Archaea 1, 223–229

Automated metabolic reconstruction for *Methanococcus jannaschii*

SOPHIA TSOKA,[1,2] DAVID SIMON[1,3] and CHRISTOS A. OUZOUNIS[1]

[1] *Computational Genomics Group, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, U.K.*

[2] *Corresponding author (tsoka@ebi.ac.uk)*

[3] *Present address: Institut Pasteur, Génopole, Plateforme Bioinformatique, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France*

Received April 8, 2003; accepted August 12, 2003; published online October 2, 2003

Summary We present the computational prediction and synthesis of the metabolic pathways in *Methanococcus jannaschii* from its genomic sequence using the PathoLogic software. Metabolic reconstruction is based on a reference knowledge base of metabolic pathways and is performed with minimal manual intervention. We predict the existence of 609 metabolic reactions that are assembled in 113 metabolic pathways and an additional 17 super-pathways consisting of one or more component pathways. These assignments represent significantly improved enzyme and pathway predictions compared with previous metabolic reconstructions, and some key metabolic reactions, previously missing, have been identified. Our results, in the form of enzymatic assignments and metabolic pathway predictions, form a database (MJCyc) that is accessible over the World Wide Web for further dissemination among members of the scientific community.

Keywords: metabolic databases, Methanocaldococcus, pathway synthesis.

Introduction

The first genome sequence became publicly available in 1995; since then, more than 100 species have been completely sequenced and the data submitted to public repositories (Janssen et al. 2003). Interest in the use of computational methods for revealing the mechanisms of cellular processes has increased because of: (1) the ability of such methods to handle large amounts of genome data and thus produce integrated views of functional processes, (2) the capacity for rapid and objective processing of data, and (3) the suitability of these methods for analysis of experimentally non-tractable organisms that cannot be cultured in the laboratory.

Prediction of the metabolic complement, or the entire set of metabolic reactions, from the genome sequence—a process known as metabolic reconstruction—has attracted significant interest (Karp and Paley 1994). Systems such as MetaCyc (Karp et al. 2002*b*), KEGG (Kanehisa et al. 2002) and WIT (Overbeek et al. 2000) offer a reference knowledge base that associates functional features of individual genes (Enzyme Commission (EC) numbers, enzyme names) with specific metabolic reactions and pathways. Given an uncharacterized genome sequence, such systems deduce a map of the entire metabolic complement of the target organism, providing an excellent basis for further computational or experimental analysis of cellular metabolism.

The advantage of computational metabolic network synthesis is that the assumed "completeness" of a genome (in terms of the sequences it encodes) or biochemical pathway (in terms of the reactions it includes) may form the basis for further functional assignment beyond sequence similarity by exploiting contextual constraints (Karp et al. 1996, Bono et al. 1998). Current computational analyses of genome sequences, which use a combination of sequence similarity and comparative genomics, are combined with high-throughput experimental methods to progress from individual functional assignments to integrated descriptions of cellular metabolic networks (Eisenberg et al. 2000).

In that sense, metabolic reconstruction for entire genomes not only compiles existing knowledge, but also assists with the analysis and delineation of protein function through the association of enzyme properties with biochemical pathways. Coupled with the ever-increasing amount of experimentally characterized protein sequences in the databases, it is possible to generate accurate metabolic maps using computational methods. The ultimate goal is to facilitate experimental analysis by organizing all available genomic information for a specific organism into an integrated, coherent and homogeneous resource (Tsoka and Ouzounis 2000). Formation of protein complexes is implicit in the identification procedure and, when multiple proteins are assigned to a single EC number and reaction, they are assumed to participate in the same reaction step. This procedure flags possible participation of proteins as subunits of the same enzymatic complex, which can in turn be established by literature searches and annotation.

The genome of *Methanococcus jannaschii* was the first archaeal genome to be sequenced (Bult et al. 1996). The sequence data have been used to support the hypothesis that the Archaea constitute a domain of life separate from Bacteria and Eukarya—a hypothesis proposed several years before the genome sequence became available (Woese and Fox 1977, Graham et al. 2000*b*). Archaea have attracted significant interest because of their extremophilic properties and their evolutionary relationships to members of the other two domains. De-

spite growth in the amount of archaeal genome sequence data, the domain is still underrepresented in protein sequence databases.

Approximately 40% of the *M. jannaschii* genome is specific either to this organism or to the archaeal kingdom (Graham et al. 2001*b*); metabolic reconstruction, therefore, is particularly challenging. The aim of this paper is to evaluate an automated procedure for detection of the reaction capacities and reconstruction of the metabolic pathways present in *M. jannaschii*. Overall, computational analyses of *M. jannaschii* are valuable because of this microorganism's evolutionary position and distinctive lifestyle. The availability of a well-annotated metabolic complement of *M. jannaschii* will facilitate comparative analyses of archaeal metabolism and evolution (Kyrpides et al. 1999) across different organisms and domains of life. Our results are compared with previous metabolic reconstructions for this species and are available to the scientific community for further analysis (accessible at: http://maine.ebi.ac.uk:1555 /server.html).

