

Improving Genome Annotations using Phylogenetic Profile Anomaly Detection

Tarjei S. Mikkelsen^{*}, James E. Galagan and Jill P. Mesirov

The Eli & Edythe L. Broad Institute

Massachusetts Institute of Technology and Harvard University

320 Charles Street, Cambridge, MA 02141, USA

Running title: Phylogenetic Profile Anomaly Detection

^{*}To whom correspondence should be addressed

Abstract

Motivation: A promising strategy for refining genome annotations is to detect features that conflict with known functional or evolutionary relationships between groups of genes. Previous work in this area has been focused on investigating the absence of “housekeeping” genes or components of well-studied pathways. We have sought to develop a method for improving new annotations that can automatically synthesize and use the information available in a database of other annotated genomes.

Results: We show that a probabilistic model of phylogenetic profiles, trained from a database of curated genome annotations, can be used to reliably detect errors in new annotations. We use our method to identify 22 genes that were missed in previously published annotations of prokaryotic genomes.

Availability: The method was evaluated using MATLAB and open source software referenced in this work. Scripts and datasets are available from the authors upon request.

Contact: tarjei@broad.mit.edu

Introduction

The first step in understanding a newly sequenced genome is to identify its genes and their putative functions. Unfortunately, current homology-based annotations of protein-coding genes remain far from perfect, even for relatively simple bacterial genomes (Brenner 1999; Devos 2001, Bocs 2002, Iliopoulos 2003). Broadly speaking, their shortcomings can be divided into two categories. First, as many as 40% of predicted genes in most bacterial genomes remain annotated as ‘hypothetical’ or ‘unknown’ proteins. Second, even for genes that do have significant database matches to known genes, systematic errors in the annotation methods themselves, such as software defects, unrealistic statistical models, and failure to distinguish functional genes disrupted by sequencing errors from pseudogenes, can lead to annotation inaccuracies. Here, we describe a statistical method designed to detect the latter type of errors.

A promising strategy for refining genome annotations is to take into account functional or evolutionary relationships between groups of genes. Genomes continually evolve through gene transfer, duplication and loss. However, the presence or absence of specific genes in a genome is generally not arbitrary, but reflects functional and structural dependencies between the proteins they encode. For example, metabolic enzymes are the building blocks of biochemical pathways that would be inactive, and perhaps even fatal to the organism, if incomplete. Other genes encode connected and interdependent subcomponents of large complexes, such as ribosomes and flagella. Because incomplete, non-functional metabolic pathways and cellular complexes confer little evolutionary benefit to an organism, groups of functionally or structurally linked genes tend to be

either present in their entirety or completely absent (Pellegrini 1999). As a consequence, unexpected co-occurrence patterns ('phylogenetic profiles' after Pellegrini *et al.*) may point towards systematic errors in an annotation. If only one gene from a group of co-occurring genes is missing in an annotation, there is reason to suspect that it is a false negative that has somehow been overlooked by the annotation system. Similarly, if only one gene from a co-occurring group is present, it may be a false positive.

Several groups have demonstrated the utility of gene dependencies in improving genome annotations, using varying degrees of phylogenetic and functional information. Methods for inferring missing genes from well-studied gene families were developed long before the first complete genomes were sequenced (Goodman 1979; Page 1994). These methods use character-based approaches to find genes whose presence or absence are inconsistent with a known phylogeny. More recently, Natale *et al.* (2000) reported finding novel genes in previously annotated bacterial genomes by examining specific unexpected phylogenetic patterns, such as a gene being present in all but one of the genomes, in the Clusters of Orthologous Groups (COG) database. Karp (2001) has developed a symbolic framework for inferring missing genes in known metabolic pathways, relying primarily on functional dependencies rather than phylogenetic information.

In this work, we present Phylogenetic Profile Anomaly Detection (PPAD), a computational framework for annotation refinement. We first generate a probabilistic model of expected phylogenetic profiles from the growing databases of existing genome annotations. The dependency model is encoded as a Bayesian network (Pearl 1988),

which has been shown to be a principled and powerful tool for integrating heterogeneous information and for modeling uncertainty and noise in biological systems (Delcher 1993; Friedman 2000; Hartemink 2001; Stuart 2003; Troyanskaya 2003). The model essentially encodes a set of probabilistic rules that predict the presence or absence of a gene, given the presence or absence of a small number of other genes. We then use this dependency model to identify statistical anomalies in new annotations that signal the presence of unexpected components in their phylogenetic profiles or, potentially, systematic errors.

