Review TRENDS in Microbiology Vol.13 No.11 November 2005

Reconstructing the metabolic network of a bacterium from its genome

Christof Francke[1,2], Roland J. Siezen[1,2,3] and Bas Teusink[1,2,3]

[1]Wageningen Centre for Food Sciences, PO Box 557, 6700AN Wageningen, the Netherlands
[2]Centre for Molecular and Biomolecular Informatics, Radboud University, PO Box 9010, 6500GL Nijmegen, the Netherlands
[3]NIZO Food Research, PO Box 20, 6710BA Ede, the Netherlands

The prospect of understanding the relationship between the genome and the physiology of an organism is an important incentive to reconstruct metabolic networks. The first steps in the process can be automated and it does not take much effort to obtain an initial metabolic reconstruction from a genome sequence. However, such a reconstruction is certainly not flawless and correction of the many imperfections is laborious. It requires the combined analysis of the available information on protein sequence, phylogeny, gene-context and co-occurrence but is also aided by high-throughput experimental data. Simultaneously, the reconstructed network provides the opportunity to visualize the 'omics' data within a relevant biological functional context and thus aids the interpretation of those data.

The usefulness of metabolic reconstruction

With the dramatic increase in the number of sequenced genomes and the profound advance of experimental high-throughput analyses, there is a clear need for computational methods in biology. These methods comprise the development of comparative tools and maintenance of databases for the analysis of genomics data (the domain of bioinformatics), as well as the construction of models for the analysis and integration of the data in terms of the system properties (known as systems biology). An important asset to such analyses is the (partial) reconstruction of cellular networks, that is the collection and visualization of all potential physiologically relevant cellular processes. The reconstruction serves to sort the individual proteins, and thus the potential molecular functions, into a context (like pathways or protein complexes) and as such enables improved functional annotation [1]. In addition, it provides a platform to visualize and analyze 'omics' data. Furthermore, the overall topology of the metabolic network gives insight in the properties of the network [2], whereas flux analysis permits predictions of phenotype on the metabolic level to guide metabolic engineering [3,4].

There are various approaches towards the reconstruction of a metabolic network

This review is written from the perspective that one wishes to reconstruct the metabolism of a single bacterium from its genome sequence to enable rational metabolic engineering and to increase an understanding of the molecular physiology of that particular organism. Such a reconstruction can be conducted in different ways. One could focus on the recovery of possible pathways from the total genome pool and then proceed to piece together the metabolism of the organism from these pathways. In general, however, researchers tend to follow a more reductive path and opt to annotate and improve the annotation of the individual proteins first and later reconstruct the metabolic network.

In our opinion, it is rewarding to start with the reconstruction of the metabolic network directly after an initial fast annotation but before refining the annotation. It has the advantage that extra information on possible reaction context (like pathways) is available (and also visually), which helps the further annotation considerably. We propose several steps in the reconstruction process (depicted in Figure 1) and will describe the process accordingly. The first steps are fast and involve an automated annotation of the genome and the generation of a putative metabolic network. The other steps are laborious and require manual curation of the assignments that were made but also of those that were missed. Therefore, here we do not focus on the function prediction of proteins per se – excellent reviews exist on that matter (e.g. [5,6]) – or on the recovery of all functions and pathways as found in the total pool of sequenced genomes (see [7]). Although the approaches differ, essentially similar methods, tools, resources and pitfalls apply, albeit at different moments in the reconstruction process.

Homology and function

The sequence, in essence, encodes the molecular properties of a protein or a gene. Therefore, most often-used function prediction methods rely on the determination of sequence similarity or homology to molecules of known function. Confidence is added to such predictions by establishing the correct evolutionary linkages [8,9]. The term 'homology' was first used in 1843 by Richard Owen to refer to structural similarities among organisms and, later, by Charles Darwin, related to descendence from a common ancestor. It was Walter Fitch who, in 1970, discerned two instances of homology for the evolutionary relationship between proteins: orthology and paralogy [10] (Box 1). We will adhere to his definitions because a clear separation of

Corresponding author: Francke, C. (c.francke@cmbi.ru.nl).
Available online 19 September 2005

