genomes. One would expect that genes with related or interdependent functions would appear or disappear in concert in what is called a similar phyletic pattern (1,2) or phylogenetic profile (3). Next higher in confidence would be the conservation of gene order, also called the neighborhood method, related to the association of genes in operons (4,5). Finally, the evidence with highest confidence is that of a pair of genes that occur as a single fused gene in other genomes (6,7).

Operons, adjacent genes transcribed into a single messenger RNA, are generally formed by genes with related functions. The finding that operons are unstable (8) and that their associations change from one genome to another suggested the uber-operon (9), and the idea that by finding operon re-associations a researcher might be able to discover functional relationships among several genes (9,10). Note that this idea is somewhat opposite to that on conservation of gene order above, since the basis for its functionality is the lack of conservation of gene order where re-associations would indicate a functional relationship. Some systems to predict functional interactions based on conservation of gene order in evolutionarily distant genomes have appeared in the literature (11–13). These methods also exploit the rearrangement of gene neighborhoods, but still their main source of important neighborhoods depends on some level of conservation. Thus, there is no method that we know about that takes advantage of overall operon predictions to assign functional links between re-arranged genes on a genome-wide scale.

Here we present a simple method and tool, Nebulon, which can be used to infer functional relationships among genes based on the rearrangement of predicted operons. It should be noted that the method to predict operons is not dependent on conservation of gene order, but on inter-genic distances (14,15). The method is well established, and has been described as having the highest accuracy for operon predictions [for a review see (16)]. For instance, a very recently published method reports accuracy values on a per predicted

*To whom correspondence should be addressed. Tel: (519) 884-0710 ext 2364; Fax: (519) 746-0677; Email: gmoreno@wlu.ca

pair of 0.85 in *Escherichia coli* K12 and of 0.83 in *Bacillus subtilis* (17), while the current accuracies with the distance-based method are of 0.84 and 0.87, respectively. We integrate this tool with the known criterion that orthologs of genes which are fused in another genome are known to functionally interact (6,7).

IMPLEMENTATION

To facilitate descriptions, we use 'problem genome' to refer to a given genome where we want to find functional relationships, and 'problem gene' to refer to any given gene whose functional relationships are to be found. To find functional associations between any pair of genes in a problem genome we need two things: (i) operon predictions across genomes and (ii) orthologs. The basic idea is that for any two genes inside a problem genome to be linked they have to be either in the same predicted operon inside the problem genome, or their orthologs have to be in the same predicted operon in any other genome. An internal link would be the prediction that the two genes are in the same operon inside the problem genome, while an external link would be the finding that the orthologs of the two genes are predicted to be in the same operon in another genome. The links for all pairs of genes within the problem genome can then be used to build a network of interactions (Figure 1).

To predict operons we used a previously published method based on inter-genic distance (14,15). Briefly, log-likelihoods (LLHs) are calculated for each inter-genic distance at 10 bp intervals by taking the base 10 logarithm of the result of dividing the fraction of known within-operon (WO) pairs of genes found at such distance by the fraction of genes at transcription-unit-boundaries (TUBs) pairs within the same inter-genic distance interval (14,15). Here, we extend the scores to include indirect LLHs for pairs of genes not immediately adjacent but present in the same operon, in which case the LLH used is the minimum required to include all intervening genes. Note again that these predictions do not depend at all on conservation of gene order. Operons were predicted across a non-redundant genome collection built according to a previous definition of genome redundancy (15).

Several kinds of thresholds can be adjusted to improve the confidence in any prediction generated by Nebulon: (i) the minimal LLH to accept an operon prediction. (ii) The number of 'shells' to reach for linkages, where the first shell would be composed of genes linked to a first problem gene or set of genes. The second shell would be composed of genes linked to the genes of the first shell and so on. (iii) The number of associations, by which we mean finding more than one example of an operon association either within a single genome due to recent duplications of operons where many copies can be orthologs of a single pair of genes in the problem genome, and/or finding the link across several genomes. For most cases the number of associations equals the number of genomes. (iv) The number of genomes where such predictions occur. As the number of genomes confirming the link increases, the method becomes more similar to those based on conservation of gene order (11–13).

Our working definition of orthology consisted of BLASTP reciprocal-best hits and fusions as described previously (15). It should be noticed that this definition does allow for more than one ortholog per genome. Fusions will consist of two genes

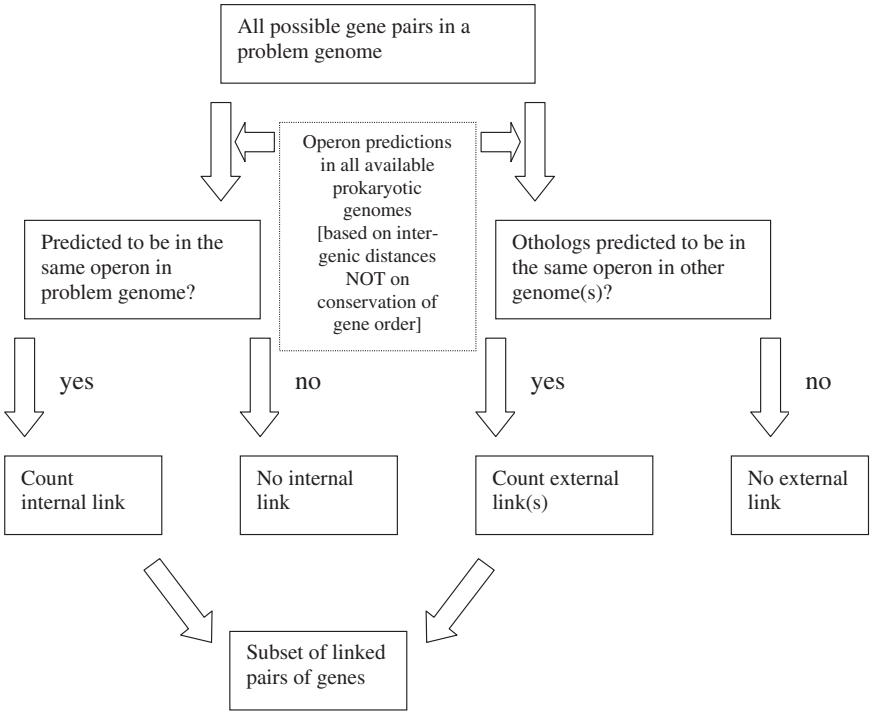

Figure 1. Finding links by operon rearrangement. Operon predictions are based on a well established method which relies solely on inter-genic distances (14,15), and not on conservation of gene order. This is the main difference with other available tools (11–13). Though we also incorporate fusions >99% of our links come from operon predictions alone.

