not exhibit large deviations from neutral predictions (Figure 1). However, in species with higher rates of gene conversion or larger effective population sizes, BGC could significantly perturb allele-frequency distributions (and thus statistics such as Tajima's *D*) from neutral expectations. This suggests that BGC should be incorporated into realistic models of neutral evolution.

Acknowledgements
We thank the four reviewers for discussions and comments on the manuscript. This work was supported by the Swedish Research Council.

References

1 Nekrutenko, A. and Li, W.H. (2000) Assessment of compositional heterogeneity within and between eukaryotic genomes. *Genome Res.* 10, 1986–1995
2 Eyre-Walker, A. and Hurst, L.D. (2001) The evolution of isochores. *Nat. Rev. Genet.* 2, 549–555
3 Green, P. *et al.* (2003) Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Genet.* 33, 514–517
4 Filipski, J. (1987) Correlation between molecular clock ticking, codon usage fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. *FEBS Lett.* 217, 184–186
5 Wolfe, K.H. *et al.* (1989) Mutation rates differ among regions of the mammalian genome. *Nature* 337, 283–285
6 Sueoka, N. (1988) Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.* 85, 2653–2657
7 Bernardi, G. (1986) Compositional constraints and genome evolution. *J. Mol. Evol.* 24, 1–11
8 Eyre-Walker, A. (1993) Recombination and mammalian genome evolution. *Proc. R. Soc. Lond. B. Biol. Sci.* 252, 237–243
9 Birdsell, J.A. (2002) Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Mol. Biol. Evol.* 19, 1181–1197
10 Marais, G. (2003) Biased gene conversion: implications for genome and sex evolution. *Trends Genet.* 19, 330–338
11 Galtier, N. *et al.* (2001) GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* 159, 907–911
12 Duret, L. *et al.* (2002) Vanishing GC-rich isochores in mammalian genomes. *Genetics* 162, 1837–1847
13 Smith, N.G.C. *et al.* (2002) Deterministic mutation rate variation in the human genome. *Genome Res.* 12, 1350–1356
14 Lercher, M.J. *et al.* (2002) The evolution of isochores: evidence from SNP frequency distributions. *Genetics* 162, 1805–1810
15 Sawyer, S.A. *et al.* (1987) Confidence interval for the number of selectively neutral amino acid polymorphisms. *Proc. Natl. Acad. Sci. U. S. A.* 84, 6225–6228
16 Kong, A. *et al.* (2002) A high-resolution recombination map of the human genome. *Nat. Genet.* 31, 241–247
17 Webster, M.T. *et al.* (2003) Compositional evolution of noncoding DNA in the human and chimpanzee genomes. *Mol. Biol. Evol.* 20, 278–286
18 Przeworski, M. *et al.* (2000) Adjusting the focus on human variation. *Trends Genet.* 16, 296–302
19 Jorde, L.B. *et al.* (2000) The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. *Am. J. Hum. Genet.* 66, 979–988
20 Tishkoff, S.A. and Williams, S.M. (2002) Genetic analysis of African populations: human evolution and complex disease. *Nat. Rev. Genet.* 3, 611–621
21 Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595
22 Reich, D.E. *et al.* (2001) Linkage disequilibrium in the human genome. *Nature* 411, 199–204
23 Nagylaki, T. (1983) Evolution of a finite population under gene conversion. *Proc. Natl. Acad. Sci. U. S. A.* 80, 6278–6281
24 Sawyer, S.A. and Hartl, D.L. (1992) Population genetics of polymorphism and divergence. *Genetics* 132, 1161–1176
25 Brown, T.C. and Jiricny, J. (1988) Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. *Cell* 54, 705–711
26 Jiricny, J. (1998) Replication errors: cha(lle)nging the genome. *EMBO J.* 17, 6427–6436
27 Svejstrup, J.Q. (2002) Mechanisms of transcription-coupled DNA repair. *Nat. Rev. Mol. Cell Biol.* 3, 21–29
28 Arndt, P.F. *et al.* (2003) Distinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. *Mol. Biol. Evol.* 20, 1887–1896
29 Fullerton, S.M. *et al.* (2001) Local rates of recombination are positively correlated with GC content in the human genome. *Mol. Biol. Evol.* 18, 1139–1142
30 Petes, T.D. (2001) Meiotic recombination hot spots and cold spots. *Nat. Rev. Genet.* 2, 360–369
31 Fay, J.C. and Wu, C.I. (2000) Hitchhiking under positive Darwinian selection. *Genetics* 155, 1405–1413

