Predicting Protein Function by Genomic Context: Quantitative Evaluation and Qualitative Inferences

Martijn Huynen, Berend Snel, Warren Lathe III, et al.

Genome Res. 2000 10: 1204-1210
Access the most recent version at doi:10.1101/gr.10.8.1204

Predicting Protein Function by Genomic Context: Quantitative Evaluation and Qualitative Inferences

Martijn Huynen,[1,2,3] Berend Snel,[1] Warren Lathe III,[1,2] and Peer Bork[1,2]

[1]*European Molecular Biology Laboratory, 69117 Heidelberg, Germany;* [2]*Max-Delbrück-Centrum for Molecular Medicine, 13122 Berlin-Buch, Germany*

Various new methods have been proposed to predict functional interactions between proteins based on the genomic context of their genes. The types of genomic context that they use are Type I: the fusion of genes; Type II: the conservation of gene-order or co-occurrence of genes in potential operons; and Type III: the co-occurrence of genes across genomes (phylogenetic profiles). Here we compare these types for their coverage, their correlations with various types of functional interaction, and their overlap with homology-based function assignment. We apply the methods to *Mycoplasma genitalium*, the standard benchmarking genome in computational and experimental genomics. Quantitatively, conservation of gene order is the technique with the highest coverage, applying to 37% of the genes. By combining gene order conservation with gene fusion (6%), the co-occurrence of genes in operons in absence of gene order conservation (8%), and the co-occurrence of genes across genomes (11%), significant context information can be obtained for 50% of the genes (the categories overlap). Qualitatively, we observe that the functional interactions between genes are stronger as the requirements for physical neighborhood on the genome are more stringent, while the fraction of potential false positives decreases. Moreover, only in cases in which gene order is conserved in a substantial fraction of the genomes, in this case six out of twenty-five, does a single type of functional interaction (physical interaction) clearly dominate (>80%). In other cases, complementary function information from homology searches, which is available for most of the genes with significant genomic context, is essential to predict the type of interaction. Using a combination of genomic context and homology searches, new functional features can be predicted for 10% of *M. genitalium* genes.

The sequencing of complete genomes has created the opportunity not only to analyze gene function in the genomic context in which it occurs, but also to exploit the information from genomic context to predict functional interactions between genes (Dandekar et al. 1998; Enright et al. 1999; Huynen and Bork, 1998; Marcotte et al. 1999a; Overbeek et al. 1999; Pellegrini et al. 1999). Context-based function prediction is complementary to homology-based function prediction (Huynen and Snel 2000). Whereas the latter in principle predicts the molecular function of a protein, the former predicts a higher order function. (e.g., in which process or pathway a particular protein plays a role, or with which other protein it interacts) However, although the correlations between various types of genomic context and functional interactions have been addressed several times (Dandekar et al. 1998; Enright et al. 1999; Marcotte et al. 1999b; Pellegrini et al. 1999), a quantification of which types of functional interactions are associated with which types of context has thus far been absent. Here we analyze these associations in the genes of *M. genitalium* and their orthologs, which have served as a benchmark for structure prediction (Teichmann et al. 1999), function prediction (Brenner 1999), and as the minimal set of genes for cellular life (Hutchison et al. 1999). In addition, we analyze the overlap of genomic context with homology-based function prediction, using a combination of homology and context to predict new functional features for *M. genitalium* proteins. Because *M. genitalium* contains a high percentage of essential genes (Hutchison et al. 1999) and shares a relatively high fraction of its genes with other genomes (Snel et al. 1999), these results are relevant outside this species.

[3]Corresponding author.
E-MAIL huynen@embl-heidelberg.de; FAX 49-6221-387517.

RESULTS

Quantitative Patterns in Genomic Context

We examined the presence of fused genes, the conservation of the local neighborhood of genes, and the co-occurrence of genes in genomes (see Methods) in a systematic comparison of *M. genitalium* with the 24 other genomes that were published up to February 1, 2000 (see http://www.TIGR.ORG/tdb/mdb/mdb.html). A graphical display of the coverage and overlap of the various categories of genomic context, and their overlap with homology-based function prediction, is given in Figure 1. There is a strong overlap between the set of proteins having homologs with a known function and the proteins for which significant context information

10:1204–1210 ©2000 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/00 $5.00; www.genome.org

