Available online at www.sciencedirect.com
ScienceDirect
Physics of Life Reviews 5 (2008) 22–49
PHYSICS of LIFE reviews
www.elsevier.com/locate/plrev

Review

Comparative genomics and functional annotation of bacterial transporters

Mikhail S. Gelfand [a,b,*], Dmitry A. Rodionov [a,c]

[a] *Institute for Information Transmission Problems, Russian Academy of Sciences, Bolshoi Karetny pereulok 19, Moscow 127994, Russia*
[b] *Department of Bioengineering and Bioinformatics, Moscow State University, Russia*
[c] *Burnham Institute for Medical Research, La Jolla, CA 92037, USA*

Received 6 October 2007; received in revised form 8 October 2007; accepted 10 October 2007

Available online 24 October 2007

Communicated by M. Frank-Kamenetskii

Abstract

Transport proteins are difficult to study experimentally, and because of that their functional characterization trails that of enzymes. The comparative genomic analysis is a powerful approach to functional annotation of proteins, which makes it possible to utilize the genomic sequence data from thousands of organisms. The use of computational techniques allows one to identify candidate transporters, predict their structure and localization in the membrane, and perform detailed functional annotation, which includes substrate specificity and cellular role.

We overview the main techniques of analysis of transporters' structure and function. We consider the most popular algorithms to identify transmembrane segments in protein sequences and to predict topology of multispanning proteins. We describe the main approaches of the comparative genomics, and how they may be applied to the analysis of transporters, and provide examples showing how combinations of these techniques is used for functional annotation of new transporter specificities in known families, characterization of new families, and prediction of novel transport mechanisms.

Keywords: Genome analysis; Transport systems; Bacteria; Comparative genomics

Contents

* Corresponding author at: Department of Bioengineering and Bioinformatics, Moscow State University, Russia.
E-mail addresses: gelfand@iitp.ru (M.S. Gelfand), rodionov@iitp.ru (D.A. Rodionov).

1571-0645/$ – see front matter
doi:10.1016/j.plrev.2007.10.003

1. Introduction

Almost thirteen hundred bacterial genomes have been sequenced already, and hundreds more are in the pipeline (according to the Genomes Online database, http://www.genomesonline.org/). Most of these genomes will never be studied experimentally at any level of detail. On the other hand, the genome sequences, supplemented by other types of high-throughput data, provide an unprecedented opportunity for applying comparative approaches for functional annotation of genes, metabolic reconstruction, filling gaps in metabolic pathways, characterization of regulatory interactions, and, eventually, reconstruction of phenotype given genome.

Traditionally, the main effort has been directed towards identification of new enzymes. Indeed, about 20–40% of genes in a bacterial genome are so-called conserved hypothetical genes whose orthologs (see Table 1 for explanation of these and other terms used in the field of comparative genomics) are present in a majority of genomes [56,148]. It is reasonable to assume that such genes encode proteins with some universal, essential functions. On the other hand, about 10% of reactions of the central metabolism are not represented by any known gene [148]; among all biochemical reactions catalyzed by proteins, this fraction reaches 40% [100,118]. Thus, the problem is to establish the correspondence between these two sets [57,100,188].

The situation with transporters is even more difficult. These non-enzyme proteins constitute a large and extremely important class of membrane proteins, which perform numerous transmembrane transport functions in the cell, such as various ions transport, transport of vitamins, amino acids, drugs, etc. They form large families with numerous duplications and horizontal transfer events (see Table 1), and transporter specificities towards their substrates are very flexible. This makes it difficult and often impossible to infer transporter specificity from homology information alone, and creates a need for application of advanced comparative genomics techniques. Indeed, in a recent survey of 87 prokaryotic genomes, three fourths of identified transmembrane proteins had functional annotation "unknown" [6].

There are many excellent reviews of genome functional annotation by comparative genomics, in particular [44,49, 55,127,148,149], and several books extensively covering this area [108,142]. There are also many surveys of transporter families and repertoires in individual genomes (see below). However, neither study specially addresses the specifics of comparative genomics of transporters. Here we will try to fill this gap.

The plan of the review is as follows. It starts with a brief description of methods used to identify transporters, transporter classifications and databases. It is followed by a discussion of the basic comparative genomic approaches to functional annotation of hypothetical genes, with a special emphasis on transporters. The last section presents recent examples of such analyses.

2. Computational studies of transporters

Transporters are identified by similarity to known transporters, presence of family-specific sequence signatures such as the pattern (Prosite PS00211, PFAM PF00005) that distinguishes ATPase components of ABC-transporters

Table 1
Glossary

ATP-binding cassette (ABC)	Transporters that use the energy of ATP hydrolysis.
Group translocators	Transporters that combine transport with a chemical reaction.
Homologs	Genes that share a common ancestor.
Horizontal (lateral) transfer	Transfer of genes between two contemporary, maybe distant genomes, as opposed to vertical descent from ancestors to descendants. This leads to incongruence of phylogenetic trees of genes and species.
Non-orthologous gene displacement	Substitution of a pre-existing gene by a *horizontally transferred non-orthologous* gene having the same function. The new gene may be *homologous* to the initial one or unrelated to it.
Non-orthologous gene displacement *in situ*	A specific case of non-*orthologous gene displacement*, when the displacing gene occupies exactly the same position in an operon as the displaced gene.
Orthologs	*Homologs* that diverged following a speciation event. Orthologs usually have the same cellular role.
Paralogs	*Homologs* that diverged after a duplication within a genome. Note that genes in two different genomes still can be paralogs (sometimes called out-paralogs), if their last common ancestor was duplicated, and the genomes diverged after that duplication. Paralogs often have similar, but not identical functions.
Phosphotransferase systems (PTS)	Phosphotransfer-driven *group translocators* that import and phosphorylate sugars.
Phyletic (phylogenetic) profile (pattern)	Separation of a group of genomes into those containing or not-containing representatives of a group of *orthologous* genes.
Primary active transporters	Transporters that use chemical, electrical, or solar energy to move compounds against a concentration gradient.
Regulon	Set of co-regulated genes and operons.
Riboswitch	Regulatory RNA structure that assumes alternative conformations in response to binding by small molecules.
T-box	Regulatory RNA structure that assumes alternative conformations in response to binding to uncharged tRNA.
Secondary porters	Transporters that use electrical potential, ion gradient, or, in case of coupled transport, concentration gradient of another substance.

from other ATPases [69], or by the analysis of structure. There are two major structural classes of transporter transmembrane domains: the main one, alpha-helical bundles, and much less prevalent beta-barrels that form pores in outer membranes, mainly in Gram-negative bacteria and organelles.

2.1. Identification of transmembrane helices

The prediction of the structure of transmembrane proteins (not necessarily transporters) is a traditional area of computational biology. The number of known distinct membrane protein structures is slightly more than one and a half hundred, which, by different estimates, constitutes 0.5–2% of all structures [30,208]. This figure is not surprising, given the fact that such proteins are difficult to overexpress [65] whereas their low concentrations in a cell make direct purification impossible [228,234]. They also notoriously resist chrystallization, a necessary step for structure determination by X-ray chrystallography [222]. The relatively small number of known structures limits the possibilities for extensive training or application of threading approaches to structure prediction, and also complicates independent testing and benchmarking of different algorithms.

There are numerous programs for identification of membrane-spanning alpha-helices (transmembrane-, or TM-segments), for recent reviews see, e.g., [25,30,81,117,150,177,208]. The prediction methods are based on several properties of transmembrane proteins, mainly the prevalence of hydrophobic residues in the TM-segments, specific amino acid patterns, so called caps, at the boundaries of TM-segments and loops [92,137], and differences in the amino acid composition of external and internal (cytoplasmic) loops [18,144,249]. In particular, the caps are enriched in aromatic amino acids tryptophan and tyrosine [239], whereas the preference for positively charged residues in the cytoplasmic loops compared to external ones, the so-called “positive inside rule” [239,249], is used to predict the transporter topology relative to the membrane.

A convenient way to take into account these statistical features is to use the language of hidden Markov models (HMMs) [37]. The most popular programs, implemented as Web servers, seem to be HMM-based TMHMM (http://www.cbs.dtu.dk/services/TMHMM-2.0/) [110,227] with its descendant Phobius (http://phobius.cgb.ki.se/) [95,97] and HMMTOP (http://www.enzim.hu/hmmtop/index.html) [235,236], and a neural network PHDhtm (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_htm.html) [205,206].

While the highest accuracy obtained by cross-validation was reported to be as high as 95% [27,120,205], benchmarking on proteins with available structural data demonstrated that even for the best methods, the prediction accuracy is in the range of 60–80%, dependent on the accuracy measure, the parameter measured (e.g. the fraction of correctly identified residues or segments), and exact definition of the "correct identification" [24,26,85,94,132,139]; the latter distinctions are important, since the average error in determining the TM-segment boundaries is about two turns of the alpha-helix [30]. One of the problems with benchmarking the programs is a very limited amount of available data: for instance, the same dataset was used by different authors in [3,91,164,205]. A specialized Web server (http://cubic.bioc.columbia.edu/services/tmh_benchmark/) for evaluation of new algorithms using a new benchmark set was created to offset some of these problems [103].

The consistency of predictions between various programs is not very high, especially on whole-genome data compared to structural benchmarks [94,208], and again this may serve as an indicator of overuse of the same training set of available structures. One of the manifestations of this lack of consistency is the discrepancy in the estimated number of transmembrane proteins encoded in complete bacterial genomes, e.g., *Escherichia coli* (21–40%), *Mycoplasma pneumoniae* (16–29%), *Synechocystis* sp. (24–41%), etc., listed in [208]. A good way to increase the quality of predictions is to combine predictions by several different methods [5,85,163,176,230]. Similarly, prediction of TM-segments in homologous proteins often produces different numbers of TM-segments, and their boundaries may not match [208]. Simultaneous prediction of TM-segments in aligned homologs leads to considerably better results [2,27,96,159,160,169,205,206,244].

An important practical problem is to distinguish transmembrane helices from signal peptides of secreted proteins [24,95,97,113,143,268]. Another problem, that has become apparent fairly recently, after several structures of multihelix transporters were solved, is that not all membrane helices are really transmembrane: some of them, so-called re-entrant helices, return to the same side of the membrane, making a sharp turn within the latter [30,114]. Some transmembrane helices were found to be very long, sometimes exceeding 40 amino acids [26].

2.2. *Structure of transmembrane proteins*

Although still relatively small, the number of solved membrane protein structures doubles every three years [26,228], and it might be expected to increase even faster due to the activity of major international projects (http://www.utoronto.ca/AlEdwardsLab/membrane_proteomics_index.html). Since the structural data are often difficult to assess directly, in particular, even the orientation of membrane protein structures from databases relative to the membrane sometimes needs to be inferred computationally [124,237,238], several databases were created that contain curated data about membrane protein structures (Table 2). This is accompanied by development of efficient proteomic tools [228,234] and solving the structure of the translocon, the protein complex mediating helix insertion into the membrane [16,20,138,242]. This yielded much better understanding of the translocation mechanism [257].

These developments pave way beyond simple prediction of transmembrane helices and their topology to more detailed structural modeling [45,51,81]. Various specific problems addressed in computational structural studies are

Table 2
Transmembrane-protein structure databases

Database	Content	URL	References
OPM	Structures and orientation in the membrane	http://opm.phar.umich.edu	[124]
PDBTM	Structures	http://www.enzim.hu	[238]
TMPDB	Structures and topologies	http://bioinfo.si.hirosaki-u.ac.jp/~TMPDB/	[86]
MPtopo	Topologies	http://blanco.biomol.uci.edu/mptopo/	[89]
TMPinGS	TM proteins predicted in prokaryotic genomes	http://bioinfo.si.hirosaki-u.ac.jp/~TMPinGS/	[4]
THGS	TM helices predicted in genomes	http://pranag.physics.iisc.ernet.in/thgs/	[48]
TMBETA-GENOME	Beta-barrel proteins predicted in genomes	http://tmbeta-genome.cbrc.jp/annotation/	[66]

determination of the rotational orientation of helices [2,170]; analysis and prediction of interhelical contacts [75, 122,133,217,224,254] and buried and lipid-exposed faces of transmembrane helices [13,161]; assembling helices into a bundle [159]; identification of re-entrant helices [114,245], helices lying on the membrane surface [147], and completely buried helices [2]; refinement of low resolution, in particular, cryoelectron microscopy structures [41,50].

2.3. Beta-barrel transmembrane proteins

A similar, though separate, problem is identification of beta-barrel pore-forming transporters. It is a less populated area, although several programs exist for identification of such proteins, listed in [208]. In addition, several new programs have appeared recently [240,253]. These newer algorithms not only predict transmembrane segments, but attempt to model the geometry of the protein, e.g., the radius of the beta-barrel and the axis inclination. The comparison of algorithms based on machine learning techniques such as neural networks and support vector machines with hidden Markov models demonstrated that the latter produce better results, and the consensus predictions are better than each individual one [7], similar to the situation with alpha-structural proteins. A database of beta-barrel transmembrane proteins, TMBETA-GENOME has been created to facilitate development and benchmarking of the algorithms [66].

2.4. Classification and genomic distribution of transporters

We now turn from the discussion of generic transmembrane protein to transporters. An extensive research over the last several years has led to the description of numerous families (Table 3) and eventually to a universal classification of transporters based mainly on sequence similarity. The IUBMB-supported transport classification database TCDB [22] contains five major functional categories based on the transport mechanism (channels and pores; primary active transporters; secondary porters driven by electrochemical gradient; group translocators; transmembrane electron flow carriers; two additional categories are auxiliary proteins and functionally uncharacterized ones; see Table 1 for definitions) and more than four hundred families based on phylogenetic analysis [213,216].

The number of transporters depends on the genome size and ecological niche of an organism. Not only the number, but also the fraction of transporters in a genome increases with the genome size [243], and it seems to be caused by simultaneous increase of the number of transporter families represented in a genome and the average number of paralogs per family [185]. However, earlier studies on more limited sets of genomes reported an approximately linear trend for the number of membrane proteins and the logarithmic dependence of the number of families on the genome size [122]. Also, for one of the largest classes containing multiple families, ATP-binding cassette (ABC) transporters, the dependence of the number of candidate transporters on the genome size is approximately linear [69, 107]. One cause for this apparent discrepancy may be that the linear trend in the distribution of ABC-transporters is observed only after discounting several outliers, specifically, rhizobial alpha-proteobacteria with large genomes [69, 107,185]. Two other groups of organisms that rely mainly on ABC transporters are photosynthesizing cyanobacteria and organisms lacking the complete citrate cycle, mainly but not exclusively pathogens [185].

More generally, clustering of genomes by their transporter content produces groups that correlate with both phylogeny and lifestyle, e.g., groups of soil or plant-associated bacteria, or small obligate symbionts and pathogens. On the other hand, even closely related organisms may have markedly different transporter content, e.g., the free-living *Corynebacterium* species and pathogenic *C. diphtheriae* [185]. Similarly, families with related functions may have different genomic distribution, e.g., the MET and ISVH families of ABC transporters that, in particular, are involved in the iron uptake. The MET family transporters (TC 3.A.1.15) import a variety of divalent metal ions (Fe^{2+}, Mn^{2+}, Zn^{2+}) and are evenly distributed in bacteria of different lifestyles, whereas the ISVH transporters (TC 3.A.1.14) import metal-containing compounds such as iron siderophores, haemin, vitamin B_{12} and are more numerous in bacteria with relatively more diverse lifestyles [69].