0168-9525/$ - see front matter
doi:10.1016/j.tig.2004.01.005

Global analysis of bacterial transcription factors to predict cellular target processes

Tobias Doerks[1,2], Miguel A. Andrade[3], Warren Lathe 3rd[1], Christian von Mering[1,2] and Peer Bork[1,2]

[1]EMBL, 69117 Heidelberg, Meyerhofstr. 1, Germany
[2]Max-Delbrueck-Centrum, Berlin, Germany
[3]Ottawa Health Research Institute, Ottawa, ON K1H 8L6, Canada

Whole-genome sequences are now available for >100 bacterial species, giving unprecedented power to comparative genomics approaches. We have applied genome-context methods to predict target processes that are regulated by transcription factors (TFs). Of 128 orthologous groups of proteins annotated as TFs, to date, 36 are functionally uncharacterized; in our analysis we predict a probable cellular target process or biochemical pathway for half of these functionally uncharacterized TFs.

Corresponding author: Tobias Doerks (doerks@email.de).

Table 1. Uncharacterized transcription factors and the cellular process that they are predicted to regulate[a]

COG number	Representative gene name	Distribution (genes in species)[b]	Orthology resolution[c]	Molecular characterization[d]	Predicted regulated process[e]
COG1318	MJ1053	11/10	Good	Predicted helix-turn-helix (HTH) transcripion factor	DNA-modification processes and general metabolism (COG0467 and COG1216)
COG1327	YBAD	50/50	Good	Zinc ribbon and ATP-cone transcription factor	Riboflavin biosynthesis[f] (COG0117, COG1985, COG0054 and COG0307)
COG1329	YDEB	21/20	Good	CarD-family transcription factors	Terpenoid biosynthesis[g] (COG0245 and COG1211)
COG1386	YPUH	56/51	Good	HTH-transcription factor	Pseudouridylate synthesis and link to cell wall formation (COG1187 and COG1686)
COG1395	MJ1164	16/16	Good	HTH-transcription factor	tRNA processing (COG1867)
COG1426	YFGA	41/39	Good	HTH-transcription factor	Terpenoid biosynthesis related (COG0821); (phospholipid biosynthesis[h]) (COG0558 and COG0575)
COG1497	MJ0558	9/9	Good	HTH-transcription factor	Nucleotide synthesis (COG0856)
COG1725	YHCF	38/23	Medium	HTH-transcription factor	ABC-transporters (COG1131)
COG1813	MJ0586	28/24	Good	HTH-transcription factor	Basal transcritption regulation (COG1675)
COG1959	YFHP	113/47	Poor	UPF0074-family putative transcription factors	Amino acid metabolism and Fe-S-cluster redox systems (COG1104, COG0822, COG0316, COG0633 and COG0820)
COG2345	BH3429	19/13	Medium	HTH-transcription factor	Enzymes related to phenylacetic acid aerobic catabolism (COG2151)
COG2378	YFJR	62/32	Medium	HTH-transcription factor	Glyoxal-like pathway (COG3324)
COG2462	AF1987	7/6	Good	Transcription factor (COG-prediction)	Helicase and/or hydrolase activity (COG2254 and COG1203)
COG2522	AF0184	19/11	Medium	HTH-transcription factor	Thiamin biosynthesis (COG1992)
COG2740	YLXR	34/34	Good	Transcription factor (COG-prediction)	General processes (e.g. translation and/or transcription) (COG0858, COG0195, COG0532, COG1358, COG2176 and COG1185)
COG3226	YBJK	14/9	Medium	HTH-transcription factor	Transmembrane transporter proteins (COG2076)
COG3655	YOZG	16/12	Medium	HTH-transcription factor	Transmembrane transporter proteins (NOG08084 and NOG19957)
NOG09448	SP1115	10/6	Medium	HTH-transcription factor	PDZ-domain-containing proteins (COG3480)

[a]The results can be reproduced at http://www.bork.embl-heidelberg.de/STRING_V3(COG database and STRING as of August 2003).
[b]Gene distribution (number of genes / number of species in which genes are found).
[c]The absence of paralogy, defined by the fraction: number of species / number of genes. Values: >0.8, good; >0.5, medium; <0.5, poor.
[d]Molecular characterization is based on Smart [19,20] and/or Pfam [21] predictions.
[e]Pathway, process or protein families that are predicted to be regulated by proteins of the corresponding cluster of orthologous group (COG). Associated COGs that are used to make predictions are shown in parenthesis.
[f]Observed previously [22], but not yet annotated in any database.
[g]Low scoring of 0.366 (based on conserved gene neighborhood in a few species).
[h]These proteins have lost their DNA-binding domain and are assumed not to be transcription factors.

