Supporting Online Material for

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

Ying Zhang, Ines Thiele, Dana Weekes, Zhanwen Li, Lukasz Jaroszewski, Krzysztof Ginalski, Ashley M. Deacon, John Wooley, Scott A. Lesley, Ian A. Wilson, Bernhard Palsson, Andrei Osterman, Adam Godzik*

*To whom correspondence should be addressed. E-mail: adam@burnham.org

Published 18 September 2009, *Science* **325**, 1544 (2009)
DOI: 10.1126/science.1174671

This PDF file includes:

Other Supporting Online Material for this manuscript includes the following: (available at www.sciencemag.org/cgi/content/full/325/5947/1544/DC1)

Tables S1 to S13 as a zipped archive: 1174671_supp_tables.zip
Metabolic reconstruction in SMBL and MATLAB formats as a zipped archive: 1174671_supp_reconstruction.zip

Supporting Online Material MS # 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

* These authors contributed equally to this work

[‖]To whom correspondence should be addressed. E-mail: adam@burnham.org

Supporting Online Material

A. Materials and Methods

1. Metabolic Reconstruction

a. Description of the iterative reconstruction procedure

The metabolic reconstruction of *Thermotoga maritima* was performed as described previously(*1, 2*) using the reconstruction software SimPheny (Genomatica, Inc., San Diego, CA). Briefly, the genome annotation was obtained from TIGR (11/21/06) and served as the basis for the gene–protein–reaction (GPR, which describes in Boolean terms (AND, OR) the required genes for a metabolic reaction) associations. Furthermore, the genome annotation, along with other databases [e.g., KEGG(*3*), SEED(*4*)], was used to identify metabolic functions in *T. maritima*. Each potential metabolic function was evaluated manually through extensive literature search. Many of the network reactions are associated with supporting literature (Table S1-S6). The literature search also led to an extension of the initial reactions list. The reconstruction process followed traditional pathway definitions, which eased the identification of missing metabolic genes and functions. This procedure led to a comprehensive reaction network, which then required computational evaluation to ensure the predictive potential of the metabolic model. To this end, we used the mathematical format of the reconstruction list in SimPheny (Genomatica, Inc., San Diego, CA). Note that the conversion from a reaction list into a mathematical, computable format can also be done using the Matlab-based COBRA toolbox (*5*). Evaluation consisted of a test to see if the model could produce all known biomass precursors under different media conditions. If the production was not possible, we evaluated manually possible reactions that needed to be included in the reconstruction and searched for supporting literature. This process was iterated until model evaluation was successful. Resources used for the reconstruction included: (i) primary literature and (ii) SEED and KEGG databases to identify candidate metabolic genes based on homology. The final reconstruction includes 478 metabolic

genes, 503 unique metabolites, and 562 internal and 83 external metabolic reactions. A total of 52 internal reactions (10.8%) were not assigned to a specific gene product, as they are catalyzed by yet-unidentified genes or they are spontaneous reactions (Table S6).

Details of the individual reconstruction steps are given in the following sections:

The metabolic reconstruction is available in four formats:

- Microsoft Excel tables (Tables S1 to S5),
- download from http://bigg.ucsd.edu
- SBML format (Supplemental File 1)
- Matlab format (Supplemental File 2)

These formats include information about the (i) metabolic reactions and their evidence, including notes and references; (ii) metabolites and their formula/charge; and (iii) metabolic genes whose products catalyze the enzymatic reactions.

b. Confidence score

Every network reaction was associated with a confidence score reflecting the quality of information and evidence currently available to support the current assignment for this reaction. The confidence score ranges from 0 to 4, where 0 is the lowest and 4 is the highest evidence score (Table S7). Note that multiple information types result in a cumulative confidence score. This means that a confidence score of 4 may represent physiological and sequence evidence. A detailed list of evidence types for each network reaction can be found at http://bigg.ucsd.edu and in Table S3. The overall confidence score for the metabolic reconstruction of *T. maritima* was 2.75. A more detailed representation of the available evidence for assignment of the different pathways in the *T. maritima* reconstruction is shown in Fig. S1.

c. Biomass formulation

The biomass reaction is normally included in a metabolic reconstruction(*1, 2*) and lists all known biomass precursors (e.g., lipids, nucleotides, amino acids) necessary to produce a new cell with their respective fractional contributions. Often this biomass reaction is used as an objective function for flux balance analysis(*6*) (FBA) (see below). No detailed information was available for the biomass composition of *T. maritima*; therefore, we adapted the biomass reaction of the *E. coli* reconstruction(*7*) (Table S2). Information about *T. maritima*–specific lipids was considered, such as polar lipids where approximately 50% of glycolipids (TM_GL1 and TM_GL2) and approximately 50% of lipids might be similar to lipids occurring in other organisms (*8*). Furthermore, *Carballeira et al.* identified hC16:0 and hC18:0 in *T. maritima* (*9*) and indicated that hydroxy fatty acids were reported in other bacteria as amide-linked fatty acids in sphingolipids (*10*). Since no further information was available, we did not include sphingolipids. Glycolipids TM_GL1 and TM_GL2 together account for about 50% of total polar lipids. Manca *et al.* suggested that at least 50% of the membrane bilayer is similar to that in other bacteria and eukarya (*8*). The peptidoglycan composition was based on *Huber et al.* and is as follows: muramic acid, N-Ac-glucosamine, glutamate, alanine, and lysine (0.41: 0.69: 1.0: 1.43: 0.89)(*11*)**.**

Lipids that were considered in the biomass reaction:

- dmosACP—13,14-Dimethyloctacosanedioic acid ACP (C30:0 ACP)
- dmtaACP—15,16-Dimethyltriacontanedioic acid ACP (C32:0 ACP)
- pa160—1-Hexadecanoyl-sn-glycerol 3-phosphate
- pgp160—Phosphatidylglycerophosphate (dihexadecanoyl, n-C16:0)
- peptido-TM—Peptidoglycan subunit of *Thermotoga maritima*
- udcpdp—Undecaprenyl diphosphate
- TM-GL1—*Thermotoga maritima* glycolipid 1

- TM-GL2—*Thermotoga maritima* glycolipid 1

Another modification to the *E. coli* biomass reaction (*7*) is that the *T. maritima* biomass reaction does not account for asparagine since *T. maritima* lacks an asparaginyl-tRNA synthetase (*12*). It is assumed that $tRNA^{Asn}$ is charged with aspartate and then converted into Asn-$tRNA^{Asn}$ by transamidation (*12*). This is similar to what was reported for the archaeon *Haloferax volcanii* (*13*).