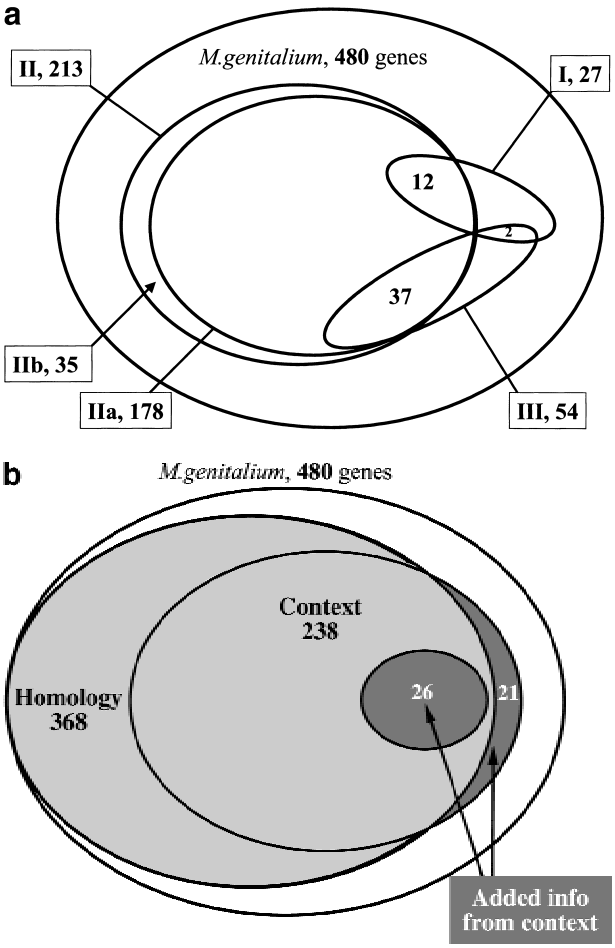

Figure 1 (a) Coverage of and overlap between various types of genomic context for *M. genitalium* genes. Type I is gene-fusion. Type II is the conservation local gene neighborhood, which is separated in type IIa (the conservation of gene order) and type IIb (the co-occurrence of genes within potential operons in absence of the conservation of gene order). Type III is the co-occurrence of genes in genomes. (b) Overlap between genes for which significant genomic context is available and genes for which functional features can be predicted by homology searches. For the latter, only genes that are homologous to genes with known molecular functions were included, which were determined by manual inspection. The dark gray areas in the figure are genes for which new functional features can be predicted by genomic context. They can be homologous to proteins with a known molecular function, in which case the context can indicate in which process this function plays a role (see text for specific examples). A complete list of genes for which new functional features could be predicted by genomic context and, if available, homology to proteins with known function, is available from http://dove.embl-heidelberg.de/MG/Context.

is available. This is, among other reasons, due to a set of 70 proteins that do not have homologs outside the Mycoplasmas, and for which only gene fusion context information is available. A complete list of *M. genitalium* proteins, for which significant genomic context was found, is available from http://dove.embl-heidelberg.de/MG/Context.

Gene Fusion

Gene fusion (Type I) is the most direct form of genomic context. The proteins encoded by genes of which homologs are fused tend to have a related function (Marcotte et al. 1999a), especially if they are orthologs of the fused genes (Enright et al. 1999; Snel et al. 2000). For a set of 27 *M. genitalium* genes, gene fusion can be observed in another published genome. Three pairs of proteins contain at least one member with a hypothetical function. In all the other cases a functional interaction between the proteins was apparent. They physically interact directly (15 proteins), physically interact indirectly by being part of the same complex (two proteins), or catalyze subsequent steps in a metabolic pathway (four proteins) (Fig. 2). There are no potential false positives (proteins with known function but unknown functional interaction) in this set.

Conservation of the Local Context of Genes

Conservation of the local genomic context can be detected via the conservation of genes as neighbors (Dandekar et al. 1998) and via the conservation of genes in runs (Overbeek et al. 1999): sets of genes with the same direction of transcription that are separated by intergenic regions of fewer than 300 bases. We separate local context into two distinct types. Type IIa refers to genes that are conserved as neighbors in phylogenetically distant genomes. Type IIb refers to genes that cooccur together in operons but are not conserved as neighbors with any genes. Type IIa can be observed for 178 genes, 12 genes of which are also involved in gene fusion events (Type I). In 142 cases the order is maintained in *M. genitalium* itself, whereas for 36 genes, although they are present in *M. genitalium*, their conserved organization as neighbors can only be observed in other species. The functional interactions between the proteins encoded by pairs of genes were divided into seven classes (Fig. 2). About 63% of the proteins encoded by gene-pairs in class IIa either directly (30%) or indirectly (33%) physically interact. This is less than the 75% reported by Dandekar et al. (1998). Note, however, that in Dandekar et al. (1998) the criteria for conservation of gene order were rather stringent: Genes were required to be neighbors in all of the three genomes compared. In the present analysis, genes are required to be neighbors in three of all the genomes compared. When the selection criteria for detection of conserved pairs are more stringent (e.g., six genomes instead of three), the fraction of the genes with known function that encode proteins that physically interact is larger than 80% (see http://dove.emblheidelberg.de/MG/Context). A similar correlation between the fraction of physically interacting proteins with the number of genomes in which their gene order is conserved was observed in a small set of genomes by Mushegian and Koonin (1996).

