Published in final edited form as:
Science. 2009 September 18; 325(5947): 1544–1549. doi:10.1126/science.1174671.

Three-dimensional Structural View of the Central Metabolic Network of *Thermotoga maritima*

Ying Zhang[1,*], **Ines Thiele**[2,*,%], **Dana Weekes**[3], **Zhanwen Li**[1], **Lukasz Jaroszewski**[3], **Krzysztof Ginalski**[4], **Ashley M. Deacon**[5], **John Wooley**[6], **Scott A. Lesley**[7], **Ian A. Wilson**[8], **Bernhard Palsson**[2], **Andrei Osterman**[9], and **Adam Godzik**[1,3,6,‖]

[1]Joint Center for Molecular Modeling, Burnham Institute for Medical Research La Jolla, CA 92037, USA [2]Department of Bioengineering, University of California at San Diego, La Jolla, CA 92093-0412, USA [3]Joint Center for Structural Genomics, Bioinformatics Core, Burnham Institute for Medical Research, La Jolla, CA 92037, USA [4]Interdisciplinary Centre for Mathematical and Computational Modelling, Warsaw University, Warsaw, Poland [5]Joint Center for Structural Genomics, Structure Determination Core, Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA [6]Joint Center for Structural Genomics, Bioinformatics Core, University of California at San Diego, La Jolla, CA 92093, USA [7]Joint Center for Structural Genomics, Crystallomics Core, Genomics Institute of the Novartis Research Foundation, San Diego, CA 92121, USA [8]Joint Center for Structural Genomics, The Scripps Research Institute, La Jolla, CA 92037, USA [9]Burnham Institute for Medical Research, La Jolla, CA 92037, USA

Abstract

Metabolic pathways have traditionally been described in terms of biochemical reactions and metabolites. Using structural genomics and systems biology, we generated a three-dimensional reconstruction of the central metabolic network of the bacterium, *Thermotoga maritima* (*TM*). The network encompassed 478 proteins of which 120 were determined by experiment and 358 were modeled. Structural analysis revealed that proteins forming the network are dominated by a small number (only 182) of basic shapes (folds) performing diverse, but mostly related functions. Most of these folds are already present in the essential core (~30%) of the network, and its expansion by nonessential proteins is achieved with relatively few additional folds. Thus, integration of structural data with networks analysis generates insight into the function, mechanism and evolution of biological networks.

The advent of genome sequencing has enabled development of computational and experimental tools to investigate complete biological systems, but has also highlighted the difficulty in integrating complex information for the hundreds to thousands of different molecules that compose even the smallest biological networks. Such integration presents many challenges, especially when assembling data from diverse fields, such as biochemistry and structural biology, which use different operational languages and conceptual frameworks. Biochemistry has traditionally focused on individual reactions and pathways, but recent advances in genomics have led to more rapid growth in the reconstruction and modeling of metabolic networks on a genome-wide scale (1–3). Thus, biochemical reactions, pathways, and networks

[‖]To whom correspondence should be addressed. adam@burnham.org.
[*]These authors contributed equally to this work
[%]Current address: Center of Systems Biology, University of Iceland, IS-101 Reykjavik, Iceland

One-line Summary
Atomic-level reconstruction of the metabolic network of a free-living bacterium gives insights into the evolution of biological systems.

can now be described in the context of entire cells, thereby enabling more realistic simulations of the behavior of metabolic networks in a growing number of organisms (4–7). Nevertheless, metabolism is still generally defined in terms of the chemical names and identity of substrates, products, and reactions. It does not explicitly consider the three-dimensional structures of its components, although such knowledge is required for a comprehensive understanding not only of the individual reactions but, more importantly, of metabolic networks as a whole. Without such knowledge, we cannot rigorously define enzyme mechanisms or predict the effects of mutations or drugs, and, on the global level, understand the evolutionary relationships between different pathways, how new metabolic capabilities are acquired, and how individual organisms adapt to their particular ecological niches and respond to environmental pressures.

Such an understanding can be provided by structural biology, which has traditionally focused on individual proteins outside of their full, system-level, biological context. The emergence of large-scale structure genomics projects, such as the Protein Structure Initiative (8), has provided an exciting new opportunity for structural biology to contribute on a scale similar to genomics.

Thermotoga maritima, one of the first discovered hyper-thermophilic bacteria (9), represents the deepest known lineage of eubacteria (9,10), has one of the smallest genomes for a free-living organism (11) and has been the subject of extensive experimental analysis (12,13), making it an ideal model organism for systems biology and for integration of biochemical and structural approaches (14).