A more detailed list of biomass precursors and their fractional contributions can be found in Table S2. Note that, despite the fact that the adapted *E. coli* biomass reaction is used for simulations and network evaluation, the qualitative predictive potential of the models is not affected. Since the fractional contributions of the individual biomass precursors may differ between *E. coli* and *T. maritima*, the quantitative predictions may not be accurate. However, gene essentiality studies, growth capability on various carbon sources, and minimal gene set analysis are not affected by this limitation.

d. Network gaps

Network gaps are missing links in the metabolic network. These gaps occur due to incomplete knowledge (e.g., missing gene annotations). During the iterative reconstruction process, network gaps are "closed" if supporting information is available or a temporary reaction is required for the metabolic capability of the network, e.g., biomass precursor synthesis. Network gaps indicate missing knowledge. Alternatively, network gaps can be a consequence of the limited scope of metabolic networks. This means that produced or consumed metabolites are involved in other cellular functions, such as DNA replication or protein synthesis, for which the

metabolic networks do not account. A complete list of the 122 network gaps in the current *T. maritima* reconstruction can be found in Table S1.

e. Metabolic map

During the reconstruction process we generated a metabolic map showing all network reactions (http://www.topsan.org/Thermotoga_maritima). This metabolic map is very helpful for visualizing the network content and simulation results. It also provides links to annotation pages for individual proteins of *T. maritima*, implemented on a wiki-like annotation platform, TOPSAN (*14*). TOPSAN, or The Open Protein Structure Annotation Network, is our experimental annotation/instant collaboration platform developed for annotating and initiating collaborations on proteins whose structure has been solved by NIH PSI production centers. For the purpose of this manuscript, TOPSAN pages have been built for all *T. maritima* proteins in the reconstruction using, where possible, predicted structural information.

f. Conversion from reaction list to mathematical model

The conversion into a mathematical, or computer-readable, format can be done automatically by parsing the stoichiometric coefficients from the network reaction list [e.g., using the COBRA toolbox(*5*)]. The mathematical format is called a stoichiometric matrix, or S-matrix, where the rows correspond to the network metabolites and the columns represent the network reactions (Fig. S2). For each reaction, the stoichiometric coefficients of the substrates are listed as negative values, while the product coefficients are positive numbers, by definition. The resulting size of the S-matrix is $m \times n$, where m is the number of metabolites and n is the number of network reactions. Mathematically, the S-matrix is a linear transformation of the flux vector

$v = (v_1, v_2, .., v_n)$ to a vector of time derivatives of the concentration vector $x = (x_1, x_2, .., x_m)$ as $\frac{dx}{dt} = S \cdot v$. At steady state, the change in concentration as a function of time is 0; hence, it follows: $\frac{dx}{dt} = S \cdot v \equiv 0$. The set of possible flux vectors v that satisfy these equality and inequality constraints might be subject to further constraints by defining $v_{i,\min} \leq v_i \leq v_{i,\max}$ for reaction *i*. For every irreversible network reaction *i*, the lower bound was defined as $v_{i,\min} \geq 0$ and the upper bound was defined as $v_{i,\max} \geq 0$. Finally, the application of constraints corresponding to different environmental conditions (e.g., minimal growth medium) or different genetic backgrounds (e.g., enzyme-deficient mutant) renders the models condition-specific. Note that the metabolic network reconstruction is unique to the target organism as defined by its genome while it can give rise to many different models by applying condition-specific constraints. All flux rates, v_i, except biomass formation, are given in mmol/gDW/h.

g. Flux balance analysis (FBA) and linear programming (LP)

FBA is a formalism in which a metabolic network is framed as a linear programming problem. The principal sets of constraints in FBA are those imposed by steady-state mass conservation of metabolites in the system.

Numerous mathematical tools have been developed to interrogate the metabolic network properties *in silico* and have been reviewed(*15*). Furthermore, many of these methods have been encoded in Matlab format [e.g., COBRA toolbox(*5*)]. A large subset of these tools relies on linear programming to find a solution to an optimization problem (e.g., maximal possible growth rate of

a metabolic network under a given set of environmental constraints). The linear programming problem is formulated as follows:

maximize $c^T v$ (objective function)

subject to $S \cdot v = 0$

$v_{i,min} \leq v_i \leq v_{i,max}$ for all $i \in n$ reactions,

where S is the stoichiometric matrix ($m*n$), c is the objective function vector, v is a vector of reaction fluxes, $v_{i,max}$ is the maximal capacity for reaction i, and $v_{i,min}$ is the minimal capacity for reaction.

All FBA calculations were carried out using SimPheny (Genomatica, Inc., San Diego, CA) or the COBRA toolbox (*5*).

h. Simulation conditions

The computational minimal medium is defined in Table S8. The flux rates are listed in mmol/gw/h, and the unit of the biomass flux rate is 1/h. Note that uptake of metabolites by the model is defined as a negative flux value ($v_i \leq 0$) and byproduct secretion is defined as a positive flux value ($v_i \geq 0$). For simulations where different carbon sources were tested, the lower bound of glucose exchange was set to $v_{i,min} = 0$ and the lower bound value of an alternative carbon source was set to $v_{j,min} \leq 0$.

i. Carbon sources

The *T. maritima* reconstruction accounts for the metabolism of 46 carbohydrates. Table S9 shows details of these carbohydrates, including their exchange reactions, names, and formula, etc.

j. Network validation and evaluation

As mentioned above, the iterative reconstruction process includes different simulations, e.g., the production of biomass precursors in known growth conditions. We tested the network's capability to growth on different carbon sources added to the minimal medium. These carbon sources include the listed carbohydrates (Table S9) and other carbon-containing metabolites for which an exchange reaction has been found (Table S2). Growth rates on larger carbohydrates (e.g., starch) were much higher than biologically feasible (Fig. S3). The reason for these high growth rates was that the simulation was only carbon limited, while other constraints may limit the growth *in vivo*.

We also investigated whether the *T. maritima* model is able to produce hydrogen from various carbon sources (Table S5). Schröder *et al.* reported that 1 mol glucose is converted into 2 mol acetate, 2 mol CO_2, and 4 mol H_2 (without the presence of elemental sulfur, S°)(*16*). The model required the presence of S°, suggesting that there is another sulfur source *T. maritima* can use for which we do not account (e.g., SO_4). In fact, Schroeder *et al.* stated that growth on glucose was dependent on yeast extract (*16*), which may contain an alternate sulfur source.

When we optimized for biomass production, 10 mol glucose resulted in a growth rate of 0.365 1/h (113 min doubling time). This result is comparable with experimental data considering that we used minimal medium conditions (Table S10). In this simulation condition, the model produced 15.38 mmol/gDW/h acetate, 16.46 mmol/gDW/h CO_2, and 33.22 mmol/gDW/h H_2. These values are less than the reported 2 mol acetate, 2 mol CO_2, and 4 mol H_2 from 1 mol glucose since we required optimal growth. However, the ratio of acetate:CO_2:H_2 agrees with the reported ratio of 1:1:2. Further analysis showed that the model could produce 19.898

mmol/gDW/h acetate from 10 mol glucose when requiring the growth rate to be 25% of the maximal possible growth rate (Table S5).

Finally, we compared the biomass yield and hydrogen yield for the different carbon sources and found that glycerol had the highest yield in both cases (Table S5).

k. Minimal set gene calculation

The minimal set genes were determined as illustrated in the algorithm below. This algorithm is based on the minimal set approach(*17, 18*). The analysis has been carried out using the optimization and single gene deletion function of the COBRA toolbox (*5*). Simulated conditions include minimal medium + glucose and rich medium.

Algorithm:

Input parameters: Metabolic model M_0 (including the growth environment descriptions), coefficient c ($0 < c < 1$) that defines the minimum acceptable growth rate.

