BRIEFINGS IN BIOINFORMATICS. VOL 9. NO 1. 34–45
Advance Access publication November 17, 2007
doi:10.1093/bib/bbm051

Current trends in the bioinformatic sequence analysis of metabolic pathways in prokaryotes

Matteo Brilli, Renato Fani and Pietro Liò

Submitted: 26th July 2007; Received (in revised form): 10th October 2007

Abstract

The study of metabolic pathways is becoming increasingly important to exploit an integrated, systems-level approach for optimizing a desired cellular property or phenotype. In this context, the integration of genomics data with genetic, metabolic and regulatory models is essential because the systematic design of artificial, biological systems requires the identification of robust building blocks like gene promoters, metabolic pathways or genetic circuits taken from natural organisms, and manipulated to develop *ad hoc* features. Computational tools allowing precise descriptions of natural pathways might thus allow improving the performance of artificial routes. In this review, we introduce the most recent bioinformatics tools enabling detailed characterizations of metabolic pathways in bacteria from different perspectives.

Keywords: *metabolic engineering; pathway; bioinformatics*

INTRODUCTION

Metabolic pathways have evolved to execute their function efficiently, while tolerating perturbations, such as changes in environmental parameters or in the physiological status of the cell. They are constrained by physical and chemical requirements (e.g. conservation of energy, redox state, pH and so on), but are remarkably robust and evolvable; mainly for this reason, the features of a metabolic route can be considered as the end-products of co-evolutionary processes between organisms and physicochemical agents during Earths history; this, in turn, might suggest that metabolic features reflect different selective forces, which are environment dependent (i.e. antibiotic resistance genes are widespread in pathogens, but not in free-living organisms). Moreover, this cross-relationship is emphasized by the presence of co-evolving competitors, all optimizing their own metabolic strategies. Environmental pressures might promote the maintenance or the loss of metabolic pathways, or they might simply modulate gene expression through a cascade of signalling molecules causing a pathway to be turned on or off: organisms respond to the quantity and quality of nutrients in the environment by adjusting their transcriptional and metabolic profiles to make optimum use of the available nutrients and by selecting alternative strategies for survival in harsh conditions (i.e. spore formation, quiescence and so on).

Obtaining descriptions of the global organization of metabolic pathways and their relationships is thus fundamental to clarify how genomes evolve and how life reached the contemporary complexity level; moreover, functional prokaryotic genomics is becoming increasingly important in the emerging field of metabolic engineering. Metabolic

Corresponding author. Matteo Brilli, Department of Animal Biology and Genetics, University of Florence, Via Romana 17/19, 50125 Firenze, Italy. Tel: +39052288244; Fax: +390552288250; E-mail: matteo.brilli@dbag.unifi.it

Matteo Brilli is a postdoc at the University of Florence; he works on computational molecular evolution and metabolic pathways characterization with a specific interest in the evolution of regulatory mechanisms.

Renato Fani is professor of Genetics at the University of Florence. In the last 20 years, he has studied the evolution of metabolic pathways both from the experimental and bioinformatics perspective.

Pietro Liò is Senior Lecturer at the Computer Laboratory of the University of Cambridge, where he undertakes research and teaching in Bioinformatics, Computational Biology and System Biology with particular interests in algorithms for multiscale phenomena, networks dynamics, sequence and microarray data analysis.

engineering in the XXI century will be heavily used for a systematic design of artificial, biological systems by using robust building blocks like gene promoters, metabolic pathways or genetic circuits taken from natural organisms, and manipulated to develop *ad hoc* features [1]. In summary, metabolic engineering exploits an integrated, systems-level approach for optimizing a desired cellular property or phenotype [2]. In this context, the biotechnology industry is interested in novel ways to improve the performances of microbial strains for production or bioremediation purposes. The integration of genomics data with genetic, metabolic and regulatory models will be essential to move forward because we cannot rely on our intuition to grasp the complexity of the biological systems involved [3]. Moreover, state-of-art genomics tools can be combined with metabolic profiling to identify key genes that could be engineered for the production of improved crop plants [4] or micro-organisms (e.g. for bioremediation, industrial purposes and drug production).

Artificially designed pathways or genetic circuits might have drastic effects on hosts physiology, even in the simplest prokaryote; this might decrease pathways efficiency, e.g. by parallel processes taking place inside the cell (i.e. co-expression with catabolic routes consuming intermediates, feedback loops and so on). It follows that a global understanding of the metabolic network of an organism might improve the performance of its engineered counterparts; moreover, a deep understanding of regulatory mechanisms might allow engineering pathways with predetermined expression patterns (i.e. expression is activated by a given compound or in a specific environmental or physiological condition).

Here, we illustrate and discuss available genomics tools for the study of metabolic pathways in a metabolic engineering context. In a concluding section, we discuss the need for integrating different sources of information in metabolic pathway analyses. High-throughput techniques produce results quantifying the global processes taking place in a living cell, from gene expression to protein–protein, protein–DNA and DNA–DNA interactions; however, these different levels are not independent and a deep understanding of biology depends on our ability of dissecting and predicting such relationships to realize both a static and a dynamic view of cell networks [5].

CHARACTERIZING METABOLIC PATHWAYS FOR METABOLIC ENGINEERING

The importance of ortholog identification

A correct identification of orthologous proteins is at the basis of functional assignment, one of the main goals of functional genomics. Moreover, because of the rapid accumulation of data from various high-throughput technologies, comprehensive gene or protein classification is one of the central issues in bioinformatics.

Although classification schemes based on functional roles, molecular interactions or reaction networks have recently attracted growing interest, those based on sequence or structural similarities are still the most used. Algorithms using motifs or profiles to characterize protein families are very popular, such as Pfam [6–10] and SMART [11, 12].

The identification of orthologs between two genomes often relies on the so-called bi-directional best-hit (BBH) criterion: two proteins, *a* and *b* from genomes *A* and *B*, respectively, are orthologs if *a* is the best-hit of *b* in genome *A* and vice versa. For three or more genomes, groups of orthologous sequences can be constructed by extending the BBH relationships with a clustering algorithm. The COG database, a widely used resource for ortholog grouping, was constructed basically using this approach [13], and a variety of methods has recently been developed for this task [14–21]. As COG, KEGG-related systems have growing popularity: recent developments are KOBAS (KEGG Orthology-Based Annotation System, [22]) and KAAS (KEGG Automatic Annotation Server) that are devoted to functional annotation of complete genomes.