pointing to the very same gene as their ortholog in another genome, and the protein sequences of several gene pairs can sometimes have the very same BLAST scores, most of them due perhaps to recent duplications. We ran BLASTPGP (18) to compare all the proteins annotated within the current collection of genomes as found at the Entrez Genome Database (19) (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). The *E*-value cut-off was 1×10^{-6} with a database size fixed at 5×10^8 ($-z\ 5 \times 10^8$), soft filtering of low information content sequences (the $-F$ 'mS' option of the NCBI BLASTPGP program), and a final Smith–Waterman alignment ($-sT$) (20). We also required coverage of at least 50% of any of the protein sequences in the alignment.

NEBULON RECOVERS REAL FUNCTIONAL LINKS (BENCHMARKING)

To evaluate the content of the networks, we compared all the Nebulon links for *E.coli* K12 with the metabolic pathway map of the same organism obtained from KEGG (21,22), where genes in the same metabolic pathway have a related function. We compared the number of links found to a set of 1000 random networks generated maintaining the same connectivity and shuffling the gene labels $n \times (n - 1)/2$ times where n is the number of edges in the original network. Figure 2a shows the distribution of recovery of KEGG functional interactions in random networks. The average random recovery was 105.71, while in the real network we recovered 1384 interactions. The z-score of real versus randomized results was 103.30.

We also obtained the set of experimentally known interactions in *E.coli* from the Database of Interacting Proteins (DIP) (23,24). We found 516 interactions with evidence from at least one experimental study in DIP. After filtering the data to exclude self-interacting genes and to keep only those with a GenBank identifier, the dataset reduced to 238. We recovered 97 of the 238 interactions (41%) in the real network, while with the randomized networks we recovered an average of 1.7 interactions (Figure 2b). The z-score was 78.06. From these two examples, KEGG and DIP, we can conclude that the links found by Nebulon are far from random and highly significant.

False positives are very hard to define. However, it is possible to give a sense of the increase in the quality of predictions with thresholds. In Figure 3, we show how the fraction of predictions confirmed by being within the same KEGG pathway increases as we increase the LLH threshold and with the number of associations. This is the same

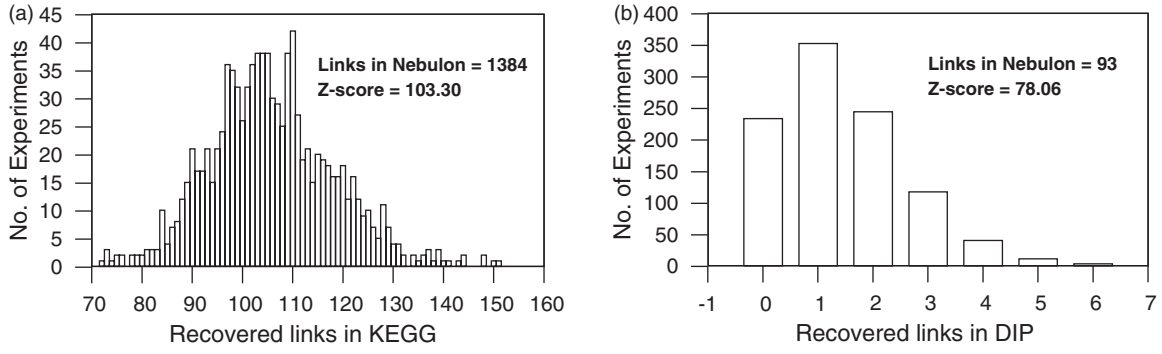

Figure 2. (**a**) Distribution of KEGG links recovered in 1000 randomly shuffled networks keeping the connectivity fixed in *E.coli* K12. (**b**) Distribution of DIP links obtained in same set of 1000 random networks.

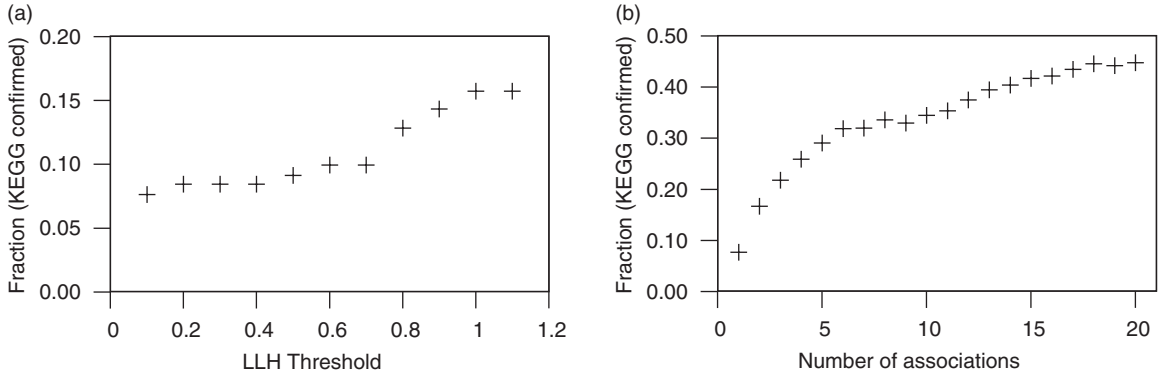

Figure 3. Effect of increasing thresholds on the quality of predictions. We used the fraction of predicted links whose products work within the same KEGG metabolic pathway as a measure of quality. (**a**) Effect of increasing the LLH to accept an operon prediction. (**b**) Effect of increasing the number of associations (number of times the genes are found in the same operon). The measure is far from perfect, but it does give a sense of what happens as thresholds increase. The apparently slow growth in quality with increasing LLH is due to the 0.0 threshold being high to start with. Operon predictions have a positive predictive value (true positives divided by the sum of true positives and false positives) of 0.86 at a 0.0 LLH, and of 0.93 at 1.0 LLH in *E.coli* K12.

kind of graph used for the same purpose in the STRING implementations (25).