1. Take metabolic model M_0 (e.g., in minimal medium + glucose)
2. Determine maximal growth capacity of M_0 by optimizing for biomass production, $\mu 0$
3. Remove all model-associated genes from the model as follows:
    a. Randomly pick one gene
    b. Remove gene and its associated reaction(s) from the network, resulting in model M_1
    c. Determine the growth capacity of M_1 by optimizing for biomass production, μ_1
    d. If μ_1 is larger than $c \cdot \mu_0$, the gene is not necessary to sustain a growth rate larger than $c \cdot \mu_0$

e. If μ_1 is smaller than $c \cdot \mu_0$, put the gene and its reaction(s) back into the model as the gene is necessary to maintain the defined growth threshold ($c \cdot \mu_0$)

f. Continue from Step 3a until all genes have been tested for essentiality

4. Repeat Step 3 for 1,000 times, i.e., create 1,000 distinct, possible minimal models
5. For each gene count in how many of the 1,000 minimal models they appear
   a. "**Core-essential**" genes are those that appear in all experiments
   b. "**Synthetic lethal**" or "conditional essential" genes are those that appear in less than 1,000 but at least 1 experiment
   c. "**Non-essential genes**" are those that do not contribute to growth above the defined threshold $c \cdot \mu_0$

Matlab (The MathWorks Inc., Natick, MA) was used as the programming environment. Tomlab (Tomlab Optimization Inc., San Diego, CA) was used for linear programming. In both medium conditions, the number of minimal set genes converged readily (Fig. S4).

2. Structural Assignments

Structures of 120 proteins from the *T. maritima* metabolic reconstruction have been determined experimentally (Table S11) to date and for the purpose of this analysis were downloaded from the Protein Data Bank (*19*). Structural models for another 336 proteins were built using the automated procedure that is described in the next paragraph. Table S12 lists the detailed information of the structural templates for structural modeling, and Table S13 provides all the details of the template selection and modeling, including the PDB id of the template, score, and length of the alignment. In addition, Table 13 provides two types of classifications of model

quality. The first classifies models into four classes, based on the target–template similarity and alignment quality:

a) "Molecular Replacement Targets": FFAS_score better than –15, coverage better than 50%; sequence identity no less than 30%;

b) "Possible Molecular Replacement Targets": FFAS score better than –15, coverage better than 50%; sequence identity no less than 20% and not in a);

c) "Fold Prediction Only Targets": FFAS score better than –9.5 and not in a) or b);

d) Models build manually when automated modeling protocol failed.

In addition, we report a PSQS score of the model. PSQS is an energy-like measure of the quality of protein structure, based on the statistical potentials of mean force describing interactions between residue pairs and between single residues and solvent (*20-22*). Negative PSQS scores generally indicate reasonable model quality. Positive PSQS scores usually signal some problems with the model.

Manual modeling was used for three soluble proteins (TM0788, TM1444, TM0540) that failed in the automated modeling pipeline. In each case, series of fold recognition algorithms were applied, and the selection of the template was done by human experts after considering scores of all the methods, consistency of template selection, functional similarity between possible templates and the target and structural conservation among analogous operons. Similar procedures, although not formalized, were used and benchmarked in CASP meetings (*23*) [n.b. The three proteins for which automated structure prediction failed are encoded by genes: *tm1444, tm0788 and tm0540*. TM1444 and TM0788 can now be modeled due to availability of newly solved structures (http://ffas.burnham.org/protmod-cgi/tmModels.pl?type=id&id=TM1444, and http://ffas.burnham.org/protmod-cgi/tmModels.pl?type=id&id=TM0788); TM0540 is now the only remaining manually build model in the reconstruction (http://ffas.burnham.org/protmod-

cgi/tmModels.pl?type=id&id=TM0540].The mrTM set also contained 19 transmembrane proteins that could not be matched to any of the available experimental structures of transmembrane proteins .These proteins were analyzed by the TMHMM 2.0 (*24*) algorithm that predicts positions and directions of transmembrane helices, followed by manual construction of low-resolution topological models.

All calculations described here have been performed using databases and tools current as of December 2007. Structures solved and deposited after that date were not used in the analysis presented in the paper. We periodically repeat the automated modeling steps, and the latest results are available on TOPSAN pages for each model. Links to these pages are provided in the last column of Table S13.

The automated procedure for building structural models consists of the following steps (n.b., this procedure is now implemented at the Joint Center for Molecular Modeling website http://jcmm.burnham.org, modeling described here was performed by a beta version of the JCMM modeling server):

(1) Find a structural template

In this step, we first used a profile–profile alignment algorithm, FFAS(*25*), to compare the target protein against the database of protein structures deposited in PDB(*19*). Proteins with FFAS scores below –9.5 were considered significantly similar to the target and added to the possible template list. In the next step, we examined all the possible templates to choose the optimal one. If multiple templates were available, a structure that had a better FFAS score, higher sequence identity, and longer alignment to the target protein was selected. All borderline modeling templates with scores between –12.5 and –9.5 were submitted to the fold-recognition metaserver at BioInfoBank Institute (*26*) and crosschecked with other homology and fold-recognition algorithms available at that site. In all cases, the metaserver assignments agreed with

FFAS assignments. Additional verification was provided by comparing the functions of proteins in the *T. maritima* metabolic network with the functions of their assigned templates. The pie chart in Fig. S5 shows that 97% of the fold assignments are based on proteins with the same or closely related biochemical functions.

(2) Build the alignment

We used FFAS(*25*) with standard parameters to calculate the sequence alignment between the target protein and the chosen template.

(3) Build the three-dimensional structural models

We used three different programs for building models of mrTM proteins: MODELLER (*27*), Jackal (*28*), and SCWRL (*29*). Each program was used to create one model for each protein so that each protein had three models.

3. Fold Assignments

The protein folds are assigned simultaneously with the modeling template assignments as described above using template protein fold assignments in the SCOP database.

4. Classification of Pairwise Relationships between Enzymes in the Network

The question we are trying to answer here concerns a dominant model of pathway evolution. Two models have been proposed in literature. One is the retrograde model (*30*), which suggests that pathways evolve by duplication and then functional differentiation of enzymes catalyzing adjacent reactions within pathways. If this model is correct, folds of enzymes catalyzing adjacent reactions should be similar, suggesting also that sharing of metabolites is the driving force of fold conservation. The other model is the patchwork model (*31*), which argues that pathways evolve by recruiting enzymes catalyzing required reactions and the protein folds

are evolutionarily conserved to catalyze similar biochemical reactions. If this model is correct, folds of enzymes catalyzing similar reactions should be similar, suggesting also that sharing of a reaction mechanism is the driving force of fold conservation. In a review by Rison *et al.* (*32*), the two models were compared statistically in a set of *E. coli* small-molecule metabolism pathways, and the results support the patchwork model. In addition, Caetano-Anollés *et al*. did an extensive study on the evolution of modern metabolism combining evolutionary and structural genomics information (*33, 34*).

Instead of using a series of disconnected pathways, as in other studies, we based our analyses on the complete and self-sustaining model of the central metabolic network in *T. maritima,* developed as part of this work. To perform the analysis, we developed the classification and statistical analyses of pairwise relationships between proteins in the reconstruction. Each protein, in association with its fold assignment, is called a functional domain (FD). To avoid ambiguity, proteins with multiple domains and reactions catalyzed by multiple proteins are excluded from this analysis. Within these limitations, our dataset contains 234 reactions catalyzed by 178 FDs. The pairwise relationships of FDs include fold relationships (same fold or different fold), as well as reaction relationships (sharing chemical mechanisms, adjacent, or unrelated). The fold relationships are easy to determine, since the fold assignment for every protein in the network is obtained from mapping a given FD to a SCOP classification, while there is at present no common scheme available to identify the pairwise relationship of chemical reactions. The Enzyme Classification (EC) numbers may be used to define chemical similarities between reactions. However, some reactions in the reconstructed network have no clearly assigned EC numbers; besides this, research by Glasner *et al.* showed that some different overall reactions share common partial reactions, and this is not captured by EC numbers (*35*). The adjacency of reactions is not clearly defined in the metabolic network either. Pathways or

subsystems may be used to define the adjacent reactions, but are not enough to solve the ambiguity for pairs of borderline reactions that are present in multiple pathways. Therefore, we designed an automated procedure to align two reactions and assign the pairwise relationships.

