Supporting Appendix.

I. MAPPING ON THE GRAPH

The *Escherichia.coli* metabolic reactions database(1) Version 1.01 contained 739 reactions as of May 2003. One reaction is one node on the graph. There are two types of edges on the graph.

A. Metabolic edges

An edge between two reactions exists if one of the reactions utilizes a metabolite produced by the other. (i) if both reactions are irreversible, the metabolite must be a product of one and an educt of the other reaction. (ii) if at least one of the two reactions is reversible, the above condition for the existence of an edge is satisfied if the two reactions share at least one metabolite, regardless of it being a product or an educt.

Some compounds or molecules, such as O_2 or CO_2 are nonspecific and may be present in many reactions, thus creating spurious edges. We excluded these nonspecific molecules when drawing edges between nodes. The following is the list of excluded metabolites in the descending order of the frequency of their appearance in the metabolic reactions database.

ATP — Adenosine triphosphate
PI — Phosphate (inorganic)
ADP — Adenosine diphosphate
HEXT — External H^+
CO_2 — Carbon dioxide
PPI — Pyrophosphate
PYR — Pyruvate
NAD — Nicotinamide adenine dinucleotide
GLU — Glutamate
NADH — Nicotinamide adenine dinucleotide reduced
NADP — Nicotinamide adenine dinucleotide phosphate
NH_3 — Ammonia
NADPH — Dihydronicotinamide adenine dinucleotide phosphate reduced
CoA — Coenzyme A
AMP — Adenosine monophosphate
O_2 — Oxygen

B. Edges based on the functional assotiation between enzymes

A weighted edge is drawn between two reactions if any enzyme catalyzing one reaction has a non-zero functional assotiation score with any enzyme catalyzing the other reation. If one or both reactiona are catalyzed by multiple enzymes, the weight of the edge is the highest score between any pair of enzymes catalyzing the reactions.

II. METHODS

A. Macroscale analysis

Cross-clustering coefficient

An important question about the macroscale level of organization is whether genomic association brings some clustering to the metabolic network. The degree of clustering or "cliquishness" in the network is commonly estimated by the local clustering coefficient. Clustering coefficient C_i of node i is defined as the number of edges among its neighbors over maximal number of possible edges among them:

$$C_i = \Sigma_{j(i),k(i)} \Delta_{jk} / d_i(d_i - 1)$$

where summation is over all neighbors $j(i)$ of node i, Δ_{ij} is the adjacency matrix of the graph, and d_i is the degree of i, i.e. number of its neighbors. Here, however, we explore a graph that has two types of edges (metabolic and associations ones) and we cannot readily apply clustering coefficient. Instead, we introduce a new quantity of cross-clustering coefficient $C(X,Y)_i$. Consider a graph with two types of edges X and Y, and define neighbors of i connected by X-edges $j(i,X)$ (X-neighbors of i). The cross-clustering coefficient is defined as the number of Y-edges between X-neighbors of i, over the total possible number of edges among X-neighbors:

$$C(X,Y)_i = \Sigma_{j(i,X),k(i,X)} \Delta(Y)_{jk} / d(X)_i(d(X)_i - 1)$$

where $\Delta(\square)_{ij}$ is the adjacency matrix of Y-edges and $d(X)_i$ is the X-degree of node i. Note that $C(X,Y)_i \neq C(Y,X)_i$. Averaging cross-clustering coefficient over all nodes gives the mean cross-clustering C(X,Y) of X and Y edges on the graph (again, $C(X,Y) \neq C(Y,X)$).

The cross-clustering coefficient for the native and randomly assigned metabolic networks is shown in Fig. 4.

Correlation between pathway distance and functional association score

Here we want to estimate how likely it is for the enzymes catalyzing two reactions short metabolic distance apart to be functionally related as measured by the functional association score.

Functional association score indicates three levels of confidence(2, 3). Score greater than 700 indicates high confidence of functional association, score between 400 and 700 — medium confidence, and score between 100 and 400 — low confidence.

Fig. 6.

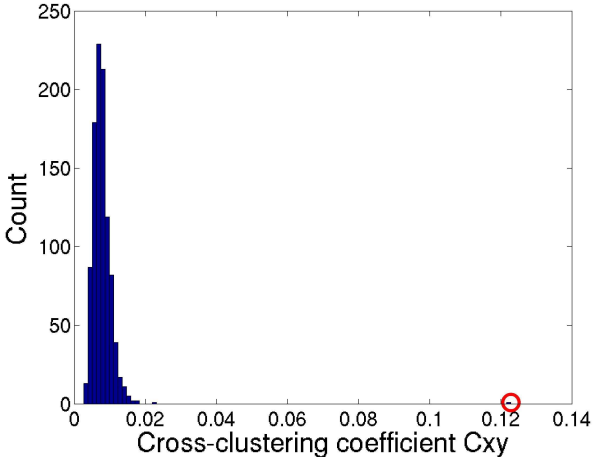

FIG. 4. Distribution of the cross-clustering coefficient for randomly reassigned (blue bars) and the native metabolic-genomic network for the association cutoff of 400.