EXTERNAL OPERON PREDICTIONS CONTRIBUTE THE MOST TO LINK RECOVERY

In Figure 4, we show the fraction of interactions recovered from internal operon predictions as compared with those obtained from external operon predictions from the perspective of *E.coli*. As mentioned in the Implementation, internal means those links recovered from operon predictions inside the problem genome, and external are those recovered from operon predictions in other genomes. As seen in Figure 4a, ~0.8 of the total recovered DIP links can be found by internal operon predictions, and almost all of them can be recovered from operon predictions elsewhere. In KEGG just 0.4 of the total recovered links can be found with internal predictions, while external predictions can find 0.9. Particular pathways in KEGG vary from almost none to almost all links recoverable from internal predictions, while recoveries by external predictions keep close to the total (Figure 4b). These results show the value of using operon predictions across genomes, at the same time they show that several operon rearrangements are indeed conservative in the sense that they occur among genes with related functions as suggested previously (9,10).

The same power of external predictions is obvious across genomes as most of the linkage generated for any genome can be obtained from external operon predictions (0.87 on average), while the linkage generated from internal operon predictions is generally poor (0.21 on average), except for *Borrelia burgdorferi* followed by *Mycoplasma pneumoniae* and *Mycoplasma genitalium* (see Supplementary Figure).

EXAMPLES

In this section we exemplify the use of Nebulon as applied to different problems around different kinds of functional gene modules. We chose the examples below mainly inspired by the interests of close collaborators. It is expected that different functional gene modules will have different 'strengths' of association, and thus, parameters should be adjusted on a per-problem basis.

NEBULON DIFFERS FROM OTHER GENOMIC CONTEXT TOOLS

The ArgR transcription factor and its binding site are universally conserved in bacterial genomes (26). In Figure 5, we present the recovery of the ArgR regulon and other likely interacting partners using Nebulon as applied from the perspective of *E.coli* K12. We used a minimum LLH of 0.4. As can be seen, we recover most of the transcription units known to be regulated by this transcription factor as reported in RegulonDB (27), namely *argA*, *argCBH*, *argD*, *argF*, *argG* and *argI*, except *argE*, *carAB* and *artPIQM*. There are some apparently new interactions that have not been reported in RegulonDB (Table 1). Among these new interactions, the genes *gmk*, *ychE* and *yfjB* have previously been predicted to be regulated by ArgR (28).

Since we wanted to know whether Nebulon, being a different concept, can find links not previously found by the current types of genomic context analyses, we also used STRING (25) to find the ArgR regulon. STRING puts together genomic context tools, computational analyses of high-throughput experiments and text mining, to find probable functional associations of genes. It includes implementations of the genomic context analysis tools mentioned at the