Classification of enzymes and reconstruction of metabolic pathways from genomic data has led to the development of metabolic databases such as MetaCyc [23] (http://metacyc.org), KEGG [98] (http://www.genome.jp/kegg/) and BRENDA [221] (http://www.brenda-enzymes.info). In contrast to metabolic pathways, much less effort has been expended on reconstructions of transport reactions from genomic studies (see Table 4 for a list of microbial transporter databases). The TransportDB database collects known and predicted transport systems encoded in complete microbial

Table 3
Some large-scale studies and recent reviews of transporters families and their representation in genomes

Transporter family	Genomes	References
All families	many genomes	[165,166]
All families	many genomes	[211]
All families	many genomes	[185,187]
All families	Spirochetes	[214]
All families	*Mycobacterium* spp.	[232]
All families	selected Gram-positive bacteria	[125]
All families	*Bacillus subtilis*	[215]
Many families	many genomes	[210]
ABC	many genomes	[69]
ABC	many genomes	[84]
ABC	many genomes	[33]
ABC	many genomes	[220]
ABC	many genomes	[233]
ABC	*Mycobacterium tuberculosis*	[19]
ABC	*Escherichia coli*	[32]
ABC	*Escherichia coli*	[121]
ABC	*Burkholderia* spp.	[70]
ABC	*Bacillus subtilis*	[178]
ABC efflux	many genomes	[104]
Drug efflux systems	many genomes	[167]
ABC transporters of amino acids (2 subfamilies)	many genomes	[78]
Opp/Dpp family of ABC transporters involved in carbohydrate uptake	*Thermotoga maritima*	[29]
ABC and TRAP transporters involved in carbohydrate uptake	*Sinorhizobium meliloti*	[131]
ABC and other transporters involved in carbohydrate uptake	*Streptomyces coelicolor*	[12]
PTS	many genomes	[8]
PTS	*Escherichia coli*	[231]
PTS	*Bacillus subtilis*	[184]
PTS	*Mycoplasma genitalium*	[183]
TonB-dependent outer membrane receptors	*Xanthomonas* spp.	[14]
SbtA family of bicarbonate transporters	cyanobacteria	[251]
TTT (tripartite tricarboxylate transporter) family	many genomes	[259]
MOP (multidrug/oligosaccharydyl-lipid/polysaccharide) exporter superfamily (4 families)	many genomes	[83]
UT (urea transporter) family	many genomes	[135]
Ion transporter superfamily (12 families)	many genomes	[173]
TheR family of amino acid efflux transporters	many genomes	[266]
ENT (equilibrative nucleoside transporter) family	many genomes	[1]
DMT (drug/metabolite transporter) superfamily (14 families)	many genomes	[88]
TRAP (tripartite ATP-independent periplasmic) transporters	many genomes	[101]
PET (putative efflux transporter) family	many genomes	[71]
MFS (major facilitator superfamily)	many genomes	[209]
APC (amino acid/polyamine/organocation) superfamily (10 families)	many genomes	[87]
FATP (fatty acid transport protein) family	many genomes	[76]
LysE (L-lysine exporter) family	many genomes	[252]
MIP (major intrinsic protein) family	many genomes	[158]

genomes and annotated based on a series of experimental and bioinformatic evidences [186]. However, the larger fraction of potential transport systems are still annotated as hypothetical and need to be characterized.

2.5. Origin of (new) transporter genes

Uneven distribution of transporter families in bacteria reflects their fast adaptation to new environments with diverse chemical composition, and its main source seems to be horizontal gene transfer. Indeed, although the exact rates of horizontal transfer in different functional groups have not been compared (and it is not clear whether such comparison can be made in a meaningful manner), anecdotal evidence shows frequent lateral transfer of transporter genes and frequent non-orthologous gene displacements. For example, the gene cluster *yiaJKLMNOPQRS* responsible for the

Table 4
Transporter databases

Database	Content	URL	References
TCDB	All transporters	http://www.tcdb.org	[216]
TransportDB	All transporters	http://www.membranetransport.org	[186]
ABCdb	ABC transporters	http://www-abcdb.biotoul.fr/	[179]
ABSCISSE	ABC transporters	http://www.pasteur.fr/recherche/unites/pmtg/abc/database.html	[33]

L-xylulose catabolism in *Escherichia coli* contains genes *yiaMNO* encoding a binding protein-dependent secondary transporter (TRAP, tripartite ATP-independent periplasmic transporter), whereas homologs of the enzyme-encoding *yiaQRS* genes in other genomes are found in tentative operons with secondary transporters (*Klebsiella oxytoca*), ABC transporters (*Yersinia pestis*), and PTS systems (*Vibrio* spp., Firmicutes) [171]. Similarly, numerous discrepancies in the phylogenetic trees for the components of PTS systems and deviations of these trees from the species phylogeny demonstrate that gene shuffling plays an important role in the evolution of these systems [212,272].

An even more frequent type of events in the history of transporters is duplications (again, a necessary caveat is that it has not been measured accurately, so this statement is based mainly on our personal impressions from analysis of dozens of diverse metabolic systems). In many cases the transporters are duplicated within larger loci containing transporters, enzymes and regulators: such behavior is especially common among sugar catabolism systems (O.N. Laikova, personal communication). A disproportionally large family of ABC transporters in *Thermotoga maritima* (TC 3.A.1.5) was likely generated by lateral gene transfer from archaea and subsequent duplication and divergence events [29]. On the other hand, duplications within operons are more common for ATPase and permease components of ABC systems, yielding homologous subunits within the complex [105,233].

2.6. *Evolution of the domain structure of transporters*

Domain duplication, domain shuffling and gene fusions play a relatively smaller role in the evolution of transporters [123]. There are, however, important exceptions. A relatively rare occasion is the domain exchange between transporters and other types of proteins. Probably, the best known example is apparent co-optation of transporter components as the sensor domains of regulatory proteins, e.g., the periplasmic ligand-binding component of ABC transporters in transcription factors from the LacI family [145] or PTS-regulated domains IIA and IIB with intact phosphorylation sites in transcription factors [64] or antiterminators from the BglG/SacY family [11,54,63].

Within-transporter events are more frequent. Some ABC-transporters contain fused ATPase and permease domains in various combinations [105]. Of these, the permease fusion proteins are the most interesting. A relatively simple case is the one when the fused domains contain an even number of transmembrane helices and are inserted in the membrane in a conformation that is roughly symmetrical relative to an axis perpendicular to the membrane plane. However, there is a more interesting situation when domains with an odd number of helices have an arrangement with approximate two-fold symmetry relative to an axis lying in the membrane plane [158]. This means that in each pair of homologous loops, one loop occurs on the cytoplasmic side, and the other one, on the external side of the membrane. The fact that the topology of homologous domains may be completely different seems to show that it is impossible to predict the topology from sequence. In particular, this may contradict the “positive inside” rule. However, in such cases the bias for positive amino acids in cytoplasmic loops is rather weak [180,225]. Interestingly, when such homologous domains are encoded by different genes and thus form separate proteins, the lysine+arginine bias is quite pronounced and opposite in homologous loops, whereas when such permeases form homodymers, the bias is non-existent, as expected [180].

Somewhat more generally, the duplication seems to be one of the major mechanisms of the permease evolution [212,225]. At that, relatively short groups of transmembrane helices (three to six) not capable of forming a transporter structure by themselves, duplicate or even triplicate, yielding fully formed permeases; in some cases partial duplication [225] or emergence of a new transmembrane segment [212] may be observed. After formation of transmembrane permeases, capable of secondary transport, they start to interact with other proteins, yielding transporters of the ABC, TRAP and PTS types [36,212]; the latter, as mentioned above, may contain non-homologous components performing the same function.

3. Comparative genomics and gene annotation

3.1. Homology

All above analyses were based on homology, or, more exactly, sequence similarity. Indeed, the simplest way to annotate a new gene is to compare it with a database of already annotated genes, and if a sufficiently similar relative is found, to transfer its annotation to the gene in question. A problem with this simplistic description is in the words "sufficiently similar". There are no universal thresholds that would tell when the similarity is sufficient to infer the functional equivalence. Another problem is that the functional equivalence itself also is a somewhat fuzzy notion. Technically speaking, no two different proteins may be considered as absolutely equivalent, since their biochemical parameters, such as binding constants towards various ligands, will be different. More relevant is the cellular role, but it is harder to define formally.

A more robust way than simple use of the closest relatives is to construct phylogenetic trees. This allows one to distinguish between orthologs and paralogs (Table 1): the former may be assumed to have a common role as they basically are what is commonly called "the same gene in different genomes", whereas the latter are likely to assume different roles after the duplication event: otherwise they would be redundant. Moreover, genes with similar roles will cluster and form branches, and it becomes possible to assign a common role to all genes in a branch from the experimental data about some representatives. However, this also may be difficult: a branch may have no experimentally studied representatives, and even if it does, it is not always clear at what depth the branch may be considered as having a single common role. This is especially true for genes from large families with numerous paralogs (transporters, transcription factors, some classes of enzymes). The roles of even closely related (that is, sharing high sequence similarity) transporters may be different, as exemplified by a variety of metal transporters: for instance, the nickel and cobalt transporters often belong to the same gene families and in many cases cannot be differentiated on a phylogenetic tree [201]; this is not surprising, as many of those transporters are capable of both nickel and cobalt import, albeit with different affinity [39], see a more detailed discussion in Section 4.

3.2. Co-localization

An indirect way to assign a cellular role to an uncharacterized gene is to take into account its chromosomal localization [31,53,82,151,229,260]; reviewed in [142,148,153,203]. Indeed, the operon organization of prokaryote genomes has a consequence that functionally related genes occur in close proximity in many bacterial genomes [269]. Examples of such genes are enzymes from one pathway or enzymes utilizing a compound with importers of this compound. The suggested reasons for the existence of operons range from an obvious need to co-regulate genes in functional subsystems [72,174] to a disputed theory of selfish operons propagating by horizontal gene transfer [77,115,116,154, 174]. Of course, the operon structure is rather dynamic, not all genes from functional subsystems form operons, and sometimes operons contain functionally unrelated genes [175]. On the other hand, the eventual necessity of functional coherence of operons is underscored by the phenomenon of non-orthologous gene displacement *in situ* (Table 1), when a horizontally transferred gene occurs exactly in the same operon as the displaced one [146,262]; such displacements are especially frequent in carbohydrate catabolic operons (O.N. Laikova, personal communication). Finally, it should be noted that co-localization of functionally linked genes is not limited to operons. For example, co-regulation may be also provided by sharing a regulatory region by divergently transcribed genes or operons [255].

Whatever the reason, the consistent chromosomal co-localization is a strong indication of functional linkage used in the analysis of particular systems [148,260] (see also examples in the next section), large-scale analyses [19], and, recently, in automated systems of gene annotation [40,204,241].

An important issue here is the resolution of orthology relationships, which may be not trivial; otherwise the evolutionary signal may be overwhelmed by numerous positional linkages between members of large families of paralogs. A recently appeared complication arises from multiple strains with nearly identical gene order: in naïve implementations this artificially enhances the scores of gene clusters derived from such strains. To avoid this problem, various weighting procedures are employed, based on weighting the genomes according to the phylogenetic tree [271].

This approach was used for systematic analysis of ABC systems [105]. In many cases, genes encoding components of these systems form operons; however, in some cases they are encoded in different chromosomal regions. This leads to the appearance of "orphan" ATPase and permease genes that cannot be linked to each other by traditional

techniques. However, if such orphans, or, more exactly, their orthologs, are observed in close proximity (preferably, but not necessarily, in a candidate operon, that is, a string of genes transcribed in the same direction and separated by very short intergenic spacers) in related species, this provides the necessary linkage.

An extreme case of co-localization is gene fusion. This is a relatively rare occurrence in prokaryotic genomes (unlike eukaryotic ones), and also rare in transporters. The typical examples have been described in the previous section.

3.3. Co-regulation and co-expression

However, operons may not cover entire functional subsystems, say, metabolic pathways, that are still co-regulated. Thus, one more powerful method of linking together functionally related genes is the analysis of regulatory interactions. Again, it is based on a simple premise that if several genes are co-regulated by the same transcription factor or another regulatory mechanism (e.g., similar T-boxes or riboswitches, Table 1), they are likely to be functionally linked. In other words, the conjecture says that regulons are functionally meaningful. Indeed, transporters of catabolized compounds are often co-regulated with genes encoding the respective catabolic pathways (co-induced by the presence of the compound). In contrast, both transporter operons and operons encoding enzymes necessary for biosynthesis of a compound are switched on when the compound concentration drops below an acceptable level. Since in most cases the predictions of transcription factor binding sites are not very reliable, such analyses require careful manual re-examination of the results using comparative genomics [202]; see also examples presented in the next section.

An obvious consequence of co-regulation is co-expression that has became to be easily measurable in high-throughput experiments. Convenient centralized depositories of expression array data are available, such as GEO at NCBI [10] and ArrayExpress at EBI [162]. However, it should be noted that the expression data are rather noisy due to limitations of experimental techniques and because they usually do not distinguish between consequences of co-regulation and additional, distant correlations (modulation) such as major changes in the metabolic state of the cell [267]. Compared to simple two-conditions experiments, clustering of gene expression profiles across the entire spectrum of tested conditions using machine learning approaches offers a large potential to systematically predict transcriptional regulatory networks [46,58] and operons in prokaryotes [15,35,112,207].

3.4. Co-occurrence

One more indication that genes belong to the same functional system may be the fact that they appear in genomes together, that is, have the same phyletic (or phylogenetic) profiles (or patterns) (Table 1). This technique has been validated and benchmarked in a number of studies [43,93,102,168,226,261]. Again, an important technical issue is taking into account uneven density of sequenced genomes in different taxa [28]; in fact, selection of a correct set of genomes is essential for this type of analysis [93]. One may also consider phyletic profiles based on phenotypic properties [52,128] or regulatory motifs [198]. In the former case, a set of genes responsible for a given trait is selected, whereas in the latter one, the relevant regulator has been found. Again, resolution of orthology, which is especially difficult for large transporter families, is essential for this type of analysis.

It should be noted, however, that co-occurrence may be expected only if the genes are absolutely dependent upon each other for forming an actively functioning subsystem. In the context of this review, such correlated genes are substrate transporters and utilization pathways, e.g., sugar operons and sugar catabolism enzymes. However, the mosaic structure of such loci, the fact that the same transport role may be assumed by completely unrelated proteins, and difficulties with establishment of orthology obscure phylogenetic patterns. The situation with enzymes in many cases is less complicated, and several cases of missing enzymes or non-orthologous displacement were successfully analyzed by a combination of position analysis and phyletic distribution (e.g., [34,62,119,140]; reviewed in [108,142,148]). More exactly, for missing enzymes (that is, cases when a known biochemical activity is not represented by a single known gene) the positional and phyletic analyses work in parallel, reinforcing each other. For non-orthologous displacements the situation is more interesting [108,140,142]. The gene is assigned to a functional subsystem using positional clustering or other available evidence, whereas the fact that its phyletic distribution is exactly complementary to the distribution of another gene yields hypothesis that these genes perform the same function and any of them is sufficient (of course, the situation is rarely that equivocal).

Table 5
Comparative genomics resources for prokaryotes

Database	Content	URL	References
STRING	co-localization, fusions, co-expression, co-occurrence, text mining, high-throughput experiments	http://string.embl.de	[250]
IMG	co-localization, co-occurrence	http://img.jgi.doe.gov	[130]
SEED	co-localization, co-occurrence, annotation and metabolic reconstruction	http://theseed.uchicago.edu	[152]
Prolinks	co-localization, fusions, co-occurrence	http://dip.doe-mbi.ucla.edu/pronav	[17]
Phydbac	co-localization, fusions	http://igs-server.cnrs-mrs.fr/phydbac/	[40]

For transporters, a variation of this idea proved to be fruitful. Again, a transporter is initially assigned to a pathway using positional and regulatory analysis. Then its specificity is predicted based on the following reasoning: for a catabolic pathway we expect co-occurrence of transporters and enzymes (with the above caveat for non-orthologous displacement); for a biosynthetic pathway, a transporter of the end product or an intermediate compound may substitute for the complete pathway (respectively, the upstream part of the pathway). Thus, if we observe a genome with the transporter but no biosynthetic enzymes, we can more or less confidently predict that it is the transporter of the end product of the pathway. Some examples of such predictions are listed in the next section.

3.5. Combined evidence

As usual in bioinformatics, the best results are obtained when diverse types of evidence are combined. In addition to research projects where it was demonstrated that a combination of co-localization, co-expression, and co-occurrence allows one to reproduce physical and/or functional interactions between proteins (e.g., [265]), there are several resources that provide both pre-computed databases and tools for the analysis of new genomes (Table 5). The most popular of these resources seems to be STRING [250]. The SEED genomic platform includes a number of tools for comparative genomic analysis, including metabolic subsystem encoding and genome context analysis (chromosomal clustering and phylogenetic profiling) [152].

4. Examples

4.1. Vitamin and carbohydrate transporters

Comparative genomics techniques provided for substantial progress in functional annotation of hypothetical transporter genes. A collection of 36 solute uptake transporters, whose substrate specificities were tentatively predicted by genomic analysis, is presented in Table 6. It should be noted that usually such predictions do not automatically apply to an entire family, but only to certain groups of orthologous proteins whose specificity may be deduced by various types of genomic evidence. However, in some cases the family seems to contain only transporters with one assigned (and/or tested) specificity, for instance, the BioY family of biotin transporters [74], or the YpaA/RibU family of riboflavin transporters [247].