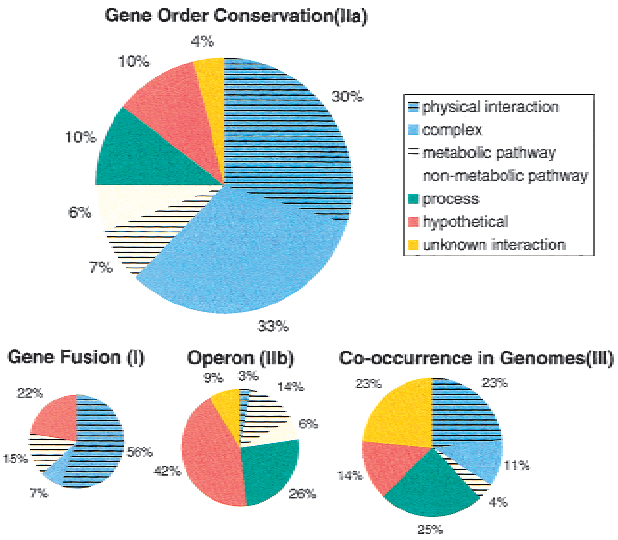

Figure 2 The types of functional interactions between *M. genitalium* proteins for the different types of genomic context. The surface areas of the circles are proportional to the number of genes for which the techniques apply. Classification was done by manual inspection, allowing detection of all possible described functional interactions between proteins. Subsequently, the functional interactions were divided along the following hierarchical classification:

1: direct physical interaction between the proteins
2: indirect physical interaction (i.e., the proteins are part of the same protein complex, but there is no evidence that they interact directly with each other)
3: the proteins are part of a single metabolic pathway
4: the proteins are part of a non-metabolic pathway, either regulatory or otherwise
5: the proteins take part in the same process
6: pairs of proteins of which at least one is hypothetical
7: proteins with known functions between which no functional interactions are known

Class 5 was only considered if the functional interactions between the proteins did not fall in classes 1–4. Types of functional interactions can be counted in two ways: per gene and per interaction. In general, the number of interactions is smaller than the number of genes, as two interacting genes only represent a single interaction. However, a single gene can have multiple genomic associations: In those cases they were normalized per gene. The results in the figure are based on a per gene count. The frequencies of the different classes of functional interactions did not alter significantly upon counting each interaction.

For an additional 35 genes, conservation of local neighborhood could only be detected under the criteria of Type IIb: For example, these genes co-occur repeatedly with specific genes in potential operons, but they are not conserved as neighbors with any genes. The types of functional interactions in this set are less direct: Physical interaction can only be observed for 3%, while the categories pathway or process included 20% and 26%, respectively. For 51% of the genes, their functional interaction was not known, because they are hypothetical proteins (42%) or because they have known functions but a functional interaction is not apparent (9%). The latter is a maximum estimate of the fraction of false positives.

Co–occurrence of Genes in Genomes

We find significant co-occurrence of genes in genomes (phylogenetic profiles) for 45 gene pairs, containing a total of 54 genes (see Methods). This set has a substantial overlap with the above categories: 37 out of the 54 genes fall into type IIa. This is not surprising, as genes that are shared between genomes tend to be clustered on them. The functional interactions between these genes are less dominated by physical interaction than in types I and IIa, and was observed at 34% (Huynen and Bork 1998). The fraction of proteins with known functions but unknown functional interactions, which is a maximum estimate of the false positives, is relatively high (23%). It is, however, lower than the previous estimate of 29.5% (Marcotte et al. 1999b), which was not restricted to orthologous relations, and in which phylogenetic patterns in shared gene content were not filtered out.