Methods and Results

Microorganism

Methanococcus jannaschii is a hyperthermophilic methanogenic archaeon (Bult et al. 1996). It was isolated from surface material collected at a "white smoker" chimney at a depth of 2600 m in the East Pacific Rise near the western coast of Mexico. Its extreme habitat (48–94 °C, 200 atm) suggests that it possesses adaptations for growth at high temperature, high pressure, and moderate salinity. It is an autotrophic methanogen capable of nitrogen fixation (Wolfe 1992, Deppenmeier 2002). Cells are irregular cocci possessing polar bundles of flagella. The cell envelope consists of a cytoplasmic membrane and a protein surface layer (Bult et al. 1996).

Automated metabolic reconstruction

To establish the functional attributes of *M. jannaschii* protein sequences, the entire genome was searched against the nonredundant protein database (NRDB version of October 2002, 999845 entries) using BLAST (e-value 10^{-10}) (Altschul et al. 1997), filtered for composition bias using CAST (Promponas et al. 2000). Among the 1792 sequences that make up the *M. jannaschii* genome, sequence similarity searches against the non-redundant database identified a total of 587 proteins likely to have enzymatic activity, as indicated by the -ase suffix in the function annotation string. These proteins corresponded to 376 potential enzymes (21% of the total number of sequences) and were assigned to any of 246 unique EC numbers. Functional annotations for each protein, including EC numbers where available, and the genome sequence were used as inputs to the Pathway Tools software (Karp et al. 2002*a*), which incorporates the PathoLogic algorithm for inference of metabolic pathways of a genome given its sequence and functional annotations (Paley and Karp 2002).

The PathoLogic protocol uses MetaCyc, a manually curated collection of metabolic pathways and reactions from a multitude of organisms (Karp et al. 2002*b*), as a reference database upon which reconstruction for the query genome is performed. Given the functional annotations of the genome, the system associates EC numbers and enzyme names with reference reactions, to infer all reactions present in the organism of interest. After the reaction detection stage, the reaction-to-pathway associations of the reference state are used to identify candidate metabolic pathways in the target organism.

For each candidate pathway, three parameters are evaluated: the number of reactions in the reference pathway (X), the number of reactions identified in the query genome (Y), and the number of these reactions that are shared across more than one pathway (Z). The values for X, Y and Z provide a measure of confidence for the existence of the pathway in the reconstruction outcome. A pathway is considered to be present if, for example, at least half of the pathway reactions are found in the organism of interest and not all of them are shared. A number of pathways defined in the reference set are highly related to one another (termed variants); in this case, PathoLogic selects from among these pathways the one for which a unique enzyme has been identified. We predicted 113 pathways as well as an additional 17 pathways (termed super-pathways) each comprising several other smaller pathways.

The metabolic reconstruction as implemented by Pathway Tools is largely performed automatically; however, some manual intervention is required at the final stage, namely in the following two steps: (1) the user can inspect instances of enzymes identified by the suffix -ase for which PathoLogic has not identified an associated reaction (probable enzymes), and manually associate them with a specific reaction; and (2) the user may eliminate pathways for which there is little evidence. Because one of the main purposes of this study was hypothesis-generation of the metabolic complement of *M. jannaschii*, we aimed to obtain the maximum possible number of pathways and thus decided to retain all predicted pathways.

Pathway Tools employs frame representation technology (Karp and Paley 1994), a type of object-oriented database schema, so the terms *object* and *frame* are used interchangeably here. To construct the metabolic database for *M. jannaschii*, whenever an association between a genome protein and a MetaCyc pathway is made (through the EC number or the name-matching procedure), the relevant reaction and pathway frames corresponding to that pathway are copied to the genome database. Database objects are also created for each genetic element and gene, with appropriate pointers between each gene and the protein product it encodes, as well as between each product and the reactions that the product catalyzes. The knowledge base for *M. jannaschii* (MJCyc) built by this process contains 1792 gene and protein frames, 130 pathway frames, 609 reaction frames and 461 enzymatic-reaction frames. A summary of the reconstruction assignments is shown in Table 1.

Following the manual curation steps that augment the initial input information, 436 proteins from among the total number of gene products (1792) were associated with a reaction. Of these, 418 contained EC information and mapped to a total of 266 unique EC numbers. In most cases there was a one-to-one

Table 1. Overall outcome of metabolic reconstruction for *Methanococcus jannaschii*.

Assignment	Number
Proteins	1792
Enzymes	436
Enzymatic reactions	609
Missing reactions	312
Pathways	130
Complete pathways	22

Table 2. Pathways predicted to be complete in *Methanococcus jannaschii*. The total number of reactions in each pathway is shown.