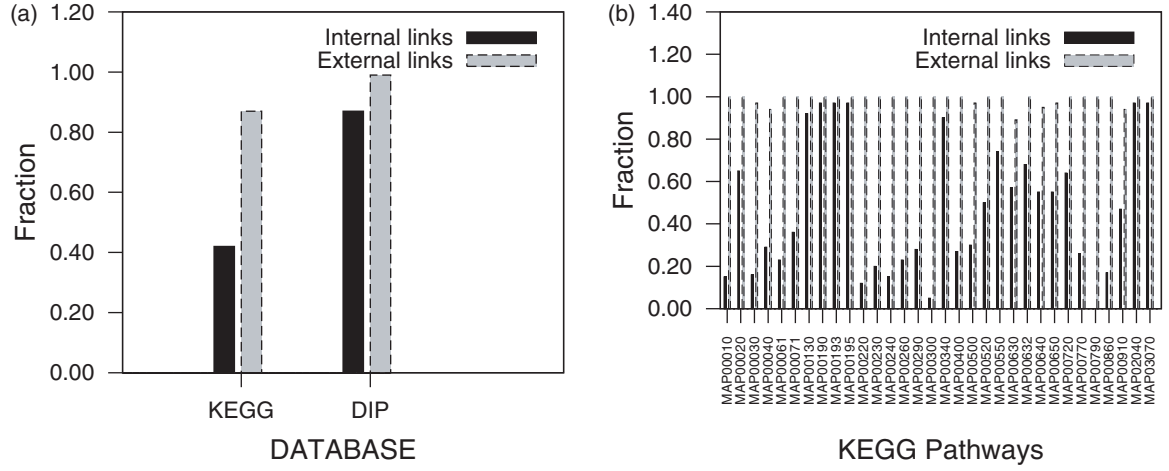

Figure 4. Fraction of internal versus external links found in *E.coli* K12 network. (**a**) KEGG and DIP datasets. (**b**) Fraction of internal versus external links found in *E.coli* K12 network for each pathway in KEGG. Pathway identifiers mean MAP00193: ATP synthesis; MAP00632: Benzoate degradation via CoA ligation; MAP00650: Butanoate metabolism; MAP00020: Citrate cycle (TCA cycle); MAP00061: Fatty acid biosynthesis (path 1); MAP00071: Fatty acid metabolism; MAP02040: Flagellar assembly; MAP00790: Folate biosynthesis; MAP00260: Glycine, serine and threonine metabolism; MAP00010: Glycolysis/Gluconeogenesis; MAP00630: Glyoxylate and dicarboxylate metabolism; MAP00340: Histidine metabolism; MAP00300: Lysine biosynthesis; MAP00910: Nitrogen metabolism; MAP00520: Nucleotide sugars metabolism; MAP00190: Oxidative phosphorylation; MAP00770: Pantothenate and CoA biosynthesis; MAP00040: Pentose and glucuronate interconversions; MAP00030: Pentose phosphate pathway; MAP00550: Peptidoglycan biosynthesis; MAP00400: Phenylalanine, tyrosine and tryptophan biosynthesis; MAP00195: photosynthesis; MAP00860: porphyrin and chlorophyll metabolism; MAP00640: propanoate metabolism; MAP00230: purine metabolism; MAP00240: pyrimidine metabolism; MAP00720: reductive carboxylate cycle (CO2 fixation); MAP00500: starch and sucrose metabolism; MAP03070: type III secretion system; MAP00130: ubiquinone biosynthesis; MAP00220: Urea cycle and metabolism of amino groups; and MAP00290: valine, leucine and isoleucine biosynthesis.

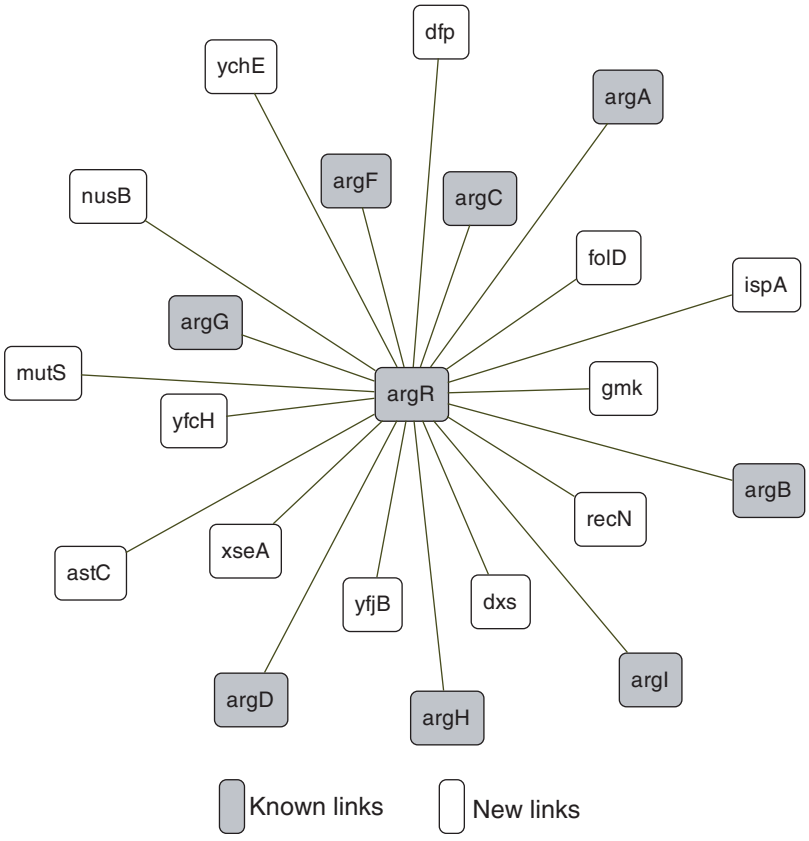

Figure 5. Links to the *argR* gene coding for the ArgR transcription factor in *E.coli* K12 using a LLH threshold of 0.4 and associations found in at least one genome.

Table 1. Details of the newly found links in the recovery of the ArgR regulon

Gene	No. of associations and genomes in which the evidence is found	No. of intervening genes and LLHs	Function of protein
recN[a]	5—*Bacillus halodurans*, *B.subtilis*, *Oceanobacillus iheyensis*, *Staphylococcus aureus* Mu50, *Thermoanaerobacter tengcongensis*	0 (0.4291), 0 (0.4291), 0 (0.5067), 0 (0.8840), 0 (0.8840)	Protein used in recombination and DNA repair
astC[a]	3—*Corynebacteriumefficiens* YS-314, *Mycobacterium avium paratuberculosis*, *Streptomyces Coelicolor*	1 (0.8840), 1 (1.1343), 0 (0.7944)	Amino acid biosynthesis, arginine acetylornithine delta-aminotransferase
mutS[a]	2—*Streptococcus agalactiae* 2603, *Streptococcus pneumoniae* R6	0 (0.1721), 0 (0.8840)	DNA-replication, repair. Methyl-directed mismatch repair
yfcH[a]	2—*Haemophilus ducreyi* 35000HP, *Pasteurella multicoda*	0 (1.1343), 0 (0.5067)	Putative enzyme
dfp[a]	1—*Thermus thermophilus* HB27	0 (0.7944)	DNA-replication, repair. Flavoprotein affecting synthesis of DNA and pantothenate metabolism
gmk	1—*T.thermophilus* HB27	2 (0.7944)	Purine ribonucleotide biosynthesis, guanylate kinase
ychE[a]	1—*T.thermophilus* HB27	0 (0.7944)	Putative transport
dxs	1—*T.tengcongensis*	2 (0.4291)	Central intermediary metabolism, 1-deoxyxylulose-5-phosphate synthase
yfjB[a]	1—*T.tengcongensis*	0 (0.8840)	Hypothetical protein
folD[b]	1—*O.iheyensis*	4 (0.5067)	Biosynthesis of cofactors, folic acid 5,10-methylene-tetrahydrofolate dehydrogenase
ispA	1—*O.iheyensis*	1 (0.5067)	Biosynthesis of cofactors, geranyltransferase
nusB[b]	1—*O.iheyensis*	5 (0.5067)	RNA synthesis, transcription termination, L factor
xseA[b]	1—*O.iheyensis*	3 (0.5067)	Degradation of DNA