Recent advancements showed that clustering techniques applied to matrices storing pair-wise similarities [23–25] perform quite well; these algorithms focused on either the grouping of weakly similar homologs or the identification of protein domains. Recent software includes orthoMCL [19], which is based on the Markov Clustering algorithm [26] previously implemented by Enright *et al*. [27] in tribeMCL. TribeMCL was designed to identify groups of evolutionarily related sequences (*tribes*), while orthoMCL is designed to identify clusters composed of *true* orthologous sequences. Ortholuge [28] aims at evaluating clusters of homologous

sequences to identify *bona fide* orthologs by comparing proteins and species phylogenetic trees. InParanoid [29] performs a similar task, and it has also been recently implemented to allow multiple proteomes comparisons (MultiParanoid, [30]).

Phylogenetic fingerprinting

The BBH criterion for orthologs identification is an approximate method because it relies on pair-wise similarities only. For small-scale analysis, the better is to choose phylogenetics methods, whose fundamental role in bioinformatics is well-established [31, 32]. Parsimony and distance-based methods are widely used but the most statistically robust approach is to consider the problem in a likelihood framework and use accurate models of evolution. Maximum likelihood (ML) takes the hypothesis (the tree topology) that maximizes the likelihood of the data (the sequence alignment) in the light of an evolutionary model.

A great attraction of this approach is the ability to perform robust statistical hypothesis tests and to use modern statistical techniques such as hidden Markov models, Markov chain Monte Carlo and Bayesian inference [33, 34]. The ML framework also allows each site of the alignment to evolve with different replacement patterns, and with different substitution rates in all branches of the tree [31] as in real proteins, where slowly evolving sites are generally functionally or structurally constrained, while variable sites tend to be less important for protein function. ConSurf [35, 36] takes advantage of the 3D structure of a protein to obtain a graphical mapping of site-specific evolutionary rates. The rate of evolution at each site is calculated using either an empirical Bayesian [37] or a ML [38] method. A similar perspective has been implemented in ConSeq [39] for proteins with unknown structure. In this case, functionally or structurally important sites are identified on the basis of a neural network for calculating site accessibility using the methods described in [35, 40].

Similarly, when considering a pair of paralogous proteins that underwent evolutionary divergence (i.e. they recognize different substrates, they interact with different partners and so on), it is possible to identify those residues plausibly involved in divergence by studying substitution rates differences [41–43]. The identification of functionally or structurally important residues in proteins is fundamental for protein function improvement in a directed way.

Operon prediction

Genes of a metabolic pathway can be organized in operons; the evolutionary origin of operons and the selective forces promoting or demoting it are still a matter of debate [44–50], but one of the major benefits of an operon is the co-expression of component genes. Genes belonging to the same operon are often involved in the same metabolic pathway; it follows that if an unknown gene is found in operon with genes of a specific process, it might be involved in the same or a related one, especially if this association is evolutionary conserved.

For example, the *Lactococcus lactis* histidine biosynthetic operon contains genes not found in the *Escherichia coli* operon. Two of these genes have been identified as being involved in histidine biosynthesis: they are *hisZ*, previously known as ORF3 [51], and *hisJ*, previously known as ORF13 [52]. The former encodes a regulatory subunit of the first enzyme of the pathway (HisG); the latter is instead the Histidinol-phosphate phosphatase, performing the eighth step of the route. Comparative analyses failed in recognizing these genes because the first one is missing in *E. coli* and the second belongs to a different protein family [52, 53].

This case shows that the development of computational tools for the prediction of operons, coupled with comparative genomics, might help in assigning gene functions, which represents one of the most important goals in the genome era.

Most predictions have focused on *E. coli* and *Bacillus subtilis*, and were trained and validated on databases of experimentally identified transcripts [54–59]. Unfortunately, these resources are available just for a few organisms (cultivable and well-studied); therefore, unsupervised methods for operon prediction are essential.

Some of the proposed algorithms are based on functional classification of the genes and codon usage similarities to form a numerically weighted set for gene pair scoring [60] and can be taken as data sources for machine-learning, such as Bayesian network and joint probabilistic distribution [57, 58, 61, 62]. They can also be converted into fuzzy values for genetic algorithms [63].

The most comprehensive online tools are Operon DataBase [64] and MicrobesOnline [65] (Table1). The first one provides a data retrieval system of operons in many genomes; information from comparative genomics, metabolic networks and expression has been integrated in the