We constructed a metabolic model of *T. maritima* by a “bottom-up” approach, which first identified all known biochemical reactions and substrates from almost 150 publications (Table S3), providing direct biochemical, genomic, and physiological evidence for more than 50% of the metabolic reactions. The remaining reactions were then identified from high confidence, homology-based prediction annotation databases (15,16) and from experimental or modeled protein structures (see below). Flux Balance Analysis (17) was used to test the completeness of the network, revealing gaps, such as missing enzymes or redundant functional assignments, which were then resolved by manual curation for individual cases. Iterative evaluation was continued until the performance reproduced, in silico, the experimentally determined metabolic capabilities of *T. maritima* (18, tables S9, S10).

Our resulting metabolic reconstruction included 478 metabolic genes, 503 unique metabolites and 562 intracellular and 83 extracellular metabolic reactions (18), and reproduced *T. maritima's* ability to grow on diverse carbohydrates (Table S9) and to produce known metabolic by-products, e.g., acetate and hydrogen. The overall scope, content, and quality of this metabolic reconstruction were comparable with state-of-the-art reconstructions for other model organisms (Table S6). Although the current model does not yet provide an exhaustive description of *T. maritima* metabolism, it represents a major step in an iterative process of annotation and modeling of this organism.

The *T. maritima* metabolic reconstruction (mr) defines a specific set of proteins (mrTM) that carry out the biochemical functions that comprise a self-sustaining, metabolic network. Of 478 proteins in this mrTM set, structures of 120 proteins have been determined experimentally (12) and 358 were predicted and modeled using a variety of computational approaches (18). The quality of the modeled structures spans the spectrum from those comparable to low-resolution, experimental structures (190 were built on templates with over 30% identity to the targets) to very approximate (52 were based only on fold predictions). For three (TM1444, TM0788 and TM0540), the automated structure prediction approach failed and approximate structures were modeled by combining several different fold prediction algorithms with manual refinement (18). Quality control, as based on public benchmarks in modeling and fold recognition, suggests high confidence that all models are correct at the fold-assignment level (18). Thus, these combined approaches allowed us to achieve complete structural coverage for the mrTM set (Fig. 1).

The information from structural determination of *T. maritima* proteins and their homologs provided additional support for functional assignment of 181 individual genes. A total of 41 experimental structures of *T. maritima* proteins contained relevant metabolites and 140 crystal structures, used as templates for homology modeling, were also determined as complexes with ligands, all of which support the functional assignment in the reconstruction. In at least two cases, TM0449 (19–22) and TM1643 (23), structural analysis was critical for identification of enzymatic function and, in many other cases, substantially contributed to assignment of function.

Metabolic reconstruction can be described not only in a mathematical format of a matrix that can be used for metabolic simulations to predict essential genes or growth rates, but also can be represented as a graph. Because the reconstruction represents a fully functional, cell-level model of a metabolic network, analysis of the topology of this graph allows us to answer many interesting questions, especially when combined with knowledge of structures or models for all proteins in the network. For instance, what is the dominant mechanism for expansion of a metabolic network in a single organism? In the “patchwork” hypothesis (24), network expansion is driven by recruitment of proteins that perform similar reactions, but are present in distinct pathways. Conversely, in the “retrograde” hypothesis (25), novel proteins are recruited to perform dissimilar reactions, but reside in the same pathway or neighboring parts of the network. Analysis of fold conservation as a function of network topology, therefore, addresses this key issue. Similar analyses have been performed previously on a small set of known pathways (26,,27), but our integrated approach allowed us to analyze the complete set of pathways that form the fully functional, self-sustained metabolic network of a single organism.

We then established an automated protocol to classify metabolic reactions into three categories: Similar (S), Connected (C), and Unrelated (U) (Fig. S6 and Fig. 2). Enzymes that catalyze similar types of reactions have a six-fold higher probability of having the same fold than enzymes catalyzing connected reactions (Fig. 2c), supporting the “patchwork” hypothesis (24). However, it should be noted that proteins catalyzing connected (C) reactions still have a higher chance of having the same fold as those catalyzing unrelated reactions (U), suggesting a role for gene duplication within pathways during pathway evolution, i.e. the retrograde model. More importantly, the “patchwork” hypothesis can account for only 11% of the observed structural similarity between mrTM proteins of similar function, indicating that convergent evolution of similar reaction mechanisms [i.e. non-homologous gene displacement (28), where two non-homologous proteins perform the same or similar metabolic function] is not a rare event and significantly contributes to evolution of the central metabolic network.