 0966-842X/$ - see front matter doi:10.1016/j.tim.2005.09.001

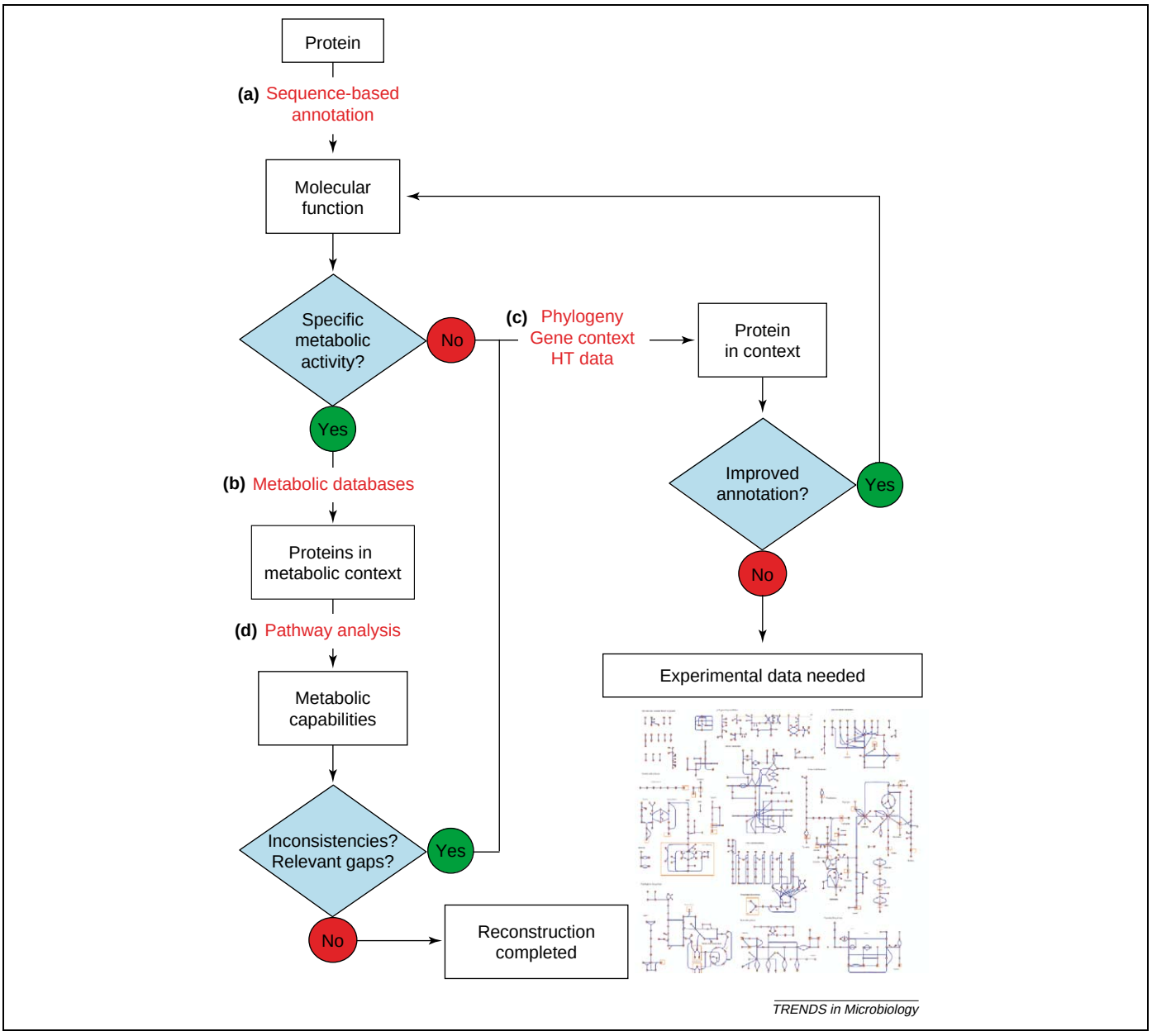

Figure 1. Flow chart of the metabolic reconstruction process. **(a)** Automatic assignment of molecular function on the basis of sequence homology and domain profiles. **(b)** When a specific metabolic activity is defined for the protein, an association is made with the corresponding reaction. **(c)** In case the molecular function is not specified, the assigned reaction is inconsistent with the metabolic context, or when specific reactions are needed to close gaps in pathways, comparative genomics approaches are applied to put the protein into a context that might provide information on the molecular function. **(d)** Pathway analysis of the metabolic reconstruction leads to predictions of missing reactions in certain pathways and metabolic capacities that can be checked with experimental data.

homologous genes into orthologous and paralogous genes is important to functional annotation. It is anticipated that orthologous genes will in most cases (but certainly not in all [11,12]) carry out identical functions, whereas paralogous genes will have similar but possibly distinct functions [8,9].

Like homology, function is a central concept in the process of annotation but its use also causes much confusion. The ambiguity of the attribute 'function' can be compared to that of the attribute 'meaning' in relation to a word. An individual word might have several meanings, which can be looked up in a dictionary but what the word means and what role it has (like subject or direct object) will depend on the context (e.g. sentence) it is in. Similarly, a protein can have various context-independent functions, referred to as molecular functions [5] (for example, catalyze certain reactions, bind to specific proteins or bind to a specific sequence of DNA) and have different context-dependent roles, mostly also referred to as functions (for example, act in different pathways, recruit another protein to a protein complex or act as a transcriptional activator). For the purpose of the reconstruction of metabolism, the molecular function, or to be more precise the catalytic function of a protein as represented by EC numbers [13], is the aspect of function that should be assigned.

Several classification schemes have been developed that capture a particular aspect of function and/or role (for a review, see [14]). In addition to EC numbers [13] these include: TC numbers (molecular transport [15]), SCOP classifications (protein structure [16]) and gene

Box 1. The evolutionary relationship between genes

The lineage of a gene is often supposed to relate to its function. In other words, function should be tractable through phylogeny. To aid the phylogenetic analyses, unambiguous terminology to describe the relationship between genes was proposed by W. Fitch [10]:

analogy is the relationship between two genes that have descended from unrelated ancestral genes but have converged to acquire similar functionality

homology is the relationship between two genes that have descended from a common ancestral gene and have diverged on the sequence level

orthology is the relationship between two homologous genes that originate from the same gene in the most recent common ancestor of the species that are compared

paralogy is the relationship between two homologous genes that arose from a gene duplication.

ontologies (GO [17]), amongst others. Nonetheless, awareness is increasing that it is inherently impossible to describe the complete functionality of a protein without describing the network because the network (context) defines the role of the protein [3,5,18]. Therefore, we will hereafter use the term 'annotation' only to refer to molecular function. However, one can, perhaps somewhat confusingly, use the metabolic or genomic context of a protein to increase specificity of the assigned molecular function, as explained in later sections.