[a]Cases where we expect the genes to be linked functionally because the LLH scores are high and the orthologs are conserved with no intervening genes in the genome of evidence. The genes *gmk*, *ychE* and *yfjB* have been predicted to be regulated by ArgR (28). It can also be noticed that in all these cases the genes are either putative, hypothetical or poorly annotated indicating the possibility of these associations to be real.
[b]In all 13 of these links we only expect the links marked (3 in number) to be false positives because of the high number of intervening genes. Such links could serve as a guide for future refinements in Nebulon. Complete genome names can be found in the website.

introduction: conserved gene neighborhood, gene fusions and co-occurrence (phylogenetic profiles). Searching with default parameters, and a maximum number of 200 hits, for ARGR_ECOLI allowed us to find seven members of the regulon: ArgA, ArgC, ArgD, OTC2 (ArgF), ASSY (ArgG), OTC1 (ArgI) and CarA. STRING found all of these links from text mining with no contribution from genomic context tools. It is expected that Nebulon will recover links not previously found by other genomic context tools, as in Nebulon there is no requirement for the genes to be kept adjacent in evolutionarily distant genomes. It is enough to have them predicted in operons within the LLH threshold in any genome to obtain associations. Other links found by STRING through text mining were AmpA (PepA), BtuB, RL13 (RplM), TyrR, XerC and XerD.

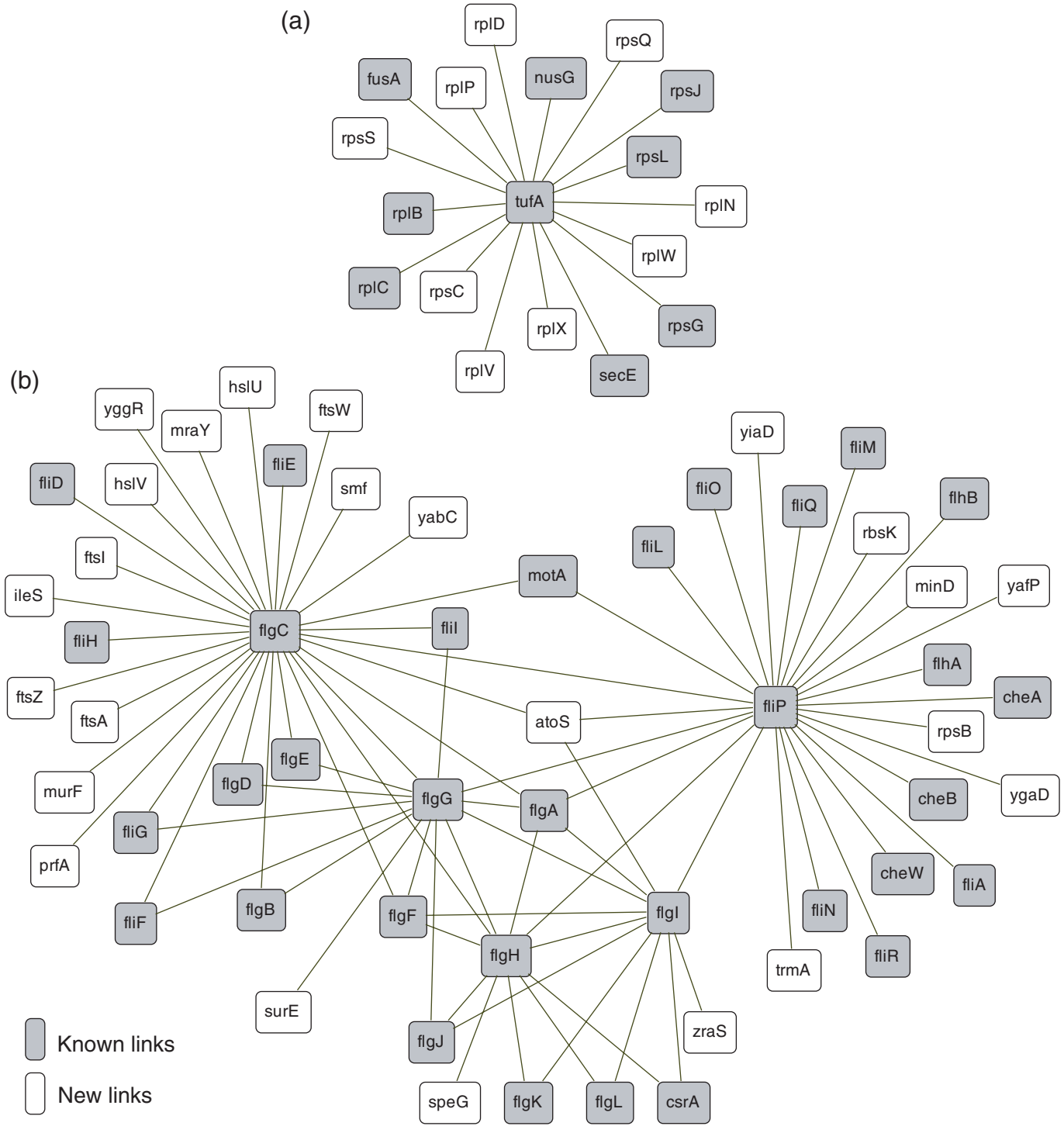

Figure 6. Uber-operon recovery. (**a**) Links to *tuf*A in Nebulon with a LLH threshold of 0.4. (a) Minimum number of evidences set to 1. (**b**) Minimum number of evidences set to 2. (b) Two shells of links to *flgA* in Nebulon showing known and predicted associations.

The genes MetJ, MutH, YciS and YejL were found through co-occurrence analysis (phylogenetic profiles), and TrpR was found by both text mining and co-occurrence analysis.

Interestingly, Nebulon found several links to genes coding for proteins involved in DNA recombination and repair (Table 1). At first, this result surprised us, but the STRING results show also proteins involved in recombination and repair. Reading through the abstracts used by the text-mining tool in STRING we learned that ArgR has been found to be involved in site-specific recombination [see for instance (29,30)]. Thus, despite a lack of precise coincidence between the Nebulon and STRING findings, both tools linked ArgR to this kind of activity.