Table I: List of selected resources for functional genomics

Resource	Web Address	Reference
Orthologs identification and protein analysis		
Pfam	http://www.sanger.ac.uk/Software/Pfam/	Finn *et al*. [10]
SMART	http://smart.embl-heidelberg.de/	Schultz *et al*. [12]; Letunic *et al.*[11]
ProtoNet	http://www.protonet.cs.huji.ac.il/	Sasson *et al*.[23]
MCL	http://micans.org/mcl/	Enright *et al*.[26]
RIO	http://www.rio.wustl.edu/	Zmasek and Eddy [15]
OrthoMCL	http://www.cbil.upenn.edu/gene-family/	Li *et al.* [19]; DB: Chen *et al.*. 2006 [20]
BranchClust	http://bioinformatics.org/branchclust/	Poptsova and Gogarten, 2007 [140]
MultiParanoid	http://multiparanoid.cgb.ki.se/	Alexeyenko *et al*. [30]
CDD	http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd	Marchler-Bauer *et al*., 2005 [141]
PROSITE	http://hits.isb-sib.ch/cgi-bin/PFSCAN	Hulo *et al*., 2004 [142]
PSORT	http://psort.nibb.ac.jp/	Wu *et al.* [22]
SignalP	http://www.cbs.dtu.dk/services/SignalP/	Bendtsen *et al*., [132]
Phylogenetic inference		
Phylip	http://evolution.gs.washington.edu/phylip.html	
PAUP*	http://paup.csit.fsu.edu/	
MEGA	http://www.megasoftware.net/mega.html	Tamura *et al.*, 2007 [136]
PAML	http://abacus.gene.ucl.ac.uk/software/paml.html	Yang [137]
Gene assignments		
KOBAS	http://kobas.cbi.pku.edu.cn/	Wu *et al.* [22]
KAAS	http://www.genome.jp/kegg/kaas/	Moriya *et al*., 2007 [143]
Operon prediction		
ODB: Operon Database	http://odb.kuicr.kyoto-u.ac.jp/	Okuda et al.[64]
MicrobesOnline	http://www.microbesonline.org/	Price et al. [62]
InterPro	http://www.ebi.ac.uk/interpro/	Mulder *et al.*, 2007 [66]
KEGG	http://www.genome.jp/kegg/	Kanehisa *et al.* [67]
Gene Ontology (GO)	http://www.geneontology.org/	The Gene Ontology Consortium, 2000 [144]
Regulatory motifs identification		
MotifRegressor	http://www.math.umass.edu/~conlon/mr.html	Conlon *et al*. [72]; Liu *et al*. [75]
Reduce	http://bussemaker.bio.columbia.edu/reduce/	Roven and Bussemaker [73]
MDscan	http://ai.stanford.edu/~xsliu/MDscan/	Liu *et al*. [75]
PhyloGibbs	http://www.imsc.res.in/~rsidd/phylogibbs/	Siddharthan *et al*. [76]
PhyloCon	http://ural.wustl.edu/~twang/PhyloCon/	Wang and Stormo [77]
FootPrinter	http://wingless.cs.washington.edu/htbin-post/unrestricted/FootPrinterWeb/FootPrinterInput2.pl	Blanchette and Tompa [78]
MotifScorer	http://www.dbag.unifi.it/renatofanilab/motifscorer.htm	Brilli *et al*. [74]
AlignACE	http://atlas.med.harvard.edu/cgi-bin/alignace.pl	Roth *et al*., 1998 [145]
Structural analysis of DNA		
Mfold	http://www.bioinfo.rpi.edu/applications/mfold/	Zuker 2003 [130]
DNA analysis	http://hydra.icgeb.trieste.it/dna/	Vlahovicek *et al*. [131]
DNA structural Atlas	http://www.cbs.dtu.dk/services/GenomeAtlas/	Pedersen *et al*. [90]
Protein–Protein Interaction		
DIP	http://dip.doe-mbi.ucla.edu	Salwinski *et al*. [133]
MINT	http://cbm.bio.uniroma2.it/	Chatr-aryamontri *et al*.[134]
PREDICTOME - VisANT	http://predictome.bu.edu/	Hu *et al*.[135]
STRING	http://string.embl.de/newstring.cgi/show.input.page.pl	Von Mering *et al*. [115]
Gene expressivity		
CodonW	http://bioweb.pasteur.fr/seqanal/interfaces/codonw.html	Peden J, PhD thesis [146]
EvolvingCode.net	http://www.evolvingcode.net/codon/cai/cai.php	
CAIAP	www.unifi.it/scibio/bioinfo/caiap/html/	Ramazzotti *et al*.[109]
Protein variability		
ConSeq	http://conseq.bioinfo.tau.ac.il/	Berezin *et al*. [39]
ConSurf	http://consurf.tau.ac.il/	Landau *et al*. [36]
General purpose resources		
MIPS	http://mips.gsf.de	Mewes *et al*. [138]
PUMA2	http://compbio.mcs.anl.gov/puma2/	Maltsev *et al*. [139]

prediction algorithm. MicrobesOnline offers tools to analyse the potential functions of genes in predicted operons, including protein family analyses derived from InterPro [66] and COG [13], metabolic maps from KEGG [67] and links to research papers. Genomic context has been associated to phylogenetic methods allowing studies of phylogenetic distribution, allowing an approximate assessment of the statistical significance of gene clusters. Moreover, this database includes transcriptomics data to track expression changes of genes belonging to the operons of interest. This algorithm accounts for the number of base pairs separating two genes, how often their orthologs are contiguous in other genomes, whether their predicted functions are in the same COG [13] and the similarity of their codon adaptation index (CAI, see subsequently), which is a measure of synonymous codon usage bias [68] and then outputs the probability of two or more genes being in the same operon.

Regulatory motifs identification

Although operons constitute an important portion of many prokaryotic genomes, co-regulation might be achieved if different genes are targets of a common transcription factor, through the presence of similar regulatory motifs in their promoter regions. A DNA sequence motif is a relatively well-conserved pattern that has a biological significance; in this case, it is the target of a transcription factor.

The automatic identification of functional motifs is fundamental to understand the intricate network of regulatory relationships taking place in the cell; however, this issue continues to be a challenging problem in computational biology because regulatory motifs are short (5–25 nucleotides), degenerated and embedded in a noisy background DNA.

On a biotechnological perspective, it is important to characterize those natural promoters having the desired properties, such as activating transcription of downstream genes in response to specific signals (i.e. the [un]availability of a given compound) or constitutively at a defined level. It follows that exploiting the regulatory network of an organism will not only allow a deeper understanding of its biology, but also the identification of promoters representing good starting points for expression pattern engineering; the use of finely tuned promoters allows engineering organisms expressing the alien pathway in specific conditions and in predefined amounts.

For this reason, in the last years, we have experienced an exponential increase in the number of available resources for motif identification (motif finding algorithms, MFA).

Most MFA describe motifs probabilistically, i.e. using position weight matrices (PWM, an $m \times n$ matrix, where $m = 4$ and $n =$ motif length). Each entry $j_{m,n}$ of the matrix is the probability of having nucleotide m at position n, allowing to model and take into account the intrinsic variability of DNA motifs.

By considering the different positions of a motif as if they evolved independently, the score of a sequence given a PWM can be calculated using probability rules. Compositional features of the background DNA must be considered, allowing to weight motif scores; this is often done using Markov chains, where empirical frequencies of single bases, dinucleotides or higher order oligonucleotides are used to calculate the probability that a DNA string has emerged by chance in that genome. High-order model show a better fit to empirical data, but calculations might be affected by finite size effects; most used are Markov models of order 5–7.

One of the ways to obtain the score of a motif is by making the ratio of its probabilities calculated from the PWM and the background frequencies.

However, even with very accurate backgrounds most of the high-scoring motifs are not true regulatory sites, therefore the new generation of regulatory motif identification have been developed to take into account additional information; we might identify two major categories of MFA: the first one can be defined as the '*single species multiple genes*' approach and it is based on the identification of unusually common motifs in upstream sequences of co-regulated genes; these are identified from available expression data or functional classification. A complete survey of existing tools is out the scope of the present review and the interested reader will find very accurate comparisons of most popular MFAs in [69–71].

A relatively novel approach in this category combines MFA and transcriptomics or proteomics data to reduce false positives; it has been implemented e.g. in MotifRegressor [72], REDUCE [73] and MotifScorer [74]. In MotifRegressor, candidate motifs are collected with MDscan [75], which implements a Gibbs sampler and returns the motif PWM and the list of motif occurrences; these information, together with the background model,

are used to score each motif, obtaining an $m \times n$ matrix (the *scores matrix*), where m and n are the number of genes and motifs, respectively; each column of the scores matrix is first used to fit a simple linear regression model, where the dependent variable is the log transformed expression ratio of genes. Candidate motifs with a significant *P*-value are retained and used in the following stepwise regression procedure [72]. The regression coefficient of each significant motif indicates if it is correlated with enhanced or inhibited expression of the corresponding gene, allowing the identification of important regulatory motifs.