Pathway	No. of reactions
Arginine biosynthesis II	9
Glycolysis	9
Methanogenesis from CO_2	9
TCA cycle, variation 1	8
Pyrimidine biosynthesis	6
Tryptophan biosynthesis	5
PRPP biosynthesis	5
Glutamate degradation VIII	5
Glyceraldehyde-3-phosphate catabolism	5
Leucine biosynthesis	4
Valine biosynthesis	4
Homoserine biosynthesis	3
UDP-glucose conversion	3
Polyamine biosynthesis, *Bacillus subtilis*	3
Threonine biosynthesis from homoserine	2
Asparagine biosynthesis and degradation	2
Glycine biosynthesis I	2
Aspartate biosynthesis II	2
Asparagine biosynthesis I	2
Ammonia assimilation cycle	2
Methionine and *S*-adenosylmethionine synthesis	2
Glutamine biosynthesis I	1

correspondence between proteins and their EC number assignments; however, we have identified six proteins with more than one EC number, and 67 EC numbers that map to more than one protein (from 2 to a maximum of 8 proteins). Proteins that are assigned more than one EC number may be paralogous enzymes, whereas proteins that share the same EC number may be subunits of the same enzyme complex. For example, the formylmethanuran dehydrogenase enzyme that catalyzes the first step of methanogenesis from CO_2 (EC 1.2.99.5) contains seven subunits (see http://maine.ebi.ac.uk:1555/MJNRDB2/new-image?type=REACTION-IN-PATHWAY&object=FORMYLMETHANOFURAN-DEHYDROGENASE-RXN).

The 609 reaction frames comprising the pathway database for *M. jannaschii* correspond to all of the reactions present in the 113 predicted metabolic pathways. The group of 609 reaction frames consists of 297 reactions that have been identified in *M. jannaschii* and 312 reactions that are currently not associated with any gene product. Of the 297 reactions that are present in *M. jannaschii*, 231 have been assigned to a pathway, and 66 reactions remain unassigned.

Analysis of the functional attributes of all 609 reactions (including reactions currently absent) revealed that 541 reactions map to a pathway, whereas 68 reactions remain unassigned. Furthermore, 511 reactions have at least one EC number and map to a total of 467 unique EC numbers, whereas the remaining 98 reactions have no associated EC number. The 461 enzymatic-reaction and 609 reaction objects correspond to 436 unique enzymes, due to enzymes involved in more than one enzymatic reaction. Finally, the metabolic reconstruction predicts the existence of 510 compounds within the metabolic network.

Pathways in which all reactions have been identified represent the most strongly supported assignments and are summarized in Table 2. The majority of pathways found to be complete in *M. jannaschii* are involved in either amino acid biosynthesis or degradation. These pathways are highly conserved across different taxonomic domains (Peregrin-Alvarez et al. 2003) and thus represent cases where pathway prediction has been particularly successful.

Also predicted in its entirety is the methanogenesis pathway. In contrast with the amino acid biosynthesis pathways, methanogenesis is a highly specific biochemical conversion cascade present in only a few organisms capable of producing methane from CO_2. Nevertheless, this pathway is well characterized biochemically (Ferry 1992, Deppenmeier 2002), and because of its specialized nature, has not diverged much. Other pathways related to energy metabolism, such as glycolysis and variants of the TCA cycle, are also well characterized and highly conserved and are thus strongly predicted (Tsoka and Ouzounis 2001).

Comparison with previous metabolic reconstructions

Previously, the most recent and most extensive analysis of the metabolic complement of *M. jannaschii* was the reconstruction reported by Selkov et al. (1997). This work was part of a wider database project for metabolic analysis (Overbeek et al. 2000), which has placed particular emphasis on *M. jannaschii* reconstruction (Selkov et al. 1997). The approach consisted of identification of enzymatic functions through sequence similarity searches and analysis of biochemical and phenotypic data (Selkov et al. 1997). Although sequence similarity searches routinely serve as the basis for computational assignments of protein function, and the results can be subjected to closer scrutiny by different researchers, phenotypic characteristics used by authors have generally not been made explicitly available and are therefore not reproducible. In contrast, MJCyc contains only computational function assignments. These assignments can be supplemented with experimental data as it becomes available, and amended as necessary.

The present reconstruction (MJCyc), with 436 proteins assigned to 266 EC numbers and 113 pathways, may be a significant improvement upon the previous metabolic reconstruction, which consisted of 245 proteins assigned to a total of 171 EC numbers and 97 pathways (Selkov et al. 1997). The two re-

constructions had 184 assignments in common; 61 enzymes were found only by Selkov et al. (1997) and 252 assignments were unique to MJCyc (the agreement between annotations is shown schematically in Figure 1). There were 149 EC numbers common to the two annotation sets, whereas 22 EC numbers were unique to the previous reconstruction and 117 EC numbers were identified only in MJCyc. It is important to emphasize that the majority of the new assignments in MJCyc are a result of updates of database records with proteins characterized since publication of the first metabolic reconstruction for this species (Selkov et al. 1997). Nevertheless, the use of Pathway Tools assists in the elimination of problems such as false positives that arise from weak sequence similarities to the query proteins, paralogous families or unclear function assignments in database entries.