Homology-based function annotation is an established approach and has set the standard for assigning function to novel proteins during the past decades. Homology searches often assign molecular features such as catalytic or DNA-binding activity to novel proteins but often fail to provide information about cellular processes involving the proteins. To provide such information, several complementary methods have been developed recently; these predict functional associations among protein-coding genes, based on their genomic context [1–6]. These methods consider the conservation of gene neighborhood, gene fusion events or the significant co-occurrence of genes in different species, and predict associations that comprise physical interactions, related functional roles or similar pathway memberships [7]. In our study, we systematically applied these methods to regulators of gene expression; we analyzed their conservation in the genomic context (based mainly on conserved gene neighborhood) to predict the cellular process that they participate in (which they most probably control transcriptionally). Recent research in *Escherichia coli* has elucidated the design principles of the transcription-regulatory network [8,9] and suggested that transcriptional control is at the heart of organismal complexity. Probabilistic methods for identifying genes and their regulators from gene-expression data have been used, for example, in yeast [10], but bacterial TFs whose target processes are unknown are awaiting precise annotation.

Identification of bacterial transcription factors

In this article, we describe the use of the conserved genomic associations of genes that encode bacterial TFs to predict the cellular and biological process that they potentially regulate.

To define the starting set of TFs, we extracted clusters of orthologous groups (COGs) from the COGs database [11] (http://www.ncbi.nlm.nih.gov/COG/), which were annotated directly as TFs (or regulators).

In addition, we considered COGs of which at least 30% of the members were assigned to relevant Gene Ontology (GO) terms [12,13] (GO:0006355, biological process: 'regulation of transcription, DNA-dependent'; GO:0030528, molecular function: 'transcription regulator'; and subcategories of these).

Following this procedure we retrieved 278 COGs believed to be involved in transcription regulation; from these we removed RNA-polymerase subunits, transposases, restriction enzymes and other DNA-modification