REDUCE [73] is based on a multivariate model in which upstream motifs contribute additively to the log-expression level of a gene. While MDscan needs some knowledge on coregulation of input genes this method requires a single genome-wide set of expression levels and the upstream sequence for each gene, and outputs statistically significant motifs.

MotifScorer [74] has been designed to work with several motif finding programs, and it allows merging the results obtained with different programs. Tompa *et al*. [69, 70] showed that different MFAs suffer distinct problems; MotifScorer helps overcoming these limits allowing the use of several MFAs. It implements Partial Least Squares regression techniques and could be used with multiple expression conditions in a novel and ambitious strategy: motifs are searched separately in groups of surely coregulated genes and the PWM found used to score the upstream sequence of each gene in the genome. The resulting scores matrix is used to perform advanced regression techniques against a collection of expression data i.e. an expression compendium. The output of the regression virtually allows deriving the regulatory network and its changes in the conditions tested, along with its time-dependent dynamics using time-series.

The second category of MFAs is often indicated as *phylogenetic footprinting* or '*single gene, multiple species*' approach, in contrast with the previous one. It relies on the higher conservation of regulatory sequences across species over non-functional intergenic regions at close evolutionary distances. In phylogenetic footprinting, promoter sequences are not regarded as evolving independently but as orthologous sequences linked by a phylogenetic tree and deriving from a common promoter ancestor. By comparative genomic analysis, motifs emerge as unusually well-conserved substrings. This strategy has been implemented, e.g. in PhyloGibbs [76], PhyloCon [77] and FootPrinter [78], the last being accessible online (Table 1).

Structural properties of DNA regulatory regions

The characterization of the structural properties of DNA regions surrounding a promoter is an important step towards engineering efficient promoters with optimized recognition by a given transcription factor. Transcription begins with the RNA polymerase (RNAP) binding to DNA in the promoter region, close to the transcription start site. How RNAP exactly locates specific binding sites in a large excess of non-promoter DNA is a field of intense investigation [79, 80] but it is difficult to believe that regulatory motifs are completely responsible for RNAP–promoter finding. In fact, surrounding sequences are also important, and the structural properties of promoters might play a (fundamental) role. An important step during transcription is a local separation of the two strands with formation of the open complex [81–86]; accordingly, the low stability of promoter regions assists in initial melting [87–90]. Many studies show that promoters are often significantly curved if compared to the bulk DNA [91–95]. Curvature can be an inherent, sequence-dependent property of a DNA molecule; another curvature-related property of DNA is bendability, which is defined as the ease with which the molecule can be made to curve by an external factor, i.e. binding of a protein. Experiments have supported a functional role for bendability suggesting that promoter DNA wraps around the RNA polymerase [96, 97] during the first steps of transcription process. Available tools for characterizing several structural properties of DNA sequences are listed in Table 1.

The importance of the structural properties of promoter regions has motivated the use of a measure of DNA anisotropic flexibility to predict highly expressed genes in microbial genomes [98]. The *position preference* measure was originally derived for eukaryotes [99] and it is in the form of a trinucleotide model describing the preference of a given trinucleotide for being in a nucleosomal region. This reasoning applies to eukaryotes only since prokaryotes do not have nucleosomes; however, the position preference value is also an index of DNA flexibility and it might be used to describe how easily

a DNA region can be wrapped around chromatin proteins. Regions of DNA that are not condensed into chromatin are more accessible to the RNA polymerase suggesting a possible correlation between position preference values and expressivity of genes [100].

Predicting gene expression

The work of Willenbrock and Ussery [98] represents a novel way to predict gene expression with a good correlation with experimental data in the six organisms considered.

The first approach for such a prediction was the Codon Adaptation Index (CAI) presented by Sharp and Li [68]; CAI takes advantage of the genetic codes degeneracy and followed the observation that highly expressed genes prefer some of the synonymous codons. The evolutionary trend towards biased relative synonymous codon usages (RSCU) has been related to the different expression levels of synonymous tRNAs and very probably evolved to face tRNA depletion when the rate of translation of a mRNA is high and/or the misincorporation of amino acids due to the use of codons with rare cognate tRNAs [101, 102].

Taking these measures into account, biases have been used for estimating the 'expressivity' of genes [103–109]. The CAI quantifies the similarity of the RSCU of a gene and that calculated for a set of highly expressed genes [68] and it allows inferring a putative 'expressivity' value for each gene. A second approach is based on the idea that tRNA copy number is linearly correlated with tRNA abundance; genes which are enriched in codons recognized by abundant tRNAs are likely to have high translational rates (tRNA Adaptation Index or tAI [110]). Given the tAI and the CAI are based on the translation process they allow studying protein coding genes (i.e. not genes encoding ribosomal RNA), while the approach based on position preference is more general.

All these predictors of expressivity might be used to optimize genes' sequences to be inserted in an organism allowing achieving the best possible translational efficiency, as implemented in OPTIMIZER [111], which is specifically devoted to codon usage adjustment for heterologous gene expression.

Some of the available tools for analysing gene expressivity are listed in Table 1.

Predicting protein–protein interactions

Most biological functions are regulated by protein–protein interactions and the advent of several high-throughput analytical methods has brought to a rapid popularity increase for large scale protein–protein interaction research. There are various high-throughput methods for detecting protein interactions such as yeast two-hybrid, affinity purification/mass spectrometry, protein chips, phage display and synthetic lethality [112]; these techniques generate vast amounts of data but with many false positives. For this reason, computational methods have been developed to assess the quality and validate interaction data (i.e. using phylogenetic conservation, the subcellular localization of interaction partners, their degree of co-expression, and network topology; see [113] and reference therein for a detailed description).

Small-scale analyses of characterized proteins suggest that proteins involved in the same metabolic process are often interacting partners [114, 115] and accurate maps of interaction may therefore help to assign protein functions. However, these techniques give static views, while single proteins can interact with diverse partners under different conditions, resulting in different biological outcomes depending on what protein partner(s) is (are) present.

In recent times, scientists have coped with the complexity of high-throughput methods outputs applying graph theory to their study. In a protein interaction network, nodes are proteins and links connect interacting partners. The functional role of a protein can be viewed as its 'position' within the cellular interaction network stimulating the idea that in the post-genomic era, we need an 'expanded' view of protein function, to take into account that proteins are embedded in a meta-network made by physical and chemical interactions [116].