In addition to identifying new reactions in metabolic pathways known to be present in *M. jannaschii*, we were able to detect enzymes involved in sulfate assimilation (phosphoadenosine phosphosulfate reductase; EC 1.8.4.8; MJ0406) and methionine synthesis from homocysteine (EC 2.1.1.14; MJ1473), as well as most reactions involved in cobalamin biosynthesis, the mevalonate pathway, and synthesis of polyamines. A complete list of pathway assignments can be found in the web-accessible MJCyc database.

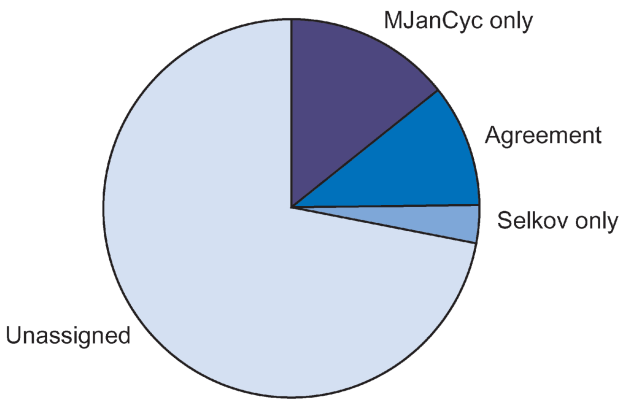

Figure 1. Agreement between enzyme annotation in MJCyc and a previous metabolic reconstruction (Selkov et al. (1997)).

Selkov et al. also identified a total of 65 EC numbers missing from the 97 pathways (i.e., these enzyme activities would complete the 97 reported pathways). We were able to identify several of these missing activities, some of which are reported in Table 3. Some of these activities are associated with reactions of key metabolic pathways such as the glycolytic cascade. These newly discovered enzymes have been identified by direct biochemical experiments either in *M. jannaschii* or in other related archaeal species. Some of these enzymes are discussed below to illustrate that *M. jannaschii* (and possibly archaea in general) employ enzymes different from those of bacteria and eukarya to catalyze some well known reactions, and that these cases may represent complex patterns of evolutionary divergence from a common ancestral sequence.

First, glycolytic and gluconeogenic reactions in *M. jannaschii* are catalyzed by phosphoglycerate mutase. This enzyme is encoded by sequences MJ0010 and MJ1612 and catalyzes the conversion of 3-phosphoglycerate to 2-phosphoglycerate. Generally, two structurally distinct forms of the enzyme are known: a cofactor-dependent and a cofactor-independent form. In archaea, phosphoglycerate mutase is only distantly related to the cofactor-independent form of the enzyme (van der Oost et al. 2002). The gene encoding the enzyme belongs to a family that is widely distributed among archaea and some bacteria, but is significantly different in sequence in eukarya and most bacteria. Phylogenetic analyses have indicated that the family arose before divergence of the archaeal lineage (Graham et al. 2002*a*).

Fructose-1,6-biphosphate aldolase is another archaeal enzyme whose sequence differs markedly from those of eukaryal and most bacterial Class I and Class II aldolases. It has been

Table 3. Reactions identified in this analysis that were previously reported as missing (Bult et al. 1996, Selkov et al. 1997). References indicate the source of annotation.

Protein	EC	Function	Reference	Pathway
MJ0010	5.4.2.1	Phosphoglycerate mutase	Graham et al. 2002*a*, van der Oost et al. 2002	Glycolysis, gluconeogenesis
MJ1612	5.4.2.1	Phosphoglycerate mutase	Graham et al. 2002*a*, van der Oost et al. 2002	Glycolysis, gluconeogenesis
MJ1585	4.1.2.13	Fructose bisphosphate aldolase	Siebers et al. 2001	Glycolysis, gluconeogenesis
MJ0316	4.1.1.19	Pyruvoyl dependent arginine decarboxylase	Graham et al. 2002*b*	Polyamine and arginine metabolism
MJ0422	1.3.1.26	Dihydrodipicolinate reductase		Lysine biosynthesis
MJ0721	2.6.1.11	Acetylornithine aminotransferase		Glutamate and arginine metabolism
MJ0862	5.3.3.2	Isopentenyl-diphosphate delta-isomerase	Kaneda et al. 2001	Lipid and polyisoprenoid biosynthesis
MJ1054	1.1.1.22	UDP-glucose 6-dehydrogenase	Kereszt et al. 1998	Carbohydrate metabolism, cell surface structure
MJ1486	2.1.2.–	Phosphoribosylglycinamide formyltransferase	Marolewski et al. 1994, Pomper and Vorholt 2001	Purine biosynthesis
MJ0656	2.7.4.14	Cytidylate kinase		Nucleotide conversion
MJ0667	2.4.2.4	Thymidine phosphorylase		Nucleotide conversion

suggested that the Class I aldolase genes found in several archaea, including *M. jannaschii*, separated very early from the gene lineages of the classical Class I and Class II aldolases, and that subsequent evolutionary events involved gene duplications with subsequent differential loss and probably some late lateral gene transfer (Siebers et al. 2001).