Sequence based annotation

If it is of interest to reconstruct the metabolism of an organism whose genome has already been sequenced and annotated by others, the sequences and coupled annotation can be recovered from public resources, like that of NCBI [19] and many others (Table 1). The most common methods to obtain a rapid functional annotation of a new genome require sequence comparisons with already annotated genomes [20]. In their simplest form, these methods compare, using for instance BLAST [21] or FASTA [22], the sequences of the predicted gene products to all sequences, or a selection thereof [23], deposited in the reference databases linked to the Uniprot resource [24]. The annotations connected to the most similar sequences, known as 'best hits', are then transferred. Confidence is added to the prediction by requiring bi-directionality between the best hits (like in the tool INPARANOID [25]) or by categorization (and selection) of the deposited sequences into so-called clusters of orthologous groups (COGs) and requiring connectivity between the best hits [26]. Nevertheless, the members of a single COG are often not truly orthologous but simply homologous and, as a consequence, in many cases their molecular function does not concur. Therefore, methods have been developed that include automated phylogenetic analysis using sequence profiles [12]. Sequence comparison tools based on profiles employ Hidden Markov Models (HMMs [27]) to match proteins of unknown molecular function to enzyme or domain profiles of proteins of known molecular function deposited in profile databases (described in Table 1). The profile-based methods have become increasingly popular and powerful [28–30]. With the growing number of sequenced genomes, the methods become more specific as the profiles implicitly include information on the relative importance of every residue with respect to the molecular function through evolutionary conservation [20]. Simultaneously, the number of available profiles in the databases continues to grow.

Metabolic databases: coupling gene content with metabolic reaction information

Association of the annotated genes and proteins with a reaction is usually done by searching a pathway and/or reaction database with the protein name, EC number or another identifier to which it was linked. Useful resources of functional information on enzymes are Brenda [31], ENZYME [32] and the databases owned by the International Union of Biochemistry and Molecular Biology (IUBMB), which provide not only reaction stoichiometries but also information on enzyme properties, substrate specificities and cofactor usage, including variations thereof among different organisms. Membrane proteins and extracellular proteins can be identified as such, using transmembrane helix [33] and signal peptide [34] prediction tools. In addition, TransportDB [35] can be used to retrieve the transport functions encoded by many already sequenced genomes.

There are several excellent resources that contain a comprehensive amount of reaction information that can be coupled to genome data and, at the same time, enable visualization of the data within the biochemical framework of pathways. An example of such a resource is KEGG [36]. The KEGG resource is easily accessible and represents the most extensive collection of combined information on genes, metabolites, reactions and pathways publicly available. The KEGG pathway maps for a specific organism represent automatic metabolic reconstructions of the publicly available sequenced and annotated genomes (identification on basis of corresponding EC number) and have been adopted by many researchers and resources as a standard.

Another comprehensive and excellent resource is MetaCyc [37], which contains an inventory of reaction and pathway information for various organisms. With the aid of Pathway Tools [38] and using the information stored in MetaCyc, the metabolism of any organism can be reconstructed automatically, provided that the annotation file is available. The information is extracted from the annotation file (GenBank format) based on EC numbers and name matching. The resulting reconstruction can be edited and curated by the user (see [39] for an example). BioCyc (http://www.biocyc.org/) contains the collection of these reconstructions and includes the most advanced one publicly available: EcoCyc [40].

There are numerous other initiatives and tools dedicated to the integration of metabolic and genome information in the public and the private domain (Table 1). These resources are also useful to extract metabolic reactions and related pathways. However, it remains to be seen which of these initiatives will gain enough momentum to become comprehensive and to be maintained in the long term.

Table 1. Selection of resources databases and tools that support metabolic reconstruction

Main resources of genomic information		
DDBJ (Japan)	www.ddbj.nig.ac.jp (p)[a]	
EBI - EMBL (Europe)	www.ebi.ac.uk (p)	
NCBI (USA)	www.ncbi.nlm.nih.gov (p)	
Large sequencing projects		
DOE	Doegenomes.org (p)	Includes the Joint Genome Institute (www.jgi.doe.gov), which has sequenced and annotated 117 microbial genomes (32 complete). Evidence supporting the annotation is provided.
MIPS	mips.gsf.de (p)	Supports and maintains a set of generic databases and tools related to the systematic comparative analysis of several microbial, fungal and plant genomes.
TIGR	www.tigr.org (p)	Includes functional annotation of 80 genomes. Several analysis tools (like Genome Properties) were developed as well as an in-house classification of protein families.
Main sources of proteins and enzyme information		
IntEnz	www.ebi.ac.uk/intenz (p/np)	Links to the main collections of enzyme functional data: Brenda (www.brenda.uni-koeln.de), ENZYME (ca.expasy.org/enzyme) and IUBMB (www.iubmb.unibe.ch). These databases store information on reaction stoichiometries, kinetics, protein stability and more. Classification on basis of EC-number.
Uniprot (Swissprot)	www.expasy.uniprot.org (p)	This resource links to all main resources related to functional information on proteins (i.e. through EBI, EXPASY and PIR).
Enzyme profile databases		
Blocks	blocks.fhcrc.org (p)	
Interpro	www.ebi.ac.uk/interpro (p)	Links to profile databases.
PFAM	www.sanger.ac.uk/Software/Pfam (p)	
PRIAM	genopole.toulouse.inra.fr/bioinfo/priam (p)	
PROSITE	www.expasy.org/prosite (p)	
SMART	smart.embl.de (p)	
Automated annotation tools (software)		
Gamola	www.cals.ncsu.edu:8050/food_science/KlaenhammerLab/GAMOLA (p)	
GeneQuiz	jura.ebi.ac.uk:8765/ext-genequiz (p)	
Pedant	pedant.gsf.de (p)	
Automated annotation and comparative genomics resources and tools		
ERGO	ergo.integratedgenomics.com (c)	Provides specific annotation and useful tools to analyze the conservation of gene context.
MBGD	mbgd.genome.ad.jp (p)	A resource for comparative analysis of completely sequenced microbial genomes.
Prolinks	dip.doe-mbi.ucla.edu/pronav (p)	Prediction of functional linkages between proteins (straightforward interface).
STRING	www.bork.embl-heidelberg.de/STRING (p)	Powerful resource for the retrieval of interacting proteins (straightforward interface).
the SEED	theseed.uchicago.edu/FIG/index.cgi (p)	Large resource for genome annotation and analysis.
Pathway resources, databases and tools		
Amaze	www.amaze.ulb.ac.be (p)	Database on reactions and pathways that can be queried (no overviews of the content).
KEGG	www.genome.ad.jp/kegg (p)	The main resource for pathway information and much more (like metabolites).
BioCyc	www.biocyc.org (p)	Important resource for pathway information. Includes the best reconstruction: EcoCyc.
PATIKA	www.patika.org (np)	Environment for construction of cellular pathways. Focus on human pathway data.
Phylosopher	www.genedata.com (c)	Information management system for genomics data.
PUMA2	compbio.mcs.anl.gov/puma2/cgi-bin/index.cgi (p)	Contains pathway information.
SimPheny	www.genomatica.com (c)	Knowledge management system that enables the construction and simulation of a metabolic model.