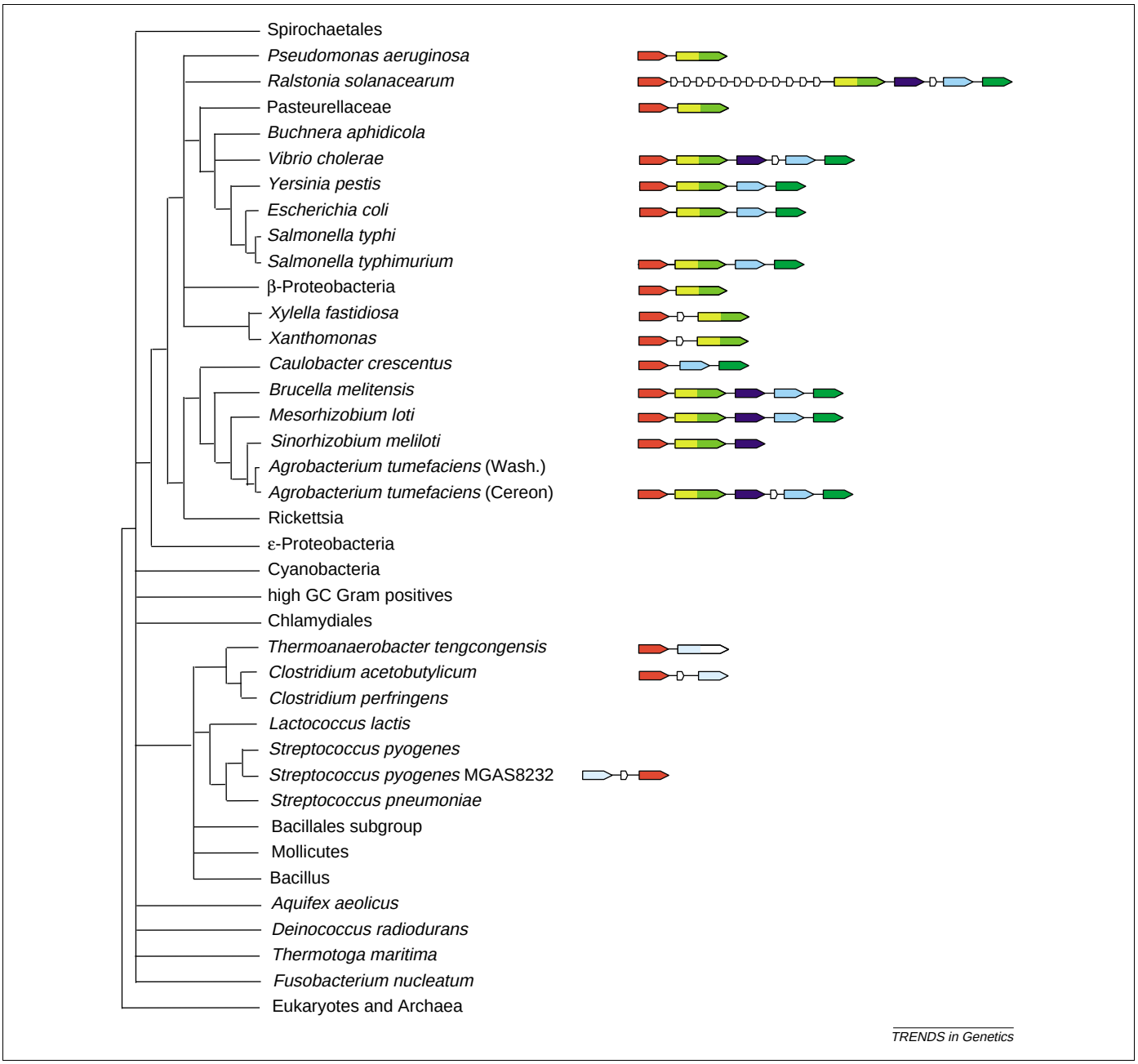

Figure 1. Conserved gene neighborhood in clusters of orthologous group (COG) 1327 [from Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) analysis]. The species tree with family representatives is shown on the left and corresponding operon architecture is shown on the right. Genes that are connected by black lines are in the immediate neighborhood on the genome (i.e. within 300 bp on the same strand). Genes that have multiple colors are members of several orthologous groups, which is indicative of putative fusion events. The genes depicted in white are neighbors, but do have not a sufficiently high score. The color scheme corresponds with the COGs that are depicted in Figure 2: red unit, query gene that encodes a hypothetical transcription factor (COG1327); yellow unit, gene that encodes pyrimidine deaminase (COG0117); green unit, gene that encodes pyrimidine reductase (COG1985); dark green unit, gene that encodes a transcription termination factor (COG0781); light blue unit, gene that encodes an uncharacterized response regulator (COG0745); sky blue unit, gene that encodes the riboflavin synthase β chain (COG0054); dark blue unit, gene that encodes the riboflavin synthase α chain (COG0307).

enzymes, in addition to proteins involved in DNA stability, repair and replication. All groups contained orthologous TFs derived from several genomes. Some groups also included duplicated genes (paralogs); these sequences could not be divided into separate COGs by the COG annotators. The final list consisted of 128 COGs of known and putative TFs. This list summarizes TFs from a wide variety of bacterial species; when projected to the K12-strain of *E. coli,* it covers 229 proteins (85%) of a recently assembled list of 268 *E. coli* TFs [14]. Of the remaining 39 *E. coli* proteins, eight represent non-conserved TFs, which are not assigned to orthologous groups, and the remaining 31 reflect differences in the retrieval strategies. Of the 128 COGs, we then subtracted 34 'inclusive COGs' (26.6%), which do not provide sufficient orthology resolution for genomic-context methods because of a prevalence of recent gene duplications. The remaining 94 COGs consist of 58 COGs that are already functionally described (45.3% of the dataset) and 36 COGs of putative TFs with an unknown functional role (28.1%).