Surprisingly, one of the major sources of specificity annotations is the relatively less popular analysis of regulatory motifs (see the 'R' evidence in Table 6). It was used to assign specificity to candidate uptake transporters for amino acids arginine ArgP, lysine LysW, methionine MetT, and glycine GlyP in *S. oneidensis* (and, of course, their orthologs in other species), and for vitamins niacin YceI/NiaP, riboflavin YpaA/RibU, biotin BioY, and thiamin YuaJ in *Bacillus subtilis* (again, with orthologs). All these predictions were based on co-regulation with the respective amino acid or vitamin biosynthetic genes by a specific metabolite-responsive riboswitch or a transcription factor. A distinctive feature of these transporters is that they rarely co-localize with the biosynthetic genes and because of that their positional analysis was not sufficient.

For instance, the functional role of the BioY family of hypothetical transmembrane proteins in biotin uptake was tentatively suggested based on three combined types of evidence: transcriptional co-regulation, co-occurrence and co-localization with the biotin biosynthesis genes [192]. The *bioY* genes in prokaryotes are co-regulated with biotin biosynthesis genes by at least three different regulators, the widespread biotin repressor BirA [192], and two

Table 6
Specificities of solute uptake transporters predicted by the comparative genomic analysis

Name[*#]	Substrates	Family[1]	Example[2]	Phylogenetic distribution[3]	Evidence[4]	Reference[5]
YicJ	α-xylosides	GPH	YicJ_Ecoli	Ent	R (XylR), C (*xylL*)	*[111]*
YagG	β-xylosides	GPH	YagG_Ecoli	Ent	R (XylR), C (*xynB*)	*[111]*
TogMNAB[#]	oligogalacturonides	ABC	ECA2403	Ent	R (KdgR), C (*pelW*)	*[191,196]*; **[79]**
NagP[*#]	N-acetylglucosamine	MFS	SO3503	Alt, Xan	R (NagR), C (*nagABK*)	*[264]*
RhiT[*#]	rhamnogalacturonides	MFS	ECA3560	Ent	R (RhaS, KdgR), C (*rhiN*)	*[80,196]*
RhiABC[*#]	rhamnogalacturonides	ABC	STY3817	Ent, BCl	R (RhaS, KdgR), C (*rha*)	*[80,196]*
GlyT[*]	D-glycerate	GntP	SO1771	Alt, Vib, Pse,	R (SdaR), C (*garK*)	*U*
ScrT[*]	sucrose	FucP	Sfri_3989	Alt	R (ScrR), C (*scrPK*)	*U*
NanP[*]	sialic acid	MFS	VF0668	Ent, Vib	R (NanX), C (*nanAKE*)	*U*
LysW[*#]	lysine	NhaC	SO1007	BCl, Alt, Vib, Pas	R (*LYS*)	*[194]*
TyrT[*#]	tyrosine	NhaC	EF0402	BCl	R (T-box)	*[157,197]*
MetT[*#]	methionine	NhaC	SO1087	BCl, Alt, Vib	R (MetJ, S-box)	*[197]*
MetNIQ	methionine	ABC	YusCBA_Bacsu	BCl, Ent, Vib, Pas	R (S-box, T-box, MetJ)	**[197], [270]**
GlyP[*]	glycine	AgcS	SO0858, Spy1270	Alt, Vib, Pas, BCl	R (*GLY*)	*[129]*
SteT	threonine	APC, LAT	YkbA_Bacsu	BCl	R (T-box), C (*tdcB*)	*[246]*, **[182]**
TrpP	tryptophan	YhaG	YhaG_Bacsu	BCl, Arc	R (TRAP, T-box)	*[157,218]*, **[263]**
TrpXYZ[*]	tryptophan	ABC	Spy1016	BCl	R (T-box), C (*kynU*), O	*[157]*
ArgP[*]	arginine	PF00860	SO1245	Alt, Vib	R (ArgR)	*U*
YqiXYZ	arginine	ABC	YqiX_Bacsu	BCl	R (ArgR, T-box)	*[126]*; **[223]**
YvsH	lysine	APC	YvsH_Bacsu	Bacillales	R (*LYS*)	*[194]*
LysXY[*]	lysine	ABC	SPy0277	Lactobacillales	R (*LYS*), O	*[194]*
BioY[#]	biotin (vitamin H)	BUT	YuiG_Bacsu	BCl, α, Arc, Act, Cya	R (BirA, BioR, BioQ), C, O	*[192,199]*; **[68,74]**
PanP[*]	pantothenate (vitamin B_5)	COG4684	Spy1223	BCl	C (*dfp*), O	*U*
NiaP[*]	niacin (vitamin B_3)	MFS	YceI_Bacsu	BCl, Act, TM, α, β, γ	R (NiaR, NrtR, NadR)	*[189]*
NiaT[*]	niacin (vitamin B_3)	–	Spy1425	BCl	R (NiaR), O	*[190]*
YuaJ[#]	thiamin (vitamin B_1)	COG3859	YuaJ_Bacsu	BCl	R (*THI*), O	*[193]*
PnuT[*#]	thiamin (vitamin B_1)	PnuC	SO2713	α, β, γ, ε, CFB	R (*THI*), C (*tnr3*)	*[193]*
YpaA/RibU[#]	riboflavin (vitamin B_2)	RFT	YpaA_Bacsu	BCl	R (*RFN*), O	*[60,247]*; **[21,248]**
PnuX[*]/RibM[#]	riboflavin (vitamin B_2)	PnuC	SCO1442	Act	R (*RFN*), C (*ribABEH*)	*[247]*
PnuN[*#]	deoxynucleotides	PnuC	EF0739	Lactobacillales	R (NrdR), C (*dgk*), O	*[198]*
ThiXYZ[*#]	hydroxymethylpyrimidine	ABC	HI0354-357	BCl, α, Pas, Vib, TM	R (*THI*), C (*thiMDE*), O	*[90,193]*
CytX[*#]	hydroxymethylpyrimidine	NCS1	NMB2067	BCl, β, Pas, Pse, Arc	R (*THI*), C (*thiMDE*), O	*[193]*
YicE	xanthine	NCS2	YicE_Ecoli	Ent, Pas, Vib	R (PurR)	*[136,181]*; **[99]**

Table 6 (*continued*)

Name[*#]	Substrates	Family[1]	Example[2]	Phylogenetic distribution[3]	Evidence[4]	Reference[5]
CbiMNQO[#]	cobalt	ABC	SCO5961	α, γ, Act, BCl, Cya, Arc	R (*B12*), C (*cbi*)	*[195]*; **[201]**
NikMNQO[#]	nickel	ABC	SCO3159	α, γ, δ, Act, BCl, Cya, Arc	R (NikR), C	**[201]**
HoxN[#]	cobalt/nickel	NiCoT	SA2489	γ, β, Act, BCl, Arc	R (NikR, *B12*), C	*[195]*; **[73]**

[*] Tentatively suggested transporter names are marked by asterisks.

[#] Transporters discussed in the text.

[1] Transport protein families are given according to the transport classification system of the TransportDB database. For uncharacterized protein families, PFAM or COG identification numbers are given.

[2] Standard gene/protein identificators are either from Genbank or Swiss-Prot. Genome abbreviations are: Ecoli, *Escherichia coli*, SO, *Shewanella oneidensis*, Bacsu, *Bacillus subtilis*; SCO, *Streptomyces coelicolor*; EF, *Enterococcus faecalis*; SA, *Staphylococcus aureus*; Spy, *Streptococcus pyogenes*; ECA, *Erwinia carotovora*; VF, *Vibrio vischeri*; HI, *Haemophilus influenzae*; BME, *Brucella melitensis*; NMB, *Neisseria meningitidis*.

[3] Abbreviations of taxonomic groups of microorganisms: α, β, γ, δ, and ε correspond to α-, β-, γ-, δ-, and ε-proteobacteria; Ent, Enterobacteriales; Alt, Altermonadales; Xan, Xanthomonadales; Vib, Vibrionales; Pse, Pseudomonadales; Pas, Pasteurellales; BCl, *Bacillus/Clostridium* group; Act, Actinobacteria; Arc, Archaea; CFB, *Chlorobium/Bacteroides*; Cya, Cyanobacteria; TM, Thermotogales.

[4] Genome context evidences are: R, co-regulation by a conserved regulatory motif which is either a candidate transcription factor-binding site or a metabolite-sensing RNA structural element (the name of the transcription factor or the RNA element is given in parenthesis); C, conserved gene clustering with functionally related metabolic genes (gene abbreviation is given in parenthesis); O, co-occurrence profile (see the text for explanation).

[5] References for the papers describing predictions and validations are in *italic* and **bold**, respectively. "*U*" stands for unpublished results.

specialized transcription factors from different families, BioR in α-proteobacteria [199] and BioQ in Actinobacteria [202]. The predicted biotin specificity of BioY transporters was experimentally confirmed by gene complementation in *Rhizobium etli* [68] and biotin uptake assays in *Rhodobacter capsulatus* [74].

The riboflavin (vitamin B_2) transporter YpaA was first identified by comparative analysis of *RFN* regulatory elements (FMN-specific riboswitches) that co-regulate *ypaA* with the riboflavin synthesis *rib* genes [60]. Analysis of co-occurrence of the riboflavin genes in Gram-positive bacteria provided with an additional evidence: *ypaA* compensates for the absence of *rib* genes in *Streptococcus pyogenes* [247]. The predicted function of YpaA in *B. subtilis* and its ortholog RibU in *Lactococcus lactis* was confirmed by direct measurements of riboflavin uptake in the wild-type and knock-out strains [21,109,248].

The gene neighborhood technique was extensively used to assign specificity to various carbohydrate uptake transporters (see the 'C' evidence in Table 6) that are in most cases co-localized with sugar catabolism genes and form extensive carbohydrate utilization gene clusters [9,29,59]. Combining co-localization and co-regulation evidence is an even more powerful approach used to identify candidate transporters for oligogalacturonides TogMNAB and rhamnogalacturonides RhiT [191,196], N-acetylglucosamine NagP [264], glycerate GlyT, sucrose ScrT, and sialic acids NanP (Fig. 1(A)). The predicted oligogalacturonide uptake system TogMNAB from *Erwinia chrysanthemi* was later shown to provide *E. coli* with the ability to transport pectic oligomers [79].

Similarly, analysis of regulation and positional clustering was used to predict specificity of various metal ion transporters [155,156,200,201]. For example, genes encoding nickel transporters are regulated by the NikR repressor and/or co-localized with genes encoding nickel-dependent enzymes, whereas homologous cobalt transporters are regulated by the cobalamin riboswitch and co-localized with cobalamin biosynthesis genes (see below and Fig. 2(B)).

Though in most of these cases the evidence was sufficient to link a transporter to a metabolic pathway, the exact specificity could not be established without further analysis. With sugar and oligosaccharide transporters, the specificity could be predicted based on the metabolic context: for instance, if the transporter gene cluster contains genes for cytosolic sugar hydrolases, it is likely that the transporter imports oligosaccharides. For example, many members of Opp/Dpp ABC transporter family of *Thermotoga maritima* presumably involved in oligosaccharide transport are co-localized with intracellular oligosaccharide hydrolases [29]. The oligogalacturonate transporter TogMNAB and intracellular oligogalacturonate lyase PelW are encoded by the same KdgR-regulated operon in enterobacteria (Fig. 1) [196].

4.2. Thiamin-related transporters

For vitamin transporters (also for amino acid transporters), the specificity towards the end product or an intermediate compound could be assigned by the analysis of phyletic patterns. For example, comparative analysis of the thiamin regulatory *THI* elements (TPP riboswitches) resulted in identification of numerous candidate thiamin-related transporters, including YuaJ, PnuT, ThiXYZ and CytX (see Fig. 1(B)) [193]. Before this analysis, the ABC cassette ThiBPQ was the only thiamin transporter characterized in bacteria. Analysis of the distribution of genes involved in the thiamin biosynthesis pathway in bacteria has revealed that *yuaJ* is the only thiamin-regulated gene in several *Streptococcus* species that have no thiamin biosynthesis genes (e.g., *S. pyogenes*), suggesting that it is involved in the thiamin uptake. Two other hypothetical thiamin-regulated transporters, *thiXYZ* and *cytX*, were never found in the genomes without thiamin biosynthesis genes, but sometimes they occur in the genomes with an incomplete thiamin biosynthesis pathway lacking the hydroxymethylpyrimidine (HMP) synthesis gene *thiC*. Other metabolic pathways for the HMP synthesis are not known. In many genomes, the *thiXYZ* genes are clustered with the HMP kinase gene *thiD*. Based on these observations it was proposed that ThiXYZ and CytX are involved in the uptake of HMP, a metabolic precursor of thiamin [193]. Additional supporting evidence came from the analysis of distant protein similarities. The ThiY proteins constitute putative substrate-binding components of ABC transporters and are predicted to have an N-terminal transmembrane segment. The C-terminal soluble domain of ThiY appears to be weakly similar to the HMP biosynthesis enzyme Thi5 from yeasts [258], suggesting that ThiY is able to bind HMP analogs. Another predicted HMP transporter, CytX, belongs to the NCS1 family of nucleobase transporters. Recently one of these predictions was confirmed in an experimental study that demonstrated that ThiXYZ is involved in the HMP salvage pathway and ThiY binds an analog of this thiamin precursor, formyl aminopyrimidine [90].

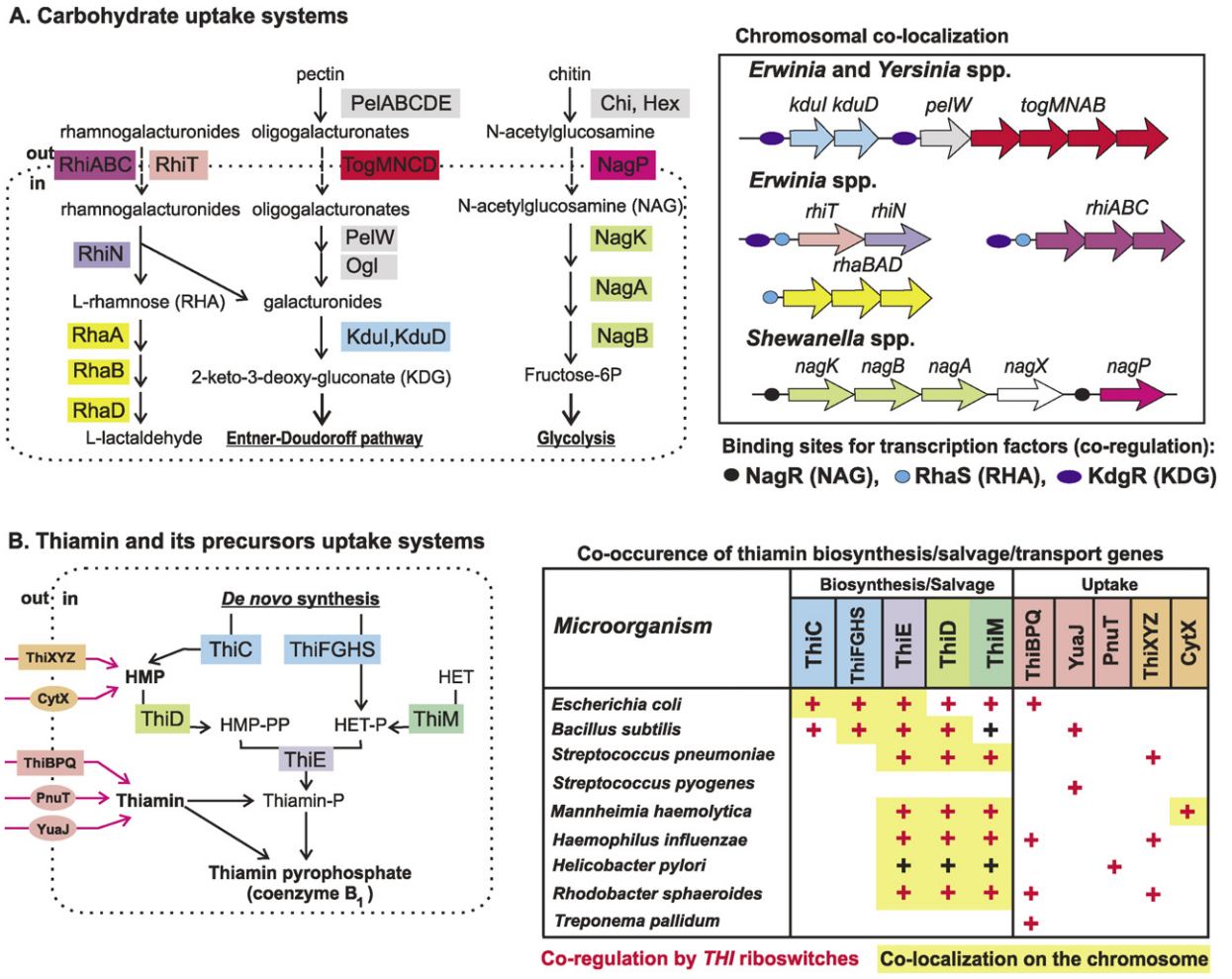

Fig. 1. Genome context techniques for prediction of transporter specificities in prokaryotes. (A) Carbohydrate utilization subsystems. Left, metabolic pathways for catabolism of pectin, chitin, and rhamnogalacturonides. Right, chromosomal co-localization and co-regulation of genes involved in these three metabolic pathways. Genes are shown by arrows with colors according to the functional role on the metabolic map. (B) Thiamin synthesis subsystem. Left, metabolic pathway for thiamin biosynthesis and uptake. Right, table of gene co-occurrence with indication of gene co-regulation by *THI* riboswitches (red font) and/or clustering on the chromosome (yellow background). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.3. NhaC family