[a]The accessibility of the various websites is indicated between brackets (p=public; np=non-profit; c=commercial).

The necessity of manual curation

The process as described previously (Figure 1) provides only the first steps towards a reconstruction of the metabolic network from a sequenced genome. Unfortunately, automatic procedures will produce a reconstruction that is neither complete (in fact far from it [41]) nor (fully) appropriate [42,43]. Manual curation is therefore a necessity and includes:

(i) Inspection of the annotation present in the source databases because many entries are incorrect [42]. This is basically done by tracing back the available experimental evidence, although it should be kept in

mind that the supposed evidence also sometimes should be evaluated (as illustrated by [44]);
(ii) Resolving inconsistencies between protein and function identifiers in different databases. Due to these inconsistencies, different annotations published for the small genome of *Mycoplasma genitalium* deviated for ~8% of the gene products [45].
(iii) Addition of new and/or organism-specific reactions or pathways that are absent in the queried databases;
(iv) Judging the correctness of the coupling between query sequence and the sequence in the resource database that has identical molecular function because homology- and profile-based methods do not always yield a correct coupling [5,11,43]. The methods involved in that process will be discussed in the next section; and finally
(v) Evaluation of the coupling between the function identifiers and the retrieved reactions. The reconstruction of folate biosynthesis in the lactic acid bacteria *Lactococcus lactis* and *Lactobacillus plantarum* WCFS1 (AL935263) exemplifies one of the problems associated with point (v). When KEGG is consulted, the folate biosynthesis pathway appears complete except for the biosynthesis of the precursor 4-aminobenzoate. However, the conversion of dihydroneopterin triphosphate to dihydroneopterin is connected to an incomplete EC number (3.6.1.-) and, although the map suggests that the appropriate enzyme is present, in fact none of the enzymes coupled to this EC number in the database catalyzes the specific conversion. Thus, the use of unspecific functional identifiers (like incomplete EC numbers) could lead to false reaction associations, which have to be checked manually. A potential candidate pyrophosphatase was identified in the *L. lactis* genome by bioinformatic study of the folate gene cluster and its function was proven experimentally [46].

The use of phylogeny, gene context and high-throughput data

Curation requires the reconsideration of all individual proteins within the context of the initial reconstruction. Comparative genomics can be applied to generate additional data to support or reconsider the functional attributes of individual proteins [47]. This might involve the analysis of phylogeny [48], gene fusions [49], gene order [50], co-occurrence [51], regulatory motifs [52] or experimental evidence. Not one indicator, but several aspects, should be weighed.

In most bacteria, several functional classes represent a large number of homologous proteins and for these proteins orthology is quite often not properly assigned using bidirectional best hits. These classes include transport (e.g. ABC transporters and phosphotransfer systems) and signalling (e.g. two-component systems and transcription factors) and also processes related to carbohydrate and amino acid conversion. In such cases, phylogeny should be used to determine orthology [8,53]. Figure 2a illustrates the use of phylogeny in the process of curation of the metabolic reconstruction of *L. plantarum* WCFS1. Automatic annotation methods yielded four proteins (Map1–4) that are homologous to maltose phosphorylase of *L. lactis* (function extracted from the literature [54]) and phylogeny was used to distinguish between the homologs. The phylogenetic tree indicated that Map2 and Map3 are closely related and that either one could be orthologous to the maltose phosphorylase of *L. lactis* and thus might have identical molecular function.

However, although sequence similarity and orthology are strong indicators of functional identity, it is definitely not the case that they always concur [5,11,12]. Therefore, in addition the conservation of gene context between different organisms should be used as an important indicator of functional equivalence. Comparison of gene context renders information on the functional context the various gene-products have to operate in and thereby can provide a means to improve the annotation of the individual proteins [49,55]. For example, in the case of the maltose phosphorylases Map2 and Map3 of *L. plantarum*, two distinct evolutionary conserved functional contexts can be discerned (Figure 2b); one associated with active, the other with passive, sugar transport. The active transport system is orthologous to the well-characterized maltose ABC transporter of *Escherichia coli*, therefore, the bioinformatic evidence for the proposed molecular function of Map2 is convincing. At the same time, the preservation of context for the cluster containing the passive transport system suggests that the system could be a maltose transporter. Presumably, Map2 and Map3 have retained identical molecular function but could have an essentially different physiological role. One hypothesis would be that the former initiates maltose catabolism at low external concentrations of maltose (when active transport is needed), whereas the latter operates at high external concentrations.