This example does show that operon rearrangements link ArgR-regulated genes with the *argR* gene, while no other genomic context tool, at least as implemented in STRING, does the job. However, we are not claiming operon rearrangements to be better than other genomic context tools, but that they recover new links, and thus they can be a complement. We downloaded the dataset of linkages with medium or better confidence values from the latest version of STRING (version 6.0) to have a view as of how much of a complement Nebulon might be. In *E.coli* K12 STRING reports 109 710 links. We compared these with 17 358 links found by operon rearrangements at 0.4 LLH, which have a positive predictive value [true positives/(true positives + false positives)] of ~0.9 for operon predictions in both *E.coli* K12 and *B.subtilis* (data not shown). The intersection with the STRING links is 10 718, leaving 6640 new links obtained with Nebulon. We consider 6640 links a very good contribution where ~38% of all Nebulon links in *E.coli* K12 would be new and the rest might be considered confirmations that can give more confidence to those links within STRING. Unfortunately, the STRING links file does not have details on the evidence of the functional linkage, which, again, can come from either a single or a combination of genomic context, high-throughput experiments, or text mining. Thus, the contribution to genomic context functional inference from Nebulon might be much higher. In *B.subtilis* we found 47 330 STRING links and 16 224 Nebulon links. The intersection is 6168,

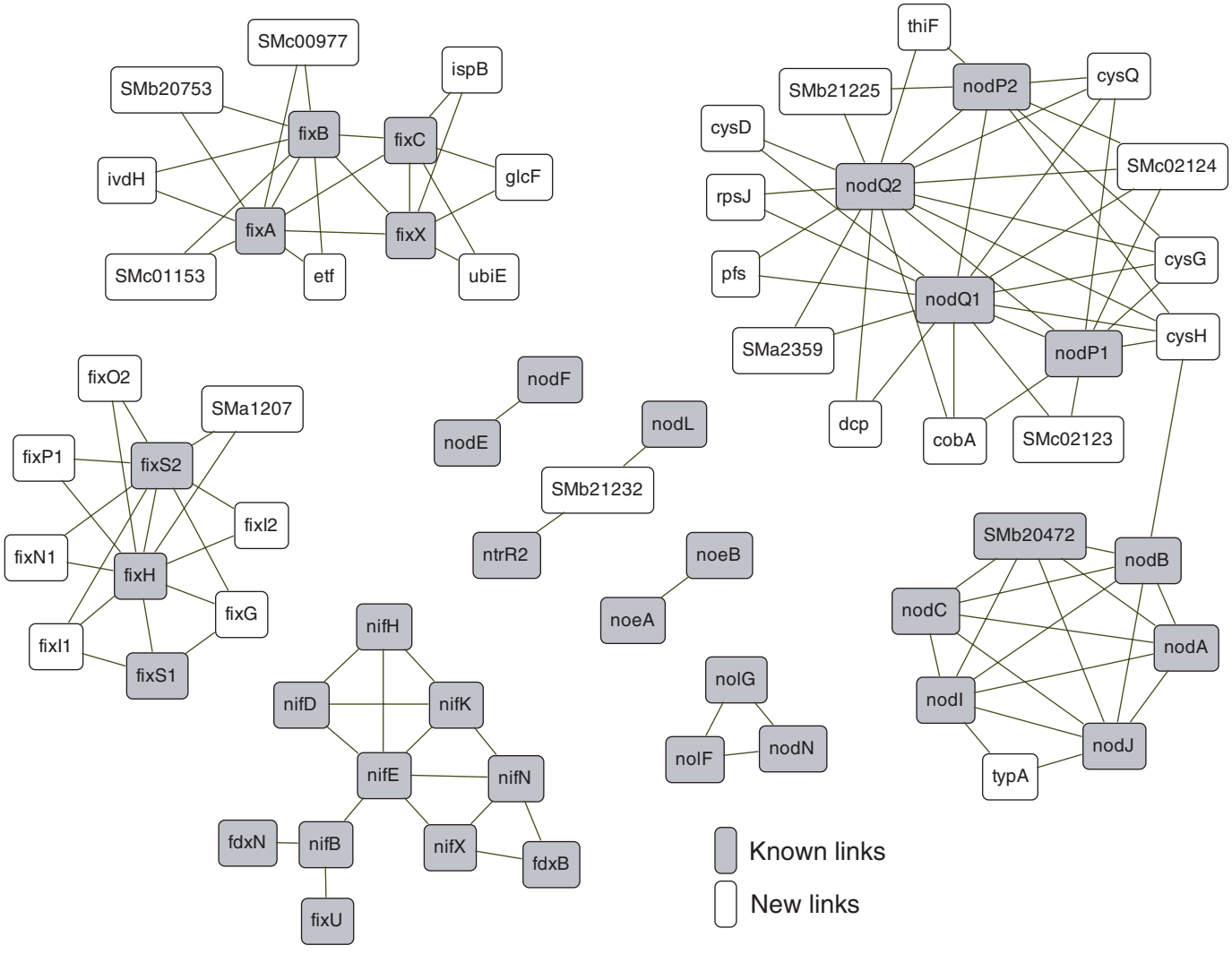

Figure 7. Links among genes involved in Nitrogen fixation and in Nodulation of *S.meliloti*. Core genes refer to genes annotated as involved in these activities in *S.meliloti* (31), while non-core are other linked genes found by Nebulon.

and thus 10 056 (62%) Nebulon links would be new. In *Sinorhizobium meliloti* we found more interesting numbers, 59 640 STRING links, 20 656 Nebulon links, and an intersection of 1686. Thus, we had 18 970 (92%) new links. On the website, we provide a table comparing Nebulon and STRING for the complete list of prokaryotic genomes that can be mapped from STRING to the Entrez genome database. A more comprehensive and detailed analysis of links found by any genomic context tools, their intersections and differences should be the focus of future work.