Twenty-three of the enzyme function assignments reported by Selkov et al. (1997) were not reproducible by sequence similarity searches. Although the origin of annotation cannot be traced, it is likely that in these cases the assignment was based on phenotypic rather than genome sequence data. For 17 of these 23 cases, the EC number assigned by Selkov et al. (1997) has been assigned in our reconstruction to a different gene product, so no significant gain in pathway prediction would be expected if these annotations were obtained.

Discussion

Despite the continuously increasing variety and sophistication of high-throughput genome analysis methods, sequence similarity remains the most widely used means of assessing genome-wide features. By implication, successful reconstruction of the entire metabolic complement of a particular species depends on (1) the specificity and sensitivity of database search algorithms through which gene products can be associated with particular reactions according to their similarity to known enzyme sequences, and (2) the phylogenetic distance from previously characterized species that form the reference set of reaction to pathway associations. For a species that is phylogenetically distant from well characterized species, it is likely that both the annotation and the pathway detection procedure will be particularly challenging. However, even in such cases, prediction of metabolic pathways for the entire genome can significantly enhance contextual analysis by directing experimental and computational approaches (e.g., remote homology searches) toward enzymatic functions that are expected to be present but have not yet been detected. Setting aside the possibility of erroneous gene prediction, these missing activities may be attributed to the use of unique biochemical routes by the target organism that are significantly different from the reference pathways used for the metabolic reconstruction. We have predicted 312 as-yet-undiscovered (missing) reactions in 113 predicted metabolic pathways that may serve as potential targets for functional genomics experiments, and which illustrate how metabolic reconstruction can enhance genome annotation.

Having identified enzymatic activities that have not yet been attributed to a particular gene through metabolic reconstruction, subsequent application of computational protocols for function assignment, which rely on contextual information rather than sequence similarity, can identify candidate genes that perform these previously absent functions. For example, all reactions in chorismate biosynthesis (implicated in biosynthesis of aromatic amino acids) have been well characterized in model genomes (De Feyter 1987). In archaea, most enzymes in this pathway have been identified through sequence homology to bacteria and eukaryotes, with the exception of shikimate kinase (EC 2.7.1.71), which catalyzes the fifth of the seven reactions in this pathway. Analysis of gene clustering on the *M. jannaschii* chromosome (Overbeek et al. 1999) ascribed the shikimate kinase activity to gene MJ1440, and this was later verified experimentally (Daugherty et al. 2001). Sequence and structural analyses have revealed that archaeal shikimate kinase has no sequence similarity to any bacterial or eukaryotic shikimate kinases and is distantly related to members of the GHMP-kinase superfamily (Daugherty et al. 2001). This was the first time that any member of this family or fold type was associated with this particular biochemical activity, revealing complex patterns of protein and metabolic network evolution.

The metabolic reconstruction in MJCyc may facilitate similar scrutiny of other currently absent biochemical activities or entire metabolic pathways (for example, cysteine biosynthesis has been reported to be absent from *M. jannaschii* and *Methanosarcina barkeri* (Kitabatake et al. 2000)). The advantage of automated approaches to metabolic reconstruction is that the results form a database of functional properties that is reproducible and directly comparable across different species and time points (Karp et al. 1996). We plan to use the PathoLogic protocol to obtain similar metabolic reconstructions for more archaeal species, in order to shed light on archaeal metabolism through comparative analyses.

Another important feature of the MetaCyc metabolic pathway knowledge base and all metabolic reconstructions generated with the PathoLogic protocol is the availability of structured and flexible query capabilities that are particularly well suited to large-scale computational analyses (Paley and Karp 2002). These features are enabled by the object-oriented nature of the knowledge base (Karp and Paley 1994) and by the formal ontology specifying the relationships between biological objects of the knowledge base (Karp 2000). Two examples of specific biological queries are provided here. First, we have classified all *M. jannaschii* pathways according to their wider biochemical role by retrieving all 130 pathways and assigning them to one of 17 pathway functional classes (e.g., energy metabolism, amino acid biosynthesis, etc.). The result is presented schematically in Figure 2. A second example is the determination of whether each reaction is present in *M. jannaschii* (i.e., if it is one of the 312 missing reactions); if it is not, the associated pathway is retrieved. The result, shown in Figure 3, indicates the number of missing reactions in each pathway.