Qualitative Inferences

Function and functional interaction are concepts that can be described at many levels (Bork et al. 1998). Therefore, the functional predictions based on genomic context span a range of possibilities, depending on what other information can be obtained from homology searches or experimental data, and on the type of genomic context in which a gene occurs. In the following sections, we predict new functional features of *M. genitalium* proteins based on the genomic context of their genes and, if available, other sources of information.

Gene Fusion

One example of an *M. genitalium* gene pair that is fused in at least one other genome and contains at least one hypothetical protein is MG259-MG347. They are fused in the *Rickettsia prowazekii* gene RP847. The N-terminal domain of the protein encoded by MG259 is a SAM-dependent methyl transferases (Koonin et al. 1995). The conservation of the local context of MG259 supports a role in translation: MG259 is located immediately 3′ of *prfA*, coding for peptide chain release factor 1, in six taxa. MG347 is also homologous to methyl transferases (E-value 2e-6, using PSI-BLAST with one iteration). Therefore, the fusion protein in *R. prowazekii* actually consists of two methyl-transferases domains that are predicted to play a role in translation.

Conservation of Gene Order

Detection of Homology and Orthology

Detection of homology can be hampered by sequence divergence, especially in the case of short sequences or biases in amino acid composition. In such cases the

conservation of gene order can help the detection of homology: Because the number of genes that are candidates for homology (one per genome) is much smaller than the complete gene database, one can effectively raise the allowable E-values in PSI-BLAST from the standard 0.001 or 0.01 by several orders of magnitude. One example is MG233, which codes for a 100 amino acid protein and is located between the genes for ribosomal proteins l27 and l21. Sequences with barely significant levels of sequence similarity (E-values > 0.1 in PSI-BLAST) to MG233 could be detected at identical locations (between l27 and l21) in *Bacillus subtilis*, *Treponema pallidum*, *Borrelia burgdorferi,* and *Thermotoga maritima*. The location of the genes in this family suggests that they code for proteins that interact with the ribosome.

Physical Interaction

For two hypothetical *M. genitalium* genes, we predict physical interactions with other proteins with known function. The first is MG230 (*nrdI*). The association of this gene with genes in the nucleotide reductase operon (MG231/*nrdE* and MG229/*nrdF*) has been observed before, and it has been shown to have a stimulatory effect on ribonucleotide reduction (Jordan et al. 1997). The conservation of *nrdI* with *nrdE* (alpha subunit) is specifically strong. The gene order is conserved in all the published *nrdI* genes, including, for example, bacteriophage SPBc2. We therefore postulate a physical interaction between *nrdI* and *nrdE*.

Physical interaction is also predicted between the protein encoded by MG134 and either of the DNA polymerase III subunits gamma and tau (encoded by one gene), with which orthologs of MG134 are conserved as neighbors in six taxa. MG134 appears to be indispensable for *M. genitalium* (Hutchison et al. 1999), and has orthologs in virtually all sequenced bacterial genomes.

Conservation of Genes in Runs

Substrate Specificity

Substrate specificity is a volatile aspect of predicting enzymatic function, not only because the evolutionary signal can be obscured by sequence divergence, but also because proteins can change substrate specificity over relatively short evolutionary distances (Wu et al. 1999). The (conserved) operon context of a gene can suggest different substrate specificity than homology searches. MG053 is homologous to phosphomannomutases and phosphoglucomutases. It is encoded in a potential operon consisting of five genes, of which four genes encode enzymes for a nucleoside salvage pathway (Fig. 3). A fifth enzyme of this pathway, a phosphoribomutase, is missing. This suggests that the protein encoded by MG053 acts as a phosphoribomutase.

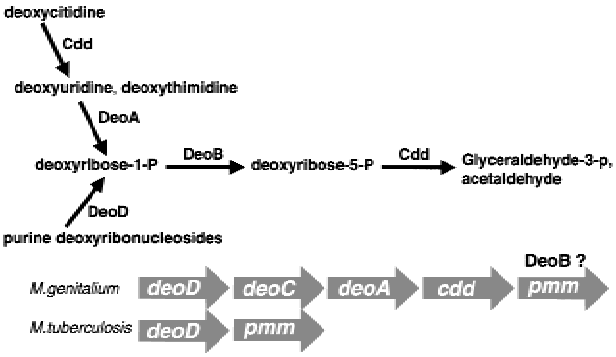

Figure 3 Genomic context predicts substrate specificity of proteins involved in a nucleoside salvage pathway in *M. genitalium*. A cluster of five genes in *M. genitalium* encodes four genes of a nucleoside salvage pathway. The "standard" gene for this fifth reaction in the pathway, phosphoribomutase (*deoB*), is absent. The fifth gene in the operon is homologous to phosphomannomutases and phosphoglucomutases. *M. genitalium* does not contain any other candidate for a phosphoribomutase. The most likely candidate for the phosphoribomutase is thus MG053. The significance of the location of a homolog of MG053 in a run with *deoD* is supported by the location of a homolog of the *M. genitalium* gene MG053 beside *deoD* in *Mycobacterium tuberculosis*.