NEBULON RECOVERS COMPLETE UBER-OPERONS

Uber-operons are defined as context conservation of groups of genes at a higher level than operons (9). Such higher level refers to the expectation that the gene members of a given operon will re-associate in a conservative manner. In other words, they will be found in other genomes in operons with other genes related to the function of all the genes in the first operon. Thus, in a strict sense, Nebulon would be an uber-operon recovery tool. We show here two examples of recovered uber-operons again from the perspective of *E.coli* K12 using Nebulon with high LLH thresholds (Figure 6).

In Figure 6a, we show the set of genes linked to the translation associated gene *tufA*. For the purpose of evaluating uber-operon recovery, we show the genes mentioned by Lathe *et al*. (9) in gray. Other genes linked to the original set (shown in white) are all known to be associated with translation indicating the ability of Nebulon to recover context sensitive links at high thresholds while keeping false positives to a minimum.

In the case of flagella-related machinery, we used *flgA* as input and obtained two shells of links. As can be seen in Figure 6b, a large fraction of recovered genes are those mentioned previously by Lathe *et al*. (9). Genes not mentioned by Lathe *et al*. comprise ~40% of the genes in the network. Detailed analysis of these outliers indicates that ~25% of them are annotated either as putative or as hypothetical while the rest seem to be loosely associated with the flagella-related genes, like those involved in transport, membrane and cell division.

NEBULON RECOVERS NITROGEN FIXATION-RELATED GENES IN *S.MELILOTI*

To exemplify a possible role for Nebulon in annotation projects, we took a core gene set consisting of the 47 genes annotated as related to 'Nitrogen fixation' and 'Nodulation' in *S.meliloti* (31). This functional gene module required us to play more with the parameters. We started by obtaining the first shell for each of these genes. Increasing the minimal number of associations to accept a link from 1 to 2 decreased the overall number of links from 288 to 95 and participation of

Table 2. Genes having at least two links with genes related to nitrogen fixation in *S.meliloti*

Gene	No. of links to core	Function of protein
cysH	5	Probable thioredoxin dependent padops reductase 3′-phosphoadenylylsulfate sulfotransferase cysteine biosynthesis protein
cysG	4	Probable siroheme synthase protein
cysQ	4	Putative transmembrane protein
SMc02124	4	Putative nitrite reductase protein
cobA	3	Probable uroporphyrin-III C-methyltransferase protein
fixG	3	Iron sulfur membrane protein
fixI1	3	Copper transport ATPase
cysD	2	Putative sulfate adenylate transferase subunit 2 cysteine biosynthesis protein
dcp	2	Probable peptidyl-dipeptidase A protein
etf	2	Probable electron transfer flavoprotein-ubiquinone oxidoreductase
fixI2	2	E1–E2 type cation ATPase
fixN1	2	Heme b/copper cytochrome *c* oxidase subunit
fixO2	2	cytochrome *c* oxidase
fixP1	2	Di-heme cytochrome *c*
glcF	2	Probable glycolate oxidase iron-sulfur subunit protein
ispB	2	Putative octaprenyl-diphosphate synthase protein
ivdH	2	Putative isovaleryl-CoA dehydrogenase protein
pfs	2	Putative MTA/SAH nucleosidase P46 includes: 5′-methylthioadenosine nucleosidase and *S*-adenosylhomocysteine nucleosidase protein
rpsJ	2	Probable 30S ribosomal protein S10
SMa1207	2	FixK-like regulatory protein
SMa2359	2	Conserved hypothetical protein
SMb20753	2	Putative acyl-CoA dehydrogenase protein
SMb21225	2	Putative inositol monophosphatase, possibly involved in PAPS metabolism protein
SMb21232	2	Putative nucleotide sugar epimerase dehydratase protein
SMc00977	2	Putative acyl-COA dehydrogenase protein
SMc01153	2	Probable enoyl COA hydratase protein
SMc02123	2	Conserved hypothetical protein
thiF	2	Putative Thiamine biosynthesis transmembrane protein
typA	2	Probable GTP-binding protein
ubiE	2	Probable ubiquinone/menaquinone biosíntesis methyltransferase protein

the core genes in finding new links decreased from 43 to 41, resulting in the loss of interconnectivity between smaller modules. Thus, we tried to work with the one-association network. We excluded non-core genes that had fewer than two links to any core-genes to minimize the number of false positives. Doing so not only reduced the number of links to 119 but also gave a high-quality set of probable interacting partners (Figure 7 and Table 2). Some of the linked genes are obviously involved in nitrogen fixation as revealed from the gene root name '*fix*,' or as deduced from some of those lacking a name being homologs to other '*fix*' genes. STRING also found *cysD* and *cysH*. Exploring through the text-mining data in STRING we found that several '*cys*' genes are involved in nodulation in *S.meliloti* and other Rhizobia (32–34). Thus, it might be worth exploring the possible role of other Nebulon-linked genes in nitrogen fixation and nodulation.

CONCLUSIONS

In this work we have shown that Nebulon recovers real functional links as defined in the KEGG metabolism database and in the DIP, as well as revealed in the specific examples. Most of the recovered links found across genomes are due to the use of operon predictions in genomes other than the problem genome, revealing that much of the rearrangement of genes into different operons across genomes actually happens among genes with related functions. As with other tools, false positives are not easy to estimate, a reason for us to prefer to call these tools 'hypothesis generators' rather that 'prediction tools'. As we have shown, different problems might require different kinds of thresholds to reduce the rate of false positives and generate the best hypotheses for interactions to be tested in the laboratory. Overall, we have demonstrated that functional linkages can be recovered with high confidence using distance-based operon predictions and that such linkages complement what other tools find. Thus, 'comparative operomics' might be the D'Artagnan to 'The Three Musketeers' (phylogenetic profiles, conservation of gene order and fusions) currently in use for the inference of functional interactions from genomic context.