There are many examples of metabolic rarities in *M. jannaschii*. First, it has been shown that *M. jannaschii* has a dual specificity tRNA synthetase: MJ1238 encodes both prolyl (EC 6.1.1.15) and cysteinyl-tRNA synthetase (EC 6.1.1.16) activities (Bunjun et al. 2000, Lipman et al. 2000, Stathopoulos et al. 2000); the cysteinyl-tRNA synthetase activity had not been previously identified (Selkov et al. 1997). Second, it has been shown that polyamines are present at high concentrations in *M. jannaschii* (Graham et al. 2000*a*, Kim et al. 2000, Sekowska et al. 2000). MJCyc includes four of the five enzymes in the polyamine biosynthesis pathway (MJ0315, MJ0316, MJ0309 and MJ0313), with ornithine decarboxylase currently

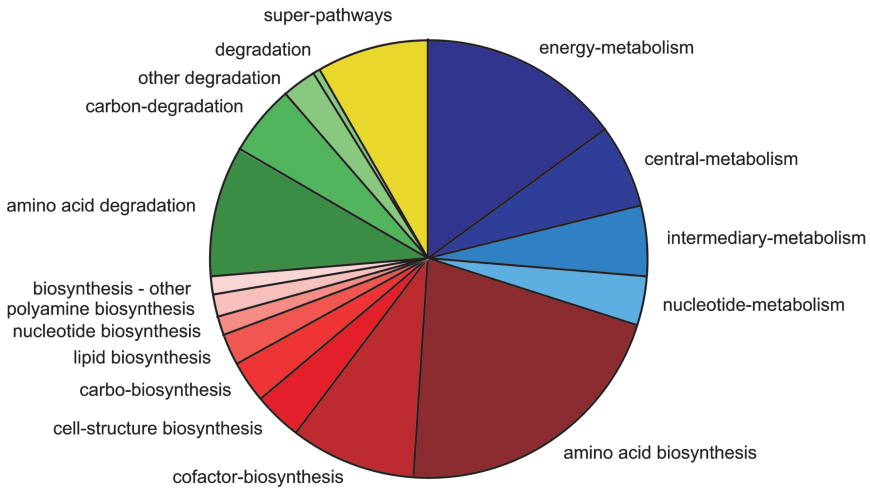

Figure 2. Classification of predicted metabolic pathways according to the type of biochemical function they perform. Shades of blue, red, green and yellow indicate central metabolism, biosynthetic functions, degradation pathways and super-pathways, respectively.

missing. *S*-Adenosylmethionine (AdoMet) decarboxylase (MJ0315) activity in *M. jannaschii* was difficult to identify because of a high degree of divergence of the sequence encoding this enzyme from the sequences of other enzymes already known to possess this activity. However, since detection of this activity, it has been possible to show that spermidine biosynthesis in Gram-positive bacteria and in archaea is similar to the related pathway in Gram-negative bacteria and in eukarya (Sekowska et al. 2000). Finally, MJCyc provides strong evidence for the presence of the pathway for biosynthesis of Coenzyme M, an important cofactor in methanogenesis: the first two reactions have been identified in *M. jannaschii* and attributed to this pathway (MJ0255, 2*r*-phospho-3-sulfolactate synthase and MJ1140, 2-phosphosulfolactate phosphatase) (Graupner et al. 2000, Graham et al. 2001*a*).

In conclusion, the metabolic reconstruction of the entire genome of *M. jannaschii* is presented, and significant improvements have been noted compared with previous analyses. Comparative analyses of metabolic reconstruction aid in the refinement of these methods and contribute to a deeper understanding of the process of genome annotation (Paley and Karp 2002). We have illustrated the potential of this approach for providing a comprehensive view of what is known about the metabolism of a given organism, as well as for supporting the search for unknown metabolic components. Given that functional genomics projects aim to increase the coverage of functional types encoded in genomes, the iterative procedure of continuously improving genome annotation and metabolic pathway prediction is expected to yield complete metabolic maps in the future.

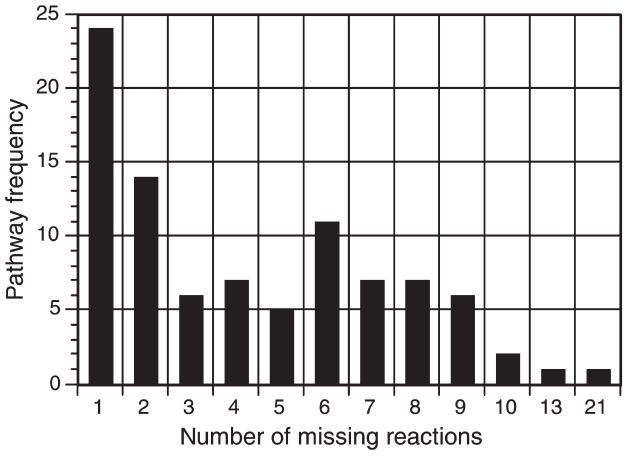

Figure 3. Frequency distribution of missing reactions with respect to pathways. This figure demonstrates an example of use of the MJCyc database. Overall, 312 missing reactions were linked to a total of 91 pathways. Details are available at http://maine.ebi.ac.uk:1555/server.html.