The reliability of attributed molecular functions can be improved further by considering experimental evidence for functional connections like those provided by high-throughput experiments, such as transcriptome analysis and protein–protein binding studies. Information on these connections between two proteins, as evident from similar gene expression patterns or physical associations, is provided by STRING [56]. STRING is an extremely useful resource that combines the information on protein associations as manifest in genome or metabolic context, phylogenetic or literature co-occurrence and in experimental data, and provides confidence scores for the recovered associations. The associations are transferred across, at least, the 180 different organisms within the database on basis of sequence homology.

Pathway analysis: filling gaps and completing the network

A metabolic reconstruction, like annotation, is never really 'completed' but evolves as insight grows. Nevertheless, depending on its purpose, there are a few minimal quality requirements that should be fulfilled for the first 'definite' version. First, the metabolic capabilities represented by the reconstructed network should be consistent with the physiology of the organism. For example, the presence or absence of certain pathways or reactions will affect the availability of substrates for other reactions, as

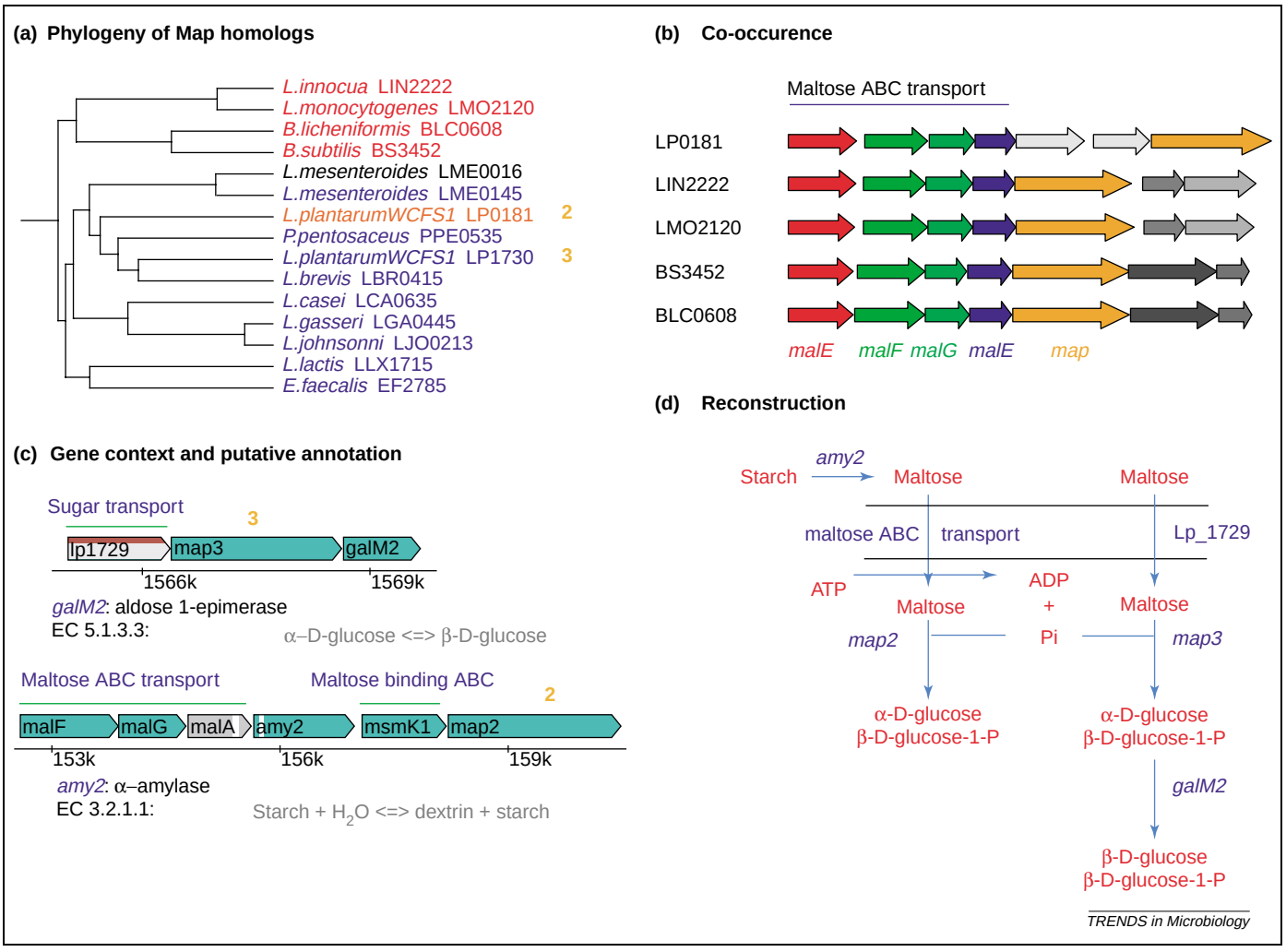

Figure 2. Annotation of the maltose phosphorylase homologs of *Lactobacillus plantarum* WCFS1 (AL935263) using phylogeny, gene context and functional context. *L. plantarum* WCFS1 contains four homologs of a putative maltose phosphorylase (*map*) gene. Identification of the potential function of the related proteins within the metabolic network of the bacterium was conducted as follows. **(a)** Homologous sequences from other Gram-positive bacteria* and *E. coli* were retrieved from the ERGO database [67]; they were aligned using MUSCLE [64] and a phylogenetic tree was made through Neighbor-Joining in CLUSTAL [68] and visualized using LOFT (Rene van der Heijden, unpublished). The proteins clustered in three distinct putative orthologous groups: Map1 with Map of *E. coli*, Map2 (Lp0181) and Map3 (Lp1730) with Map of *L. lactis* [54] and Map4 in a separate group. Because Map2 and Map3 appear in the same cluster, this cluster could be composed of two orthologous groups. **(b)** Comparison of the various gene contexts in ERGO and of the phylogenetic trees made for the neighbouring genes (not shown), showed that the cluster indeed consists of two orthologous groups. In the first group [indicated in red under (a)], the gene is clustered on the genome with a maltose ABC transporter. In the second group [indicated in blue under (a)] the gene is coupled to a transporter that was annotated as facilitating the diffusion of sugar. There is a high probability that both Map2 and Map3 have identical molecular function. However, at a higher level they have functionally diversified because one is associated with active, and the other with passive, transport. **(c)** Subsequently, the information obtained for the Map homologs can be used for the functional annotation of the neighbouring genes. Considering its gene context, it is probable that the transporter associated with Map3 will transport maltose. Similarly, the gene *amy2* probably encodes a protein with a slightly different function than originally annotated, namely one related with the production of maltose instead of the larger dextrins. **(d)** Finally, the new assignments can be incorporated in the network. * The species in the phylogenetic tree include: *Lactococcus lactis, Enterococcus faecalis, Lactobacillus gasseri, L. johnsonni, L. casei, L. plantarum, L. brevis, Pediococcus pentosaceus, Leuconostoc mesenteroides, Bacillus licheniformis, Bacillus subtilis, Listeria innocua* and *Listeria monocytogenes*.