Our PPAD method extends and generalizes the methods mentioned above in the following ways:

- The inputs and outputs are probabilistic, allowing explicit modeling of uncertainties.
- The probabilistic dependency model is better suited for modeling noisy biological correlations than logical rules or *ad hoc* pattern matching.
- All or part of the dependency model can be efficiently learned directly from existing annotations and other relevant data sets.

Unlike the true phylogenetic methods of Goodman *et al.* (1979) and Page *et al.* (1994), the version of PPAD presented here does not consider the phylogenetic relationships between the species used to learn the dependency model. This has the potential advantage that lateral gene transfers, which appear to be common in the bacterial kingdom (Mirkin 2003; Peregrin-Alvarez 2003; Hao 2004), are not considered anomalies as long as they fit an expected co-occurrence pattern.

We show through simulations that our method can reliably improve on the accuracy of genome annotations to which artificial errors have been added. We also identify 22 previously missed genes by applying the PPAD method to bacterial genome annotations from GenBank.

Methods

Problem domain and definitions

We define the genome annotation task as the problem of assigning descriptive labels to every predicted gene in a genome. We denote the set of available labels $G = \{G_1, G_2, \dots, G_N, NA\}$. The NA label is used to designate “no assignment” for genes for which no other label in G is applicable. In the context of homology based annotations, the labels are the names of known genes or gene families that have been identified or characterized previously.

In this work, we use the database of Clusters of Orthologous Groups (COG) as our reference source of descriptive labels (Tatusov 2003), although the method applies equally well to any other protein classification scheme. A COG is defined as a set of predicted orthologous genes identified in three or more species, where at least one species is a prokaryote. The database contains COGs covering 25 major functional categories, as well as curated annotations of 50 bacterial genomes. We focus on the annotation of genes with information processing and storage functions, such as transcription, translation and replication. There are 731 such COGs, and thus in our notation $N = 731$ and G_i

corresponds to the i th COG in functional categories A, B, J, K and L, as defined in the COG database (<http://www.ncbi.nlm.nih.gov/COG/new/>).

For a genome we wish to annotate, we define a binary vector of length N , where the i th entry is 1 if at least one gene in the genome is best described by label G_i , and 0 otherwise. This vector, denoted \bar{g} , is equivalent to a phylogenetic profile, as first introduced by Gaasterland (1998) and Pellegrini (1999). A genome annotation, which assigns labels to each gene in the genome, implicitly defines a phylogenetic profile indicating which of the available labels in G have been used in that annotation. This phylogenetic profile can be evaluated against a probabilistic model of dependencies between genes to identify potential errors and inconsistencies.

A Bayesian network model of gene dependencies

We utilize a Bayesian network (Pearl 1988; reviews in Charniak 1991 and Cowell 1998) to capture and represent dependencies between gene families as we observe them in a set of trusted phylogenetic profiles, D . The nodes in the Bayesian network are a set of binary random variables $\hat{G} = \{\hat{G}_1, \hat{G}_2, \dots, \hat{G}_N\}$, corresponding directly to the entries of the phylogenetic profiles introduced above (Figure 1a). Edges between nodes describe a directed acyclic graph and define a set of conditional independence relations between the binary variables. The probability distribution associated with each random variable depends only on the parent nodes in the graph (Figure 1b). The complete Bayesian network B_D therefore represents a probability distribution over all possible phylogenetic profiles, given D :

$$B_D \equiv P(\hat{G}) = \prod_{i=1, \dots, N} P(\hat{G}_i | Parents(\hat{G}_i))$$