Involvement in Pathways and Processes

Generally, genomic context provides information about the process in which a gene is involved. MG009 is a hypothetical protein that has orthologs in all sequenced genomes except *Archaeoglobus fulgidus*. It occurs in potential operons with the gene for thymidilate kinase in four taxa. Homology searches using PSI-BLAST (Altschul et al. 1997) reveal this gene to be part of a large family of TIM-barrel proteins that are involved in deaminase, dehydratase, and phosphohydrolase reactions, with a substrate that is generally a nucleotide or a precursor of a nucleotide (Holm and Sander 1997). The co-occurrence in potential operons of MG009 and its orthologs with thymidilate kinase suggests that MG009 is involved in the generation of a precursor of deoxyribonucleotides, specifically of deoxythymidine 5′ triphosphate (dTTP). A potential function for the protein encoded by this gene, which is consistent with the homology information and with the context information, is that of dCTP-pyrophosphatase (EC 3.6.1.12). dCTP-pyrophosphatase catalyzes the dephosphorylation of dCTP to dCMP and of dCDP to dCMP. It has been measured in a close relative of *M. genitalium*, *Mycoplasma mycoides*, where it is involved in the biosynthesis of dTTP (Neale et al. 1983). A pathway for dTTP biosynthesis involving dCTP-pyrophosphatase has also been proposed in *E. coli* (Krogan et al. 1998), based on the observation that the mutation of one of the genes in this pathway is a necessary condition for creating thymidine auxotrophy. No gene for dCTP-pyrophosphatase has so far been identified in prokaryotes.

Genomic context can also refine the function de-

scription from a phenotypic one to one that specifies the process in which a protein is involved. An ortholog of MG008 in *E.coli* has been shown to be essential for the oxidation of thiophene and furan (Alam and Clark 1991). The conserved location of this protein in a potential operon, with ribosomal protein L34 in four taxa, suggests that it is involved in protein synthesis. This is supported by experiments with *mss1*, an ortholog of MG008, that codes for a nuclear encoded, mitochondrial GTPase in *Saccharomyces cerevisiae*. Mss1p appears to interact with SSU mitochondrial RNA and to be involved in translation (Decoster et al. 1993). Furthermore, Mss1p has been shown to interact with Mto1p (Colby et al. 1998) which is also involved in mitochondrial protein synthesis. Orthologs of *mss1* and *mto1* are neighbors in the genomes of *B. subtilis* and *B. burgdorferi*. The effect of the deletion of MG008 on the phenotype might be caused by the inability of the cell to synthesize certain proteins in the absence of MG008.

Finally, by combining context information and sensitive homology searches, we predict that MG246 and MG130 are part of a ribonucleic acid processing pathway. Orthologs of MG130 and MG246 are neighbors in the genomes of *B. burgdorferi* and *B. subtilis* (Fig. 4). MG130 has been shown to contain a KH RNA-binding domain (Musco et al. 1996) and a HD phosphohydrolase domain, and it has been proposed to play a role in nucleotide metabolism (Aravind and Koonin 1998). Reciprocal PSI-BLAST searches with a 5′ nucleotidase (5′NT) from *E. coli* (*ushA*) from the list of borderline hits of MG246, showed MG246 to be homologous to the catalytic domain of 5′NT (E-value 5e-11, 5 iterations). This makes MG246 a candidate for 5′ nucleotidase activity, which has been shown to be present in *M. genitalium* (Hamet et al. 1979), but for which no gene had been assigned. Note, however, that MG246 and its orthologs lack one of the conserved residues (D/E) in the motif GNH(D/E), which might be required for catalytic activity. The latter might be compensated by an aspartic acid that is located at position 163 in MG246, and that is conserved among all its orthologs. Juxtaposing the sequence of MG246 over the 3D structure of the *E. coli* 5′NT shows this to be located close to the catalytic site GNH motif.