Another interesting question is the distribution and frequency of protein folds in this mrTM set. The 478 proteins contain 714 domains, but only 182 distinct folds, which are substantially fewer than would be expected (~300) for an equivalent random set of proteins with known structures (Fig. S8). The surprisingly small number of folds arises from the fact that the most popular folds (e.g., the P-loop, TIM barrel, and Rossmann folds) are overrepresented compared to their frequency in the general protein population (Fig. 3). Some relatively rare folds, abundant in the mrTM set, such as the biotin synthetase and the thiamin diphosphate binding folds, include groups of enzymes that perform specific, but essential functions, such as tRNA aminoacylation or carbon metabolism.

The most obvious interpretation of this skewed fold distribution is that the mrTM set, which covers the most fundamental protein functions, consists of the most ancient and, thus, the most abundant protein families. To probe this interpretation further, we analyzed the fold distribution for the core of the *T. maritima* metabolic network, as represented by the set of essential proteins. We identified essential proteins by a reductive evolution simulation approach (18,29) where iterative simulations are performed to identify a minimal network by randomly eliminating genes from the model until additional elimination would result in a non-viable network. Each simulation led to a different minimal network, of size anywhere between 200 and 300 genes, i.e. corresponding to 42–63% of the mrTM set. Statistical analysis of 1,000 such minimal networks in independent simulations in glucose minimal medium (18) allowed classification of genes from the mrTM set into three categories: (I) core- or unconditional-essential genes that are always present; (II) non-essential genes that never appear; and (III) “synthetic lethal” or “conditional-essential” genes (30) that appear only in some simulations, but not in others, depending on which other genes are removed or retained in a particular network minimization. For example, if two genes have the same, essential function, deletion of either would not be lethal, but at least one has to be present in the minimal network. The frequency of such genes in multiple simulations reflects the topology of the network and the relative redundancy of gene functions in the network. It is important to emphasize that the core-essential genes would not be sufficient to maintain a viable metabolic network, as all of the many possible minimal networks contain constant (core-essential genes) and variable (subset of conditionally-essential genes) components. The mrTM set consists of 177 core-essential, 203 non-essential, and 98 conditional-essential genes. Proteins in these three sets have very different fold distributions (Fig. 4). The number of folds in the core-essential group is surprisingly large for its sample size (111 folds for 177 proteins) compared to the non-essential group, which contains more proteins, but a smaller number of folds (92 folds for 203 proteins). This trend is inverse to that observed when mrTM is compared with non-redundant sequences in the NCBI database (33) (Fig. S8), where the mrTM set was more abundant in popular folds. These analyses suggest that core-essential proteins perform unique chemical functions that are strongly associated with specific folds and are so fundamental that their deletion would result in a non-viable network.

In summary, we present here integration of a metabolic and structural view of the central metabolic network of a thermophilic bacterium, *T. maritima*. Achieving a complete description on these two levels is an important milestone that now enables large-scale analyses, such as network-scale comparison of correlations between fold conservation and biochemical function. From our study, we can provide not only a quantitative estimate of the dominance of the patchwork model (24) versus the retrograde model (25) of metabolic evolution, but also illustrate the importance of convergent or parallel evolution in proteins carrying out similar biochemical functions. We further show that the set of proteins responsible for the central metabolism in *T. maritima* is highly non-random and dominated by a small number of folds that significantly exceed their already dominant distribution in the protein universe suggesting that the central metabolism network has evolved mainly from a set of the most ancient proteins that have had sufficient time to develop divergent functionalities and, hence, expand into the very large and very diverse protein families that we observe today. At the same time, the subset of core-essential proteins reverses this trend and is relatively more diverse than an equivalent subset of non-essential proteins. This counterintuitive situation is due to the presence of some specific folds with functions that are so unique that it is impossible to replace them with other existing folds.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References and Notes