Nebulon can be accessed online at http://tikal.cifn.unam.mx/nebulon/. Supplementary information and a command-line version of Nebulon are available at: http://tikal.cifn.unam.mx/nebulon/nebulon.html.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Warren F. Lamboy for critical reading of the manuscript. We also thank Heladia Salgado for help with database design, and Edgar Peredo Díaz for designing a web interface for Nebulon. We also acknowledge two anonymous referees for comments and suggestions. G.M.-H. acknowledges funds from WLU and support from SHARCNET as chair in Biocomputing. Funding to pay the Open Access publication charges for this article was provided by WLU (Wilfrid Laurier University).

Conflict of interest statement. None declared.

REFERENCES

1. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
2. Gaasterland,T. and Ragan,M.A. (1998) Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microb. Comp. Genomics*, **3**, 199–217.
3. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
4. Dandekar,T., Snel,B., Huynen,M. and Bork,P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci*., **23**, 324–328.
5. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.
6. Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
7. Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.
8. Itoh,T., Takemoto,K., Mori,H. and Gojobori,T. (1999) Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol. Biol. Evol*., **16**, 332–346.
9. Lathe,W.C.,III, Snel,B. and Bork,P. (2000) Gene context conservation of a higher order than operons. *Trends Biochem. Sci*., **25**, 474–479.
10. Galperin,M.Y. and Koonin,E.V. (2000) Who's your neighbor? New computational approaches for functional genomics. *Nat. Biotechnol*., **18**, 609–613.
11. Rogozin,I.B., Makarova,K.S., Murvai,J., Czabarka,E., Wolf,Y.I., Tatusov,R.L., Szekely,L.A. and Koonin,E.V. (2002) Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res*., **30**, 2212–2223.
12. Snel,B., Bork,P. and Huynen,M.A. (2002) The identification of functional modules from the genomic association of genes. *Proc. Natl Acad. Sci. USA*, **99**, 5890–5895.
13. Yanai,I., Mellor,J.C. and DeLisi,C. (2002) Identifying functional links between genes using conserved chromosomal proximity. *Trends Genet*., **18**, 176–179.
14. Salgado,H., Moreno-Hagelsieb,G., Smith,T.F. and Collado-Vides,J. (2000) Operons in *Escherichia coli*: genomic analyses and predictions. *Proc. Natl Acad. Sci. USA*, **97**, 6652–6657.
15. Moreno-Hagelsieb,G. and Collado-Vides,J. (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, **18** (Suppl. 1), S329–S336.
16. Stormo,G.D. and Tan,K. (2002) Mining genome databases to identify and understand new gene regulatory systems. *Curr. Opin. Microbiol*., **5**, 149–153.
17. Price,M.N., Huang,K.H., Alm,E.J. and Arkin,A.P. (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res*., **33**, 880–892.
18. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
19. Tatusova,T.A., Karsch-Mizrachi,I. and Ostell,J.A. (1999) Complete genomes in WWW Entrez: data representation and analysis. *Bioinformatics*, **15**, 536–543.
20. Schaffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res*., **29**, 2994–3005.
21. Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H. and Kanehisa,M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*., **27**, 29–34.
22. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.
23. Xenarios,I., Rice,D.W., Salwinski,L., Baron,M.K., Marcotte,E.M. and Eisenberg,D. (2000) DIP: the database of interacting proteins. *Nucleic Acids Res*., **28**, 289–291.

24. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*., **32**, D449–D451.
25. von Mering,C., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P. and Snel,B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res*., **31**, 258–261.
26. Makarova,K.S., Mironov,A.A. and Gelfand,M.S. (2001) Conservation of the binding site for the arginine repressor in all bacterial lineages. *Genome Biol*., **2**, RESEARCH0013.
27. Salgado,H., Gama-Castro,S., Martinez-Antonio,A., Diaz-Peredo,E., Sanchez-Solano,F., Peralta-Gil,M., Garcia-Alonso,D., Jimenez-Jacinto,V., Santos-Zavaleta,A., Bonavides-Martinez,C. *et al*. (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res*., **32**, D303–D306.
28. Robison,K., McGuire,A.M. and Church,G.M. (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J. Mol. Biol*., **284**, 241–254.
29. Sherratt,D.J., Arciszewska,L.K., Blakely,G., Colloms,S., Grant,K., Leslie,N. and McCulloch,R. (1995) Site-specific recombination and circular chromosome segregation. *Philos. Trans. R. Soc. Lond. B Biol. Sci*., **347**, 37–42.
30. Colloms,S.D., Alen,C. and Sherratt,D.J. (1998) The ArcA/ArcB two-component regulatory system of *Escherichia coli* is essential for Xer site-specific recombination at psi. *Mol. Microbiol*., **28**, 521–530.
31. Galibert,F., Finan,T.M., Long,S.R., Puhler,A., Abola,P., Ampe,F., Barloy-Hubler,F., Barnett,M.J., Becker,A., Boistard,P. *et al*. (2001) The composite genome of the legume symbiont *Sinorhizobium meliloti*. *Science*, **293**, 668–672.
32. Schwedock,J. and Long,S.R. (1990) ATP sulphurylase activity of the nodP and nodQ gene products of *Rhizobium meliloti*. *Nature*, **348**, 644–647.
33. Schwedock,J.S., Liu,C., Leyh,T.S. and Long,S.R. (1994) *Rhizobium meliloti* NodP and NodQ form a multifunctional sulfate-activating complex requiring GTP for activity. *J. Bacteriol*., **176**, 7055–7064.
34. Snoeck,C., Verreth,C., Hernandez-Lucas,I., Martinez-Romero,E. and Vanderleyden,J. (2003) Identification of a third sulfate activation system in *Sinorhizobium* sp. strain BR816: the CysDN sulfate activation complex. *Appl. Environ. Microbiol*., **69**, 2006–2014.