Along with experimental techniques, several approaches for the identification of interacting partners from a computational perspective have been developed. The methods can be classified into three categories:

- Phylogenetic profiling: analysis of co-occurrence of genes within entire genomes, used as an indicator of interactions between those gene products [117, 118].
- Gene fusions: if two stand-alone proteins in some organism are fused into a single protein in other, they are likely to interact [119, 120].

- Gene neighbourhood: in bacteria, genes encoding functionally related proteins tend to be localized in the same operon and contiguous genes are indeed more likely to interact than genes that are not neighbourhood in any genomes [121–123].

Moreover, as in the case of functional transfer by sequence homology, it is also possible to make inferences on the interaction partners of a protein by looking for documented interactions involving its orthologs. Huynen and colleagues [124] have shown that current methods allow the prediction of functional links with 80% confidence for the majority of prokaryotic proteomes; they also implemented a web server, STRING [115] allowing the identification and characterization of such interactions. Using computational methods, we can thus obtain a view on the network of interactions within a cell, even if we do not know the functions of its components.

CONCLUSION AND PERSPECTIVES

Comparative genomics suggests that variability is a major feature of living organisms. In fact, model organisms (e.g. *E. coli* and *B. subtilis*) often show features that are not (fully) conserved in other; the same is true for regulatory mechanisms controlling gene expression [44, 125–129]. Experimental approaches give insight into these variations (e.g. for regulatory divergence); however, novel computational approaches are required to extract information from sequence data only, and allow studying organisms that are difficult to be manipulated in the laboratory; computational analyses might suggest possible biologically interesting genes, allowing to focus efforts on specific targets. The tools described in the previous sections might be used to design artificial pathways with the desired properties and, more importantly, able to function efficiently in a specific organism or condition.

Artificial pathways might be viewed as perturbations for intracellular systems and their design must be carefully calibrated to minimize the risk of unwanted and deleterious effects. Obtaining a deeper knowledge of the properties of natural pathways is the basis for a rational design of novel pathways fulfilling particular tasks *in vivo*.

The reductionist approach of the last century has provided useful information about individual or small-scale cellular components; however, the plethora of modern high-throughput approaches has clarified that discrete functions can be attributed to individual molecules in few cases with most biological characteristics emerging from the complex interactions between cells constituents (proteins, DNA, RNA and small molecules) [5]. The different levels that make up the living cell are not independent entities but the modules of a *network of networks* that is responsible for the complex behaviour of the cell. The same idea applies to an ecological perspective, because each organism is not independent but integrated in biological communities where the relationships between individuals and with the surrounding environment are responsible for much of the changes taking place inside the cell.

It follows that one of the major challenges of contemporary biology is to embark on integrated theoretical and experimental programs to map out, understand and model different networks that control the behaviour of the cell [5], allowing inferences on the structure and the dynamics of the complex intracellular webs that contribute to the structure and function of a living cell.

Key Points

- Metabolic engineering allows designing organisms with specific properties using available efficient building blocks identified in existing metabolic pathways.
- A correct identification of orthologs is at the core of metabolic pathway analysis.
- Analysis of gene organization helps assigning gene functions and allows the identification of missing genes in metabolic pathways.
- Studying natural promoters and their structural properties is fundamental to design organism expressing the engineered pathway in specific conditions.
- Codon usage can be studied to develop gene sequences optimized for high-level translation.
- The integration of different sources of information will allow depicting a global view of the 'network of networks', which makes up the living cell.

Acknowledgements

The authors wish to thank two anonymous reviewers whose criticism and suggestions helped to notably improve the article.

References

1. Pleiss J. The promise of synthetic biology. *Appl Microbiol Biotechnol* 2006;**73**:735–9.
2. Tyo KE, Alper HS. Stephanopoulos GN. Expanding the metabolic engineering toolbox: more options to engineer cells. *Trends Biotechnol* 2007;**25**:132–7.

3. Smid EJ, Molenaar D, Hugenholtz J, *et al*. Functional ingredient production: application of global metabolic models. *Curr Opin Biotechnol* 2005;**16**:190–7.
4. Oksman-Caldentey KM, Saito K. Integrating genomics and metabolomics for engineering plant metabolic pathways. *Curr Opin Biotechnol* 2005;**16**:174–9.
5. Barabasi AL, Oltvai ZN. Network biology: understanding the cells functional organization. *Nat Rev Genet* 2004;**5**: 101–13.
6. Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* 1997;**28**:405–20.
7. Bateman A, Birney E, Durbin R, *et al*. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res* 1999;**27**:260–2.
8. Bateman A, Coin L, Durbin R, *et al*. The Pfam protein families database. *Nucleic Acids Res* 2004;**32**(Database issue): D138–41.
9. Finn RD, Marshall M, Bateman A. iPfam: visualization of protein–protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics* 2005;**21**:410–2.
10. Finn RD, Mistry J, Schuster–Bockler B, *et al*. Pfam: clans, web tools and services. *Nucleic Acids Res* 2006;**34**(Database issue):D247–51.
11. Letunic I, Copley RR, Pils B, *et al*. SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res* 2006;**34**(Database issue):D257–60.
12. Schultz J, Milpetz F, Bork P, *et al*. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci USA* 1998;**95**:5857–64.
13. Tatusov RL, Natale DA, Garkavtsev IV, *et al*. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* 2001;**29**: 22–8.
14. Abascal F, Valencia A. Clustering of proximal sequence space for the identification of protein families. *Bioinformatics* 2002;**18**:908–21.
15. Zmasek CM, Eddy SR. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics* 2002;**3**:14.
16. Storm CE, Sonnhammer EL. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics* 2002;**18**:92–9.
17. Cotter PJ, Caffrey DR, Shields DC. Improved database searches for orthologous sequences by conditioning on outgroup sequences. *Bioinformatics* 2002;**18**:83–91.
18. Cannon SB, Young ND. OrthoParaMap: distinguishing orthologs from paralogs by integrating comparative genome data and gene phylogenies. *BMC Bioinformatics* 2003;**4**:35.
19. Li L, Stoeckert CJ, Jr., Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 2003;**13**:2178–89.
20. Chen F, Aaron JM, Stoeckert CJ, Jr., *et al*. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* 2006;**34**:D363–8.
21. Storm CE, Sonnhammer EL. Comprehensive analysis of orthologous protein domains using the HOPS database. *Genome Res* 2003;**13**:2353–62.
22. Wu J, Mao X, Cai T, *et al*. KOBAS server: a web-based platform for automated annotation and pathway identification. *Nucleic Acids Res* 2006;**34**:W720–4.
23. Sasson O, Vaaknin A, Fleischer H, *et al*. ProtoNet: hierarchical classification of the protein space. *Nucleic Acids Res* 2003;**31**:348–52.
24. Servant F, Bru C, Carrere S, *et al*. ProDom: Automated clustering of homologous domains. *Briefings in Bioinformatics* 2002;**3**:246–51.
25. Yona G, Linial N, Linial M. ProtoMap: automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. *Proteins* 1999; **37**:360–78.
26. van Dongen S., *Graph Clustering by Flow Simulation*, PhD thesis, University of Utrecht. 2000.
27. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 2002;**30**:1575–84.
28. Fulton DL, Li YY, Laird MR, *et al*. Improving the specificity of high-throughput ortholog prediction. *BMC Bioinformatics* 2006;**7**:270.
29. O'Brien KP, Remm M, Sonnhammer EL. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* 2005;**33**(Database issue):D476–80.
30. Alexeyenko A, Tamas I, Liu G, *et al*. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* 2006;**22**:e9–15.
31. Whelan S, Liò P, Goldman N. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet* 2001;**17**:262–72.
32. Liò P, Goldman N. Models of molecular evolution and phylogeny. *Genome Res* 1998;**8**:1233–44.
33. Shoemaker JS, Painter IS, Weir BS. Bayesian statistics in genetics: a guide for the uninitiated. *Trends Genet* 1999;**15**: 354–58.
34. Ewens WJJ, Grant GR. *Statistical Methods in Bioinformatics: An Introduction*. New York: Springer, 2001Inc.
35. Glaser F, Pupko T, Paz I, *et al*. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* 2003;**19**:163–4.
36. Landau M, Mayrose I, Rosenberg Y, *et al*. ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res* 2005; **33**(Web Server issue):W299–302.
37. Mayrose I, Graur D, Ben–Tal N, *et al*. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol Biol Evol* 2004;**21**:1781–91.
38. Pupko T, Bell RE, Mayrose I, *et al*. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* 2002;**18**:S71–7.
39. Berezin C, Glaser F, Rosenberg J, *et al*. ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics* 2004;**20**:1322–4.
40. Pollastri G, Baldi P, Fariselli P, *et al*. Prediction of coordination number and relative solvent accessibility in proteins. *Proteins* 2002;**47**:142–53.
41. Dermitzakis ET, Clark AG. Differential selection after duplication in mammalian developmental genes. *Mol Biol Evol* 2001;**18**:557–62.
42. Gu X. Statistical methods for testing functional divergence after gene duplication. *Mol Biol Evol* 1999;**16**: 1664–74.