The compounds in metabolic reactions can be clearly divided into two sets: cofactors, such as ATP, and intermediate metabolites, such as Mannose. In most cases, reactions with shared chemical mechanisms have common cofactors, while adjacent reactions have common intermediate metabolite(s). Exceptions are isomerase reactions (and other reactions), in which chemical mechanisms are shared, but cofactors are not present, as well as cofactor biosynthesis reactions, in which cofactors act as metabolites. The automated procedure gives no consideration to these special cases, and these were analyzed by hand. The major steps in the automated reaction classification are as follows:

1. Reaction reformulation

Every reaction can be presented in the form: {a} + {B} = {c} + {D}, where {a}, {c} are sets of intermediate metabolites on each side of the reaction (ignore directionality), which usually have low connectivity, and {B}, {D} are sets of cofactors, which usually have high connectivity.

2. Reaction alignment

For every pair of reactions: R1 ≡ {a1} + {b1} = {c1} + {d1}; R2 ≡ {a2} + {b2} = {c2} + {d2}: let two sets, e.g., {a1} and {a2}, match if these sets have at least one identical intermediate metabolite. We adjust the order of the sets so that if there is only one match, the corresponding sets are positioned in the left-hand side of the reactions, and all the matched sets are properly aligned.

3. Decision making

We use the decision tree below to assign reaction relationships (Fig. S6), where a letter a, B, c, D reflects a match, and x, X reflects a mismatch. The decision tree has four possible outputs:

S-A, where the two reactions share cofactors in both the left- and right-hand sides and share intermediate metabolite in at least one side, so that the two reactions are both similar and adjacent; S-N, where the two reactions share cofactors in both the left- and right-hand sides and have no shared intermediate metabolite, so that the two reactions are chemically similar, but not adjacent; D-A, where the two reactions share an intermediate metabolite in at least one side, so that the two reactions are not similar and adjacent; and D-N, where none of the previous criteria are fulfilled, so that the reactions are neither similar nor adjacent. S-A is a very rare case, appearing only once in our dataset. S-N, D-A, and D-N correspond to the similar (S), connected (C), or unrelated (U) categories.

We also compared our classification to the EC classifications and SEED subsystem classification for a subset of annotated FDs. The result shows that the adjacent reactions are positively correlated with reactions in the same SEED subsystem, and the chemically similar reactions are positively correlated with reactions that share the first three digits of their EC numbers (Fig. S7).

5. Random Sampling of Non-Redundant Database

We examined the fold distribution of the proteins in mrTM by comparing its number of representatives in each fold family with a randomly selected protein set of the same size. The random set was created through sampling of the NCBI non-redundant (NR) database(*36*). Fold assignment of NR proteins was done following the procedure described in Section II. We constructed 100 random sets with the same number of fold domains as the mrTM and calculated their average and standard deviations. The comparison of genome coverage by the mrTM set versus average of the randomly selected sets of proteins is shown in Fig. S8.

B. Figures

Fig. S1. Heat map depicting the experimental evidence available for different metabolic pathways. For each metabolic pathway, the percentage of reactions with a confidence score of 4, 2, and 1 was determined. Red corresponds to 100%, blue to 0% of all the reactions in a pathway. Thus, this heat map highlights pathways that are well or less-well studied. As expected, many of the carbon-related catabolic pathways have reactions with a high confidence score (biochemical evidence). Note that the metabolic reconstruction of *T. maritima* does not contain any reactions with a confidence score of 3 or 0.

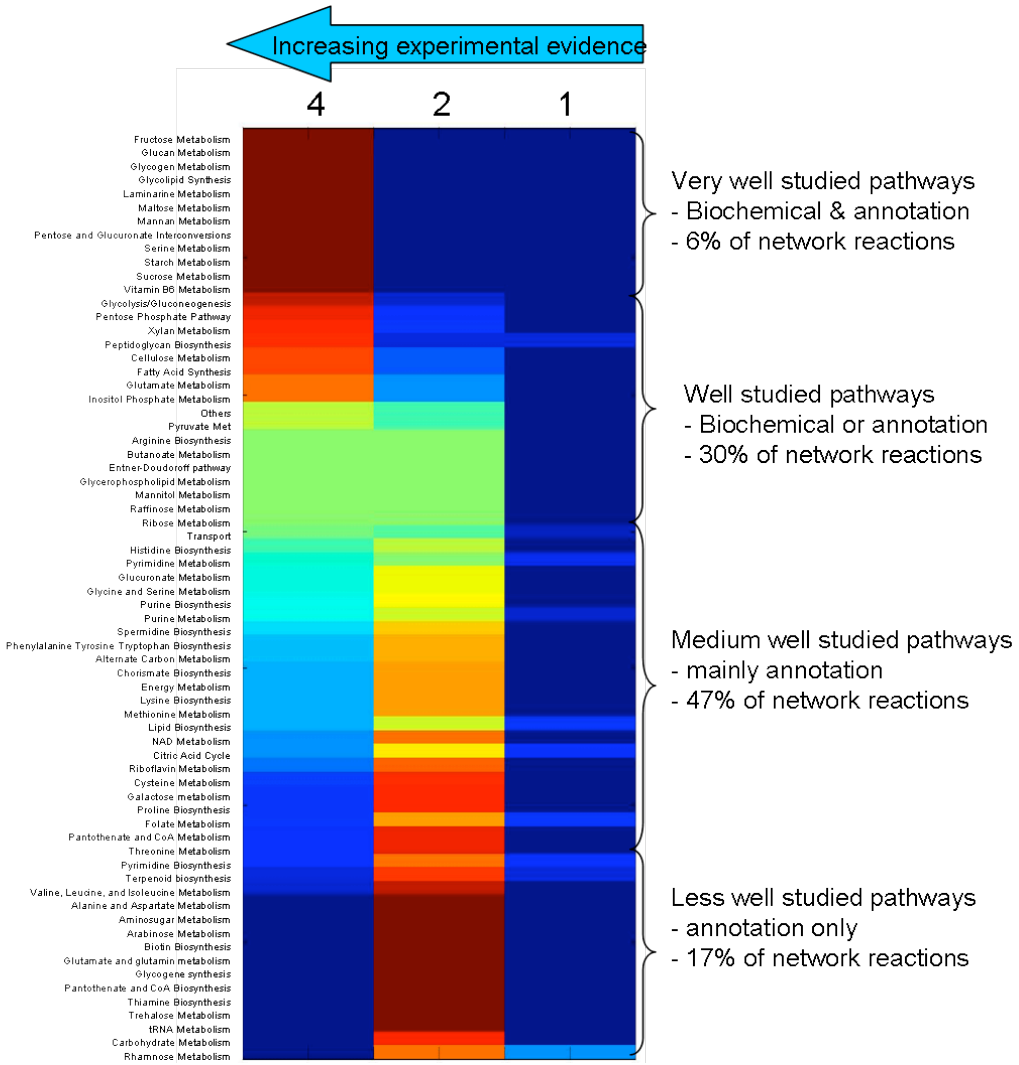


Fig. S2. Schematic representation of the conversion from reaction list to mathematical model.