illustrated in Figure 3. An initial reconstruction of the metabolism of *L. plantarum* suggested that succinyl-CoA serves as a reactant in one of the reactions involved in methionine biosynthesis. However, succinyl-CoA is not produced by the organism because of an incomplete TCA-cycle. After study of the phylogeny (as detailed in the figure legend), it could be concluded that the appropriate substrate in this reaction is likely to be acetyl-CoA [39]. Inconsistencies such as these can only be identified by pathway analysis of the whole metabolic network. In addition, it requires good knowledge of the physiology of the organism. For example, it could have been the case that the apparent inability to synthesize succinyl-CoA, as indicated by the gap in the pathway producing succinyl-CoA (where a gap refers to a reaction in a pathway that is not coupled to a gene-product), arose because the appropriate enzymes were present but could not be recovered by the annotation procedure. Alternatively, if *L. plantarum* was not able to synthesize methionine, the gap could be a potential reason for this inability.

Furthermore, when the reconstruction is used to produce a genome-scale metabolic model with the purpose of yield and flux predictions [57], reactions should be elementary balanced and essential pathways should be complete. When the reconstruction serves as a framework to provide a metabolic context in annotation and data analysis, unbalanced reactions or gaps in pathways are not disastrous. Nonetheless, a reconstruction that is more complete clearly provides more insight (and satisfaction).

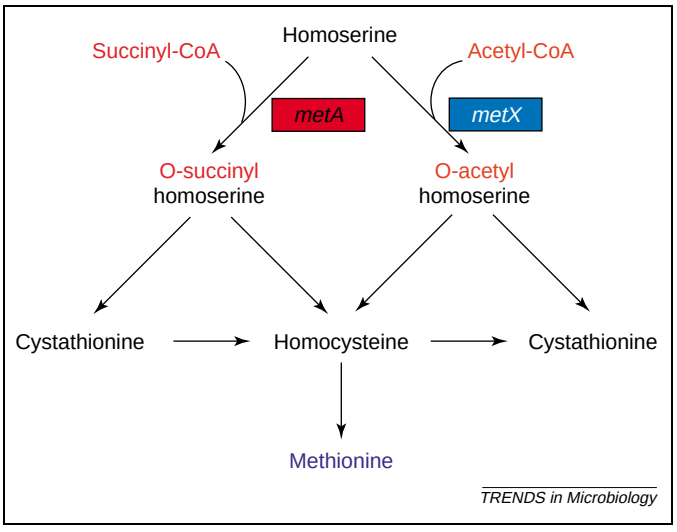

Figure 3. Proteins in a metabolic context: the case of methionine biosynthesis. The first step in the biosynthesis of methionine in many bacteria is the transfer of succinyl-CoA or acetyl-CoA to homoserine. The reactions are catalyzed by a homoserine O-succinyltransferase and by a homoserine O-acetyltransferase, respectively. Two distinct orthologous clusters of genes exist that are associated in KEGG and other resources to either homoserine O-succinyl- or homoserine O-acetyltransferase. Interestingly, the corresponding genes, most often designated as *metA* and *metX*, respectively, show phylogenetic anticorrelation (Teusink and Francke, unpublished), which is an indicator of analogy [69]. In *L. plantarum* WCFS1 a protein is present belonging to the orthologous cluster of *metA* [and has been annotated as a succinyltransferase in UniProt (Q88UF5) and all other databases]. From the metabolic context this was surprising because the organism can make methionine but lacks a TCA-cycle and, therefore, cannot make succinyl-CoA. Moreover, the orthologous MetA protein from *Bacillus subtilis* has specificity for acetyl-CoA [70]. These findings suggest that MetA of *L. plantarum* should be able to use acetyl-CoA as substrate. This is a case, therefore, where the existence of two distinct protein families does not correspond to distinct molecular functions and where metabolic context and experimentation was crucial for elucidating this information.

The analysis and closure of gaps in pathways has been excellently reviewed previously [58]. The search for proteins that correspond to pathway gaps, in principle, involves the same tools and resources that are used in molecular function prediction of individual proteins (described in the previous section) but new fully automated integrative tools have also been developed [59]. Despite the fact that later tools are useful to fill any gaps, many of them remain. Moreover, it is not always clear whether the gap should be filled at all or whether, in fact, the pathway should be redefined [60]. Knowledge on the composition of similar pathways in closely related organisms [1,58] and knowledge on their physiology can help in making that decision. In addition, comprehensive metabolomics studies are helpful in revealing compounds whose presence require corresponding conversion routes to be active and, hence, closed.