43. Gu X. Maximum-likelihood approach for gene family evolution under functional divergence. *Mol Biol Evol* 2001; **18**:453–64.
44. Fani R, Brilli M, Liò P. The origin and evolution of operons: the piecewise building of the proteobacterial histidine operon. *J Mol Evol* 2005;**60**:378–90.
45. Price MN, Alm EJ, Arkin AP. The histidine operon is ancient. *J Mol Evol* 2006;**62**:807–8.
46. Fani R, Brilli M, Liò P. Inference from proteobacterial operons shows piecewise organization: a reply to Price et al. *J Mol Evol* 2006;**63**:577–80.
47. Lawrence JG, Roth JR. Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* 1996;**143**: 1843–60.
48. Svetic RE, MacCluer CR, Buckley CO. A metabolic force for gene clustering. *Bull Math Biol* 2004;**66**:559–81.
49. Omelchenko MV, Makarova KS, Wolf YI. Evolution of mosaic operons by horizontal gene transfer and gene displacement *in situ*. *Genome Biol* 2003;**4**:R55.
50. Tamames J. Evolution of gene order conservation in prokaryotes. *Genome Biol*. 2001;**2**RESEARCH0020.
51. Sissler M, Delorme C, Bond J, *et al*. An aminoacyl-tRNA synthetase paralog with a catalytic role in histidine biosynthesis. *Proc Natl Acad Sci USA* 1999;**96**:8985–90.
52. le Coq D, Fillinger S, Aymerich S. Histidinol phosphate phosphatase, catalyzing the penultimate step of the histidine biosynthesis pathway, is encoded by *ytvP* (*hisJ*) in *Bacillus subtilis*. *J Bacteriol* 1999;**181**:3277–80.
53. Brilli M, Fani R. Molecular evolution of *hisB* genes. *J Mol Evol* 2004;**58**:225–37.
54. Salgado H, Moreno–Hagelsieb G, Smith TF, *et al*. Operons in *Escherichia coli*: genomic analyses and predictions. *Proc Natl Acad Sci USA* 2000;**97**:6652–7.
55. Zheng Y, Szustakowski JD, Fortnow L, *et al*. Computational identification of operons in microbial genomes. *Genome Res* 2002;**12**:1221–30.
56. Sabatti C, Rohlin L, Oh MK, *et al*. Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res* 2002;**30**:2886–93.
57. Bockhorst J, Craven M, Page D, *et al*. A Bayesian network approach to operon prediction. *Bioinformatics* 2003;**19**: 1227–35.
58. Bockhorst J, Qiu Y, Glasner J, *et al*. Predicting bacterial transcription units using sequence and expression data. *Bioinformatics* 2003;**19**(Suppl 1):I34–43.
59. de Hoon MJL, Imoto S, Kobayashi K, *et al*. Predicting the operon structure of *Bacillus subtilis* using operon length, intergene distance, and gene expression information. *Pac Symp Biocomput* 2004;**9**:276–87.
60. Wang L, Trawick JD, Yamamoto R, *et al*. Genome-wide operon prediction in *Staphylococcus aureus*. *Nucleic Acids Res* 2004;**32**:3689–702.
61. Chen X, Su Z, Xu Y, *et al*. Computational prediction of operons in *Synechococcus* sp. WH8102. *Genome Inform* 2004; **15**:211–22.
62. Price MN, Huang KH, Alm EJ, *et al*. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res* 2005;**33**:880–92.
63. Jacob E, Sasikumar R, Nair KN. A fuzzy guided genetic algorithm for operon prediction. *Bioinformatics* 2005;**21**: 1403–7.
64. Okuda S, Katayama T, Kawashima S, *et al*. ODB: a database of operons accumulating known operons across multiple genomes. *Nucleic Acids Res* 2006;**34**(Database issue): D358–62.
65. Alm EJ, Huang KH, Price MN, *et al*. The MicrobesOnline Web site for comparative genomics. *Genome Res* 2005;**15**: 1015–22.
66. Mulder NJ, Apweiler R, Attwood TK, *et al*. New developments in the InterPro database. *Nucleic Acids Res* 2007;**35**(Database issue):D224–8.
67. Kanehisa M, Goto S, Hattori M, *et al*. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 2006;**34**(Database issue):D354–7.
68. Sharp PM, Li WH. The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 1987;**15**: 1281–95.
69. Tompa M, Li N, Bailey TL, *et al*. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* 2005;**23**:137–44.
70. Li N, Tompa M. Analysis of computational approaches for motif discovery. *Algorithms Mol Biol* 2006;**1**:8.
71. MacIsaac KD, Fraenkel E. Practical strategies for discovering regulatory DNA sequence motifs. *PLoS Comput Biol* 2006;**2**:e36.
72. Conlon EM, Liu XS, Lieb JD, *et al*. Integrating regulatory motif discovery and genome-wide expression analysis. *Proc Natl Acad Sci USA* 2003;**18**:3339–44.
73. Roven C, Bussemaker HJ. REDUCE: An online tool for inferring cis-regulatory elements and transcriptional module activities from microarray data. *Nucleic Acids Res*. 2003;**31**: 3487–90.
74. Brilli M, Fani R, Liò P. MotifScorer: using a compendium of microarrays to identify regulatory motifs. *Bioinformatics* 2007;**23**:493–5.
75. Liu XS, Brutlag DL, Liu JS. An algorithm for finding protein–DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol* 2002;**20**:835–9.
76. Siddharthan R, Siggia ED, van Nimwegen E. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol* 2005;**1**:e67.
77. Wang T, Stormo GD. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* 2003;**19**:2369–80.
78. Blanchette M, Tompa M. FootPrinter: a program designed for phylogenetic footprinting. *Nucleic Acids Res* 2003;**31**: 3840–2.
79. Halford SE, Marko JF. How do site-specific DNA-binding proteins find their targets? *Nucleic Acids Res* 2004;**32**: 3040–52.
80. Stanford NP, Szczelkun MD, Marko JF, *et al*. One- and three-dimensional pathways for proteins to reach specific DNA sites. *EMBO J* 2000;**19**:6546–57.
81. Buckle M, Buc H. Fine mapping of DNA single-stranded regions using base-specific chemical probes: study of an open complex formed between RNA polymerase and the lac UV5 promoter. *Biochemistry* 1989;**28**:4388–96.
82. Chen YF, Helmann JD. DNA-melting at the *Bacillus subtilis* flagellin promoter nucleates near–10 and expands unidirectionally. *J Mol Biol* 1997;**267**:47–59.