1. Karp PD. Science 2001;293:2040. [PubMed: 11557880]
2. Ideker T, Lauffenburger D. Trends Biotechnol 2003;21:255. [PubMed: 12788545]
3. Joyce AR, Palsson BO. Prog. Drug. Res 2007;64:265. [PubMed: 17195479]
4. Duarte NC, et al. Proc. Natl. Acad. Sci. U.S.A 2007;104:1777. [PubMed: 17267599]
5. Feist AM, et al. Mol. Syst. Biol 2007;3:121. [PubMed: 17593909]
6. Durot M, et al. BMC Syst. Biol 2008;2:85. [PubMed: 18840283]
7. Senger RS, Papoutsakis ET. Biotechnol. Bioeng 2008;101:1036. [PubMed: 18767192]
8. Matthews BW. Nat. Struct. Mol. Biol 2007;14:459. [PubMed: 17549078]
9. Huber R, et al. Arch. Microbiol 1986;144:324.
10. Woese CR. Microbiol. Rev 1987;51:221. [PubMed: 2439888]
11. Nelson KE, et al. Nature 1999;399:323. [PubMed: 10360571]
12. Lesley SA, et al. Proc. Natl. Acad. Sci. U.S.A 2002;99:11664. [PubMed: 12193646]
13. Conners SB, et al. FEMS Microbiol. Rev 2006;30:872. [PubMed: 17064285]
14. "Integrative Biology of *Thermotoga maritima*". San Diego, CA: Workshop; 2007 Jul 9–10. 2007 http://metagenomics.calit2.net/2007/thermotoga/
15. Okuda S, et al. Nucleic Acids Res 2008;36:W423. [PubMed: 18477636]
16. Overbeek R, et al. Nucleic Acids Res 2005;33:5691. [PubMed: 16214803]
17. Schilling CH, et al. Biotechnol. Bioeng 2000;71:286. [PubMed: 11291038]
18. Materials and methods are available as supporting material on *Science* Online.
19. Kuhn P, et al. Proteins 2002;49:142. [PubMed: 12211025]
20. TM0449 is a novel, FAD-dependent thymidylate synthase and our structure has contributed to new developments in functional studies of this and related proteins (see refs. 21,22 and references therein).
21. Murzin AG. Science 2002;297:61. [PubMed: 12029066]
22. Koehn EM, et al. Nature 2009;458:919. [PubMed: 19370033]
23. Yang Z, et al. J. Biol. Chem 2003;278:8804. [PubMed: 12496312]
24. Jensen RA. Annu. Rev. Microbiol 1976;30:409. [PubMed: 791073]
25. Horowitz NH. Proc. Natl. Acad. Sci. U.S.A 1945;31:153. [PubMed: 16578152]
26. Holliday GL, et al. Nat. Prod. Rep 2007;24:972. [PubMed: 17898893]
27. Rison SC, Thornton JM. Curr. Opin. Struct. Biol 2002;12:374. [PubMed: 12127458]
28. Koonin EV, Mushegian AR, Bork P. Trends Genet 1996;12:334. [PubMed: 8855656]
29. Pal C, et al. Nature 2006;440:667. [PubMed: 16572170]
30. Hartman JL, Garvik B, Hartwell L. Science 2001;291:1001. [PubMed: 11232561]
31. Yang C, et al. J. Bacteriol 2008;190:1773. [PubMed: 18156253]
32. Murzin AG, Brenner SE, Hubbard T, Chothia C. J. Mol. Biol 1995;247:536. [PubMed: 7723011]
33. Pruitt KD, Tatusova T, Maglott DR. Nucleic Acids Res 2007;35:D61. [PubMed: 17130148]

34. We specifically acknowledge the invaluable work of individual crystallographers at the JCSG and other PSI centers, as well as individual research groups, who have solved structures analyzed here, either directly or that we used as modeling templates. The full list of these proteins is provided in the Supporting Online Material. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIGMS. This work was supported by the NIH Protein Structure Initiative grants P20 GM076221 (JCMM) and U54 GM074898 (JCSG) from the National Institute of General Medical Sciences, grant DE-FG02-08ER64686 from the Office of Science (BER) U.S. Department of Energy and by the Gordon and Betty Moore Foundation CAMERA project.