The structure and parameters of a Bayesian network that captures the dependencies in D can be estimated using a variety of learning algorithms. Here, we use a greedy hill-climbing approach, as implemented by the *LibB* software package (Friedman and Elidan, <http://www.cs.huji.ac.il/labs/compbio/LibB/>). Starting from an unconnected network, the learning algorithm computes which single structural modification (edge addition, removal or inversion) will optimize the Bayesian Information Criterion (BIC) score. The BIC score of a network is the log-likelihood of the training data, given the network, minus a complexity penalty to prevent over-fitting. The penalty depends on d , the number of parameters in the model:

$$\text{BIC}(B) = \log P(D | B) - \frac{d}{2} \log N$$

The learning algorithm iteratively applies the best modification to the Bayesian network until the BIC score converges. Although this greedy approach is not guaranteed to find a globally optimal model, it tends to perform well in practice (Heckerman 1995). The end result of this learning procedure is a Bayesian network that assigns low probabilities to phylogenetic profiles that are inconsistent with the gene dependencies observed in the training set. Thus, if the resulting phylogenetic profile for a newly annotated genome is deemed highly unlikely by our trained Bayesian network, then either the new genome is very different from anything we have seen before, or the genome annotation is flawed.

Note that we explicitly assume that each phylogenetic profile in the training set is equally informative, and that the training set is a random sampling from some underlying probability distribution of correct annotations. This is a naïve assumption that simplifies

the learning procedure, but which may result in uninformative predictions if the species included in the training set are not sufficiently diverse. More complex variants of PPAD can be developed to take full advantage of known phylogenetic relationships between the species included in the training set.

Detecting potential errors in a phylogenetic profile

To detect potential errors in a genome annotation, we consider the phylogenetic profile \bar{g}_I implied by the annotation. We then make explicit our confidence in the entries of this initial phylogenetic profile in terms of probabilities. In general, each assignment in \bar{g}_I can have its own degree of confidence, but in this work we use a simple two-parameter model where α is the probability that a label that is called present in the initial annotation is actually absent (the assumed false positive rate), and β is the probability that a label that is called absent is actually present (the assumed false negative rate):

$$\begin{aligned}\bar{g}_{I,i} = 1 &\rightarrow P(\hat{G}_i = 0) = \alpha \\ \bar{g}_{I,i} = 0 &\rightarrow P(\hat{G}_i = 1) = \beta\end{aligned}$$

By incorporating our confidence in the initial phylogenetic profile as virtual evidence (Pearl 1988), we can compute the most probable (MPE) phylogenetic profile \bar{g}_{MPE} given this information and B_D as the Bayesian prior probability distribution over possible phylogenetic profiles:

$$\bar{g}_{MPE} = \arg \max_{\bar{g}} P(\bar{g} | B_D, \bar{g}_I, \alpha, \beta)$$

We find an exact solution to this optimization problem using the junction tree inference algorithm (Dawid 1992), as implemented in the Bayesian Network Toolbox for Matlab (Murphy 2001) and then compare each entry in \bar{g}_{MPE} to the corresponding entry in \bar{g}_I .

Discordances between the most probable and the initial phylogenetic profiles represent features in the initial profile that are statistically anomalous according to our Bayesian network model of gene dependencies, and may therefore correspond to annotation errors. For example, the presence of a label in the MPE profile, but not in the initial profile, signals that the annotation pipeline may have missed that label. However, it is important to note that it is not possible to distinguish between annotation errors and unexpected evolutionary events *a priori*. The statistically anomalous presence or absence of that particular label may also be due to a recent gene loss event in the species or strain we are annotating, or it may simply be a result of the species being significantly different from the set of species from which our Bayesian network was learned. Any discordance is therefore a hypothesis that can be tested by re-evaluating the relevant evidence, and, if necessary, follow-up experiments.

Results

Improvement over simulated annotations

We evaluated the ability of the PPAD method to improve upon inaccurate initial annotations by performing a series of simulated "leave-one-out" tests. We induced a Bayesian network from the curated annotations in the COG database, leaving one species out of the training set. If the database contained annotations of more than one strain of the

same species, we removed all such strains to avoid over-fitting. We then simulated inaccurate annotations by randomly adding and removing entries from the phylogenetic profiles of the species that were left out. Finally, treating the original COG annotations as "correct", we calculated the pair-wise number of differences (the Hamming distance) between them and the simulated profiles and the computed MPE profiles. We repeated this process 20 times for each species and for each set of parameters.

The PPAD method improved upon the simulated annotations in 582 of 600 trials (97%) where either the false positive or the false negative rate of the initial profile was greater than 0.01. For each of the species listed in Table 1, the first row of column A shows the mean increase in the number of correct COG calls in the MPE profiles relative to the simulated initial profiles. With false positive and false negative rates of 0.1 in the initial profiles, we saw roughly 40 corrected COG calls for all five species, which is 5.5% of the 731 COGs we considered. All positive differences are statistically significant ($p < 10^{-3}$, Mann-Whitney U). We obtained similar results for the species not shown here.