83. Craig ML, Suh WC, Record MT, Jr.. HO. and DNase I probing of E sigma 70 RNA polymerase–lambda PR promoter open complexes: Mg2+ binding and its structural consequences at the transcription start site. *Biochemistry* 1995;**34**:15624–32.
84. Sasse–Dwight S, Gralla JD. $KMnO_4$ as a probe for lac promoter DNA melting and mechanism in vivo. *J Biol Chem* 1989;**264**:8074–81.
85. Siebenlist U, Simpson RB, Gilbert W. *E. coli* RNA polymerase interacts homologously with two different promoters. *Cell* 1980;**20**:269–81.
86. Suh WC, Ross W, Record MT, Jr.. Two open complexes and a requirement for Mg^{2+} to open the lambda PR transcription start site. *Science* 1993;**259**:358–61.
87. Nakata K, Kanehisa M, Maizel JV, Jr.. Discriminant analysis of promoter regions in Escherichia coli sequences. *Comput Appl Biosci* 1988;**4**:367–71.
88. Vollenweider HJ, Fiandt M, Szybalski W. A relationship between DNA helix stability and recognition sites for RNA polymerase. *Science* 1979;**205**:508–11.
89. Margalit H, Shapiro BA, Nussinov R, *et al*. Helix stability in prokaryotic promoter regions. *Biochemistry* 1988;**27**: 5179–88.
90. Pedersen AG, Jensen LJ, Brunak S, *et al*. A DNA structural atlas for *Escherichia coli*. *J Mol Biol* 2000;**299**:907–30.
91. Kozobay–Avraham L, Hosid S, Bolshoy A. distribution in prokaryotic genomes. *In Silico Biol* 2004;**4**:29.
92. Jauregui R, Abreu–Goodger C, Moreno–Hagelsieb G, *et al*. of DNA curvature signals in regulatory regions of prokaryotic genes. *Nucleic Acids Res* 2003;**31**:6770–7.
93. Kalate RN, Kulkarni BD, Nagaraja V. Analysis of DNA curvature distribution in mycobacterial promoters using theoretical models. *Biophys Chem* 2002;**99**:77–97.
94. Gabrielian AE, Landsman D, Bolshoy A. Curved DNA in promoter sequences. *In Silico Biol* 2000;**1**:183–96.
95. Tosato V, Gjuracic K, Vlahovicek K, *et al*. The DNA secondary structure of the Bacillus subtilis genome. *FEMS Microbiol Lett* 2003;**218**:23–30.
96. Rivetti C, Guthold M, Bustamante C. Wrapping of DNA around the *E. coli* RNA polymerase open promoter complex. *EMBO J* 1999;**18**:4464–75.
97. Cheetham GM, Jeruzalmi D, Steitz TA. Structural basis for initiation of transcription from an RNA polymerase–promoter complex. *Nature* 1999;**399**:80–3.
98. Willenbrock H, Ussery DW. Prediction of highly expressed genes in microbes based on chromatin accessibility. *BMC Mol Biol* 2007;**8**:11.
99. Satchwell SC, Drew HR, Travers AA. Sequence periodicities in chicken nucleosome core DNA. *J Mol Biol* 1986; **191**:659–75.
100. Dlakic M, Ussery D, Brunak S. DNA bendability and nucleosome positioning in transcriptional regulation.. *In DNA Conformation in Transcription*. Ohyama T: Landes Bioscience, 2004.
101. Rocha EP. Codon usage bias from tRNAs point of view: redundancy, specialization, and efficient decoding for translation optimization. *Genome Res* 2004;**14**: 2279–86.
102. Ikemura T. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J Mol Biol* 1981;**151**:389–409.
103. Wu G, Nie L, Zhang W. Predicted highly expressed genes in *Nocardia farcinica* and the implication for its primary metabolism and nocardial virulence. *Antonie Van Leeuwenhoek* 2006;**89**:135–46.
104. Liu Q. Analysis of codon usage pattern in the radioresistant bacterium *Deinococcus radiodurans*. *Biosystems* 2006;**85**: 99–106.
105. Jia M, Li Y. The relationship among gene expression, folding free energy and codon usage bias in *Escherichia coli*. *FEBS Lett* 2005;**579**:5333–7.
106. Das S, Ghosh S, Pan A, *et al*. Compositional variation in bacterial genes and proteins with potential expression level. *FEBS Lett* 2005;**579**:5205–10.
107. Wu G, Culley DE, Zhang W. Predicted highly expressed genes in the genomes of Streptomyces coelicolor and *Streptomyces avermitilis* and the implications for their metabolism. *Microbiology* 2005;**151**(Pt 7):2175–87.
108. Martin–Galiano AJ, Wells JM, de la Campa AG. Relationship between codon biased genes, microarray expression values and physiological characteristics of *Streptococcus pneumoniae*. *Microbiology* 2004;**150**(Pt 7):2313–25.
109. Ramazzotti M, Brilli M, Fani R, *et al*. The CAI Analyser Package: inferring gene expressivity from raw genomic data. *In Silico Biol* 2007;**7**</isb/2007/07/0036/>.
110. Man O, Pilpel Y. Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. *Nat Genet* 2007;**39**:415–21.
111. Puigbo P, Guzman E, Romeu A, *et al*. OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. *Nucleic Acids Res* 2007;**35**(Web Server issue): W126–31.
112. Cho S, Park SG, Lee DH, *et al*. Protein–protein interaction networks: from interactions to networks. *J Biochem Mol Biol* 2004;**37**:45–52.
113. Cusick ME, Klitgord N, Vidal M, *et al*. Interactome: gateway into systems biology. *Hum Mol Genet* 2002;**14**: R171–81.
114. von Mering R, Krause B, Snel M, *et al*. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* 2002;**417**:399–403.
115. von Mering C, Jensen LJ, Kuhn M, *et al*. STRING 7-recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res* 2007;**35**(Database issue):358–62.
116. Eisenberg D, Marcotte EM, Xenarios I, *et al*. Protein function in the post-genomic era. *Nature* 2000;**405**:823–6.
117. Huynen MA, Snel B, von Mering C, *et al*. Function prediction and protein networks. *Curr Opin Cell Biol* 2003; **15**:191–8.
118. Pellegrini M, Marcotte EM, Thompson MJ, *et al*. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA* 1999;**96**: 4285–8.
119. Enright AJ, Iliopoulos I, Kyrpides NC, *et al*. Protein interaction maps for complete genomes based on gene fusion events. *Nature* 1999;**402**:86–90.
120. Marcotte EM, Pellegrini M, Ng HL, *et al*. Detecting protein function and protein–protein interactions from genome sequences. *Science* 1999;**285**:751–3.