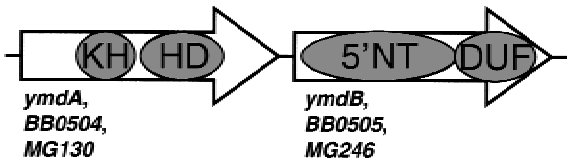

Figure 4 Domain organization of two proteins that are encoded by neighboring genes on *B. subtilis* (*ymdA* and *ymdB*) and *B. burgdorferi* (BB0504 and BB0505), and that are both present in *M. genitalium* (MG130 and MG246). The three domains that have functionally been characterized, KH, HD, and 5′NT, can all be related to ribonucleotide metabolism. KH binds (single-stranded) RNA; HD hydrolyzes phosphates from nucleotides; and 5′NT hydrolyzes NMP to nucleosides. A fourth, uncharacterized sequence domain (DUF) is present at C-terminus of MG246 and its orthologs.

A functional interaction between MG246 and MG130 thus appears likely, based on the locations of their orthologs and based on their molecular function. In addition, orthologs of MG130 occur with orthologs of MG245 as neighbors in *Aquifex aeolicus* and *Helicobacter pylori*. MG245 is homologous to 5-formyltetrahydrofolate cyclo-ligase, which is involved in the synthesis of tetrahydrofolate. The latter serves as acceptor (donor) of one-carbon units in catabolic (anabolic) reactions, among others in nucleotide metabolism, and might serve as a cofactor in the predicted pathway.

DISCUSSION

By exploiting the genomic association of genes, comparative genome analysis has provided new tools for the prediction of protein function. We have shown here that there is a correlation between the spatial proximity of genes on the genome and the directness of the interaction between the proteins they encode. In prokaryotes, physical interaction between proteins is more frequent when their genes occur fused or as conserved neighbors than when they tend to occur merely in the same operon or genome. Furthermore, the fraction of potential false positives decreases with requirements on spatial proximity on the genome. As there is a partial overlap of the different types of context, this argues for a hierarchy in the usage of genomic context to predict functional interactions.

Although the correlation of various types of genomic context with functional interactions is a fascinating aspect of computational genomics and is spurring the development of databases that combine both genomic data and interaction data like WIT (Selkov et al. 1998) and KEGG (Kanehisa and Goto 2000), the value of this information will finally be decided by the biological understanding and useful predictions they deliver. To make such predictions more specific than by merely saying that protein A is likely involved in the same process as protein B, complementary information from homology searches, when available, is invaluable.

METHODS

Orthology

Orthology is operationally defined as "bi-directional best, significant (E< 0.01), hit" based on Smith and Waterman (1981) comparisons of the complete genomes with one another, and including the possibility of gene fusion/fission (Huynen and Bork 1998). Note that the conservation of genes as neighbors, or in runs, increases the probability that they are true orthologs.

Gene Fusion / Fission

Occurrences of gene fusion and gene fission are derived from the orthology data for genes for which the orthology relationships are not "one to one" (Snel et al. 2000). A single fusion of orthologs of *M. genitalium* genes in one of the other genomes was considered a significant indication of a functional interaction between the genes.

Conservation of Gene Neighborhood

Conservation of gene order was only regarded significant for species with 87% or less SSU rRNA identity, at which gene order of non-functionally related genes is randomized (Huynen and Snel 2000). This excludes genomes from a single species or genus. In counting the number of taxa in which a given gene neighborhood was present, pairs of closely related genomes only counted for one taxon. Given the large number of genome comparisons, one also has to assess the probability that two genes occur in the same run in two genomes only by chance: All genomes were randomized, while keeping their run architecture intact, that is, the genes in each genome were randomly distributed over the loci in that genome. The co-occurrence of two genes in a single run in these randomized genomes occurred, on average, less than once per comparison of the *M. genitalium* genome with all others. In general, we use the criterion that, to infer a functional interaction between genes, they must occur as neighbors (type IIa), or if that can not be established, they must occur in a single run (type IIb), in at least three phylogenetically distant genomes. When homology-based predictions of the function of genes supported a functional interaction between them, the conservation of genes as neighbors, or in a single run in two genomes, was considered significant.