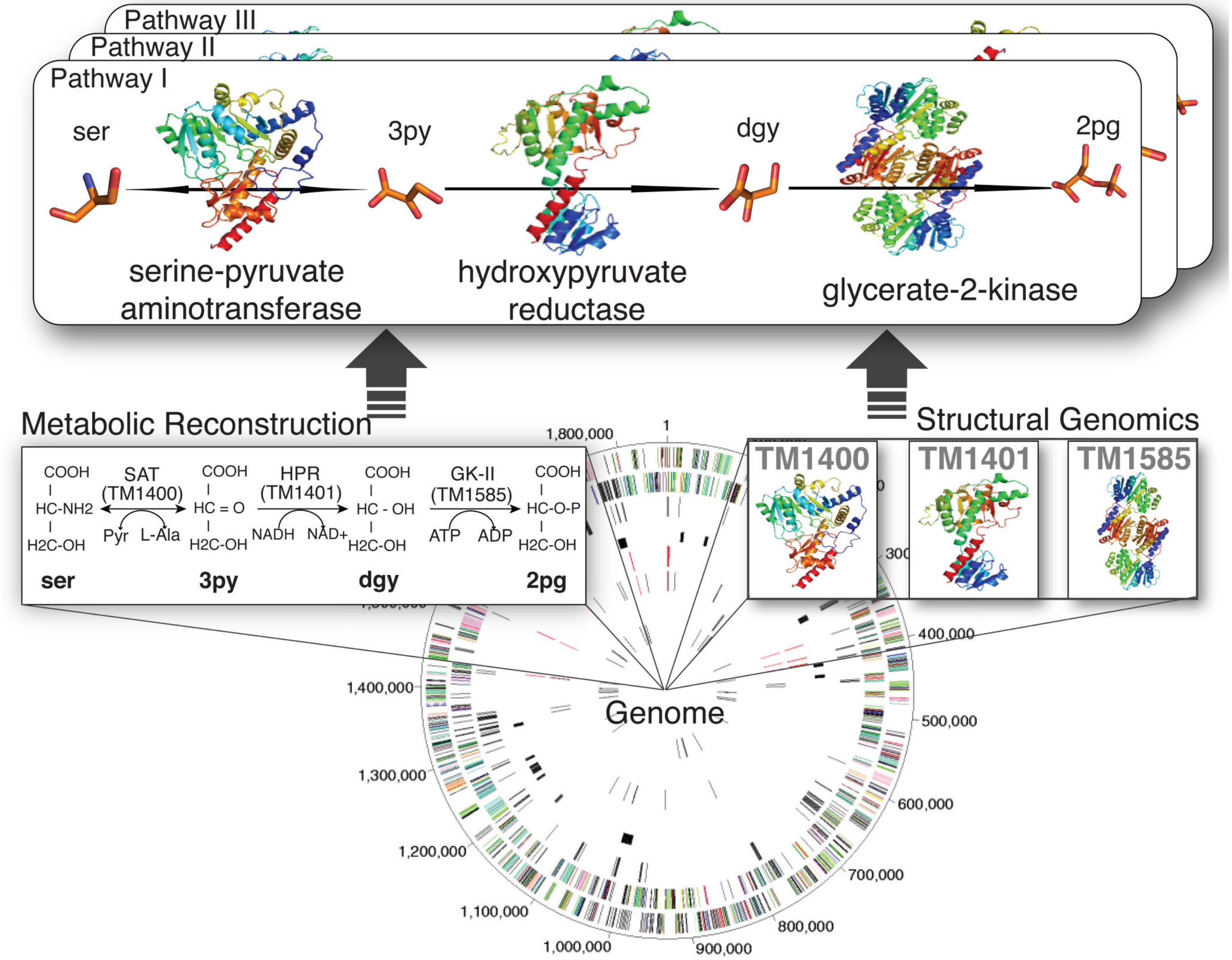
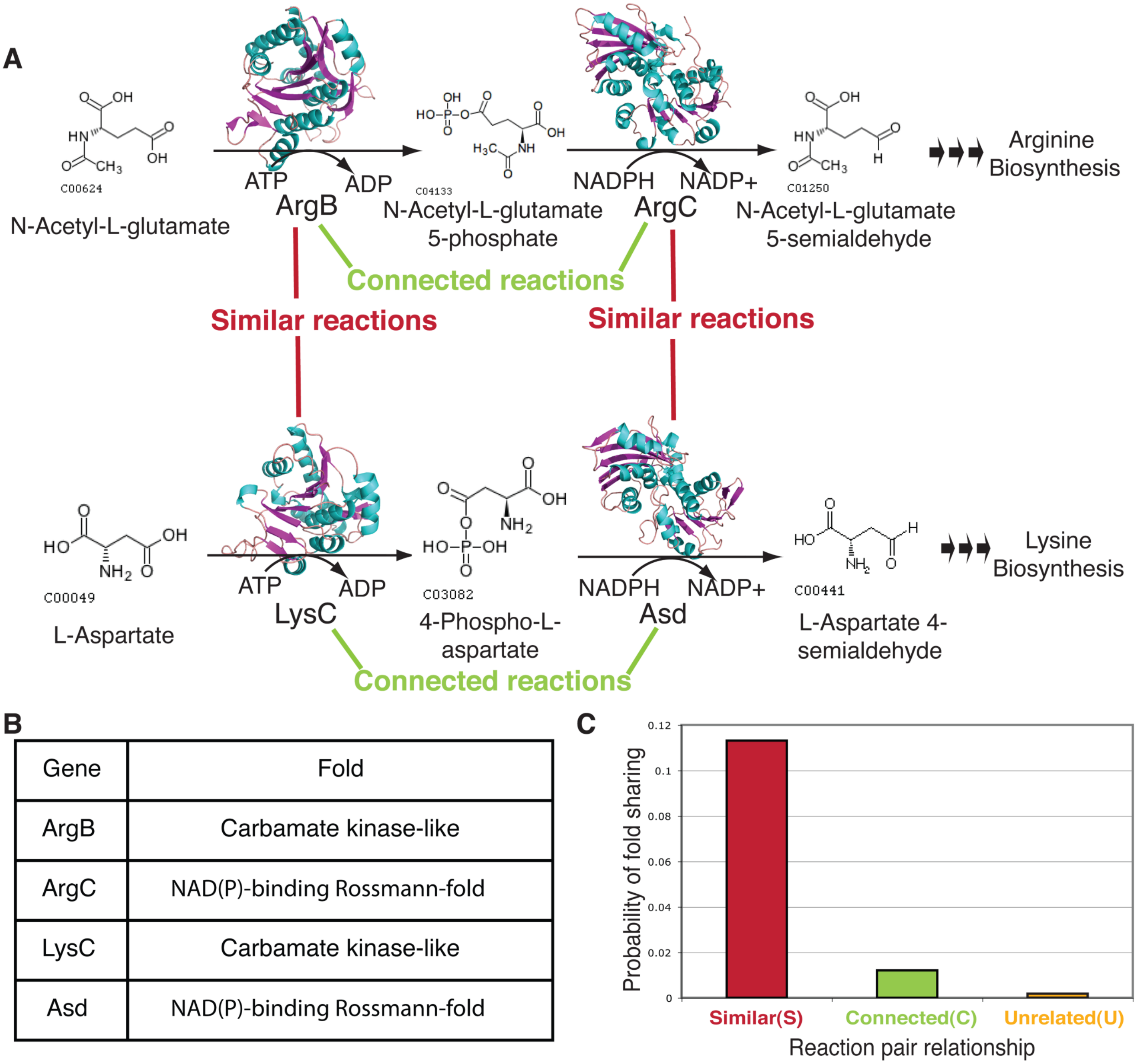