As mentioned above, the real power of the genomic techniques lies in the analysis of families containing members of different, difficult to resolve specificities. One such example is the NhaC Na^+:H^+ antiporter superfamily (Fig. 3). Two members of this superfamily have been experimentally characterized in *B. subtilis*: the Na^+:H^+ antiporter NhaC [172] and the Na^+-lactate/$2H^+$-malate antiporter MleN [256]. Comparative analysis of regulation of lysine biosynthesis and transport genes has identified a set of candidate lysine transporters from the NhaC superfamily that have been named LysW [194]. The lysine-responsive *LYS* riboswitches are present upstream of the *lysW* genes in the *Shewanella* and *Vibrio* species, the Pasteurellaceae and the *Bacillus/Clostridium* group. Another group of NhaC-like transporters (named MetT) were found to be regulated by S-adenosylmethionine (SAM) via the MetJ repressor in γ-proteobacteria and SAM riboswitch in the *Bacillus/Clostridium* group. The SAM regulons include genes for the methionine biosynthesis and transport, and thus MetT was tentatively annotated as a methionine transporter [197]. Finally, the analysis of amino acid T-box regulons identified a group of candidate tyrosine transporters in the NhaC superfamily (named TyrT) that are controlled by the tyrosine-specific T-box antitermination system (A.G. Vitreschak et al., submitted).

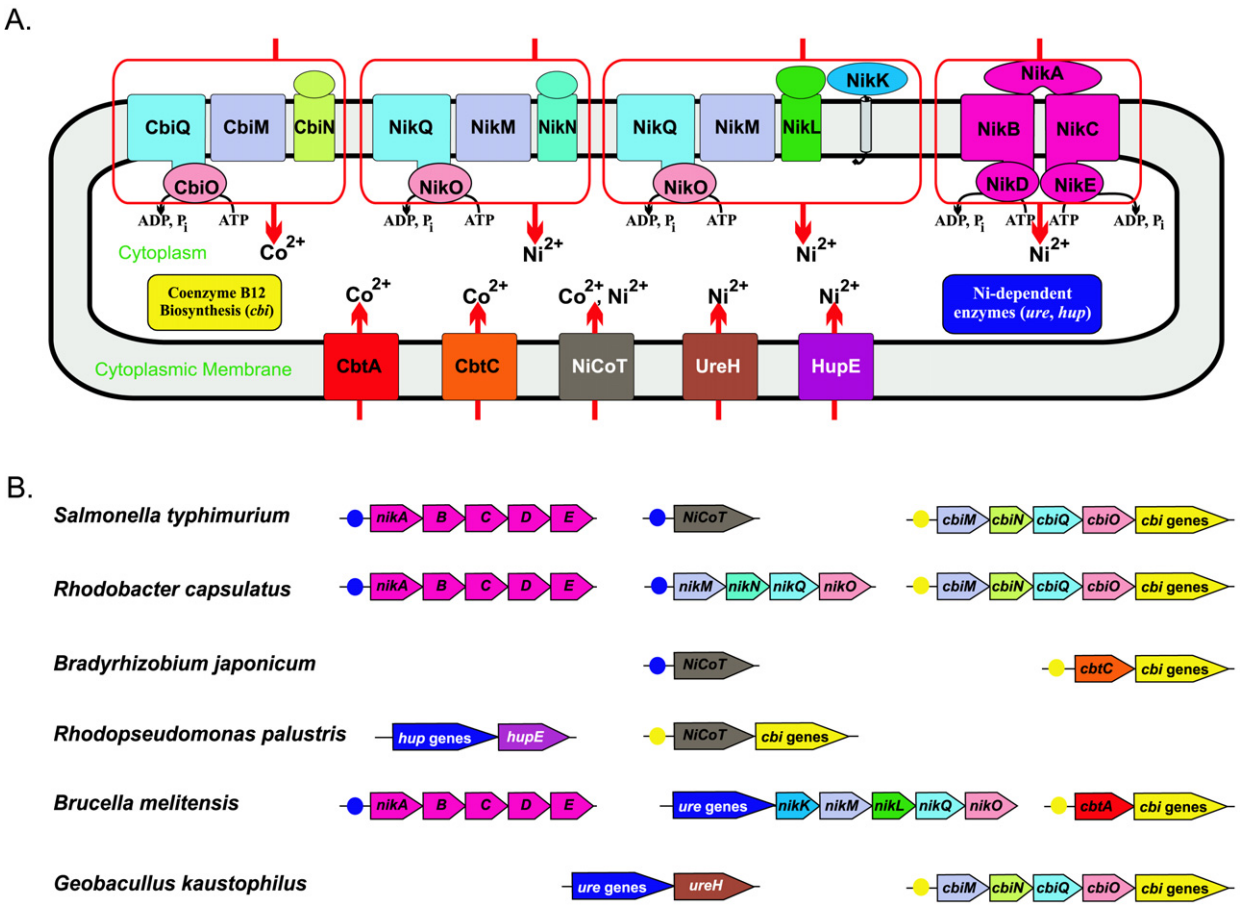

Fig. 2. Nickel and cobalt transport systems in prokaryotes. (A) properties, specificities and topology predictions. Components from different protein families are shown by different colors. (B) genome context analysis. Genes are shown by arrows with colors corresponding to the protein families. Regulatory elements, NikR-binding sites and *B12* riboswitches, are shown by dark blue and yellow dots, respectively. Positional clustering with Ni-dependent enzymes and B_{12} biosynthesis genes is shown by dark blue and yellow arrows. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.4. *PnuC family*

The PnuC family is one more example of a transporter family with high variability of specificities predicted by the comparative genomic analysis (Fig. 4). The only two experimentally characterized representatives of this family are the nicotinamide riboside permeases PnuC that are involved in the utilization of this metabolic precursor of NAD in *Haemophilus influenzae* [219] and *Salmonella enterica* [67]. In the enterobacteria, the *pnuC* gene forms an operon with the NAD biosynthesis gene *nadA*, and this operon is regulated by the NAD repressor NadR [61]. Analysis of other NAD-related regulons revealed several *pnuC*-like genes preceded by candidate binding sites of transcription factors NiaR in the *Streptococcus* spp., NrtR in *Hahella chejuensis*, and PnuR in the *Vibrio* spp. [189,190]. In addition, positional clustering of various *pnuC*-like genes with genes involved in NAD salvage was observed, e.g., in the *Pseudomonas* spp. This allows for projecting the nicotinamide riboside permease function to these genes (shown in orange in Fig. 4).

Though it was assumed that homologs of PnuC in general represent nicotinamide riboside-specific transporters [67, 219], other studies suggest a different functional role for other members of the PnuC family. A large group of *pnuC*-like transporters (called *pnuT*) was found in thiamin regulons under control of the *THI* riboswitch in the Proteobacteria and the Bacteroidetes species (shown in blue in Fig. 4). These *pnuT* genes are located in conserved gene clusters also encoding outer membrane TonB-dependent transporters (OMPs) and predicted kinases from various families (e.g.,

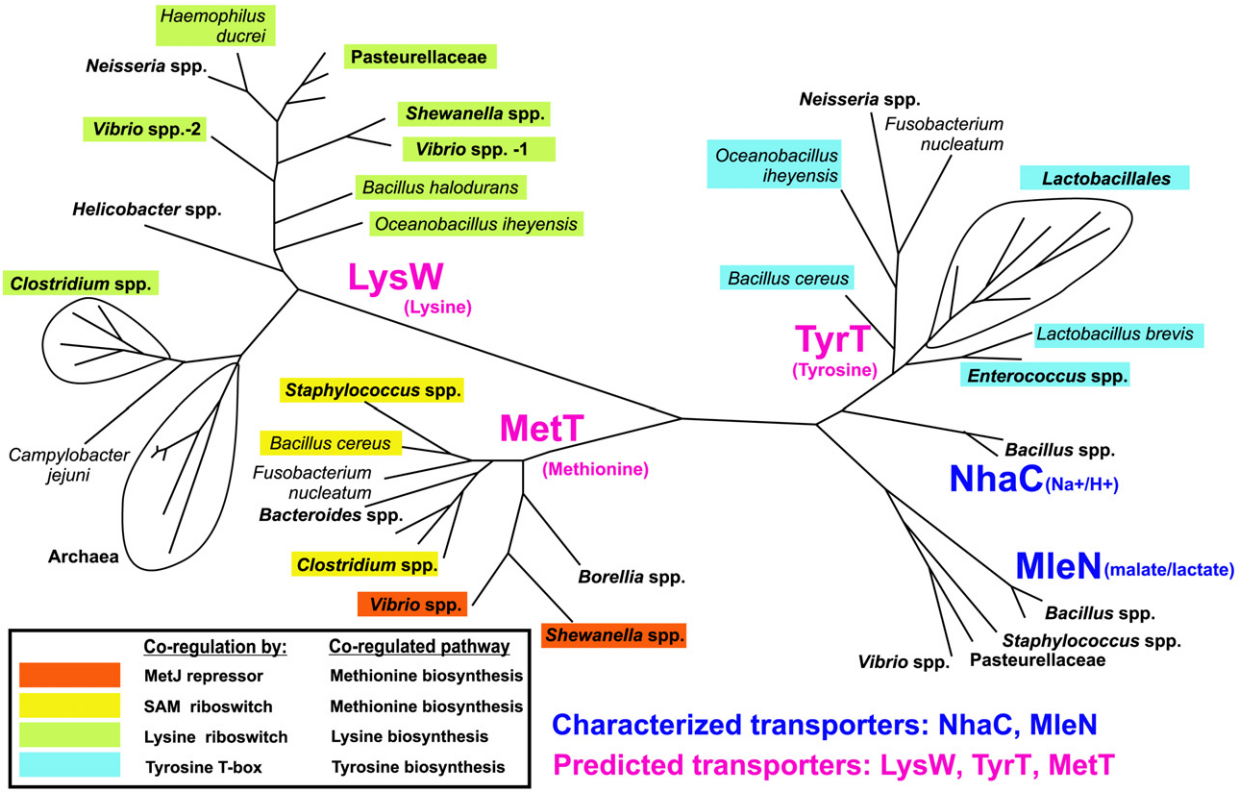

Fig. 3. Functional annotation of transporters from the NhaC family based on identification of amino acid-specific regulatory elements. Maximum likelihood phylogenetic tree of the NhaC family proteins was adopted from [197]. Background color blocks show amino acid-specific regulatory elements identified for the candidate methionine (MetT), lysine (LysW) and tyrosine (TyrT) transporters. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

a homolog of the thiamin pyrophosphokinase TNR3 from yeast). Based on these data, it was proposed that these hypothetical *THI*-regulated operons could be involved in the thiamin uptake across the outer and inner membranes (by OMP and PnuT, respectively) and its subsequent phosphorylation to form thiamin pyrophosphate coenzyme [193]. The predicted thiamin transporters PnuT are distributed among several different branches on the phylogenetic tree of the PnuC family (Fig. 4). Interestingly, one branch contains both candidate transporters PnuT and PnuC suggesting that specificity of transporters in this family is flexible in evolution.

One more branch on the PnuC family tree contains the predicted riboflavin transporters PnuX. This tentative annotation was based on the observation of *pnuX* genes preceded by *RFN* regulatory elements in the Actinobacteria [247] (shown in magenta in Fig. 4). In three Actinobacteria species, the *pnuX* genes are located in the riboflavin biosynthesis *rib* operons (Fig. 5), providing additional evidence for the predicted functional role of PnuX. Recently, the function of PnuX from *Corynebacterium glutamicum* was studied in experiment: as predicted, the transport by PnuX (re-named RibM) was not energy-dependent and had high affinity for riboflavin [248].

Finally, a group of PnuC-like transporters in the *Lactobacillus* spp. (called PnuN) is co-regulated by the NrdR repressor with ribonuclotide reductases (Nrd) and a deoxynucleoside kinase (Dgk) that are involved in the dNTP synthesis [198] (shown in green in Fig. 4). The *pnuN* and *dgk* genes in *Enterococcus faecalis* form a candidate NrdR-controlled operon. This yields tentative reconstruction of the deoxyribonucleoside salvage pathway, which involves transport and subsequent phosphorylation of deoxyribonucleosides.

Overall the PnuC-like transporters are involved in the uptake of unphosphorylated metabolites (nicotinamide riboside, thiamin, riboflavin, and deoxynucleosides) that are then phosphorylated in the cytoplasm by respective kinases, thus allowing to direct the substrate flow inside and to prevent efflux (Fig. 6). Indeed, in *H. influenzae*, the PnuC-mediated substrate flow across the membrane is coupled to the rate of nicotinamide riboside phosphorylation [134].

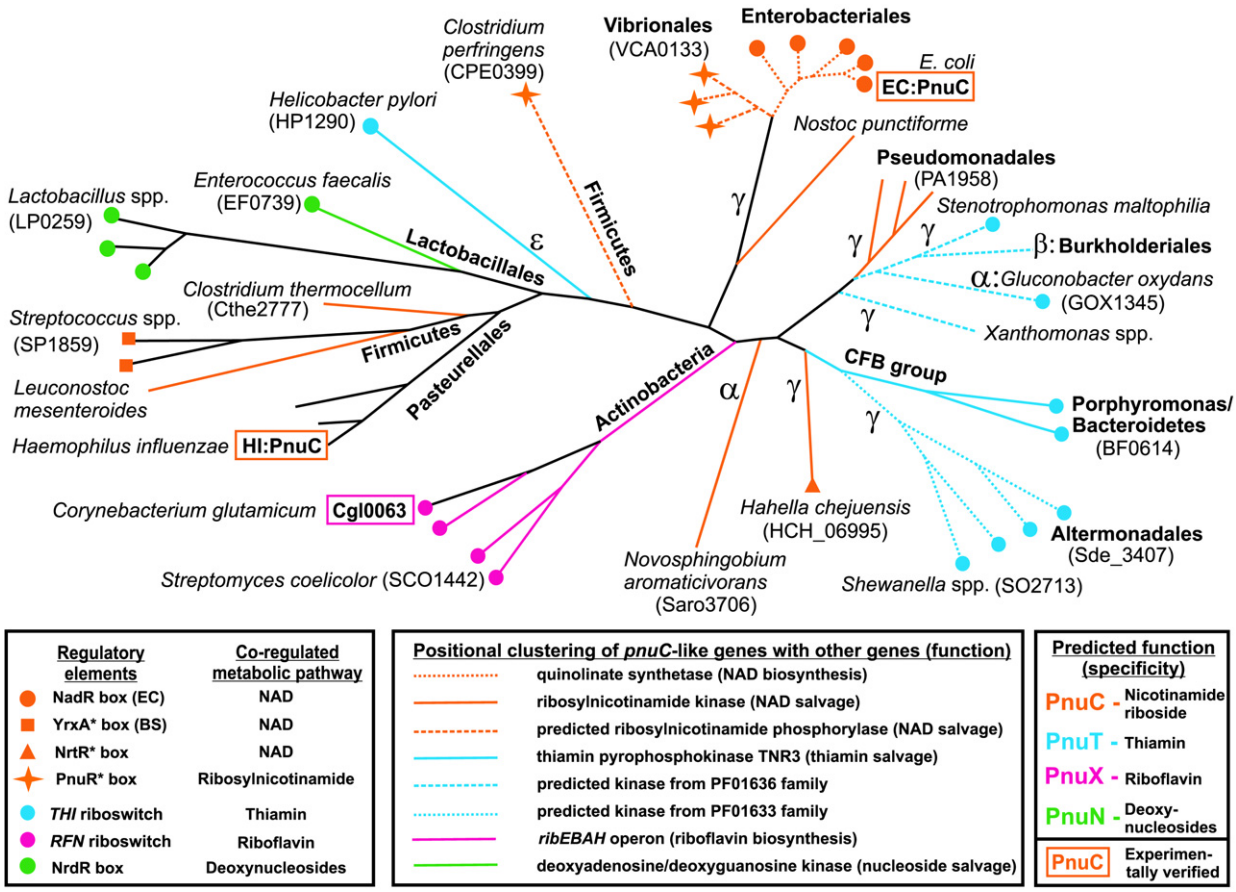

Fig. 4. Functional annotation of transporters from the PnuC family based on genome context analysis. Maximum likelihood phylogenetic tree was constructed by the Phylip package [47] and colored by predicted transporter specificity based on gene co-localization and co-regulation (see insets for details). Experimentally verified transporters are boxed. Taxonomic groups are indicated in bold. Genomic identificators for representatives from each group are given in parentheses. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Even with these examples, some might argue that the genomic analysis provides only a limited extension of existing knowledge: characterization of new specificity in an existing family or, at best, description of a new family. Our last example demonstrates that sometimes it is possible to predict new transport mechanisms using only genomic analysis.