Co-occurrence of Genes in Genomes

We quantify the co-occurrence of genes in genomes as the mutual information between genes: Specifically, what extra information we get about the probability that gene i is present in a genome, from the knowledge that another gene j is also present. The mutual information $[M(i,j)]$ between i and j is the entropy of the distributions of i $[H(i)]$ and j $[H(j)]$ minus the combined entropy of both distributions $[H(i,j)]$ (Kullback 1959). For an instructive example of the usage of mutual information in sequence analysis, see Korber et al. (1993). The mutual information is mathematically equivalent to the log-odds ratio of the expected co-occurrence of pairs of genes, based on their individual frequencies, to the observed occurrence.

$$H(i) = - \Sigma_i P_i \log P_I$$

$$H(j) = - \Sigma_j P_j \log P_j$$

$$H(i,j) = - \Sigma_{i,j} P_{i,j} \log P_{i,j}$$

$$M(i,j) = H(i) + H(j) - H(i,j) = \Sigma_{i,j} P_{i,j} / P_i P_j$$

The mutual information provides a score for the co-occurrence of two genes. It is maximal when (1) both genes occur in about 50% of the genomes (the individual entropies of the genes are maximal), and (2) the genes occur always together (the combined entropy is minimal). In principle, the combined entropy is also minimal when the genes never occur together. However, that situation does not occur when studying the genes from one genome. To eliminate correlations between genes that result from phylogenetic correlations in the gene content of the genomes, the largest sets of genes with the same phylogenetic distribution are discarded. Not unexpectedly, these large clusters reflect the phylogenetic patterns in the gene distribution (Huynen et al. 1999; Snel et al. 1999). They contain genes that have orthologs in all species (39 genes), in only the Bacteria (19 genes), in only the (low G + C) gram positives (11 genes), or only the Mycoplasmas (66 genes). By discarding these large clusters, gene pairs with an atypical pattern of co-occurrence are selected: The more atypical a pattern, the more likely that it reflects a functional constraint on the proteins rather than the phylogenetic relatedness of the genomes. Selecting small clusters has the additional advantage of increasing the probability that a cluster represents only a single functional cluster of genes, rather than multiple clusters that happen to have the same phylogenetic distribution. Decreasing the maximum cluster size allowed (10 genes) did not reduce the maximum estimate of false positives (pairs of proteins with a known function but without a known functional interaction). Gene pairs with an M threshold score of 0.5 or higher, corresponding to genes with a "perfect" co-occurrence pattern, that occur in minimal 5 and in maximal 20 genomes, were selected. Note that this M score does not require a perfect pattern of co-occurrence: For example, if one gene occurs in 12 genomes and another in 13 genomes, and the overlap of their occurrence is maximal, the M score is 0.55. This allows one to overcome (small) imperfections in orthology prediction. Lowering the M score threshold to one that reflects a perfect co-occurrence in at least four or maximal 21 genomes led to an increase in the maximum estimate of false positives. Increasing the threshold did not lower this estimate (data not shown).

ACKNOWLEDGMENTS

This work was supported by BMBF. M.H. thanks Shamil Sunyaev for useful discussions and Gerrit Lehmann for technical assistance. We thank the referees for their comments.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Alam, K.Y. and Clark, D.P. 1991. Molecular cloning and sequence of the thdf gene, which is involved in thiophene and furan oxidation by *Escherichia coli*. *J. Bacteriol.* **173:** 6018–6024.

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.

Aravind, L. and Koonin, E. 1998. The HD domain defines a new superfamily of metal-dependent phosphohydrolases. *Trends Biochem. Sci.* **23:** 469–472.

Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M., and Yuan, Y. 1998. Predicting function: From genes to genomes and back. *J. Mol. Biol.* **283:** 707–725.

Brenner, S.E. 1999. Errors in genome annotation. *Trends Genet.* **15:** 132–133.

Colby, G., Wu, M., and Tzagoloff, A. 1998. Mto1 codes for a mitochondrial protein required for respiration in paromomycin-resistant mutants of *Saccharomyces cerevisiae*. *J. Biol. Chem.* **273:** 27945–27952.

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.

Decoster, E., Vassal, A., and Faye, G. 1993. MSS1, a nuclear-encoded mitochondrial GTPase involved in the expression of COX1