Fig. 1.

Combining metabolic reconstruction and structural genomics approaches for an integrated annotation of the *T. maritima* central metabolic network. Underlying genomics information (bottom) enabled both a metabolic reconstruction (left subpanel) and an atomic-level structure determination/modeling of all *T. maritima* proteins (right subpanel). Integration of these two approaches enabled detailed information to be acquired for every reaction in the network (upper subpanel); an example from the *T. maritima* serine degradation pathway is illustrated (28).



Gene	Fold
ArgB	Carbamate kinase-like
ArgC	NAD(P)-binding Rossmann-fold
LysC	Carbamate kinase-like
Asd	NAD(P)-binding Rossmann-fold

Fig. 2.

Classification of metabolic reactions. (**A**) Examples of Similar (S), Connected (C), and Unrelated (U) reactions from the Arginine and Lysine biosynthesis pathways. ArgB and LysC share a co-substrate (ATP) that undergoes the same transformation (to ADP + Pi). Similarly, ArgC and Asd transform NADPH to NADP+. By these criteria, both pairs are classified as Similar (S). At the same time, reaction pairs ArgB/ArgC and LysC/Asd are adjacent in the pathway, since the product of the first reaction is the substrate for the next. These reaction pairs are classified as Connected (C). All other pairs of reactions (ArgB/Asd, ArgC/LysC) are classified as Unrelated (U). In this example, only the enzymes classified as similar (ArgB/LysC and ArgC/Asd) have the same fold. (**B**) Detailed information on the enzymes in subpanel A. (**C**) Bars representing the relative number of pairs with the same fold in each category of reactions.