We note that the PPAD method is dependent on accurate error rate estimates. The columns marked B in Table 1 show the results of setting α and β to values that differed significantly from the actual error rates. In these cases we saw little or no improvement in the MPE profiles. In practice, α and β can be estimated from the significance values returned from the sequence alignment tools used to compute the initial annotation. The assumed error rates can also be increased or decreased to obtain more or less conservative MPE profiles.

Improvement over nearest-relative profiles

We also compared the MPE profiles from our leave-one-out tests to the most closely related species present in the corresponding training sets. Given the amount of genomic variation across the bacterial kingdom, it is not immediately evident that comparing an initial annotation to the MPE profile is better than simply comparing it to the closest sequenced relative, which is common practice. If MPE profiles tend to make predictions consistent with “averaged genomes”, they might not fit any real species particularly well. We found that this was not the case.

The MPE profiles were more similar to the unmodified “correct” COG annotations than to the most closely related species in the training set in every one of 600 simulations where either the false positive or the false negative rate of the initial profile was less than 0.1. For all five species shown in Table 1, the difference between the MPR profile and the corresponding nearest-relative profile increased with decreasing error rates (column A). This was true even with poor error rate estimates (column B).

Identification of 22 new genes

We searched for annotation errors in the COG database by repeating our leave-one-out tests, but without adding simulated noise to the initial annotations, and without removing highly similar profiles from the training sets. Because the COGs are themselves defined as part of the annotation process (Tatusov 1997), we expected very few or no false positive assignments. However, the input peptide sets might not have been complete. To

search for evidence of missed genes we therefore assumed a very low false positive rate ($\alpha = 0.001$) and a moderately high false negative rate ($\beta = 0.1$).

The PPAD method predicted the presence of a total of 244 additional genes in the 50 curated annotations, a predicted increase of 2% for the number of genes in these species. We examined these predictions by using TBLASTN to align known members of the missing COGs to the complete genomic sequence of the genomes from which they were predicted to be missing. If a significant alignment was found, we determined whether it corresponded to a complete open reading frame, and compared its location to the GenBank annotation of the genome to see if it overlapped a known feature.

In the end, we were able to confirm 22 annotation errors across 10 bacterial genomes using TBLASTN sequence alignments alone. This set, as shown in Table 2, was dominated by ribosomal proteins. In particular, we found 9 missing ribosomal genes in *A. tumefaciens*, suggesting systematic under-detection of ribosomal genes in the original annotation of this genome (Goodner 2001). A similar pattern was found by Natale *et al.* (2000) in their analysis of an earlier version of the COG database, suggesting that under-detection of ribosomal genes may be a systematic error in several annotation systems, perhaps due to their short length.

We also found that a number of the remaining predicted genes were present in the examined genomes as pseudogenes and consequently left out of the COG database, which only catalogues functional genes. For *M. leprae*, a species known to have

undergone extensive gene loss (Cole 2001), we found that 6 out of 8 genes predicted to be present were pseudogenes annotated elsewhere. Because detectable pseudogenes are likely to have been deactivated in recent evolutionary history, they may be as statistically anomalous as annotation errors, but may also be actual genes that have been obscured by sequencing errors or strain polymorphisms. Additional laboratory work will be required to distinguish these alternatives. Finally, we note that since our Bayesian network only modeled a subset of the COGs, we expect that there may be additional missing genes in other functional categories.

The set of 244 PPAD predicted genes were significantly enriched for confirmable annotation errors. By repeating the TBLASTN analysis for 5 randomly picked sets of 244 genes absent in the initial COG annotations, we could only confirm a total of 7 missing genes ($p < 10^{-12}$ by Fisher's exact test), of which 5 were also found using PPAD. The 2 genes not found using PPAD were COG2094 and COG2167 in *C. glutamicum*.

Discussion

We have described a methodology for improving whole-genome annotations using a probabilistic model of gene co-occurrence patterns. Unlike earlier *ad hoc* analyses of unexpected phylogenetic profiles, the PPAD method is formulated within a principled probabilistic framework. PPAD can be incorporated into existing annotation pipelines to spot potential systematic errors and assist manual annotators. The fact that the annotations we examined and improved upon were published, best effort annotations shows the value of such independent quality controls.