Acknowledgments

We thank anonymous referees for comments and colleagues at the Computational Genomics Group for discussions. S.T. is supported by a Special Training Fellowship in Bioinformatics from the Medical Research Council (U.K.). C.O. acknowledges further support for his laboratory by the EMBL, the British Council and IBM Research.

References

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Bono, H., H. Ogata, S. Goto and M. Kanehisa. 1998. Reconstruction of amino acid biosynthesis pathways from the complete genome sequence. Genome Res. 8:203–210.

Bult, C.J., O. White, G.J. Olsen et al. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 273:1058–1073.

Bunjun, S., C. Stathopoulos, D. Graham, B. Min, M. Kitabatake, A.L. Wang, C.C. Wang, C.P. Vivares, L.M. Weiss and D. Soll. 2000. A dual-specificity aminoacyl-tRNA synthetase in the deep-rooted eukaryote *Giardia lamblia*. Proc. Natl. Acad. Sci. 97:12,997–13,002.

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

De Feyter, R. 1987. Shikimate kinases from *Escherichia coli* K12. Methods Enzymol. 142:355–361.

Deppenmeier, U. 2002. The unique biochemistry of methanogenesis. Prog. Nucleic Acid Res. Mol. Biol. 71:223–283.

Eisenberg, D., E.M. Marcotte, I. Xenarios and T.O. Yeates. 2000. Protein function in the post-genomic era. Nature 405:823–826.

Ferry, J.G. 1992. Biochemistry of methanogenesis. Crit. Rev. Biochem. Mol. Biol. 27:473–503.

Graham, D.E., C.L. Bock, C. Schalk-Hihi, Z.J. Lu and G.D. Markham. 2000*a*. Identification of a highly diverged class of S-adenosylmethionine synthetases in the archaea. J. Biol. Chem. 275: 4055–4059.

Graham, D.E., R. Overbeek, G.J. Olsen and C.R. Woese. 2000*b*. An archaeal genomic signature. Proc. Natl. Acad. Sci. 97:3304–3308.

Graham, D.E., M. Graupner, H. Xu and R.H. White. 2001*a*. Identification of coenzyme M biosynthetic 2-phosphosulfolactate phosphatase. A member of a new class of $Mg^{(2+)}$-dependent acid phosphatases. Eur. J. Biochem. 268:5176–5188.

Graham, D.E., N. Kyrpides, I.J. Anderson, R. Overbeek and W.B. Whitman. 2001*b*. Genome of *Methanocaldococcus* (*Methanococcus*) *jannaschii*. Methods Enzymol. 330:40–123.

Graham, D.E., H. Xu and R.H. White. 2002*a*. A divergent archaeal member of the alkaline phosphatase binuclear metalloenzyme superfamily has phosphoglycerate mutase activity. FEBS Lett. 517:190–194.

Graham, D.E., H. Xu and R.H. White. 2002*b*. *Methanococcus jannaschii* uses a pyruvoyl-dependent arginine decarboxylase in polyamine biosynthesis. J. Biol. Chem. 277:23,500–23,507.

Graupner, M., H. Xu and R.H. White. 2000. Identification of the gene encoding sulfopyruvate decarboxylase, an enzyme involved in biosynthesis of coenzyme M. J. Bacteriol. 182:4862–4867.

Janssen, P.J., B. Audit, I. Cases et al. 2003. Beyond 100 genomes. Genome Biol. 4:402.

Kaneda, K., T. Kuzuyama, M. Takagi, Y. Hayakawa and H. Seto. 2001. An unusual isopentenyl diphosphate isomerase found in the mevalonate pathway gene cluster from *Streptomyces* sp. strain CL190. Proc. Natl. Acad. Sci. 98:932–937.

Kanehisa, M., S. Goto, S. Kawashima and A. Nakaya. 2002. The KEGG databases at GenomeNet. Nucleic Acids Res. 30:42–46.

Karp, P.D. 2000. An ontology for biological function based on molecular interactions. Bioinformatics 16:269–285.

Karp, P.D. and S.M. Paley. 1994. Representations of metabolic knowledge: pathways. Proc. Int. Conf. Intell. Syst. Mol. Biol. 2: 203–211.

Karp, P.D., C. Ouzounis and S. Paley. 1996. HinCyc: a knowledge base of the complete genome and metabolic pathways of *H. influenzae*. Proc. Int. Conf. Intell. Syst. Mol. Biol. 4:116–124.

Karp, P.D., S. Paley and P. Romero. 2002*a*. The Pathway Tools software. Bioinformatics 18(Suppl 1):S225–232.