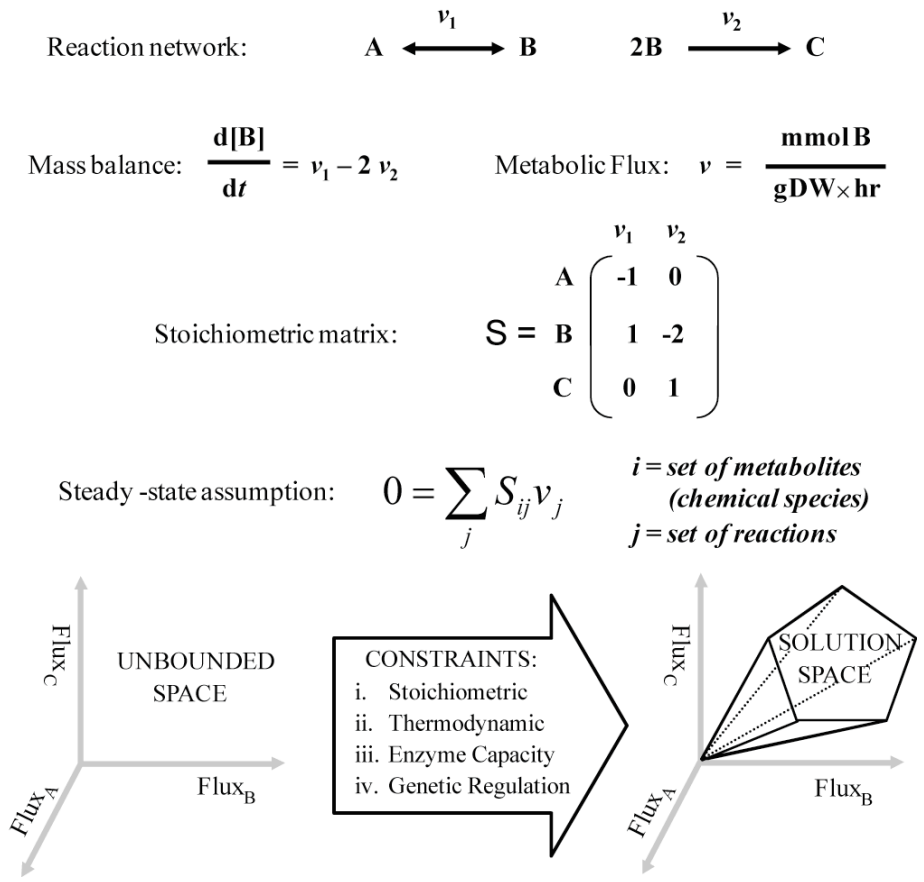


Fig. S3. All exchange reactions of carbon sources were tested to evalutate if they supported *in silico* growth. For each simulation, minimal medium was substituted with 1 mmol/gDW/h uptake of a carbon source and the model was optimized for biomass production.

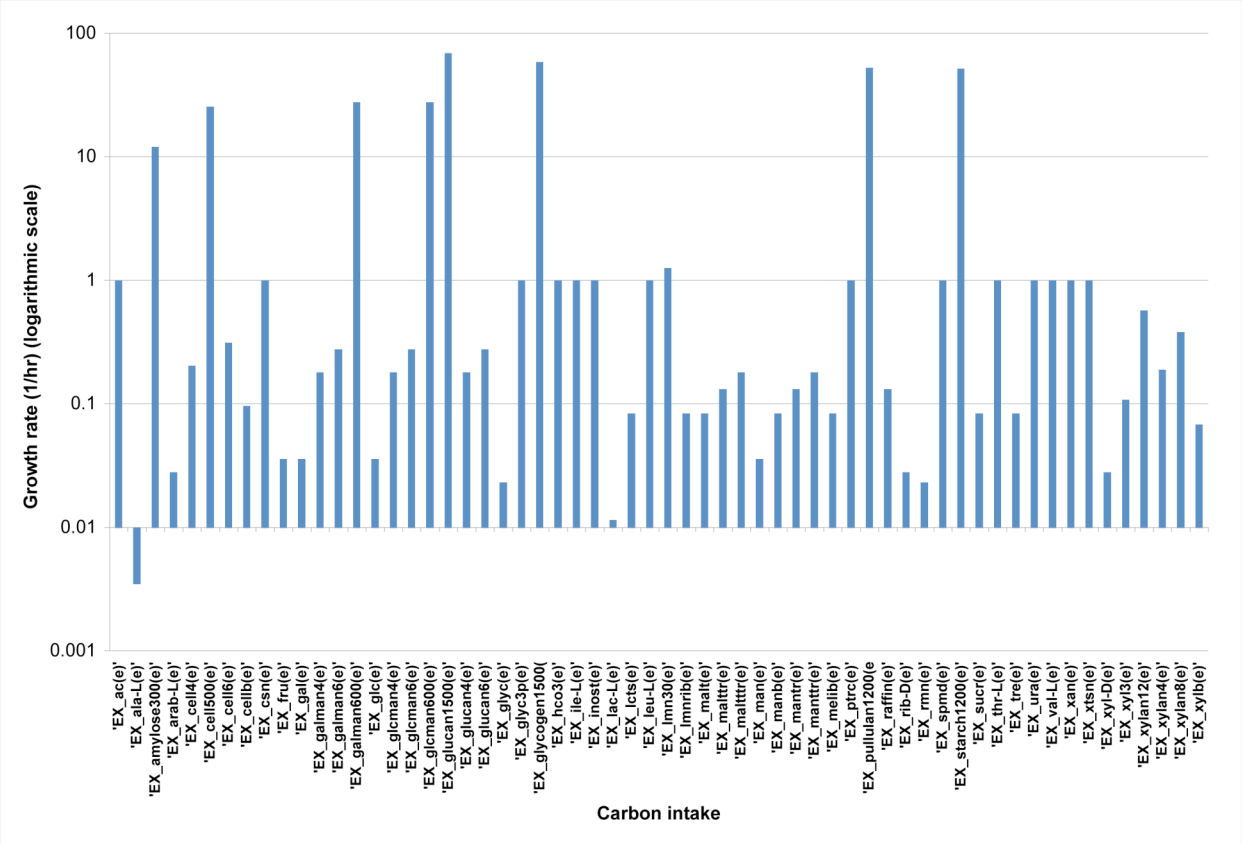
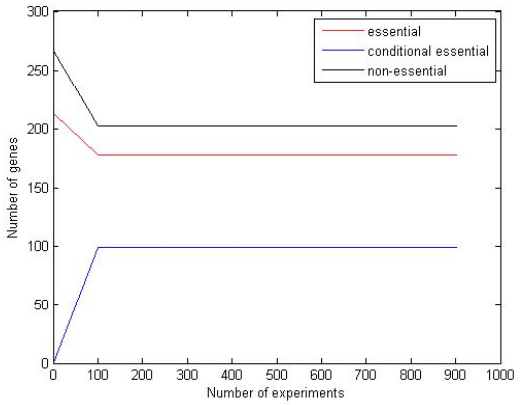


Fig. S4. Simulation of minimal networks. (A) shows that fewer than 1,000 random iterations are sufficient to identify all essential, conditional, and nonessential reactions in minimal medium + glucose. **(B)** shows that fewer than 3,000 random iterations are sufficient to identify all essential, conditional, and nonessential reactions in rich medium.