We note that there are many potential avenues for follow-up and improvement on the results presented here. For example, it would be desirable develop a better quantitative understanding of both the potential and the limitations of phylogenetic information in this context. Recent evolutionary innovations and species-specific adaptations cannot be distinguished from true errors by a method that relies on historic correlations. However, it is currently unclear whether the numbers of false predictions encountered in this and related methods (Natale 2000; Paley 2001; Zheng 2002) can be lowered by using better training sets and explicit models of phylogenetic relationships between the species in the training set, or whether they are absolute limits set by the phylogenetic and functional plasticity of gene families (Mirkin 2003; Peregrin-Alvarez 2003).

A pragmatic strategy for improvement might be to assess the statistical significance of each component of the probabilistic model and avoid making predictions based on uncertain parameters. Several such strategies have been developed in other contexts, such as gene expression analysis (Pe'er 2001). More sophisticated learning algorithms, such as simulated annealing, might also be helpful in improving the predictive power of the models.

Finally, it appears worthwhile to develop models with more problem specific semantics or constraints. One promising possibility is a hierarchical or noisy-or model (Pearl 1988) where the observable variables, which indicate the presence and absence of genes, only depend on a smaller set of hidden variables that indicate the presence or absence of

particular functional pathways, cellular complexes or other higher-level biological features. It would also be possible to use other annotation labels such as enzymatic activities, instead of protein families, analogous to the pathway inference systems developed by Karp *et al.* (2001).

Acknowledgements

We thank Nir Friedman and Gal Elidan for sharing their *LibB* software; Simon Kasif, Nick Patterson and Douglas Lauffenburger for helpful discussions and support; and two anonymous referees for helpful comments.

References

- Bocs, S., Danchin, A., Médigue, C. Re-annotation of genome microbial CoDing-Sequeces: finding new genes and inaccurately annotated genes, *BMC Bioinformatics*, **3**:5
- Brenner, S.E. (1999) Errors in genome annotation, *Trends genet*, **15**, 132-133.
- Charniak, E. (1991) Bayesian Networks without Tears, *AI Magazine*, **12**, 50-63.
- Cole, S.T., Eiglmeier, K., Parkhill, J., James, K.D., Thomson, N.R., Wheeler, P.R., Honore, N., Garnier, T., Churcher, C., Harris, D. *et al.* (2001) Massive gene decay in the leprosy bacillus, *Nature*, **409**, 1007-1011.
- Cowell, R. (1998) Introduction to Inference for Bayesian Networks. In Jordan, M.I. (eds), *Learning in Graphical Models*. Kluwer Academic Publishers, Dordrecht, pp. 9-26.

- Dawid, A.P. (1992) Applications of a general propagation algorithm for probabilistic expert systems, *Statistics and Computing*, **2**, 25-26.
- Delcher, A.S., Kasif, S., Goldberg, H. & Xsu, W. (1993) Protein Secondary-Structure Modeling with Probabilistic Networks, *International Conference on Intelligent Systems and Molecular Biology*, 109-117.
- Devos, D. & Valencia, A. (2001) Intrinsic errors in genome annotation, *Trends genet*, **17**, 429-431.
- Friedman, N., Linial, M., Nachman, I. & Pe'er, D. (2000) Using Bayesian networks to analyze expression data, *J. Comput. Biol.* **7**, 601-620.
- Gaasterland, T. & Ragan, M.A. (1998) Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes, *Microb. Comp. Genomics*, **3**, 199-217.
- Goodman, M., Czelusniak, J., Moore, G.W., Matsuda, G. (1979) Fitting the gene lineage into its species lineage: a parsimony strategy illustrated by cladograms constructed from globin sequences, *Syst. Zool*, **28**, 132-163.
- Goodner, B., Hinkle, G., Gattung, S., Miller, N., Blanchard, M., Quorollo, B., Goldman, B.S., Cao, Y., Askenazi, M., Halling, C., *et al.* (2001) Genome Sequence of the Plant Pathogen and Biotechnology Agent *Agrobacterium tumefaciens* C58, *Science*, **294**, 2323-2328.
- Hao, W., Golding, G. B. (2004) Patterns of Bacterial Gene Movement, *Mol Biol Evol*, **21**, 1294-1307.