Concluding remarks

Here, we have argued that automatic pathway reconstruction methods are prone to make mistakes of many different kinds. Thus, every automatic reconstruction of metabolism in our opinion generates only a global picture of metabolism with limited use for understanding the physiology of a particular organism. Nevertheless, these reconstructions can be used for comparative studies with the aim of extracting general properties, such as scale-free nature and connectivity [61]. For the effective application in metabolic engineering and for deeper insights into the biology of a specific organism, expert curation of the initial metabolic network and concomitant experimental support and/or data are essential. A proper reconstruction places the proteins in a higher level context, which reflects (in part) their biological role and thereby enables visualization of 'omics' data in a biologically relevant context. In addition, conversion of the reconstruction to a flux model enables the prediction of the effects of metabolic engineering and the rationalization of that process [57].

Nevertheless, it should be realized that a good metabolic reconstruction is an initial first step towards a thorough understanding of the genotype–phenotype relationship of a specific organism. For one, a metabolic reconstruction represents only a fraction (less than 30%) of all the genes, namely those genes encoding enzyme-catalyzed reactions. Furthermore, the reconstruction merely reflects the full complement of the metabolic capacities of the cell, not the capacities at a specific physiological state. For that, knowledge on the regulation of network flows is required. The main challenges in the field are, therefore, to increase the percentage of functionally annotated genes and simultaneously to integrate the reconstructed metabolic network with other types of networks, such as those related to the regulation of transcription (genetic) and the regulation of protein activity (signalling).

Finally, there are several developments that will speed up the reconstruction process and improve its accuracy considerably. These include efforts to unify nomenclature and to devise physiologically relevant functional classification schemes that enable effective coupling of stored information {[17,62,63] and Biopax (http://www.biopax.org)}, the occurrence of new tools like the fast and good multiple sequence alignment algorithm MUSCLE [64], or methods for high-throughput orthology prediction [65] and the availability of well-maintained public databases and resources, such as those linked via EBI (http://www.ebi.ac.uk) or NCBI [19], or like KEGG [36] and STRING [56]. Moreover, the moment more metabolic reconstructions of high quality become available (such as Ecocyc [40]), the process of annotation and curation will become less labour-intensive as such reconstructions themselves might be used as a source of information for comparative studies to accelerate the metabolic reconstruction of other organisms, much in the way that comparative genomics aids in gene annotation. This idea is at the heart of the SEED initiative, which aims to provide 'a suite of open source tools (and information) to enable distributed researchers to rapidly annotate new genomes' [66]. All these developments are invaluable for the advance of metabolic reconstruction, model development and hence (systems) biology.

Acknowledgements

This work was conducted within the Kluyver Centre for Genomics of Industrial Fermentations and was further supported by grant NWO-BMI # 050.50.206 and the BioRange program of NBIC through the Netherlands Genomics Initiative.