4.5. *Nickel and cobalt uptake transporters*

Transition metals nickel (Ni) and cobalt (Co) are essential components of many metalloenzymes [106,141]. The most common Ni-dependent enzymes are urease, [NiFe] hydrogenease, and CO dehydrogenase. In the form of coenzyme B_{12}, cobalt plays a number of crucial roles in many biological functions (e.g., synthesis of methionine and deoxyribonucleotides). Synthesis of [Ni] enzymes and coenzyme B_{12} requires high-affinity uptake of metal ions from natural environments where they usually are available only in trace amounts [38]. The nickel and cobalt uptake in bacteria is mediated by various secondary and primary ABC transporters (Fig. 2(A)) [39].

Search for B_{12}-specific regulatory elements (*B12* riboswitches) and candidate binding sites of the nickel repressor NikR accompanied by the analysis of co-localization with Ni-dependent and B_{12} biosynthetic genes was used to assign specificities to a large number of candidate nickel and cobalt transporters from previously characterized families of metal transporters, namely NiCoT, UreH, HupE, NikABCDE (Fig. 2(B)) [201]. Secondary transporters from the NiCoT family are capable of both nickel and cobalt uptake, or prefer only nickel ions. The NiCoT transporters are widespread among bacteria and found in some archaea and fungi. The metal ion preferences of six NiCoT trans-

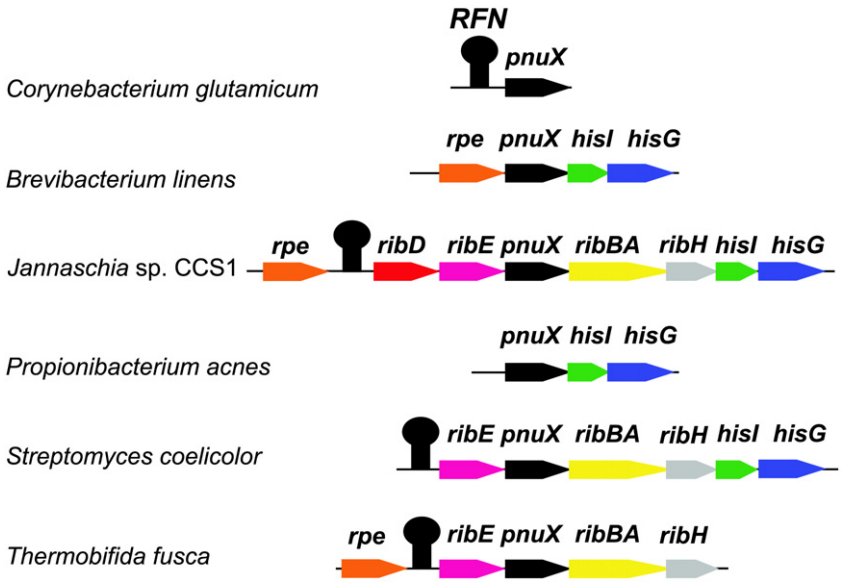

Fig. 5. Identification of riboflavin transporter PnuX in Actinobacteria. Genes and *RFN* regulatory elements are shown by arrows of various colors and by black hairpins, respectively. Functional roles of clustered genes: *ribBA*, *ribD*, *ribE*, *ribH* are involved in the riboflavin biosynthesis; *hisI* and *hisG* are from the histidine biosynthesis, and *rpe* is involved in the pentose phosphate pathway. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

porters studied in metal accumulation assays correlate with the genomic context of the respective genes [73]. The Ni-preferring NiCoT transporters are either located adjacent to genes encoding Ni-dependent enzymes or regulated by the NikR repressor. The Co-preferring NiCoT transporters are located within the B_{12} biosynthesis gene clusters that are preceded by *B12* riboswitch elements [195,201].

Secondary transporters from the UreH family are Ni-specific and are often clustered with either urease or [Ni] supeoxide dismutase [201]. Secondary transporters from the HupE/UreJ family are widespread among bacteria and encoded within [NiFe] hydrogenase and urease gene clusters. Most of them are Ni-specific transporters; however, in cyanobacteria, the *hupE* genes are preceded by *B12* riboswitches, and thus their predicted cellular role is the cobalt transport [195,201].

High-affinity Ni-specific ABC transporter NikABCDE is present in many proteobacteria and is regulated by NikR. NikA is a periplasmic substrate-binding component, NikB and NikC are permease components, and NikD and NikE are ATPases. Since NikABCDE systems belong to the large and functionally diverse nickel/peptide/opine PepT family, it is quite difficult to annotate their homologs in species distant from proteobacteria. Analysis of regulatory elements (NikR sites or *B12* riboswitches) was used to predict nickel and cobalt specificities of these transporters [201]. Diverged branches of Ni-specific systems (Nik-2, Nik-3) were detected in methanogenic archaea and some proteobacteria.

The genome context analysis also resulted in identification of new types of Ni/Co transporters, including novel cobalt transporters CbtA and CbtC [195], and Ni/Co transport systems from a novel family of ABC transporters, CbiMNQO (Fig. 2) [201]. The latter transporters consist of three conserved components (integral membrane proteins CbiM/NikM and CbiQ/NikQ; and ATPase CbiO/NikO). The predicted cobalt transport systems have a small component (CbiN) with two transmembrane segments. The predicted nickel transport systems also have additional components, either NikN or NikL, whose topology is similar to that of CbiN but the sequence does not show any detectable similarity. The CbiN/NikN/NikL proteins could be involved in the metal binding. The presence of an ATPase subunit (CbiO/NikO) suggests that these systems are energized by the ATP hydrolysis. However, an unusual feature of all ABC transporters from the CbiMNQO family is that they lack a separate substrate-binding protein, an essential component of all known ABC uptake transporters in bacteria.

Experimental characterization of the CbiMNQO and NikMNQO transporters from *Salmonella typhimurium* and *Rhodobacter capsulatus* by metal accumulation assays confirmed the substrate preferences of these transporters, as initially predicted by genomic analyses [201]. On the other hand, the metal uptake results for the *S. typhimurium*

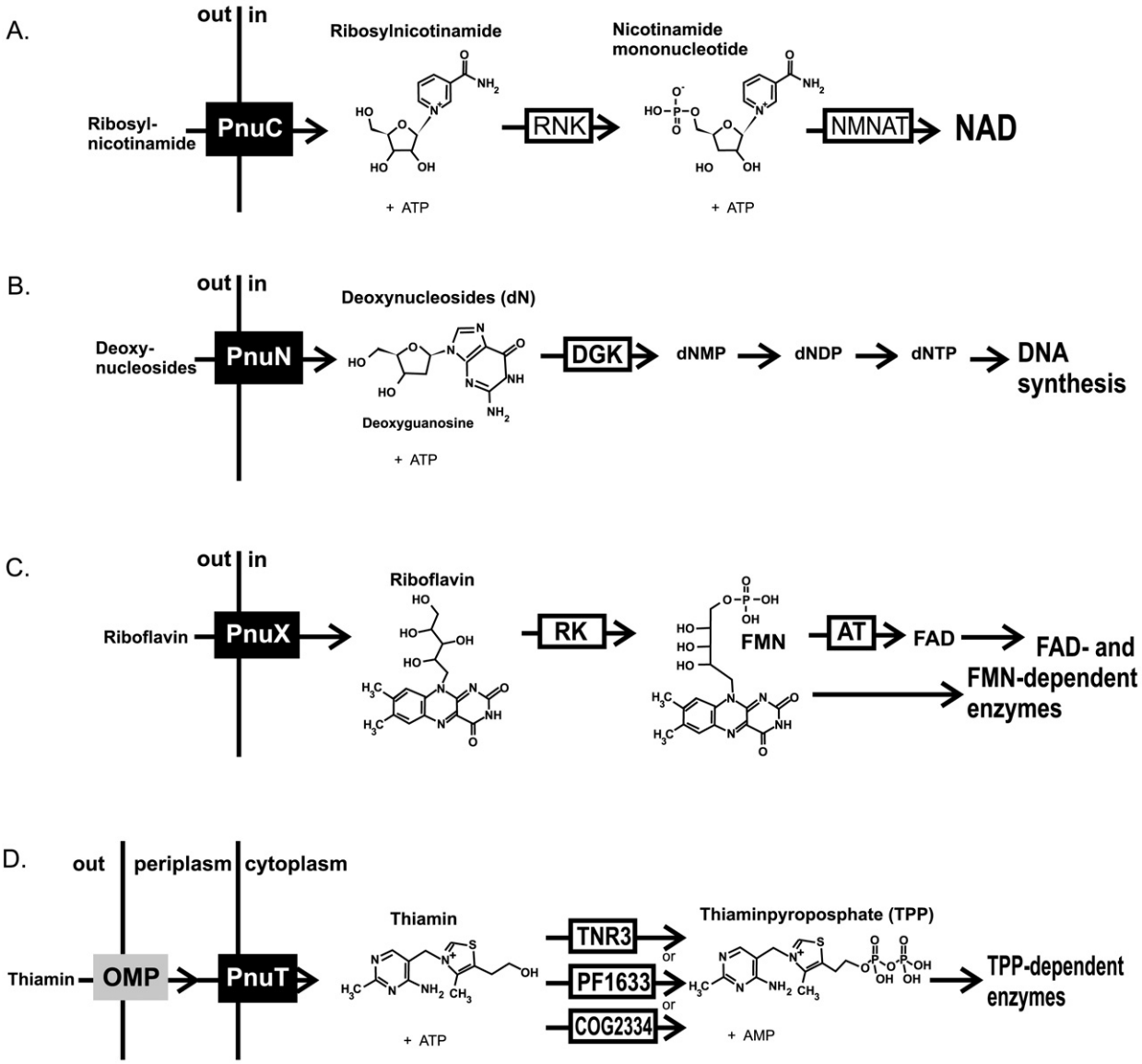

Fig. 6. Proposed salvage pathways including PnuC-like transporters. (A) Salvage of pyridine nucleosides in Enterobacteriales, Pseudomonadales, and Pasteurellales; (B) salvage of deoxynuclesides in Lactobacillales; (C) salvage of riboflavin in Actinobacteria; (D) salvage of thiamin in Proteobacteria and the Bacteroidetes. Inner membrane transporters from the PnuC familyu are shown by black boxes. Outer membrane transporter for thiamin is shown by grey box. Cytoplasmic enzymes that phosphorylate the substrate in the cell are shown by white boxes and include ribosylnicotinamide kinase (RNK), riboflavin kinase (RK), deoxynucleotide kinase (DGK); thiamin pyrophosphokinase (TNR3), and other predicted thiamin kinases from the putative choline kinase family PF01633, and the putative homoserine kinase family COG2334.

CbiMNQO cassette suggest that the transmembrane protein CbiQ and the ABC protein CbiO are not essential for function of CbiMN, which was shown to be the basic component of the cobalt transporter. Based on these data it was suggested that the CbiMNQO-like transport systems represent a mechanistically novel type of membrane transporters that are independent of solute-binding proteins and are energized by CbiQO-like ABC modules.

Similarity searches identified multiple *cbiQO*-like genes unevenly distributed in bacterial genomes. Most of these genes are co-localized with genes encoding unrelated hypothetical transmembrane proteins. The *bioM* and *bioN* genes that are clearly similar to *cbiQ* and *cbiO*, respectively, are located adjacent to the biotin transporter gene *bioY* in *Sinorhizobium meliloti* and have been reported to be required for the biotin uptake [42]. This, together with the observation of frequent co-localization of orthologous *bioMN* and *bioY* genes in prokaryotic genomes [192], suggested

that BioM, BioN and BioY may encode a tripartite, modular biotin-uptake system. Experimental characterization of the *bioYMN* cassette from *R. capsulatus* confirmed that both BioY alone and BioMNY complex are able to mediate biotin uptake albeit with different capacity and affinity [74]. In contrast to most other CbiQO-like systems that involve adjacent gene(s) for a transporter module, approximately two thirds of the *bioY* genes are not positionally linked to *bioMN* and, moreover, may occur in genomes lacking *bioMN*/*cbiQO* homologs. On the other hand, in many Firmicutes, the *cbiQOO* operons are adjacent to highly expressed genes encoding essential cellular functions (e.g., ribosomal proteins, RNA polymerease subunits and pseudouridylate synthase). There is a possibility that they serve as universal energizing components for diverse transporters found in these bacteria.

5. Conclusions

Examples presented in this review demonstrate that the bioinformatics and comparative genomic analysis of transporters is a powerful approach, allowing one to address important biological questions. Both the structure and the function may be predicted with reasonable reliability and at varying level of detail, dependent on the user's needs. The number and location of transmembrane segments and the protein topology (relative to the inner and outer side of the membrane) may be predicted using statistical analysis of the protein sequence. The simultaneous analysis of several related proteins increases the reliability of predictions.

An even more interesting area is the analysis of the transporter function. Unlike the structure prediction, the function prediction is still more art than technology, and no Internet tools are specifically devoted to this task. However, even application of more or less standard comparative genomic techniques often produces non-trivial results. The simplest situation arises when functional specificity is ascribed to a new, experimentally uncharacterized member of a known family based on co-localization, co-expression, and co-occurrence. The role of co-occurrence is somewhat more important in the transporter analysis compared to more common enzyme studies, since it allows one to determine precisely the transported compound as the entry point of the transporter in the associated metabolic pathway.

An important point here is the somewhat philosophical difference between the biochemical function and the cellular role. Many transporters are capable of importing several related compounds, albeit with different efficiency. Thus, the biochemical role of a transporter is best represented as a vector of efficiencies towards different compounds. On the other hand, the cellular role seems to be normally limited to one compound (or few related compounds) that is further metabolized by an associated pathway or is used immediately. For example, many transporters are importing both nickel and cobalt, but their role may be predicted by co-regulation and/or co-localization with nickel-dependent or cobalamin-biosynthesis enzymes. In cases when the experimental data are available, the predicted role coincides with the biochemical preference towards one of those ions, although the biochemical preference is not absolute. Notably, in some families thus predicted nickel and cobalt transporters seem to be intermixed in the phylogenetic tree, which indicates that their biochemical properties are evolving at a fast rate. The same seems to be the case with many sugar transporters.

However, the possibilities are not limited to the specificity analysis in already established families. New families of transporters may be characterized using the same comparative approaches, but in this case the analysis starts with structural characterization of the seed members: they are recognized as transporters based on the presence of multiple transmembrane segments. In this case, the comparative genomic analysis not only is necessary for the assignment of specificity, but provides with additional support for the initial annotation.