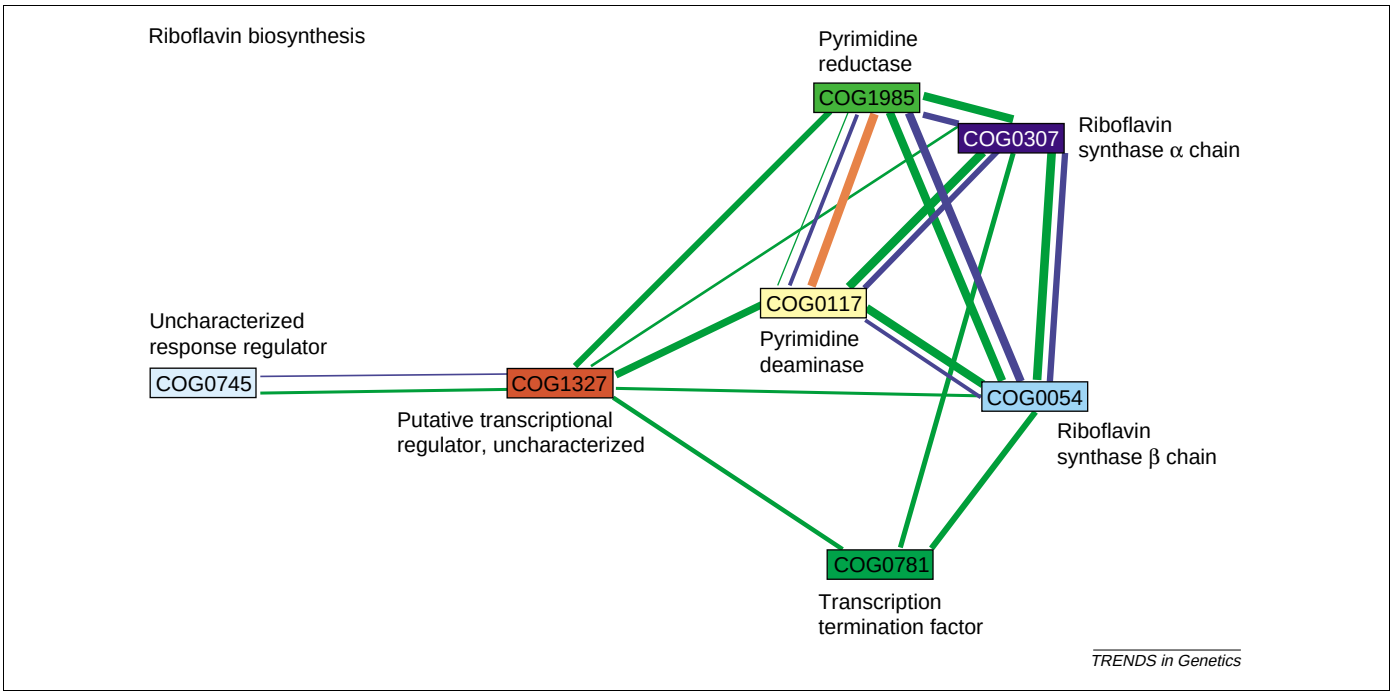

Figure 2. Network of predicted associations for a particular group of proteins [related to clusters of orthologous group (COG) 1327 (red)]. The network edges represent the predicted functional associations. An edge can be drawn with up to three differently colored lines; these lines represent the existence of the three types of evidence used in predicting the associations. A red line indicates fusion evidence; a green line indicates neighborhood evidence; a blue line indicates co-occurrence evidence. The line thickness correlates linearly with Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) scores.

We defined a COG as functionally characterized if it contained one or more genes for which the cellular context has been determined experimentally. Known and unknown COGs were analyzed using the tool STRING (Search Tool for the Retrieval of Interacting Genes/Proteins; http://www.bork.embl-heidelberg.de/STRING/) [6], applying a conservative score threshold of 0.5 (see Ref. [6] for a benchmark). STRING calculates this 'confidence score' on the basis of the three genomic-context methods: conserved gene neighborhood, gene fusion events and significant co-occurrence of the genes across a specific subset of species. We calculated the number of TFs for which genomic context is implemented in STRING, and we reviewed the predicted associations.

Prediction of cellular target processes

Of the 58 orthologous groups of TFs with an experimentally confirmed functional role, 34 retrieved significant hints to their cellular target processes. In the majority of these cases we could identify several genes that are known to be regulated by a specific TF. This shows that the method is suitable for confirming $>59\%$ of functionally described TFs; in 24% of the cases we could not predict any target process and only 10 (17%) of the predictions point to misleading, mostly general, processes.

Of the 36 COGs that contain uncharacterized TFs, we were able to suggest a cellular role or a target process for 18 groups (Table 1). The lists of associated COGs shown in Table 1 represent the functional context and describe this cellular process. The predictions are based on the genomic associations of one or more TF and apply to their orthologs. For species in which duplication events have led to multiple genes in the same group (paralogy), some copies of these genes might not be a part of the same target process. (Orthology resolution is shown in Table 1.)