References

1 Kharchenko, P. *et al*. (2004) Filling gaps in a metabolic network using expression information. *Bioinformatics* 20, i178–i185
2 Stelling, J. *et al*. (2002) Metabolic network structure determines key aspects of functionality and regulation. *Nature* 420, 190–193
3 Palsson, B.O. (2004) *In silico* biotechnology. Era of reconstruction and interrogation. *Curr. Opin. Biotechnol.* 15, 50–51
4 Smid, E.J. *et al*. (2005) Functional ingredient production: application of global metabolic models. *Curr. Opin. Biotechnol.* 16, 190–197
5 Bork, P. *et al*. (1998) Predicting function: from genes to genomes and back. *J. Mol. Biol.* 283, 707–725
6 Whisstock, J.C. and Lesk, A.M. (2003) Prediction of protein function from protein sequence and structure. *Q. Rev. Biophys.* 36, 307–340
7 Ye, Y. *et al*. (2005) Automatic detection of subsystem/pathway variants in genome analysis. *Bioinformatics* 21, i478–i486
8 Eisen, J.A. and Wu, M. (2002) Phylogenetic analysis and gene functional predictions: Phylogenomics in action. *Theor. Popul. Biol.* 61, 481–487
9 Sjölander, K. (2004) Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics* 20, 170–179
10 Fitch, W.M. (2000) Homology a personal view on some of the problems. *Trends Genet.* 16, 227–231
11 Gerlt, J.A. and Babbitt, P.C. (2001) Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu. Rev. Biochem.* 70, 209–246
12 Zmasek, C.M. and Eddy, S.R. (2002) RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics* 3, 14
13 Fleischmann, A. *et al*. (2004) IntEnz, the integrated relational enzyme database. *Nucleic Acids Res.* 32, D434–D437
14 Ouzounis, C.A. and Karp, P.D. (2002) The past, present and future of genome-wide re-annotation. *Genome Biol.* 3, COMMENT2001
15 Saier, M.H., Jr. (2000) A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiol. Mol. Biol. Rev.* 64, 354–411
16 Andreeva, A. *et al*. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.* 32, D226–D229
17 Harris, M.A. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32, D258–D261
18 Huynen, M.A. *et al*. (2005) Variation and evolution of biomolecular systems: Searching for functional relevance. *FEBS Lett.* 579, 1839–1845
19 Wheeler, D.L. *et al*. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 33, D39–D45
20 Watson, J.D. *et al*. (2005) Predicting protein function from sequence and structural data. *Curr. Opin. Struct. Biol.* 15, 275–284
21 Altschul, S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410
22 Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U. S. A.* 85, 2444–2448
23 Cummings, L. *et al*. (2002) Genomic BLAST: custom-defined virtual databases for complete and unfinished genomes. *FEMS Microbiol. Lett.* 216, 133–138
24 Bairoch, A. *et al*. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.* 33, D154–D159
25 Remm, M. *et al*. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* 314, 1041–1052
26 Tatusov, R.L. *et al*. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 29, 22–28
27 Durbin, R.E.S. *et al*. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press
28 Claudel-Renard, C. *et al*. (2003) Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res.* 31, 6633–6639
29 Pinney, J.W. *et al*. (2005) metaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of *Plasmodium falciparum* and *Eimeria tenella*. *Nucleic Acids Res.* 33, 1399–1409
30 Soding, J. *et al*. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* 33, W244–W248
31 Schomburg, I. *et al*. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.* 32, D431–D433
32 Bairoch, A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.* 28, 304–305
33 Arai, M. *et al*. (2004) Proteome-wide functional classification and identification of prokaryotic transmembrane proteins by transmembrane topology similarity comparison. *Protein Sci.* 13, 2170–2183
34 Boden, M. and Hawkins, J. (2005) Prediction of subcellular localisation using sequence-biased recurrent networks. *Bioinformatics* 21, 2279–2286
35 Ren, Q. *et al*. (2004) TransportDB: a relational database of cellular membrane transport systems. *Nucleic Acids Res.* 32, D284–D288
36 Kanehisa, M. *et al*. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32, D277–D280
37 Krieger, C.J. *et al*. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* 32, D438–D442
38 Karp, P.D. *et al*. (2002) The Pathway Tools software. *Bioinformatics* 18, S225–S232
39 Teusink, B. *et al*. *In silico* reconstruction of the metabolic pathways of *Lactobacillus plantarum*: comparing predictions of nutrient requirements with growth experiments. *Appl. Environ. Microbiol.* (in press)
40 Keseler, I.M. *et al*. (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res.* 33, D334–D337
41 Iliopoulos, I. *et al*. (2001) Genome sequences and great expectations. *Genome Bio.* 2, INTERACTIONS0001
42 Devos, D. and Valencia, A. (2001) Intrinsic errors in genome annotation. *Trends Genet.* 17, 429–431
43 Devos, D. and Valencia, A. (2000) Practical limits of function prediction. *Proteins* 41, 98–107
44 Iyer, L.M. *et al*. (2001) Quod erat demonstrandum? The mystery of experimental validation of apparently erroneous computational analyses of protein sequences. *Genome Biol.* 2, RESEARCH0051
45 Brenner, S.E. (1999) Errors in genome annotation. *Trends Genet.* 15, 132–133
46 Klaus, S.M. *et al*. (2005) A nudix enzyme removes pyrophosphate from dihydroneopterin triphosphate in the folate synthesis pathway of bacteria and plants. *J. Biol. Chem.* 280, 5274–5280
47 Huynen, M.A. *et al*. (2003) Function prediction and protein networks. *Curr. Opin. Cell Biol.* 15, 191–198
48 Pazos, F. and Valencia, A. (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng.* 14, 609–614
49 Yanai, I. *et al*. (2001) Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proc. Natl. Acad. Sci. U. S. A.* 98, 7940–7945
50 Dandekar, T. *et al*. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* 23, 324–328
51 Pellegrini, M. *et al*. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* 96, 4285–4288
52 Bulyk, M.L. *et al*. (2004) A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in *Escherichia coli*. *Genome Res.* 14, 201–208
53 Holder, M. and Lewis, P.O. (2003) Phylogeny estimation: traditional and Bayesian approaches. *Nat. Rev. Genet.* 4, 275–284
54 Nilsson, U. and Radstrom, P. (2001) Genetic localization and regulation of the maltose phosphorylase gene, *malP*, in *Lactococcus lactis*. *Microbiology* 147, 1565–1573
55 Valencia, A. and Pazos, F. (2002) Computational methods for the prediction of protein interactions. *Curr. Opin. Struct. Biol.* 12, 368–373
56 von Mering, C. *et al*. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* 33, D433–D437
57 Price, N.D. *et al*. (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.* 2, 886–897
58 Osterman, A. and Overbeek, R. (2003) Missing genes in metabolic pathways: a comparative genomics approach. *Curr. Opin. Chem. Biol.* 7, 238–251

59 Green, M.L. and Karp, P.D. (2004) A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* 5, 76
60 Cordwell, S.J. (1999) Microbial genomes and 'missing' enzymes: redefining biochemical pathways. *Arch. Microbiol.* 172, 269–279
61 Barabasi, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113
62 Hucka, M. *et al*. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 19, 524–531
63 Roux-Rouquie, M. *et al*. (2004) Using the Unified Modelling Language (UML) to guide the systemic description of biological processes and systems. *Biosystems* 75, 3–14
64 Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797
65 Sicheritz-Ponten, T. and Andersson, S.G. (2001) A phylogenomic approach to microbial evolution. *Nucleic Acids Res.* 29, 545–552
66 Overbeek, R.A. *et al*. (2004) The SEED: a peer-to-peer environment for genome annotation. *Commun. ACM* 47, 46–51
67 Overbeek, R. *et al*. (2003) The ERGO genome analysis and discovery system. *Nucleic Acids Res.* 31, 164–171
68 Thompson, J.D. *et al*. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25, 4876–4882
69 Morett, E. *et al*. (2003) Systematic discovery of analogous enzymes in thiamin biosynthesis. *Nat. Biotechnol.* 21, 790–795
70 Hacham, Y. *et al*. (2003) *In vivo* analysis of various substrates utilized by cystathionine gamma-synthase and O-acetylhomoserine sulfhydrylase in methionine biosynthesis. *Mol. Biol. Evol.* 20, 1513–1520