Finally, in at least one case the comparative genomics created a starting base for experimental studies of a new transport mechanism, where a permease capable of secondary transport may also use ATP hydrolysis to increase the affinity towards the main transported compound. This unusual ATP-dependent transport mechanism, initially predicted based on the analysis of phyletic patterns, is becoming an area of intense experimental studies.

Acknowledgements

We are grateful to Andrei Osterman and Thomas Eitinger for useful discussions, and to Maxim Frank-Kamenetskii for encouragement and patience. This study was partially supported by the Howard Hughes Medical Institute (grant 55005610 to M.G.) and the Russian Academy of Sciences (Program "Molecular and Cellular Biology").

References

[1] Acimovic Y, Coe IR. Molecular evolution of the equilibrative nucleoside transporter family: identification of novel family members in prokaryotes and eukaryotes. Mol Biol Evol 2002;19:2199–210.

[2] Adamian L, Liang J. Prediction of transmembrane helix orientation in polytopic membrane proteins. BMC Struct Biol 2006;6:13.

[3] Aloy P, Cedano J, Oliva B, Aviles FX, Querol E. 'TransMem': A neural network implemented in Excel spreadsheets for predicting transmembrane domains of proteins. Comput Appl Biosci 1997;13:231–4.

[4] Arai M, Ikeda M, Shimizu T. Comprehensive analysis of transmembrane topologies in prokaryotic genomes. Gene 2003;304:77–86.

[5] Arai M, Mitsuke H, Ikeda M, Xia JX, Kikuchi T, Satake M, et al. ConPred II: A consensus prediction method for obtaining transmembrane topology models with high reliability. Nucleic Acids Res 2004;32:W390–3.

[6] Arai M, Okumura K, Satake M, Shimizu T. Proteome-wide functional classification and identification of prokaryotic transmembrane proteins by transmembrane topology similarity comparison. Protein Sci 2004;13:2170–83.

[7] Bagos PG, Liakopoulos TD, Hamodrakas SJ. Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method. BMC Bioinformatics 2005;6:7.

[8] Barabote RD, Saier Jr MH. Comparative genomic analyses of the bacterial phosphotransferase system. Microbiol Mol Biol Rev 2005;69:608–34.

[9] Barrangou R, Azcarate-Peril MA, Duong T, Conners SB, Kelly RM, Klaenhammer TR. Global analysis of carbohydrate utilization by Lactobacillus acidophilus using cDNA microarrays. Proc Natl Acad Sci USA 2006;103:3816–21.

[10] Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, et al. NCBI GEO: Mining tens of millions of expression profiles—database and tools update. Nucleic Acids Res 2007;35:D760–5.

[11] Ben-Zeev E, Fux L, Amster-Choder O, Eisenstein M. Experimental and computational characterization of the dimerization of the PTS-regulation domains of BglG from Escherichia coli. J Mol Biol 2005;347:693–706.

[12] Bertram R, Schlicht M, Mahr K, Nothaft H, Saier Jr MH, Titgemeyer F. In silico and transcriptional analysis of carbohydrate uptake systems of Streptomyces coelicolor A3(2). J Bacteriol 2004;186:1362–73.

[13] Beuming T, Weinstein H. A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins. Bioinformatics 2004;20:1822–35.

[14] Blanvillain S, Meyer D, Boulanger A, Lautier M, Guynet C, Denance N, et al. Plant carbohydrate scavenging through tonb-dependent receptors: a feature shared by phytopathogenic and aquatic bacteria. PLoS ONE 2007;2:e224.

[15] Bockhorst J, Craven M, Page D, Shavlik J, Glasner J. A Bayesian network approach to operon prediction. Bioinformatics 2003;19:1227–35.

[16] Bostina M, Mohsin B, Kuhlbrandt W, Collinson I. Atomic model of the E. coli membrane-bound protein translocation complex SecYEG. J Mol Biol 2005;352:1035–43.

[17] Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, Eisenberg D. Prolinks: A database of protein functional linkages derived from coevolution. Genome Biol 2004;5:R35.

[18] Boyd D, Manoil C, Beckwith J. Determinants of membrane protein topology. Proc Natl Acad Sci USA 1987;84:8525–9.

[19] Braibant M, Gilot P, Content J. The ATP binding cassette (ABC) transport systems of Mycobacterium tuberculosis. FEMS Microbiol Rev 2000;24:449–67.

[20] Breyton C, Haase W, Rapoport TA, Kuhlbrandt W, Collinson I. Three-dimensional structure of the bacterial protein-translocation complex SecYEG. Nature 2002;418:662–5.

[21] Burgess CM, Slotboom DJ, Geertsma ER, Duurkens RH, Poolman B, van Sinderen D. The riboflavin transporter RibU in Lactococcus lactis: molecular characterization of gene expression and the transport mechanism. J Bacteriol 2006;188:2752–60.

[22] Busch W, Saier Jr MH. The IUBMB-endorsed transporter classification system. Methods Mol Biol 2003;227:21–36.

[23] Caspi R, Foerster H, Fulcher CA, Hopkinson R, Ingraham J, Kaipa P, et al. MetaCyc: A multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res 2006;34:D511–6.

[24] Chen CP, Kernytsky A, Rost B. Transmembrane helix predictions revisited. Protein Sci 2002;11:2774–91.

[25] Chen CP, Rost B. State-of-the-art in membrane protein prediction. Appl Bioinformatics 2002;1:21–35.

[26] Chen CP, Rost B. Long membrane helices and short loops predicted less accurately. Protein Sci 2002;11:2766–73.

[27] Claros MG, von Heijne G. TopPred II: An improved software for membrane protein structure predictions. Comput Appl Biosci 1994;10:685–6.

[28] Cokus S, Mizutani S, Pellegrini M. An improved method for identifying functionally linked proteins using phylogenetic profiles. BMC Bioinformatics 2007;8(Suppl 4):S7.

[29] Conners SB, Montero CI, Comfort DA, Shockley KR, Johnson MR, Chhabra SR, et al. An expression-driven approach to the prediction of carbohydrate transport and utilization regulons in the hyperthermophilic bacterium Thermotoga maritima. J Bacteriol 2005;187:7267–82.

[30] Cuthbertson JM, Doyle DA, Sansom MS. Transmembrane helix prediction: a comparative evaluation and analysis. Protein Eng Des Sel 2005;18:295–308.

[31] Dandekar T, Snel B, Huynen M, Bork P. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci 1998;23:324–8.

[32] Dassa E, Hofnung M, Paulsen IT, Saier Jr MH. The Escherichia coli ABC transporters: an update. Mol Microbiol 1999;32:887–9.

[33] Dassa E, Bouige P. The ABC of ABCS: a phylogenetic and functional classification of ABC systems in living organisms. Res Microbiol 2001;152:211–29.

[34] Daugherty M, Vonstein V, Overbeek R, Osterman A. Archaeal shikimate kinase, a new member of the GHMP-kinase family. J Bacteriol 2001;183:292–300.

[35] De Hoon MJ, Imoto S, Kobayashi K, Ogasawara N, Miyano S. Predicting the operon structure of Bacillus subtilis using operon length, intergene distance, and gene expression information. Pac Symp Biocomput 2004:276–87.
[36] Driessen AJ, Rosen BP, Konings WN. Diversity of transport mechanisms: common structural principles. Trends Biochem Sci 2000;25:397–401.
[37] Durbin R, Eddy S, Krogh A, Mitchison G. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press; 1999.
[38] Eitinger T, Mandrand-Berthelot MA. Nickel transport systems in microorganisms. Arch Microbiol 2000;173:1–9.
[39] Eitinger T, Suhr J, Moore L, Smith JA. Secondary transporters for nickel and cobalt ions: theme and variations. Biometals 2005;18:399–405.
[40] Enault F, Suhre K, Poirot O, Abergel C, Claverie JM. Phydbac2: Improved inference of gene function using interactive phylogenomic profiling and chromosomal location analysis. Nucleic Acids Res 2004;32:W336–9.
[41] Enosh A, Fleishman SJ, Ben-Tal N, Halperin D. Assigning transmembrane segments to helices in intermediate-resolution structures. Bioinformatics 2004;20(Suppl 1):i122–9.
[42] Entcheva P, Phillips DA, Streit WR. Functional analysis of Sinorhizobium meliloti genes involved in biotin synthesis and transport. Appl Environ Microbiol 2002;68:2843–8.
[43] Ettema T, van der Oost J, Huynen M. Modularity in the gain and loss of genes: applications for function prediction. Trends Genet 2001;17:485–7.
[44] Ettema TJ, de Vos WM, van der Oost J. Discovering novel biology by in silico archaeology. Nat Rev Microbiol 2005;3:859–69.
[45] Eyre TA, Partridge L, Thornton JM. Computational analysis of alpha-helical membrane protein structure: implications for the prediction of 3D structural models. Protein Eng Des Sel 2004;17:613–24.
[46] Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, et al. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol 2007;5:e8.
[47] Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 1981;17:368–76.
[48] Fernando SA, Selvarani P, Das S, Kumar ChK, Mondal S, Ramakumar S, et al. THGS: A web-based database of Transmembrane Helices in Genome Sequences. Nucleic Acids Res 2004;32:D125–8.
[49] Field D, Wilson G, van der Gast C. How do we compare hundreds of bacterial genomes? Curr Opin Microbiol 2006;9:499–504.
[50] Fleishman SJ, Harrington S, Friesner RA, Honig B, Ben-Tal N. An automatic method for predicting transmembrane protein structures using cryo-EM and evolutionary data. Biophys J 2004;87:3448–59.
[51] Fleishman SJ, Ben-Tal N. Progress in structure prediction of alpha-helical membrane proteins. Curr Opin Struct Biol 2006;16:496–504.
[52] Forterre P. A hot story from comparative genomics: reverse gyrase is the only hyperthermophile-specific protein. Trends Genet 2002;18:236–7.
[53] Fujibuchi W, Ogata H, Matsuda H, Kanehisa M. Automatic detection of conserved gene clusters in multiple genomes by graph comparison and P-quasi grouping. Nucleic Acids Res 2000;28:4029–36.
[54] Fux L, Nussbaum-Shochat A, Amster-Choder O. Interactions between the PTS regulation domains of the BglG transcriptional antiterminator from Escherichia coli. J Biol Chem 2003;278:46203–9.
[55] Gabaldon T, Huynen MA. Prediction of protein function and pathways in the genome era. Cell Mol Life Sci 2004;61:930–44.
[56] Galperin M. Conserved 'hypothetical' proteins: new hints and new puzzles. Comp Funct Genom 2001;2:14–8.
[57] Galperin MY, Koonin EV. 'Conserved hypothetical' proteins: prioritization of targets for experimental study. Nucleic Acids Res 2004;32:5452–63.
[58] Gardner TS, di Bernardo D, Lorenz D, Collins JJ. Inferring genetic networks and identifying compound mode of action via expression profiling. Science 2003;301:102–5.
[59] Gelfand M, Laikova O. In: Galperin M, Koonin E, editors. Prolegomena to the evolution of transcriptional regulation in bacterial genomes. Caister Academic Press; 2003.
[60] Gelfand MS, Mironov AA, Jomantas J, Kozlov YI, Perumov DA. A conserved RNA structure element involved in the regulation of bacterial riboflavin synthesis genes. Trends Genet 1999;15:439–42.
[61] Gerasimova AV, Gelfand MS. Evolution of the NadR regulon in Enterobacteriaceae. J Bioinform Comput Biol 2005;3:1007–19.
[62] Glazko GV, Mushegian AR. Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. Genome Biol 2004;5:R32.
[63] Gorke B. Regulation of the Escherichia coli antiterminator protein BglG by phosphorylation at multiple sites and evidence for transfer of phosphoryl groups between monomers. J Biol Chem 2003;278:46219–29.
[64] Greenberg DB, Stulke J, Saier Jr MH. Domain analysis of transcriptional regulators bearing PTS regulatory domains. Res Microbiol 2002;153:519–26.
[65] Grisshammer R, Tate CG. Overexpression of integral membrane proteins for structural studies. Q Rev Biophys 1995;28:315–422.
[66] Gromiha MM, Yabuki Y, Kundu S, Suharnan S, Suwa M. TMBETA-GENOME: Database for annotated beta-barrel membrane proteins in genomic sequences. Nucleic Acids Res 2007;35:D314–6.
[67] Grose JH, Bergthorsson U, Xu Y, Sterneckert J, Khodaverdian B, Roth JR. Assimilation of nicotinamide mononucleotide requires periplasmic AphA phosphatase in Salmonella enterica. J Bacteriol 2005;187:4521–30.
[68] Guillen-Navarro K, Araiza G, Garcia-de los Santos A, Mora Y, Dunn MF. The Rhizobium etli bioMNY operon is involved in biotin transport. FEMS Microbiol Lett 2005;250:209–19.
[69] Harland DN, Garmory HS, Brown KA, Titball RW. An association between ATP binding cassette systems, genome sizes and lifestyles of bacteria. Res Microbiol 2005;156:434–42.
[70] Harland DN, Dassa E, Titball RW, Brown KA, Atkins HS. ATP-binding cassette systems in Burkholderia pseudomallei and Burkholderia mallei. BMC Genomics 2007;8:83.

[71] Harley KT, Saier Jr MH. A novel ubiquitous family of putative efflux transporters. J Mol Microbiol Biotechnol 2000;2:195–8.
[72] Hazkani-Covo E, Graur D. Evolutionary conservation of bacterial operons: does transcriptional connectivity matter? Genetica 2005;124:145–66.
[73] Hebbeln P, Eitinger T. Heterologous production and characterization of bacterial nickel/cobalt permeases. FEMS Microbiol Lett 2004;230:129–35.
[74] Hebbeln P, Rodionov DA, Alfandega A, Eitinger T. Biotin uptake in prokaryotes by solute transporters with an optional ATP-binding cassette-containing module. Proc Natl Acad Sci USA 2007;104:2909–14.
[75] Hildebrand PW, Lorenzen S, Goede A, Preissner R. Analysis and prediction of helix–helix interactions in membrane channels and transporters. Proteins 2006;64:253–62.
[76] Hirsch D, Stahl A, Lodish HF. A family of fatty acid transporters conserved from mycobacterium to man. Proc Natl Acad Sci USA 1998;95:8625–9.
[77] Homma K, Fukuchi S, Nakamura Y, Gojobori T, Nishikawa K. Gene cluster analysis method identifies horizontally transferred genes with high reliability and indicates that they provide the main mechanism of operon gain in 8 species of gamma-Proteobacteria. Mol Biol Evol 2007;24:805–13.
[78] Hosie AH, Poole PS. Bacterial ABC transporters of amino acids. Res Microbiol 2001;152:259–70.
[79] Hugouvieux-Cotte-Pattat N, Reverchon S. Two transporters, TogT and TogMNAB, are responsible for oligogalacturonide uptake in Erwinia chrysanthemi 3937. Mol Microbiol 2001;41:1125–32.
[80] Hugouvieux-Cotte-Pattat N. The RhaS activator controls the Erwinia chrysanthemi 3937 genes rhiN, rhiT and rhiE involved in rhamnogalacturonan catabolism. Mol Microbiol 2004;51:1361–74.
[81] Hurwitz N, Pellegrini-Calace M, Jones DT. Towards genome-scale structure prediction for transmembrane proteins. Philos Trans R Soc Lond B Biol Sci 2006;361:465–75.
[82] Huynen M, Snel B, Lathe III W, Bork P. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res 2000;10:1204–10.
[83] Hvorup RN, Winnen B, Chang AB, Jiang Y, Zhou XF, Saier Jr MH. The multidrug/oligosaccharidyl-lipid/polysaccharide (MOP) exporter superfamily. Eur J Biochem 2003;270:799–813.
[84] Igarashi Y, Aoki KF, Mamitsuka H, Kuma K, Kanehisa M. The evolutionary repertoires of the eukaryotic-type ABC transporters in terms of the phylogeny of ATP-binding domains in eukaryotes and prokaryotes. Mol Biol Evol 2004;21:2149–60.
[85] Ikeda M, Arai M, Lao DM, Shimizu T. Transmembrane topology prediction methods: a re-assessment and improvement by a consensus method using a dataset of experimentally-characterized transmembrane topologies. In Silico Biol 2002;2:19–33.
[86] Ikeda M, Arai M, Okuno T, Shimizu T. TMPDB: A database of experimentally-characterized transmembrane topologies. Nucleic Acids Res 2003;31:406–9.
[87] Jack DL, Paulsen IT, Saier MH. The amino acid/polyamine/organocation (APC) superfamily of transporters specific for amino acids, polyamines and organocations. Microbiology 2000;146(Pt 8):1797–814.
[88] Jack DL, Yang NM, Saier Jr MH. The drug/metabolite transporter superfamily. Eur J Biochem 2001;268:3620–39.
[89] Jayasinghe S, Hristova K, White SH. MPtopo: A database of membrane protein topology. Protein Sci 2001;10:455–8.
[90] Jenkins AH, Schyns G, Potot S, Sun G, Begley TP. A new thiamin salvage pathway. Nat Chem Biol 2007;3:492–7.
[91] Jones DT, Taylor WR, Thornton JM. A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry 1994;33:3038–49.
[92] Jones DT. Do transmembrane protein superfolds exist? FEBS Lett 1998;423:281–5.
[93] Jothi R, Przytycka TM, Aravind L. Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment. BMC Bioinformatics 2007;8:173.
[94] Kall L, Sonnhammer EL. Reliability of transmembrane predictions in whole-genome data. FEBS Lett 2002;532:415–8.
[95] Kall L, Krogh A, Sonnhammer EL. A combined transmembrane topology and signal peptide prediction method. J Mol Biol 2004;338:1027–36.
[96] Kall L, Krogh A, Sonnhammer EL. An HMM posterior decoder for sequence feature prediction that includes homology information. Bioinformatics 2005;21(Suppl 1):i251–7.
[97] Kall L, Krogh A, Sonnhammer EL. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. Nucleic Acids Res 2007;35:W429–32.
[98] Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, et al. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 2006;34:D354–7.
[99] Karatza P, Frillingos S. Cloning and functional characterization of two bacterial members of the NAT/NCS2 family in Escherichia coli. Mol Membr Biol 2005;22:251–61.
[100] Karp PD. Call for an enzyme genomics initiative. Genome Biol 2004;5:401.
[101] Kelly DJ, Thomas GH. The tripartite ATP-independent periplasmic (TRAP) transporters of bacteria and archaea. FEMS Microbiol Rev 2001;25:405–24.
[102] Kensche PR, van Noort V, Dutilh BE, Huynen MA. Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. J R Soc Interface 2007.
[103] Kernytsky A, Rost B. Static benchmarking of membrane helix predictions. Nucleic Acids Res 2003;31:3642–4.
[104] Khwaja M, Ma Q, Saier Jr MH. Topological analysis of integral membrane constituents of prokaryotic ABC efflux systems. Res Microbiol 2005;156:270–7.
[105] Kihara D, Kanehisa M. Tandem clusters of membrane proteins in complete genome sequences. Genome Res 2000;10:731–43.
[106] Kobayashi M, Shimizu S. Cobalt proteins. Eur J Biochem 1999;261:1–9.