subunit of cytochrome *c* oxidase. *J. Mol. Biol.* **232:** 79–88.
Enright, A., Ilipoulos, I., Kyrpides, N., and Ouzounis, C. 1999. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402:** 86–90.
Hamet, M., Bonissol, C., and Cartier, P. 1979. Activities of enzymes of purine and pyrimidine metabolism in nine Mycoplasma species. *Adv. Exp. Med. Biol.* **122B:** 231–235.
Holm, L. and Sander, C. 1997. An evolutionary treasure: Unification of a broad set of amidohydrolases related to urease. *Proteins* **28:** 72–82.
Hutchison, C., Peterson, S., Gill, S., Cline, R.T., White, O., Fraser, C. M., Smith, H., and Venter, J.C. 1999. Global transposon mutagenesis and a minimal mycoplasma genome. *Science* **286:** 2165–2169.
Huynen, M.A. and Bork, P. 1998. Measuring genome evolution. *Proc. Natl. Acad. Sci. USA* **95:** 5849–5856.
Huynen, M.A. and Snel, B. 2000. Gene and context: Integrative approaches to genome analysis. In *Analysis of Amino Acid Sequences* (ed. P. Bork), pp. 345–379. Adv. Prot. Chem. Academic Press, San Diego, CA.
Huynen, M.A., Snel, B., and Bork, P. 1999. Lateral gene transfer, genome surveys and the phylogeny of prokaryotes. *Science* **286:** 1441a.
Jordan, A., Aslund, F., Pontis, E., Reichard, P., and Holmgren, A. 1997. Characterization of *Escherichia coli* NrdH. *J. Biol. Chem.* **272:** 18044–18050.
Kanehisa, M. and Goto, S. 2000. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **27:** 29–34.
Koonin, E.V., Tatusov, R.L., and Rudd, K.E. 1995. Sequence similarity analysis of Escherichia coli proteins: functional and evolutionary implications. *Proc. Natl. Acad. Sci. USA* **92:** 11921–11925.
Korber, B.T.M., Farber, R.M., Wolpert, D.H., and Lapedes, A.S. 1993. Covariation of mutations in the v3 loop of human immunodeficiency virus type 1 envelope protein: An information theoretic analysis. *Proc. Natl. Acad. Sci. USA* **90:** 7176–7180.
Krogan, N.J., Zaharik, M.L., Neuhard, J., and Kelln, R.A. 1998. A combination of three mutations, *dcd*, *pyrH* and *cdd*, establishes thymidine (deoxyuridine) auxotrophy in *thyA* $^{+}$ strains of Salmonella typhimurium. *J. Bacteriol.* **180:** 5891–5895.
Kullback, S. 1959. *Information theory and statistics*. Wiley, New York.
Marcotte, E.M., Pellegrini, M., Ng, H., Rice, W.D., Yeates, T.O., and Eisenberg, D. 1999a. Detecting protein function and protein-protein interactions from genome sequences. *Science* **285:** 751–753.
Marcotte, E., Pellegrini, M., Thompson, M., Yeates, T., and Eisenberg, D. 1999b. A combined algorithm for genome-wide prediction of protein function. *Nature* **402:** 83–86.
Musco, G., Stier, G., Joseph, C., Castiglione-Morelli, M., Nilges, M., Gibson, T., and Pastore, A. 1996. Three-dimensional structure and stability of the KH domain: Molecular insights into the fragile x syndrome. *Cell* **85:** 237–245.
Mushegian, A.R. and Koonin, E.V. 1996. Gene order is not conserved in bacterial evolution. *Trends Genet.* **12:** 289–290.
Neale, G.A.M., Mitchell, A., and Finch, L.R. 1983. Enzymes of pyrimidine deoxyribonucleotides metabolism in *Mycoplasma mycoides* subsp. *mycoides*. *J. Bacteriol.* BO156: 1001–1005.
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.
Pellegrini, M., Marcotte, E. M., J., Thompson, M., Eisenberg, D., and Yeats, T. O. 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* **96:** 4285–4288.
Selkov, E., Grechkin, Y., Mikhailova, N., and Selkov, E. 1998. Mpw: The metabolic pathways database. *Nucleic Acids Res.* **26:** 43–45.
Smith, T. and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **147:** 195–197.
Snel, B., Bork, P., and Huynen, M. 1999. Genome phylogeny based on gene content. *Nat. Genet.* **21:** 108–110.
Snel, B., Bork, P., and Huynen, M. 2000. Genome evolution: Gene fusion versus gene fission. *Trends Genet.* **16:** 9–11.
Teichmann, S., Chothia, C., and Gerstein, M. 1999. Advances in structural genomics. *Curr. Opin. Struct. Biol.* **9:** 390–399.
Wu, G., Fisher, A., ter Kuile, B., Sali, A., and Muller, M. 1999. Convergent evolution of *Trichomonas vaginalis* lactate dehydrogenase from malate dehydrogenase. *Proc. Natl. Acad. Sci. USA* **96:** 6285–6290.

Received March 27, 2000; accepted in revised form June 2, 2000.