121. Dandekar T, Snel B, Huynen M, *et al*. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* 1998;**23**:324–8.

122. Huynen M, Snel B, Lathe W, 3rd, *et al*. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res* 2000;**10**: 1204–10.

123. Huynen M, Snel B, Lathe W, 3rd, *et al*. Exploitation of gene context. *Curr Opin Struct Biol* 2000;**10**:366–70.

124. Huynen MA, Snel B, von Mering C, *et al*. Function prediction and protein networks. *Curr Opin Cell Biol* 2003; **15**:191–8.

125. Landry CR, Oh J, Hartl DL, *et al*. Genome-wide scan reveals that genetic variation for transcriptional plasticity in yeast is biased towards multi-copy and dispensable genes. *Gene* 2006;**366**:343–51.

126. Cummings CA, Bootsma HJ, Relman DA, *et al*. Species- and strain-specific control of a complex, flexible regulon by Bordetella BvgAS. *J Bacteriol* 2006;**188**: 1775–85.

127. Tettelin H, Masignani V, Cieslewicz MJ, *et al*. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial ''pan-genome''. *Proc Natl Acad Sci USA* 2005;**102**:13950–5.

128. Iturbe–Ormaetxe I, Burke GR, Riegler M, *et al*. Distribution, expression, and motif variability of ankyrin domain genes in Wolbachia pipientis. *J Bacteriol* 2005;**187**: 5136–45.

129. Ou K, Ong C, Koh SY, *et al*. Integrative genomic, transcriptional, and proteomic diversity in natural isolates of the human pathogen Burkholderia pseudomallei. *J Bacteriol* 2005;**187**:4276–85.

130. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 2003;**31**: 3406–15.

131. Vlahovicek K, Kajan L, Pongor S. DNA analysis servers: plot.it, bend.it, model.it and IS. *Nucleic Acids Res* 2003;**31**: 3686–7.

132. Bendtsen JD, Nielsen H, von Heijne G, *et al*. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 2004; **340**:783–95.

133. Salwinski L, Miller CS, Smith AJ, *et al*. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* 2004; **32**(Database issue):D449–51.

134. Chatr–aryamontri A, Ceol A, Palazzi LM, *et al*. MINT: the Molecular INTeraction database. *Nucleic Acids Res* 2007; **35**(Database issue):D572–4.

135. Hu Z, Mellor J, Wu J, *et al*. VisANT: data-integrating visual framework for biological networks and modules. *Nucleic Acids Res* 2004;**33**(Web Server issue):W352–7.

136. Tamura K, Dudley J, Nei M, *et al*. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 2007;**24**:1596–9.

137. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* 1997;**13**:555–6.

138. Mewes HW, Frishman D, Mayer KF, *et al*. MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res* 2005;**34**(Database issue):D169–72.

139. Maltsev N, Glass E, Sulakhe D, *et al*. PUMA2—grid-based high-throughput analysis of genomes and metabolic pathways. *Nucleic Acids Res* 2006;**34**(Database issue): D369–72.

140. Poptsova MS, Gogarten JP. BranchClust: a phylogenetic algorithm for selecting gene families. *BMC Bioinformatics* 2007;**8**:120.

141. Marchler-Bauer A, Anderson JB, Cherukuri PF, *et al*. CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res* 2005;**33**(Database issue):D192–6.

142. Hulo N, Sigrist CJ, Le Saux V, *et al*. Recent improvements to the PROSITE database. *Nucleic Acids Res* 2004;**32**(Database issue):D134–7.

143. Moriya Y, Itoh M, Okuda S, *et al*. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 2007;**35**(Web Server issue):W182–5.

144. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nat Genet* 2000;**25**:25–9.

145. Roth FP, Hughes JD, Estep PW *et al*. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol* **16**:939–945.

146. Peden J., *Analysis of Codon Usage*. PhD thesis, University of Nottingham. 1999.