A. minimal medium + glucose



B. rich medium

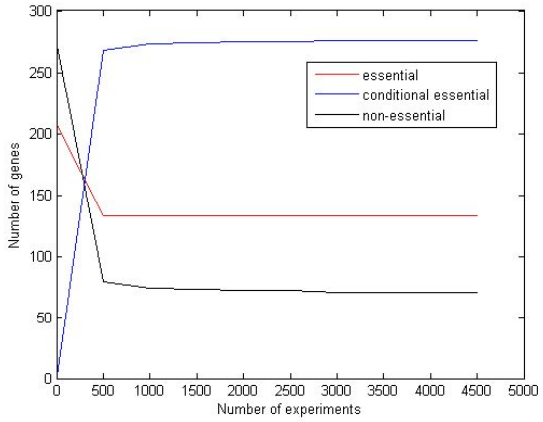


Fig. S5. Functional comparison between *T. maritima* proteins and their templates. In this analysis, 429 out of 478 *T. maritima* proteins in the metabolic reconstruction were evaluated. The percentage values show the fraction of the proteins with same, related, distant, or unknown functional relationships with their homologs in the SCOP database.

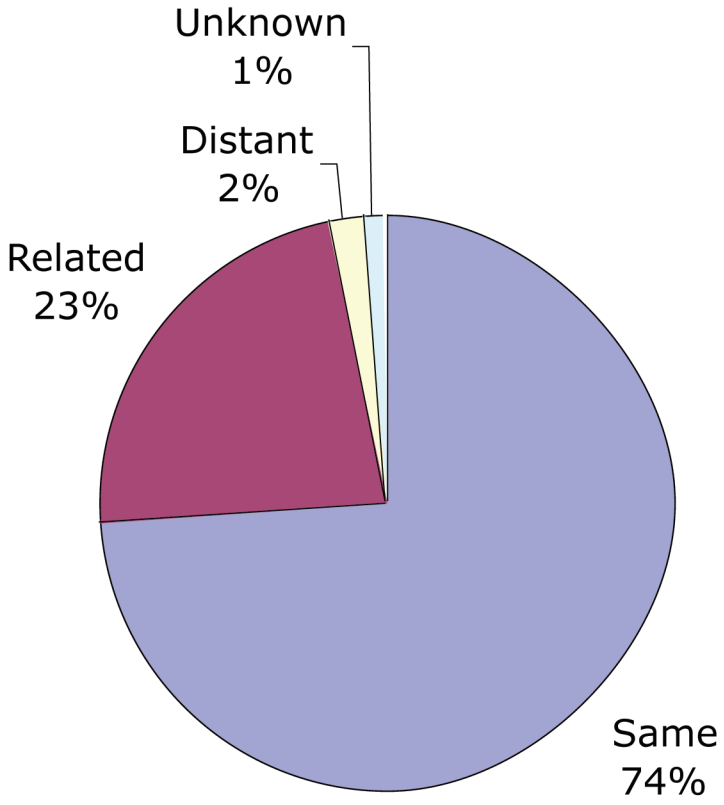


Fig. S6. Decision tree for classification of reaction pairs. LRC: low connectivity compound; HRC: high connectivity compound. A match is indicated by the letter a, B, c, or D; a mismatch is indicated by the letter x or X.

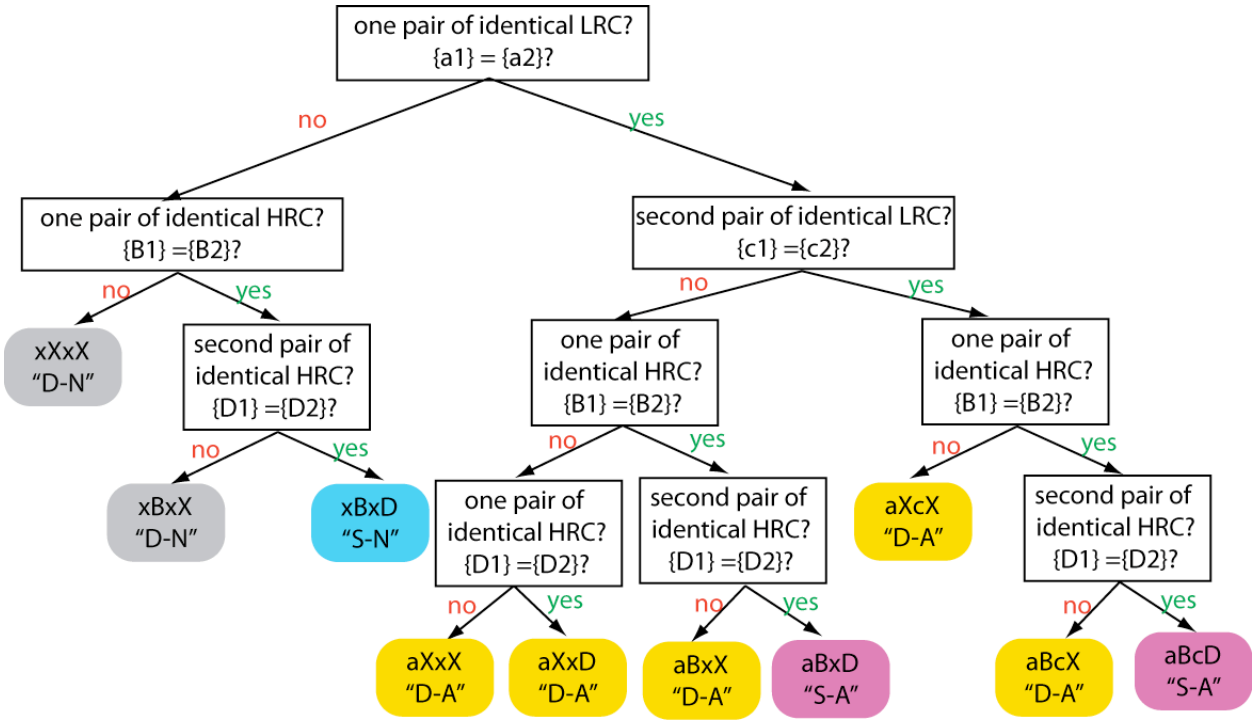


Fig. S7. Correlations of our automated classification of reaction pairs to the classification based on EC (identifies similar reactions) and subsystems (identifies connected reactions). ec_EQ/NOEQ: ratio between the number of reaction pairs with the same EC number and those with different EC numbers; subsystem_EQ/NOEQ: ratio between the number of reaction pairs with the same subsystem assignment and those with different subsystem assignments.