- Hartemink, A.J., Gifford, D.K., Jaakkola, T.S. & Young, R.A. (2001) Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks, *Pacific Symposium on Biocomputing*, **6**, 422-433.
- Heckerman, D., Geiger, D. & Chickering, D. (1995) Learning Bayesian networks: The combination of knowledge and statistical data, *Machine Learning*, **20**, 197-243.
- Iliopoulos, I., Tsoka, S., Andrade, M.A., Enright, A.J., Carroll, M., Pouillet, P., Promponas, V., Liakopoulos, T., Palaios, G., Pasquier, C. *et al.* (2003) Evaluation of annotation strategies using an entire genome sequence, *Bioinformatics*, **19**, 717-726.
- Karp, P.D. (2001) Pathway Databases: A Case Study in Computational Symbolic Theories, *Science*, **293**, 2040-2044.
- Mirkin, B.G., Fenner, T.I., Galperin, M.Y. & Koonin, E.V. (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes, *BMC Evol Biol*, **3**.
- Murphy, K.P. (2001) The Bayes Net Toolbox for Matlab, *Computing Science and Statistics*, **33**.
- Natale, D.A., Galperin, M.Y., Tatusov, R.L. & Koonin, E.V. (2000) Using the COG database to improve gene recognition in complete genomes, *Genetica*, **108**, 9-17.
- Page, R.D.M. (1994) Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas, *Syst Biol*, **43**, 58-77.
- Paley, S.M. & Karp, P.D. (2001) Evaluation of computational metabolic-pathway predictions for *Helicobacter pylori*, *Bioinformatics*, **18**, 715-724.

- Pearl, J. (1988) *Probabilistic inference in intelligent systems*. Morgan Kaufmann, San Mateo, CA.
- Pe'er, D., Regev, A., Elidan, G. & Friedman, N. (2001) Inferring Subnetworks from Perturbed Expression Profiles, *ISMB*, 215-224.
- Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. & Yeates, T.O. (1999) Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles, *Proc Natl Acad Sci USA*, **96**, 4285-4288.
- Peregrin-Alvarez, J.P., Tsoka, S. & Ouzounis, C.A. (2003) The Phylogenetic Extent of Metabolic Enzymes and Pathways, *Genome Res*, **13**, 422-427.
- Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules, *Science*, **302**, 249-255.
- Tatusov, R.L., Koonin, E.V., Lipman, D.J. (1997) A genomic perspective on protein families, *Science*, **278**, 631-637.
- Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J. & Natale, D.A. (2003) The COG database: an updated version includes eukaryotes, *BMC Bioinformatics*, **11**, 41.
- Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman, R.B. & Botstein, D. (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*), *Proc Natl Acad Sci USA*, **100**, 8348-8353.

Zheng, Y., Roberts, R.J. & Kasif, S. (2002) Genomic functional annotation using co-evolution profiles of gene clusters, *Genome Biol*, **3**.

Table 1 Mean increase in the number of correct COG annotations in the MPE profiles

	(A) Accurate α and β							(B) Inaccurate α and β¹		
	False positive rate	0.1	0.05	0.1	0.05	0.05	0.01	0.01	0.05 / 0.01	0.01 / 0.05
False negative rate	0.1	0.1	0.05	0.05	0.01	0.05	0.01	0.01 / 0.05	0.05 / 0.01	
<i>C. trachomatis</i>										
vs. simulated	43	37	25	19	14	3	-1	-2	-1	
vs. nearest relative	15	17	21	24	27	28	34	10	28	
<i>H. pylori</i>										
vs. simulated	48	43	27	19	17	2	0	0	2	
vs. nearest relative	-3	1	3	5	10	11	14	-8	10	
<i>M. tuberculosis</i>										
vs. simulated	36	30	22	14	12	-2	-3	-8	0	
vs. nearest relative	18	22	23	30	37	31	40	17	33	
<i>N. meningitides</i>										
vs. simulated	38	31	23	15	10	4	-1	-2	-4	
vs. nearest relative	15	20	26	29	35	39	42	23	30	
<i>S. pyogenes</i>										
vs. simulated	33	30	17	16	11	2	-4	-3	-3	
vs. nearest relative	-3	5	4	13	18	19	23	6	16	