The specificity of the predictions varies, ranging from exact operon matches, when, for example, the TFs in COG1327 are a part of a well-defined operon of a known biosynthesis pathway, to more general assignments, such as COG1318, which is involved in the control of DNA-modification processes and to broadly defined functions, such as the expression control of transmembrane transporters (COG3226 and COG3655) or PDZ-domain-containing proteins (NOG09448).

In some cases we observed that genes that were assigned to a single COG were apparently involved in the regulation of different processes. For example, some genes of the group COG2378 appear to be associated with an enzyme of glyoxal-like pathways in a small subset of species, whereas other genes of the same group appear to regulate the expression of a different set of target genes of unknown function (Table 1).

An example of a specific prediction is COG1327, which consists of uncharacterized TFs present in a variety of eubacterial species. As shown in Figure 1 the genes that encode these hypothetical TFs often occur (STRING scores are in the significant range between 0.509 and 0.847) next to the genes known to be part of the riboflavin biosynthesis pathway [15,16]. The network view (Figure 2) illustrates the functional association of COG1327 based on its conserved neighborhood with well-annotated genes, and the obvious connections between the other genes of this operon, which have been confirmed by conserved neighborhood, phylogenetic co-occurrence and a fusion event.

A second example represents a less-specific prediction; genes of the same orthologous group seem to be involved in different cellular processes and several associated genes