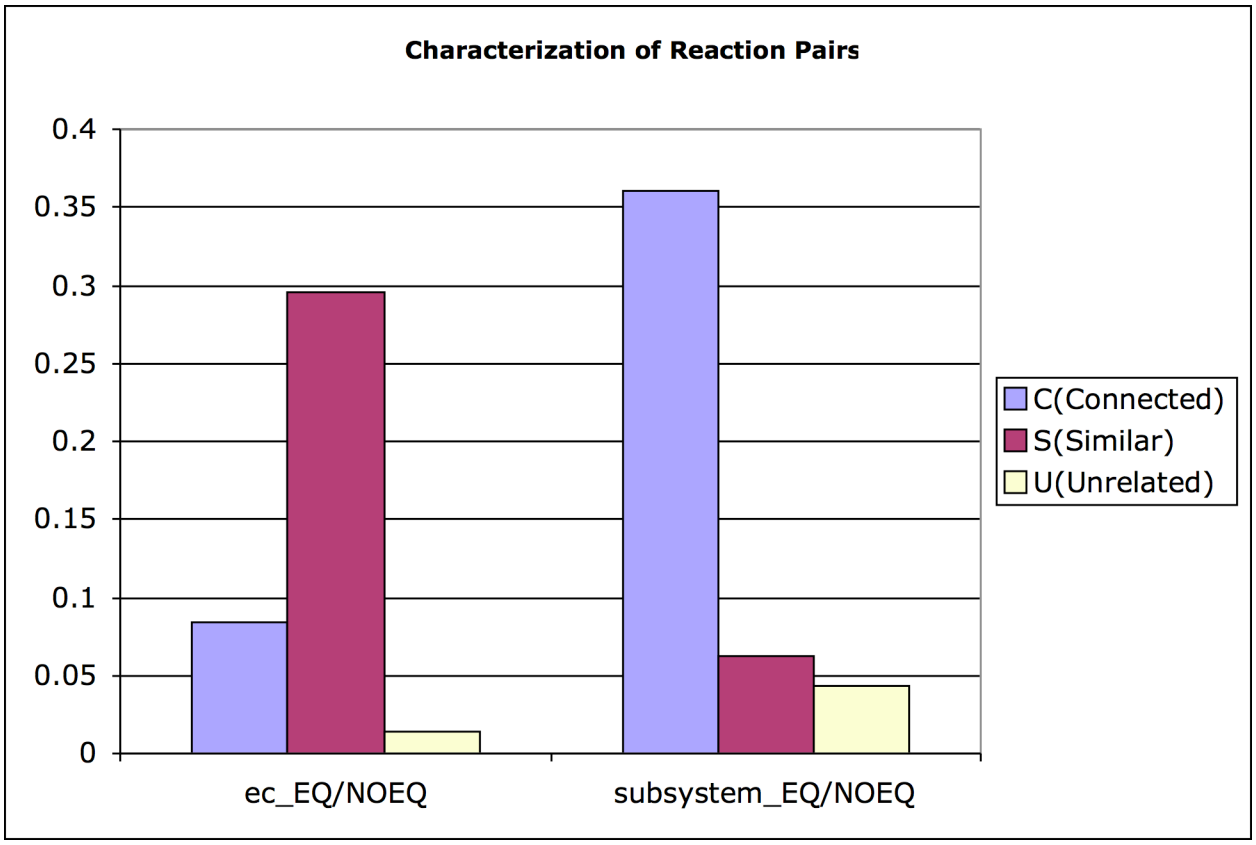


Fig. S8. Comparison of the genome coverage by the mrTM set versus randomly selected set of proteins with known structures. The x-axis represents the rank of folds as measured by number of protein domains that contain the fold. The y-axis represents the cumulative percentage coverage of proteins corresponding to the chosen set [blue: *T. maritima* metabolic network; magenta: average (and standard deviation) of 100 random protein sets from the non-redundant database (*36*).] RANGE shows the maximum and minimum number of fold families in the random protein sets; STDEV is the standard deviation of the number of fold families; 288 is the average number of fold families in the random sets.

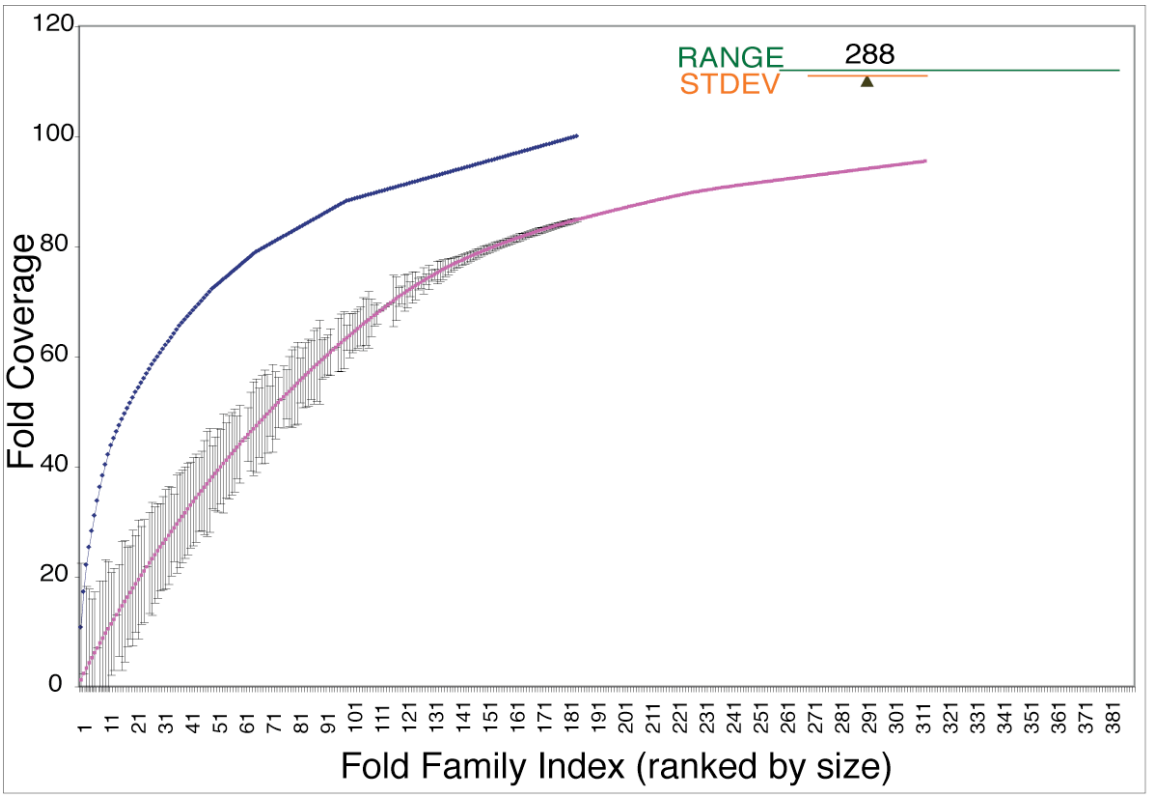
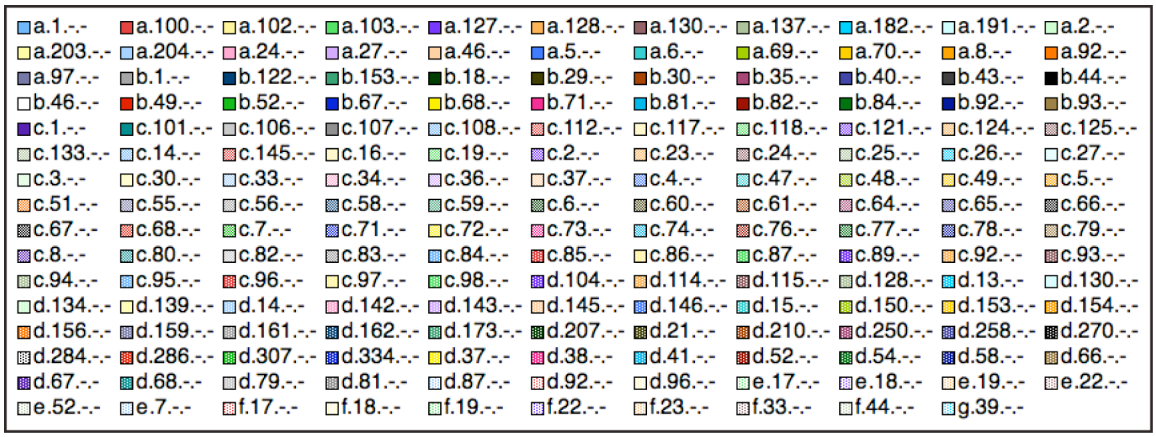


Fig. S9. Color legend for Fig. 4 - Fold composition of the non-essential, synthetic lethal, and core-essential protein sets. Each color represents different fold in the SCOP database. The folds are represented in the format of "<letter>.<number>.-.-". <letter> is the code for SCOP class. For example, "a" is the code for the class "All alpha proteins". <number> is the code for SCOP fold in a specific class. For example, "1" in class "a" stands for "Globin-like". Therefore, "a.1.-.-" codes for "All alpha Globin-like" fold in SCOP classification.