[107] Konstantinidis KT, Tiedje JM. Trends between gene content and genome size in prokaryotic species with larger genomes. Proc Natl Acad Sci USA 2004;101:3160–5.
[108] Koonin E, Galperin M. Sequence–Evolution–Function. Computational approaches in comparative genomics. Kluwer Academic Publishers; 2003.
[109] Kreneva RA, Gel'fand MS, Mironov AA, Iomantas IA, Kozlov II, Mironov AS, et al. Study of the phenotypic occurrence of ura gene inactivation in Bacillus subtilis. Genetika 2000;36:1166–8.
[110] Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 2001;305:567–80.
[111] Laikova ON, Mironov AA, Gelfand MS. Computational analysis of the transcriptional regulation of pentose utilization systems in the gamma subdivision of Proteobacteria. FEMS Microbiol Lett 2001;205:315–22.
[112] Laing E, Mersinias V, Smith CP, Hubbard SJ. Analysis of gene expression in operons of Streptomyces coelicolor. Genome Biol 2006;7:R46.
[113] Lao DM, Arai M, Ikeda M, Shimizu T. The presence of signal peptide significantly affects transmembrane topology prediction. Bioinformatics 2002;18:1562–6.
[114] Lasso G, Antoniw JF, Mullins JG. A combinatorial pattern discovery approach for the prediction of membrane dipping (re-entrant) loops. Bioinformatics 2006;22:e290–7.
[115] Lawrence J. Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. Curr Opin Genet Dev 1999;9:642–8.
[116] Lawrence JG. Selfish operons and speciation by gene transfer. Trends Microbiol 1997;5:355–9.
[117] Lehnert U, Xia Y, Royce TE, Goh CS, Liu Y, Senes A, et al. Computational analysis of membrane proteins: genomic occurrence, structure prediction and helix interactions. Q Rev Biophys 2004;37:121–46.
[118] Lespinet O, Labedan B. Orphan enzymes? Science 2005;307:42.
[119] Li H, Pellegrini M, Eisenberg D. Detection of parallel functional modules by comparative analysis of genome sequences. Nat Biotechnol 2005;23:253–60.
[120] Liakopoulos TD, Pasquier C, Hamodrakas SJ. A novel tool for the prediction of transmembrane protein topology based on a statistical analysis of the SwissProt database: the OrienTM algorithm. Protein Eng 2001;14:387–90.
[121] Linton KJ, Higgins CF. The Escherichia coli ATP-binding cassette (ABC) proteins. Mol Microbiol 1998;28:5–13.
[122] Liu Y, Engelman DM, Gerstein M. Genomic analysis of membrane protein families: abundance and conserved motifs. Genome Biol 2002;3. research0054.
[123] Liu Y, Gerstein M, Engelman DM. Transmembrane protein domains rarely use covalent domain recombination as an evolutionary mechanism. Proc Natl Acad Sci USA 2004;101:3495–7.
[124] Lomize MA, Lomize AL, Pogozheva ID, Mosberg HI. OPM: Orientations of proteins in membranes database. Bioinformatics 2006;22:623–5.
[125] Lorca GL, Barabote RD, Zlotopolski V, Tran C, Winnen B, Hvorup RN, et al. Transport capabilities of eleven gram-positive bacteria: comparative genomic analyses. Biochim Biophys Acta 2007;1768:1342–66.
[126] Makarova KS, Mironov AA, Gelfand MS. Conservation of the binding site for the arginine repressor in all bacterial lineages. Genome Biol 2001;2. RESEARCH0013.
[127] Makarova KS, Koonin EV. Comparative genomics of Archaea: how much have we learned in six years, and what's next? Genome Biol 2003;4:115.
[128] Makarova KS, Wolf YI, Koonin EV. Potential genomic determinants of hyperthermophily. Trends Genet 2003;19:172–6.
[129] Mandal M, Lee M, Barrick JE, Weinberg Z, Emilsson GM, Ruzzo WL, et al. A glycine-dependent riboswitch that uses cooperative binding to control gene expression. Science 2004;306:275–9.
[130] Markowitz VM, Korzeniewski F, Palaniappan K, Szeto E, Werner G, Padki A, et al. The integrated microbial genomes (IMG) system. Nucleic Acids Res 2006;34:D344–8.
[131] Mauchline TH, Fowler JE, East AK, Sartor AL, Zaheer R, Hosie AH, et al. Mapping the Sinorhizobium meliloti 1021 solute-binding protein-dependent transportome. Proc Natl Acad Sci USA 2006;103:17933–8.
[132] Melen K, Krogh A, von Heijne G. Reliability measures for membrane protein topology prediction algorithms. J Mol Biol 2003;327:735–44.
[133] Melnyk RA, Kim S, Curran AR, Engelman DM, Bowie JU, Deber CM. The affinity of GXXXG motifs in transmembrane helix–helix interactions is modulated by long-range communication. J Biol Chem 2004;279:16591–7.
[134] Merdanovic M, Sauer E, Reidl J. Coupling of NAD+ biosynthesis and nicotinamide ribosyl transport: characterization of NadR ribonucleotide kinase mutants of Haemophilus influenzae. J Bacteriol 2005;187:4410–20.
[135] Minocha R, Studley K, Saier Jr MH. The urea transporter (UT) family: bioinformatic analyses leading to structural, functional, and evolutionary predictions. Receptors Channels 2003;9:345–52.
[136] Mironov AA, Koonin EV, Roytberg MA, Gelfand MS. Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. Nucleic Acids Res 1999;27:2981–9.
[137] Mitaku S, Hirokawa T, Tsuji T. Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces. Bioinformatics 2002;18:608–16.
[138] Mitra K, Schaffitzel C, Shaikh T, Tama F, Jenni S, Brooks III CL, et al. Structure of the E. coli protein-conducting channel bound to a translating ribosome. Nature 2005;438:318–24.
[139] Moller S, Croning MD, Apweiler R. Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics 2001;17:646–53.
[140] Morett E, Korbel JO, Rajan E, Saab-Rincon G, Olvera L, Olvera M, et al. Systematic discovery of analogous enzymes in thiamin biosynthesis. Nat Biotechnol 2003;21:790–5.
[141] Mulrooney SB, Hausinger RP. Nickel uptake and utilization by microorganisms. FEMS Microbiol Rev 2003;27:239–61.

[142] Mushegian A. Foundations of Comparative Genomics. Elsevier Academic Press; 2007.
[143] Nakai K, Horton P. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci 1999;24:34–6.
[144] Nakashima H, Nishikawa K. The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins. FEBS Lett 1992;303:141–6.
[145] Nguyen CC, Saier Jr MH. Phylogenetic, structural and functional analyses of the LacI-GalR family of bacterial transcription factors. FEBS Lett 1995;377:98–102.
[146] Omelchenko MV, Makarova KS, Wolf YI, Rogozin IB, Koonin EV. Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. Genome Biol 2003;4:R55.
[147] Orgel JP. Sequence context and modified hydrophobic moment plots help identify 'horizontal' surface helices in transmembrane protein structure prediction. J Struct Biol 2004;148:51–65.
[148] Osterman A, Overbeek R. Missing genes in metabolic pathways: a comparative genomics approach. Curr Opin Chem Biol 2003;7:238–51.
[149] Osterman AL, Begley TP. A subsystems-based approach to the identification of drug targets in bacterial pathogens. Prog Drug Res 2007;64:131. 3–70.
[150] Ott CM, Lingappa VR. Integral membrane protein biosynthesis: why topology is hard to predict. J Cell Sci 2002;115:2003–9.
[151] Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci USA 1999;96:2896–901.
[152] Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 2005;33:5691–702.
[153] Overbeek R, Bartels D, Vonstein V, Meyer F. Annotation of bacterial and archaeal genomes: improving accuracy and consistency. Chem Rev 2007;107:3431–47.
[154] Pal C, Hurst LD. Evidence against the selfish operon theory. Trends Genet 2004;20:232–4.
[155] Panina EM, Mironov AA, Gelfand MS. Comparative analysis of FUR regulons in gamma-proteobacteria. Nucleic Acids Res 2001;29:5195–206.
[156] Panina EM, Mironov AA, Gelfand MS. Comparative genomics of bacterial zinc regulons: enhanced ion transport, pathogenesis, and rearrangement of ribosomal proteins. Proc Natl Acad Sci USA 2003;100:9912–7.
[157] Panina EM, Vitreschak AG, Mironov AA, Gelfand MS. Regulation of biosynthesis and transport of aromatic amino acids in low-GC Gram-positive bacteria. FEMS Microbiol Lett 2003;222:211–20.
[158] Park JH, Saier Jr MH. Phylogenetic characterization of the MIP family of transmembrane channel proteins. J Membr Biol 1996;153:171–80.
[159] Park Y, Helms V. Assembly of transmembrane helices of simple polytopic membrane proteins from sequence conservation patterns. Proteins 2006;64:895–905.
[160] Park Y, Helms V. How strongly do sequence conservation patterns and empirical scales correlate with exposure patterns of transmembrane helices of membrane proteins? Biopolymers 2006;83:389–99.
[161] Park Y, Helms V. On the derivation of propensity scales for predicting exposed transmembrane residues of helical membrane proteins. Bioinformatics 2007;23:701–8.
[162] Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, et al. Array Express—a public database of microarray experiments and gene expression profiles. Nucleic Acids Res 2007;35:D747–50.
[163] Parodi LA, Granatir CA, Maggiora GM. A consensus procedure for predicting the location of alpha-helical transmembrane segments in proteins. Comput Appl Biosci 1994;10:527–35.
[164] Pasquier C, Promponas VJ, Palaios GA, Hamodrakas JS, Hamodrakas SJ. A novel method for predicting transmembrane segments in proteins based on a statistical analysis of the SwissProt database: the PRED-TMR algorithm. Protein Eng 1999;12:381–5.
[165] Paulsen IT, Sliwinski MK, Saier Jr MH. Microbial genome analyses: global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities. J Mol Biol 1998;277:573–92.
[166] Paulsen IT, Nguyen L, Sliwinski MK, Rabus R, Saier Jr MH. Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. J Mol Biol 2000;301:75–100.
[167] Paulsen IT, Chen J, Nelson KE, Saier Jr MH. Comparative genomics of microbial drug efflux systems. J Mol Microbiol Biotechnol 2001;3:145–50.
[168] Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA 1999;96:4285–8.
[169] Persson B, Argos P. Prediction of transmembrane segments in proteins utilising multiple sequence alignments. J Mol Biol 1994;237:182–92.
[170] Pilpel Y, Ben-Tal N, Lancet D. kPROT: A knowledge-based scale for the propensity of residue orientation in transmembrane segments. Application to membrane protein structure prediction. J Mol Biol 1999;294:921–35.
[171] Plantinga TH, van der Does C, Driessen AJ. Transporter's evolution and carbohydrate metabolic clusters. Trends Microbiol 2004;12:4–7.
[172] Pragai Z, Eschevins C, Bron S, Harwood CR. Bacillus subtilis NhaC, an Na^+/H^+ antiporter, influences expression of the phoPR operon and production of alkaline phosphatases. J Bacteriol 2001;183:2505–15.
[173] Prakash S, Cooper G, Singhi S, Saier Jr MH. The ion transporter superfamily. Biochim Biophys Acta 2003;1618:79–92.
[174] Price MN, Huang KH, Arkin AP, Alm EJ. Operon formation is driven by co-regulation and not by horizontal gene transfer. Genome Res 2005;15:809–19.
[175] Price MN, Arkin AP, Alm EJ. The life-cycle of operons. PLoS Genet 2006;2:e96.
[176] Promponas VJ, Palaios GA, Pasquier CM, Hamodrakas JS, Hamodrakas SJ. CoPreTHi: A Web tool which combines transmembrane protein segment prediction methods. In Silico Biol 1999;1:159–62.
[177] Punta M, Forrest LR, Bigelow H, Kernytsky A, Liu J, Rost B. Membrane protein prediction methods. Methods 2007;41:460–74.