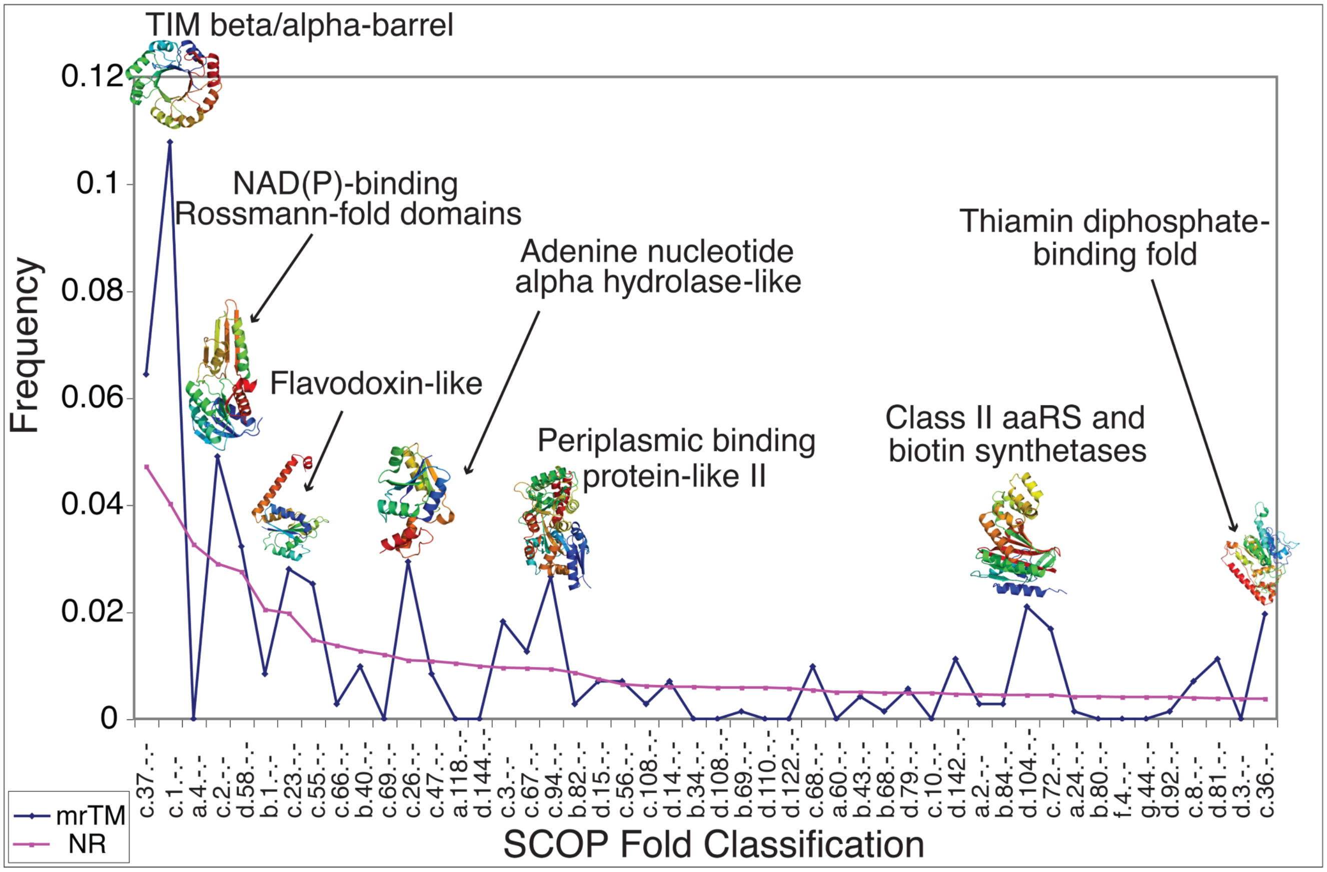


Fig. 3.

Distribution of folds in the mrTM protein set with the most overrepresented folds illustrated by structural ribbon diagrams. SCOP (32) fold codes are shown on the x-axis with the observed frequency on the y-axis. The expected frequency for each fold in the NCBI non-redundant database (33) is shown as a magenta trace.