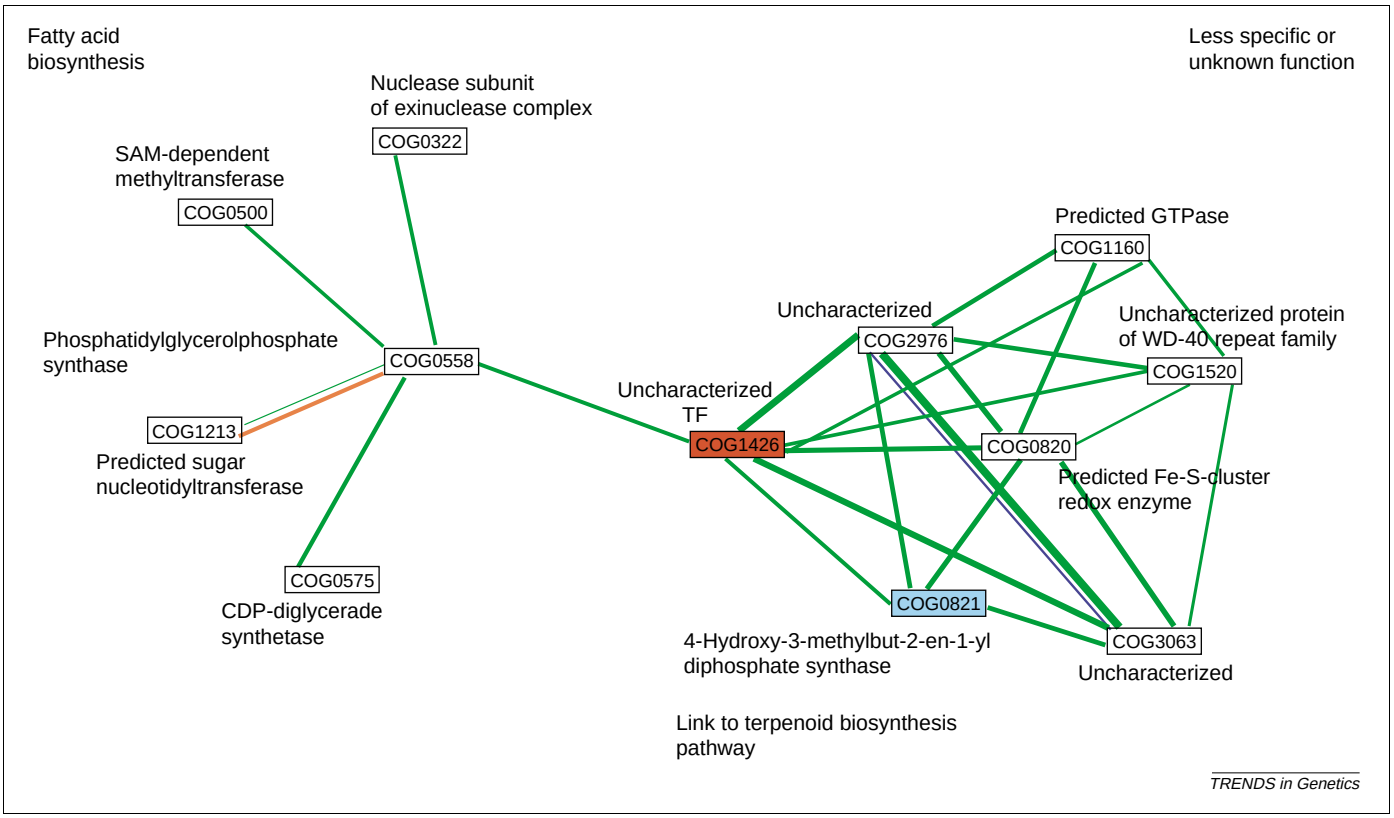

Figure 3. Network of predicted associations for a particular group of proteins [related to clusters of orthologous group (COG) 1426 (red) and COG0558 (white)]. The network edges represent the predicted functional associations. An edge can be drawn with up to three differently colored lines; these lines represent the existence of the three types of evidence used in predicting the associations. A red line indicates fusion evidence; a green line indicates neighborhood evidence; a blue line indicates co-occurrence evidence. The line thickness correlates linearly with Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) scores. COG1426 is functionally associated, in a species-specific manner, with a synthase of the terpenoid biosynthesis pathway [17] (COG0821) or proteins that are involved in fatty-acid biosynthesis [18] (COG0558).

are uncharacterized (Figure 3). In Gram-negative γ-proteobacteria only, putative TFs of COG1426 are functionally associated with a synthase of the terpenoid biosynthesis pathway [17]. The association is defined by a conserved operon, which also contains other proteins of unknown or less specific function, suggesting a cross-link to a novel undescribed pathway.

Furthermore, orthologs from COG1426 in Gram-positives and *Thermotoga maritima* are associated with genes that are known to be involved in fatty-acid biosynthesis [18]; in this example, the TFs have lost their helix-turn-helix domain and thus probably their DNA-binding capability and are instead predicted to be involved in membrane-associated processes.

Concluding remarks

Our large-scale analysis of bacterial TFs provides predictions of potentially related target processes for half of all hitherto uncharacterized TFs, which is an increase of functional knowledge of 14.1% for TFs classified in COGs. By contrast, homology searches only provide molecular characterization [e.g. evidence for a helix-turn-helix motif (and thus DNA-binding function)]. We exemplify how homology approaches can be complemented by genomic-context searches to refine the functional characterization of the protein at the cellular process level. These results indicate that this type of analysis merits extension to other protein families.

References

1 Dandekar, T. *et al.* (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* 23, 324–328
2 Overbeek, R. *et al.* (1999) The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2896–2901
3 Pellegrini, M. *et al.* (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* 96, 4285–4288
4 Marcotte, E.M. *et al.* (1999) Detecting protein function and protein–protein interactions. *Science* 285, 751–753
5 Enright, A.J. *et al.* (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402, 86–90
6 von Mering, C. *et al.* (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* 31, 258–261
7 Huynen, M. *et al.* (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.* 10, 1204–1210
8 Thieffry, D. *et al.* (1998) From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. *Bioessays* 20, 433–440
9 Shen-Orr, S.S. *et al.* (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* 31, 64–68
10 Segal, E. *et al.* (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* 34, 166–176
11 Tatusov, R.L. *et al.* (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 29, 22–28
12 Boeckmann, B. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31, 365–370
13 Camon, E. *et al.* (2003) The gene ontology annotation (GOA) project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res.* 13, 662–672
14 Babu, M.M. and Teichmann, S.A. (2003) Evolution of transcription

factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Res.* 31, 1234–1244
15 Richter, G. *et al.* (1997) Biosynthesis of riboflavin: characterization of the bifunctional deaminase-reductase of *Escherichia coli* and *Bacillus subtilis*. *J. Bacteriol.* 179, 2022–2028
16 Moertl, S. *et al.* (1996) Biosynthesis of riboflavin. Lumazine synthase of *Escherichia coli*. *J. Biol. Chem.* 271, 33201–33207
17 Hecht, S. *et al.* (2001) Studies on the nonmevalonate pathway to terpenes: the role of the GcpE (IspG) protein. *Proc. Natl. Acad. Sci. U. S. A.* 98, 14837–14842
18 Gopalakrishnan, A.S. *et al.* (1986) Structure and expression of the gene locus encoding the phosphatidylglycerophosphate synthase of *Escherichia coli*. *J. Biol. Chem.* 261, 1329–1338
19 Schultz, J. *et al.* (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. U. S. A.* 95, 5857–5864
20 Letunic, I. *et al.* (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.* 30, 242–244
21 Bateman, A. *et al.* (2002) The Pfam protein families database. *Nucleic Acids Res.* 30, 276–280
22 Wolf, Y.I. *et al.* (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res.* 11, 356–372