[178] Quentin Y, Fichant G, Denizot F. Inventory, assembly and analysis of Bacillus subtilis ABC transport systems. J Mol Biol 1999;287:467–84.
[179] Quentin Y, Fichant G. ABCdb: an ABC transporter database. J Mol Microbiol Biotechnol 2000;2:501–4.
[180] Rapp M, Seppala S, Granseth E, von Heijne G. Emulating membrane protein evolution by rational design. Science 2007;315:1282–4.
[181] Ravcheev DA, Gel'fand MS, Mironov AA, Rakhmaninova AB. Purine regulon of gamma-proteobacteria: a detailed description. Genetika 2002;38:1203–14.
[182] Reig N, del Rio C, Casagrande F, Ratera M, Gelpi JL, Torrents D, et al. Functional and structural characterization of the first prokaryotic member of the L-amino acid transporter (LAT) family: a model for APC transporters. J Biol Chem 2007;282:13270–81.
[183] Reizer J, Reizer A, Merrick MJ, Plunkett III G, Rose DJ, Saier Jr MH. Novel phosphotransferase-encoding genes revealed by analysis of the Escherichia coli genome: a chimeric gene encoding an Enzyme I homologue that possesses a putative sensory transduction domain. Gene 1996;181:103–8.
[184] Reizer J, Bachem S, Reizer A, Arnaud M, Saier Jr MH, Stulke J. Novel phosphotransferase system genes revealed by genome analysis—the complete complement of PTS proteins encoded within the genome of Bacillus subtilis. Microbiology 1999;145(Pt 12):3419–29.
[185] Ren Q, Paulsen IT. Comparative analyses of fundamental differences in membrane transport capabilities in prokaryotes and eukaryotes. PLoS Comput Biol 2005;1:e27.
[186] Ren Q, Chen K, Paulsen IT. TransportDB: A comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. Nucleic Acids Res 2007;35:D274–9.
[187] Ren Q, Paulsen IT. Large-scale comparative genomic analyses of cytoplasmic membrane transport systems in prokaryotes. J Mol Microbiol Biotechnol 2007;12:165–79.
[188] Roberts RJ. Identifying protein function—a call for community action. PLoS Biol 2004;2:E42.
[189] Rodionov DA, De Ingeniis J, Mancini C, Cimadamore F, Zhang H, Osterman A, et al. Transcriptional regulation of NAD metabolism in bacteria. NrtR family of Nudix-related regulators, submitted for publication.
[190] Rodionov DA, Li X, Rodionova I, Yang C, Gelfand M, Osterman A. Transcriptional regulation of NAD metabolism in bacteria. Genomic reconstruction of the NiaR (YrxA) regulon.
[191] Rodionov DA, Mironov AA, Rakhmaninova AB, Gelfand MS. Transcriptional regulation of transport and utilization systems for hexuronides, hexuronates and hexonates in gamma purple bacteria. Mol Microbiol 2000;38:673–83.
[192] Rodionov DA, Mironov AA, Gelfand MS. Conservation of the biotin regulon and the BirA regulatory signal in Eubacteria and Archaea. Genome Res 2002;12:1507–16.
[193] Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS. Comparative genomics of thiamin biosynthesis in procaryotes. New genes and regulatory mechanisms. J Biol Chem 2002;277:48949–59.
[194] Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS. Regulation of lysine biosynthesis and transport genes in bacteria: yet another RNA riboswitch? Nucleic Acids Res 2003;31:6748–57.
[195] Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS. Comparative genomics of the vitamin B_{12} metabolism and regulation in prokaryotes. J Biol Chem 2003;278:41148–59.
[196] Rodionov DA, Gelfand MS, Hugouvieux-Cotte-Pattat N. Comparative genomics of the KdgR regulon in Erwinia chrysanthemi 3937 and other gamma-proteobacteria. Microbiology 2004;150:3571–90.
[197] Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS. Comparative genomics of the methionine metabolism in Gram-positive bacteria: a variety of regulatory systems. Nucleic Acids Res 2004;32:3340–53.
[198] Rodionov DA, Gelfand MS. Identification of a bacterial regulatory system for ribonucleotide reductases by phylogenetic profiling. Trends Genet 2005;21:385–9.
[199] Rodionov DA, Gelfand MS. Computational identification of BioR, a transcriptional regulator of biotin metabolism in Alphaproteobacteria, and of its binding signal. FEMS Microbiol Lett 2006;255:102–7.
[200] Rodionov DA, Gelfand MS, Todd JD, Curson AR, Johnston AW. Computational reconstruction of iron- and manganese-responsive transcriptional networks in alpha-proteobacteria. PLoS Comput Biol 2006;2:e163.
[201] Rodionov DA, Hebbeln P, Gelfand MS, Eitinger T. Comparative and functional genomic analysis of prokaryotic nickel and cobalt uptake transporters: evidence for a novel group of ATP-binding cassette transporters. J Bacteriol 2006;188:317–27.
[202] Rodionov DA. Comparative genomic reconstruction of transcriptional regulatory networks in bacteria. Chem Rev 2007;107:3467–97.
[203] Rogozin IB, Makarova KS, Wolf YI, Koonin EV. Computational approaches for the analysis of gene neighbourhoods in prokaryotic genomes. Brief Bioinform 2004;5:131–49.
[204] Romualdi A, Siddiqui R, Glockner G, Lehmann R, Suhnel J. GenColors: Accelerated comparative analysis and annotation of prokaryotic genomes at various stages of completeness. Bioinformatics 2005;21:3669–71.
[205] Rost B, Casadio R, Fariselli P, Sander C. Transmembrane helices predicted at 95% accuracy. Protein Sci 1995;4:521–33.
[206] Rost B, Fariselli P, Casadio R. Topology prediction for helical transmembrane proteins at 86% accuracy. Protein Sci 1996;5:1704–18.
[207] Sabatti C, Rohlin L, Oh MK, Liao JC. Co-expression pattern from DNA microarray experiments as a tool for operon prediction. Nucleic Acids Res 2002;30:2886–93.
[208] Sadovskaya NS, Sutormin RA, Gelfand MS. Recognition of transmembrane segments in proteins: review and consistency-based benchmarking of internet servers. J Bioinform Comput Biol 2006;4:1033–56.
[209] Saier Jr MH, Beatty JT, Goffeau A, Harley KT, Heijne WH, Huang SC, et al. The major facilitator superfamily. J Mol Microbiol Biotechnol 1999;1:257–79.
[210] Saier Jr MH, Eng BH, Fard S, Garg J, Haggerty DA, Hutchinson WJ, et al. Phylogenetic characterization of novel transport protein families revealed by genome analyses. Biochim Biophys Acta 1999;1422:1–56.
[211] Saier Jr MH, Paulsen IT. Paralogous genes encoding transport proteins in microbial genomes. Res Microbiol 1999;150:689–99.
[212] Saier Jr MH. Vectorial metabolism and the evolution of transport systems. J Bacteriol 2000;182:5029–35.

[213] Saier Jr MH. A functional-phylogenetic classification system for transmembrane solute transporters. Microbiol Mol Biol Rev 2000;64:354–411.
[214] Saier Jr MH, Paulsen IT. Whole genome analyses of transporters in spirochetes: Borrelia burgdorferi and Treponema pallidum. J Mol Microbiol Biotechnol 2000;2:393–9.
[215] Saier Jr MH, Goldman SR, Maile RR, Moreno MS, Weyler W, Yang N, et al. Transport capabilities encoded within the Bacillus subtilis genome. J Mol Microbiol Biotechnol 2002;4:37–67.
[216] Saier Jr MH, Tran CV, Barabote RD. TCDB: The Transporter Classification Database for membrane transport protein analyses and information. Nucleic Acids Res 2006;34:D181–6.
[217] Sal-Man N, Gerber D, Shai Y. The identification of a minimal dimerization motif QXXS that enables homo- and hetero-association of transmembrane helices in vivo. J Biol Chem 2005;280:27449–57.
[218] Sarsero JP, Merino E, Yanofsky C. A Bacillus subtilis gene of previously unknown function, yhaG, is translationally regulated by tryptophan-activated TRAP and appears to be involved in tryptophan transport. J Bacteriol 2000;182:2329–31.
[219] Sauer E, Merdanovic M, Mortimer AP, Bringmann G, Reidl J. PnuC and the utilization of the nicotinamide riboside analog 3-aminopyridine in Haemophilus influenzae. Antimicrob Agents Chemother 2004;48:4532–41.
[220] Saurin W, Hofnung M, Dassa E. Getting in or out: early segregation between importers and exporters in the evolution of ATP-binding cassette (ABC) transporters. J Mol Evol 1999;48:22–41.
[221] Schomburg I, Chang A, Hofmann O, Ebeling C, Ehrentreich F, Schomburg D. BRENDA: A resource for enzyme data and metabolic information. Trends Biochem Sci 2002;27:54–6.
[222] Seddon AM, Curnow P, Booth PJ. Membrane proteins, lipids and detergents: not just a soap opera. Biochim Biophys Acta 2004;1666:105–17.
[223] Sekowska A, Robin S, Daudin JJ, Henaut A, Danchin A. Extracting biological information from DNA arrays: an unexpected link between arginine and methionine metabolism in Bacillus subtilis. Genome Biol 2001;2. RESEARCH0019.
[224] Senes A, Gerstein M, Engelman DM. Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. J Mol Biol 2000;296:921–36.
[225] Shimizu T, Mitsuke H, Noto K, Arai M. Internal gene duplication in the evolution of prokaryotic transmembrane proteins. J Mol Biol 2004;339:1–15.
[226] Snitkin ES, Gustafson AM, Mellor J, Wu J, DeLisi C. Comparative assessment of performance and genome dependence among phylogenetic profiling methods. BMC Bioinformatics 2006;7:420.
[227] Sonnhammer EL, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol 1998;6:175–82.
[228] Speers AE, Wu CC. Proteomics of integral membrane proteins—theory and application. Chem Rev 2007;107:3687–714.
[229] Tamames J. Evolution of gene order conservation in prokaryotes. Genome Biol 2001;2. RESEARCH0020.
[230] Taylor PD, Attwood TK, Flower DR. BPROMPT: A consensus server for membrane protein prediction. Nucleic Acids Res 2003;31:3698–700.
[231] Tchieu JH, Norris V, Edwards JS, Saier Jr MH. The complete phosphotransferase system in Escherichia coli. J Mol Microbiol Biotechnol 2001;3:329–46.
[232] Titgemeyer F, Amon J, Parche S, Mahfoud M, Bail J, Schlicht M, et al. A genomic view of sugar transport in Mycobacterium smegmatis and Mycobacterium tuberculosis. J Bacteriol 2007;189:5903–15.
[233] Tomii K, Kanehisa M. A comparative analysis of ABC transporters in complete microbial genomes. Genome Res 1998;8:1048–59.
[234] Torres J, Stevens TJ, Samso M. Membrane proteins: the 'Wild West' of structural biology. Trends Biochem Sci 2003;28:137–44.
[235] Tusnady GE, Simon I. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J Mol Biol 1998;283:489–506.
[236] Tusnady GE, Simon I. The HMMTOP transmembrane topology prediction server. Bioinformatics 2001;17:849–50.
[237] Tusnady GE, Dosztanyi Z, Simon I. Transmembrane proteins in the Protein Data Bank: identification and classification. Bioinformatics 2004;20:2964–72.
[238] Tusnady GE, Dosztanyi Z, Simon I. PDB_TM: Selection and membrane localization of transmembrane proteins in the protein data bank. Nucleic Acids Res 2005;33:D275–8.
[239] Ulmschneider MB, Sansom MS, Di Nola A. Properties of integral membrane protein structures: derivation of an implicit membrane potential. Proteins 2005;59:252–65.
[240] Valavanis IK, Bagos PG, Emiris IZ. beta-Barrel transmembrane proteins: Geometric modelling, detection of transmembrane region, and structural properties. Comput Biol Chem 2006;30:416–24.
[241] Vallenet D, Labarre L, Rouy Z, Barbe V, Bocs S, Cruveiller S, et al. MaGe: A microbial genome annotation system supported by synteny results. Nucleic Acids Res 2006;34:53–65.
[242] Van den Berg B, Clemons Jr WM, Collinson I, Modis Y, Hartmann E, Harrison SC, et al. X-ray structure of a protein-conducting channel. Nature 2004;427:36–44.
[243] van Nimwegen E. Scaling laws in the functional content of genomes. Trends Genet 2003;19:479–84.
[244] Viklund H, Elofsson A. Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. Protein Sci 2004;13:1908–17.
[245] Viklund H, Granseth E, Elofsson A. Structural classification and prediction of reentrant regions in alpha-helical transmembrane proteins: application to complete genomes. J Mol Biol 2006;361:591–603.
[246] Vitreschak A, Mironov A, Lyubetsky V, Gelfand M. Functional and evolutionary analysis of the T-box regulon in bacteria, 2007, in press.
[247] Vitreschak AG, Rodionov DA, Mironov AA, Gelfand MS. Regulation of riboflavin biosynthesis and transport genes in bacteria by transcriptional and translational attenuation. Nucleic Acids Res 2002;30:3141–51.

[248] Vogl C, Grill S, Schilling O, Stulke J, Mack M, Stolz J. Characterization of riboflavin (vitamin B_2) transport proteins from Bacillus subtilis and Corynebacterium glutamicum. J Bacteriol 2007.
[249] von Heijne G. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. J Mol Biol 1992;225:487–94.
[250] von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Kruger B, et al. STRING 7—recent developments in the integration and prediction of protein interactions. Nucleic Acids Res 2007;35:D358–62.
[251] von Rozycki T, Schultzel MA, Saier Jr MH. Sequence analyses of cyanobacterial bicarbonate transporters and their homologues. J Mol Microbiol Biotechnol 2004;7:102–8.
[252] Vrljic M, Garg J, Bellmann A, Wachi S, Freudl R, Malecki MJ, et al. The LysE superfamily: topology of the lysine exporter LysE of Corynebacterium glutamicum, a paradyme for a novel superfamily of transmembrane solute translocators. J Mol Microbiol Biotechnol 1999;1:327–36.
[253] Waldispuhl J, Berger B, Clote P, Steyaert JM. Predicting transmembrane beta-barrels and interstrand residue interactions from sequence. Proteins 2006;65:61–74.
[254] Walters RF, DeGrado WF. Helix-packing motifs in membrane proteins. Proc Natl Acad Sci USA 2006;103:13658–63.
[255] Warren PB, ten Wolde PR. Statistical analysis of the spatial distribution of operons in the transcriptional regulation network of Escherichia coli. J Mol Biol 2004;342:1379–90.
[256] Wei Y, Guffanti AA, Ito M, Krulwich TA. Bacillus subtilis YqkI is a novel malic/Na^+-lactate antiporter that enhances growth on malate at low protonmotive force. J Biol Chem 2000;275:30287–92.
[257] White SH, von Heijne G. Transmembrane helices before, during, and after insertion. Curr Opin Struct Biol 2005;15:378–86.
[258] Wightman R, Meacock PA. The THI5 gene family of Saccharomyces cerevisiae: distribution of homologues among the hemiascomycetes and functional redundancy in the aerobic biosynthesis of thiamin from pyridoxine. Microbiology 2003;149:1447–60.
[259] Winnen B, Hvorup RN, Saier Jr MH. The tripartite tricarboxylate transporter (TTT) family. Res Microbiol 2003;154:457–65.
[260] Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome Res 2001;11:356–72.
[261] Wu J, Kasif S, DeLisi C. Identification of functional links between genes using phylogenetic profiles. Bioinformatics 2003;19:1524–30.
[262] Xie G, Bonner CA, Song J, Keyhani NO, Jensen RA. Inter-genomic displacement via lateral gene transfer of bacterial trp operons in an overall context of vertical genealogy. BMC Biol 2004;2:15.
[263] Yakhnin H, Zhang H, Yakhnin AV, Babitzke P. The trp RNA-binding attenuation protein of Bacillus subtilis regulates translation of the tryptophan transport gene trpP (yhaG) by blocking ribosome binding. J Bacteriol 2004;186:278–86.
[264] Yang C, Rodionov DA, Li X, Laikova ON, Gelfand MS, Zagnitko OP, et al. Comparative genomics and experimental characterization of N-acetylglucosamine utilization pathway of Shewanella oneidensis. J Biol Chem 2006;281:29872–85.
[265] Yellaboina S, Goyal K, Mande SC. Inferring genome-wide functional linkages in E. coli by combining improved genome context methods: comparison with high-throughput experimental data. Genome Res 2007;17:527–35.
[266] Yen MR, Tseng YH, Simic P, Sahm H, Eggeling L, Saier Jr MH. The ubiquitous ThrE family of putative transmembrane amino acid efflux transporters. Res Microbiol 2002;153:19–25.
[267] Yeung KY, Medvedovic M, Bumgarner RE. From co-expression to co-regulation: how many microarray experiments do we need? Genome Biol 2004;5:R48.
[268] Yuan Z, Davis MJ, Zhang F, Teasdale RD. Computational differentiation of N-terminal signal peptides and transmembrane helices. Biochem Biophys Res Commun 2003;312:1278–83.
[269] Zaslaver A, Mayo A, Ronen M, Alon U. Optimal gene partition into operons correlates with gene functional order. Phys Biol 2006;3:183–9.
[270] Zhang Z, Feige JN, Chang AB, Anderson IJ, Brodianski VM, Vitreschak AG, et al. A transporter of Escherichia coli specific for L- and D-methionine is the prototype for a new family within the ABC superfamily. Arch Microbiol 2003;180:88–100.
[271] Zheng Y, Anton BP, Roberts RJ, Kasif S. Phylogenetic detection of conserved gene clusters in microbial genomes. BMC Bioinformatics 2005;6:243.
[272] Zuniga M, Comas I, Linaje R, Monedero V, Yebra MJ, Esteban CD, et al. Horizontal gene transfer in the molecular evolution of mannose PTS transporters. Mol Biol Evol 2005;22:1673–85.