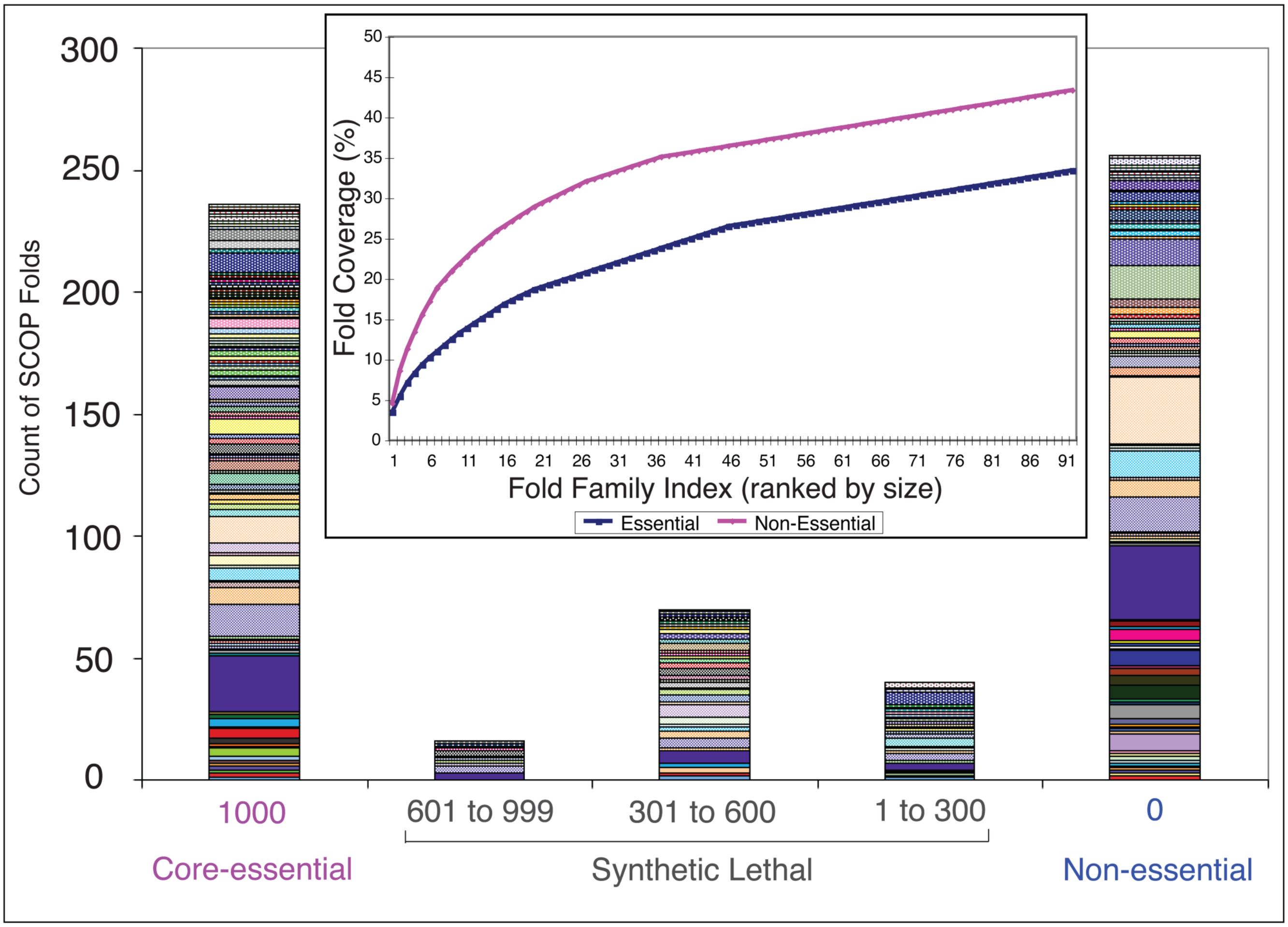


Fig. 4.

Fold composition of the non-essential, synthetic lethal, and core-essential protein sets (see text for details) illustrated by colors associated with different folds. The x-axis represents the number of simulations that resulted in identification of core-essential (1000 appearances in 1000 simulations), synthetic lethal (from 999 to 1), and non-essential genes (0), and their classification on the y-axis into SCOP fold categories. Inset: cumulative fold coverage of core-essential and non-essential protein sets (blue: core-essential; magenta: non-essential). Note the fold distribution in all three groups is different, although core-essential and non-essential have some weak similarity than either group compared to synthetic lethal.