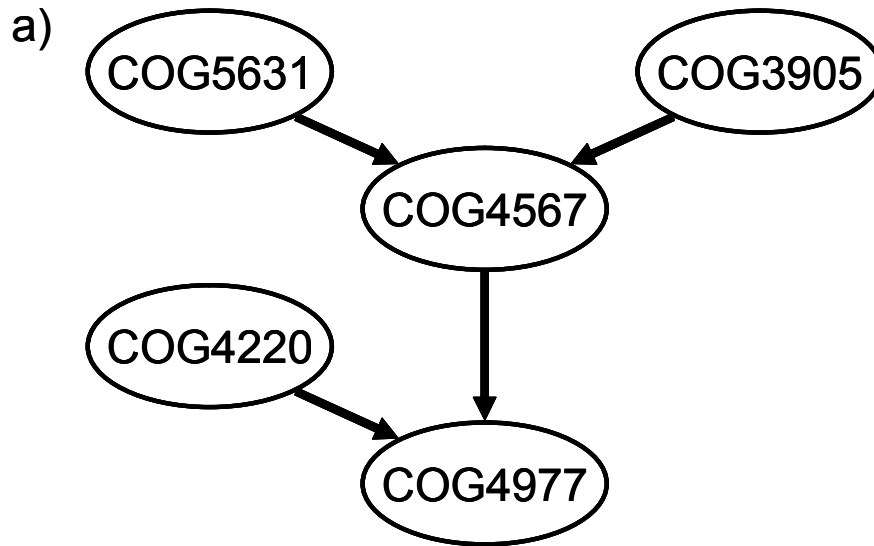
¹ (Assumed/actual)

Table 2 Genes predicted by PPAD and confirmed by manual inspection

Organism	COG	Function	RefSeq	Start	End
<i>A. tumefaciens</i>	COG0051	Ribosomal protein S10	NC_003062	1915207	1915515
	COG0186	Ribosomal protein S17	“	1910264	1910500
	COG0230	Ribosomal protein L34	“	378081	378218
	COG0238	Ribosomal protein S18	“	1081552	1081800
	COG0267	Ribosomal protein L33	“	1288873	1289040
	COG0360	Ribosomal protein S6	“	1081814	1082275
	COG0594	RNase P protein component	“	377661	378044
	COG1841	Ribosomal protein L30	“	1906183	1906389
	COG0227	Ribosomal protein L28	NC_003063	1402503	1402793
	COG0257	Ribosomal protein L36	“	1236507	1236632
<i>B. melitensis</i>	COG3905	Transcriptional regulator	NC_003317	1568957	1569217
<i>C. acetobutylicum</i>	COG0024	Methionine aminopeptidase	NC_003030	3258093	3258845
<i>F. nucleatum</i>	COG0230	Ribosomal protein L34	NC_003454	642502	642636
	COG0257	Ribosomal protein L36	“	1944372	1944485
	COG1758	RNA polymerase subunit K/ω	“	537211	537435
<i>L. innocua</i>	COG0219	rRNA methylase	NC_003213	964790	965299
<i>L. lactis</i>	COG0758	DNA uptake protein	NC_002662	1253570	1254037
<i>M. leprae</i>	COG0227	Ribosomal protein L28	NC_002677	2018906	2019100
<i>M. pulmonis</i>	COG0551	Topoisomerase Zn-domain	NC_002771	567947	569824
<i>R. prowazekii</i>	COG0317	Guanosine hydrolase/synthetase	NC_000963	382806	383495
	“	“	“	796030	796782
<i>S. pyogenes</i>	COG0590	Cytosine/adenosine deaminases	NC_002737	187765	188280

Figure Legends

Figure 1 We model evolutionary dependencies between COGs as a Bayesian network. The Bayesian network consists of a graph representation of the qualitative dependencies (a) and a set of associated probability distributions. For example, the absence or presence of COG4567 depends on COG5631 and COG3905. The conditional probability distribution associated with COG4567 (b) assigns a high probability to the presence of COG4567 in any genome that contains either COG5631 or COG3905.



b) $P(\text{COG4567} \mid \text{COG5631}, \text{COG3905})$:

COG4567	COG5631	COG3905	P
0	0	0	0.9713
0	0	1	0.1000
0	1	0	0.1923
0	1	1	0.2778
1	0	0	0.0287
1	0	1	0.9000
1	1	0	0.8077
1	1	1	0.7222