Karp, P.D., M. Riley, S.M. Paley and A. Pellegrini-Toole. 2002*b*. The MetaCyc Database. Nucleic Acids Res. 30:59–61.

Kereszt, A., E. Kiss, B.L. Reuhs, R.W. Carlson, A. Kondorosi and P. Putnoky. 1998. Novel *rkp* gene clusters of *Sinorhizobium meliloti* involved in capsular polysaccharide production and invasion of the symbiotic nodule: the *rkpK* gene encodes a UDP-glucose dehydrogenase. J. Bacteriol. 180:5426–5431.

Kim, A.D., D.E. Graham, S.H. Seeholzer and G.D. Markham. 2000. S-Adenosylmethionine decarboxylase from the archaeon *Methanococcus jannaschii*: identification of a novel family of pyruvoyl enzymes. J. Bacteriol. 182:6667–6672.

Kitabatake, M., M.W. So, D.L. Tumbula and D. Soll. 2000. Cysteine biosynthesis pathway in the archaeon *Methanosarcina barkeri* encoded by acquired bacterial genes? J. Bacteriol. 182:143–145.

Kyrpides, N., R. Overbeek and C. Ouzounis. 1999. Universal protein families and the functional content of the last universal common ancestor. J. Mol. Evol. 49:413–423.

Lipman, R.S., K.R. Sowers and Y.M. Hou. 2000. Synthesis of cysteinyl-tRNA(Cys) by a genome that lacks the normal cysteine-tRNA synthetase. Biochemistry 39:7792–7798.

Marolewski, A., J.M. Smith and S.J. Benkovic. 1994. Cloning and characterization of a new purine biosynthetic enzyme: a non-folate glycinamide ribonucleotide transformylase from *E. coli*. Biochemistry 33:2531–2537.

Overbeek, R., M. Fonstein, M. D'Souza, G.D. Pusch and N. Maltsev. 1999. The use of gene clusters to infer functional coupling. Proc. Natl. Acad. Sci. 96:2896–2901.

Overbeek, R., N. Larsen, G.D. Pusch, M. D'Souza, E. Selkov, Jr., N. Kyrpides, M. Fonstein, N. Maltsev and E. Selkov. 2000. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Res. 28:123–125.

Paley, S.M. and P.D. Karp. 2002. Evaluation of computational metabolic-pathway predictions for *Helicobacter pylori*. Bioinformatics 18:715–724.

Peregrin-Alvarez, J.M., S. Tsoka and C.A. Ouzounis. 2003. The phylogenetic extent of metabolic enzymes and pathways. Genome Res. 13:422–427.

Pomper, B.K. and J.A. Vorholt. 2001. Characterization of the formyltransferase from *Methylobacterium extorquens* AM1. Eur. J. Biochem. 268:4769–4775.

Promponas, V.J., A.J. Enright, S. Tsoka, D.P. Kreil, C. Leroy, S. Hamodrakas, C. Sander and C.A. Ouzounis. 2000. CAST: an iterative algorithm for the complexity analysis of sequence tracts. Bioinformatics 16:915–922.

Sekowska, A., J.Y. Coppee, J.P. Le Caer, I. Martin-Verstraete and A. Danchin. 2000. S-adenosylmethionine decarboxylase of *Bacillus subtilis* is closely related to archaebacterial counterparts. Mol. Microbiol. 36:1135–1147.

Selkov, E., N. Maltsev, G.J. Olsen, R. Overbeek and W.B. Whitman. 1997. A reconstruction of the metabolism of *Methanococcus jannaschii* from sequence data. Gene 197:GC11–26.

Siebers, B., H. Brinkmann, C. Dorr, B. Tjaden, H. Lilie, J. van der Oost and C.H. Verhees. 2001. Archaeal fructose-1,6-bisphosphate aldolases constitute a new family of archaeal type class I aldolase. J. Biol. Chem. 276:28,710–28,718.

Stathopoulos, C., T. Li, R. Longman, U.C. Vothknecht, H.D. Becker, M. Ibba and D. Soll. 2000. One polypeptide with two aminoacyl-tRNA synthetase activities. Science 287:479–482.

Tsoka, S. and C.A. Ouzounis. 2000. Recent developments and future directions in computational genomics. FEBS Lett. 480:42–48.

Tsoka, S. and C.A. Ouzounis. 2001. Functional versatility and molecular diversity of the metabolic map of *Escherichia coli*. Genome Res. 11:1503–1510.

van der Oost, J., M.A. Huynen and C.H. Verhees. 2002. Molecular characterization of phosphoglycerate mutase in archaea. FEMS Microbiol. Lett. 212:111–120.

Woese, C.R. and G.E. Fox. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. 74: 5088–5090.

Wolfe, R.S. 1992. Biochemistry of methanogenesis. Biochem. Soc. Symp. 58:41–49.