0168-9525/$ - see front matter
doi:10.1016/j.tig.2004.01.006

Mutational patterns correlate with genome organization in SARS and other coronaviruses

Andrei Grigoriev

GPC Biotech, Fraunhoferstr. 20, Martinsried 82152, Germany

Focused efforts by several international laboratories have resulted in the sequencing of the genome of the causative agent of severe acute respiratory syndrome (SARS), novel coronavirus SARS-CoV, in record time. Using cumulative skew diagrams, I found that mutational patterns in the SARS-CoV genome were strikingly different from other coronaviruses in terms of mutation rates, although they were in general agreement with the model of the coronavirus lifecycle. These findings might be relevant for the development of sequence-based diagnostics and the design of agents to treat SARS.

Previously, cumulative skew diagrams have been employed successfully to analyze mutational patterns in various viral genomes. They have been used to: (i) link the nucleotide content changes to the genome organization, replication and transcription of double-stranded DNA viruses [1]; (ii) correlate the transcriptional pattern of a bacteriophage T7 with its nucleotide content [2]; and (iii) associate the compositional biases with mutational pressures in retroviruses [3]. (See Box 1 on how to interpret cumulative diagrams.)

The severe acute respiratory syndrome coronavirus (SARS-CoV) plus-strand genomic RNA (plus-gRNA) consists of two distinct parts: one (comprising two thirds of the genome) encodes the replicase polyprotein and the other encodes structural and other proteins [4,5]. In this paper, these parts are referred to as the long and short arm, respectively. Strikingly, there is a change in behavior of the cumulative skew diagram at the border of the arms in all coronaviruses sequenced to date (six representatives are shown in Figure 1), indicating a lower GC skew on the short arm. This behavior suggests that biological processes that distinguish the two arms (Box 2) are responsible for the mutational pattern, rather than the fidelity of the replication machinery; the latter not would result in a constant slope of cumulative skew, as is the case in retroviruses [3]. The mutation rates (as indicated by the extent of the cumulative skew on the *y*-axis) do not appear to depend on a host organism: skews are similar in murine, avian and human 229E coronaviruses (Figure 1c,e,f) but substantially lower in SARS-CoV (Figure 1a, Table 1).

The skew diagrams support the current model of coronavirus replication and transcription (Box 2), and GC skew is particularly illustrative in this regard because in both of these processes one RNA strand is single stranded. Deamination of cytosine to uracil is >100 times faster in single-stranded DNA compared with double-stranded DNA [6], and this ratio is probably similar in

Table 1. Mean excess of guanines versus cytosines in coronavirus genomes

Virus genome[a]	Extra guanines compared with cytosines per 100 bp of genomic sequence[b]		
	L[c]	S[c]	L-S[c]
SARS-CoV	1.8	−1.7	3.5
BCoV	7.8	3.5	4.3
MHV	7.1	3.5	3.6
PEDV	4.4	1.4	3.0
HCoV	6.0	1.8	4.2
IBV	5.9	4.2	1.7

[a]Abbreviations: BCoV, enteric bovine coronavirus; IBV, avian infectious bronchitis virus; HCoV, human coronavirus (229E); PEDV, porcine epidemic diarrhea virus; SARS-CoV, severe acute respiratory syndrome coronavirus.
[b]These averages represent the trends depicted in Figure 1 but without taking into account G + C content (which ranges from 37% to 42% in *Coronaviridae*). GC content does not affect the trends observed in Figure 1.
[c]The change in number of guanines compared with cytosines is probably due to cytosine deamination in the minus strand on the short arm and reflects additional mutational pressure on that arm. Notably, this change is comparable with SARS-CoV and other coronaviruses, whereas the guanine excess on the long arm is much smaller. Definitions: L, long arm; S, short arm; L-S, change on short arm.

Corresponding author: Andrei Grigoriev (andrei.grigoriev@gpc-biotech.com).