C. References

1. A. M. Feist, M. J. Herrgard, I. Thiele, J. L. Reed, B. O. Palsson, *Nat. Rev. Microbiol.* **7**, 129 (2009).
2. J. L. Reed, I. Famili, I. Thiele, B. O. Palsson, *Nat. Rev. Genet.* **7**, 130 (2006).
3. M. Kanehisa *et al.*, *Nucleic Acids Res.* **34**, D354 (2006).
4. R. Overbeek *et al.*, *Nucleic Acids Res.* **33**, 5691 (2005).
5. S. A. Becker *et al.*, *Nat. Protoc.* **2**, 727 (2007).
6. C. H. Schilling, J. S. Edwards, B. O. Palsson, *Biotechnol. Prog.* **15**, 288 (1999).
7. J. L. Reed, T. D. Vo, C. H. Schilling, B. O. Palsson, *Genome Biol.* **4**, R54 (2003).
8. M. C. Manca *et al.*, *Biochim. Biophys. Acta.* **1124**, 249 (1992).
9. N. M. Carballeira *et al.*, *J. Bacteriol.* **179**, 2766 (1997).
10. W. R. Mayberry, P. F. Smith, T. A. Langworthy, P. Plackett, *J. Bacteriol.* **116**, 1091 (1973).
11. R. Huber *et al.*, *Arch. Microbiol.* **144**, 324 (1986).
12. K. E. Nelson *et al.*, *Nature* **399**, 323 (1999).
13. A. W. Curnow, M. Ibba, D. Soll, *Nature* **382**, 589 (1996).
14. S. Sri Krishna *et al.*, *The Second Automated Function Prediction Meeting* **Editors Ana PC Rodrigues, Barry J Grant, Adam Godzik, Iddo Friedberg**, 89 (2006).
15. N. D. Price, J. L. Reed, B. O. Palsson, *Nat. Rev. Microbiol.* **2**, 886 (2004).
16. C. Schröder, M. Selig, P. Schönheit, *Arch. Microbiol.* **161**, 460 (1994).
17. A. P. Burgard, S. Vaidyaraman, C. D. Maranas, *Biotechnol. Prog.* **17**, 791 (2001).
18. C. Pál *et al.*, *Nature* **440**, 667 (2006).
19. H. M. Berman *et al.*, *Nucleic Acids Res.* **28**, 235 (2000).
20. A. Godzik, A. Kolinski, J. Skolnick, *J. Mol. Biol.* **227**, 227 (1992).
21. A. Godzik, A. Kolinski, J. Skolnick, *Protein Sci.* **4**, 2107 (1995).
22. L. Jaroszewski, K. Pawlowski, A. Godzik, *J. Mol. Model.* **4**, 294 (1998).
23. J. Moult *et al.*, *Proteins* **69 Suppl 8**, 3 (2007).
24. A. Krogh, B. Larsson, G. von Heijne, E. L. Sonnhammer, *J. Mol. Biol.* **305**, 567 (2001).
25. L. Jaroszewski, L. Rychlewski, Z. Li, W. Li, A. Godzik, *Nucleic Acids Res.* **33**, W284 (2005).
26. K. Ginalski, *Curr. Opin. Struct. Biol.* **16**, 172 (2006).
27. N. Eswar, D. Eramian, B. Webb, M. Y. Shen, A. Sali, *Methods Mol. Biol.* **426**, 145 (2008).
28. D. Petrey *et al.*, *Proteins* **53 Suppl 6**, 430 (2003).
29. A. A. Canutescu, A. A. Shelenkov, R. L. Dunbrack, Jr., *Protein Sci.* **12**, 2001 (2003).
30. N. H. Horowitz, *Proc. Natl. Acad. Sci. U. S. A.* **31**, 153 (1945).
31. R. A. Jensen, *Annu. Rev. Microbiol.* **30**, 409 (1976).
32. S. C. Rison, J. M. Thornton, *Curr. Opin. Struct. Biol.* **12**, 374 (2002).
33. G. Caetano-Anolles, H. S. Kim, J. E. Mittenthal, *Proc. Natl. Acad. Sci. U. S. A.* **104**, 9358 (2007).

34. G. Caetano-Anolles *et al*., *Int. J. Biochem. Cell Biol.* **41**, 285 (2009).
35. M. E. Glasner, J. A. Gerlt, P. C. Babbitt, *Curr. Opin. Chem. Biol.* **10**, 492 (2006).
36. K. D. Pruitt, T. Tatusova, D. R. Maglott, *Nucleic Acids Res.* **35**, D61 (2007).
37. C. H. Schilling, B. O. Palsson, *J. Theor. Biol.* **203**, 249 (2000).
38. I. Thiele, T. D. Vo, N. D. Price, B. O. Palsson, *J. Bacteriol.* **187**, 5818 (2005).
39. A. M. Feist *et al.*, *Mol. Syst. Biol.* **3**, 121 (2007).
40. Y. K. Oh, B. O. Palsson, S. M. Park, C. H. Schilling, R. Mahadevan, *J. Biol. Chem.* **282**, 28791 (2007).
41. N. Jamshidi, B. O. Palsson, *BMC Syst. Biol.* **1**, 26 (2007).
42. M. Durot *et al*., *BMC Syst. Biol.* **2**, 85 (2008).
43. R. S. Senger, E. T. Papoutsakis, *Biotechnol. Bioeng.* **101**, 1036 (2008).
44. B. Teusink *et al*., *J. Biol. Chem.* **281**, 40041 (2006).
45. A. P. Oliveira, J. Nielsen, J. Forster, *BMC Microbiol.* **5**, 39 (2005).
46. G. J. Baart *et al*., *Genome Biol.* **8**, R136 (2007).
47. M. Heinemann, A. Kummel, R. Ruinatscha, S. Panke, *Biotechnol. Bioeng.* **92**, 850 (2005).
48. M. Riley *et al*., *Nucleic Acids Res.* **34**, 1 (2006).
49. P. Janssen, L. Goldovsky, V. Kunin, N. Darzentas, C. A. Ouzounis, *EMBO Rep.* **6**, 397 (2005).
50. S. R. Chhabra *et al*., *J. Biol. Chem.* **278**, 7540 (2003).
51. R. Hoover, *Carbohydrate Polymers* **45**, 253 (2001).
52. S. R. Chhabra, K. R. Shockley, D. E. Ward, R. M. Kelly, *Appl. Environ. Microbiol.* **68**, 545 (2002).