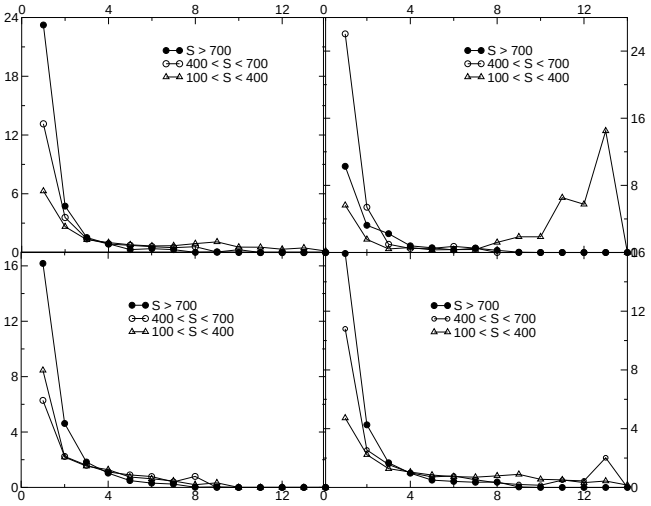

FIG. 5. Correlation between distance and functional association score — ratios of the numbers or observed reactions on the real and randomly rewired graphs. Top to bottom and left to right — functional association score based on neighborhood in the genome, domain fusion, philogenetic cooccurrence, and combined score.

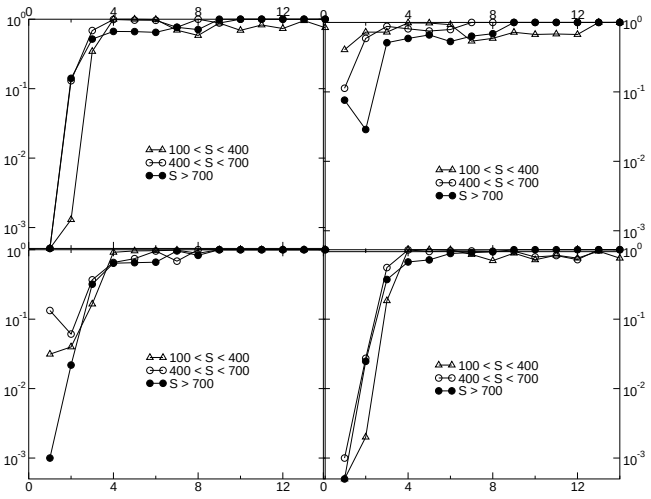

FIG. 6. Correlation between distance and functional association score — probability to observe at least the same number of reactions on a randomly rewired graph as on the real graph. Top to bottom and left to right — functional association score based on neighborhood in the genome, domain fusion, philogenetic cooccurrence, and combined score.

To investigate the correlation between metabolic distance and score, we calculate the number of reactions with a given confidence of functional association edges as a function of metabolic distance and compare this number to the one expected on a random graph.

By a random graph we mean a graph which preserves metabolic links between reactions but randomly shuffles regulatory links. We generated the random regulatory graph in two ways. One is in the table R_1 R_2 Score every reaction ID R_1, R_2, ... is replaced by another ID which is assigned at random from the list of all available IDs. This corresponds to random enzymes catalyzing chemical reactions on the graph where both the metabolic network and all the functional associations between enzymes are preserved. The other method randomly shuffles the scores around the $N \times N$ matrix of scores.

The graph was rewired 1,000 times by either method, and the histogram of correlation of distance and coregulation was averaged over these trials.

Shuffling reaction IDs and scores gives the same results within the stochastic error. For our histograms we used therefore the average of the points obtained by the two methods.

Fig. 5 shows the ratios of the number of reactions with the three levels of confidence of funcitonal association scores on the real graph to the number of reactions within the same ranges on a graph with randomly rewired regulatory connections. The X axes show metabolic distance – the number of distinct metabolic steps separating the two reactions. The ratios are on the Y axes.

Alternatively, we can calculate how likely it is to observe the same or bigger number of reactions with a certain level of confidence of functional association edges on a random graph as on the real graph. These probabilities as a function of metabolic distance are shown in

An important question to investigate is whether the fact that enzymes that belong to the same module of the metabolic network also tend to be genetically associated can be explained by operon organization of functionally linked genes.

RegulonDB database(4) contains 770 known oper-

ons as of September 2005. 356 of these have two or more genes. To investigate the effect of operon organization on our results, we excluded all functional edges between genes known to be in the same operon in our metabolic-genomic graph, and performed the analysis of this modified graph on the macro- and meso-scale.

On the original graph 1254 functional links have scores greater than 700. Removal of edges between same-operon members reduces this number to 1155. For scores between 400 and 700 the corresponding numbers are 875 and 828, and for scores between 100 and 400 — 23138 and 23047. In addition, some functional edges on the graph are due to the fact that some reactions along the pathway or in different pathways are catalyzed by the same enzyme, leading to the weight of the functional edge being 1000 and not being influenced by the edge removal. Therefore one would expect the effect of removal of edges between genes in 356 operons to be minimal. This expectation on is confirmed in Figs. 7 and 8.

A spike in ratios and drop in probabilities at some values of metabolic distances at the tails of the plots can be explained by a very small (less than one) number of reactions expected on a random graph and occasional occurrence of one or very few such reactions on the non-rewired graph.

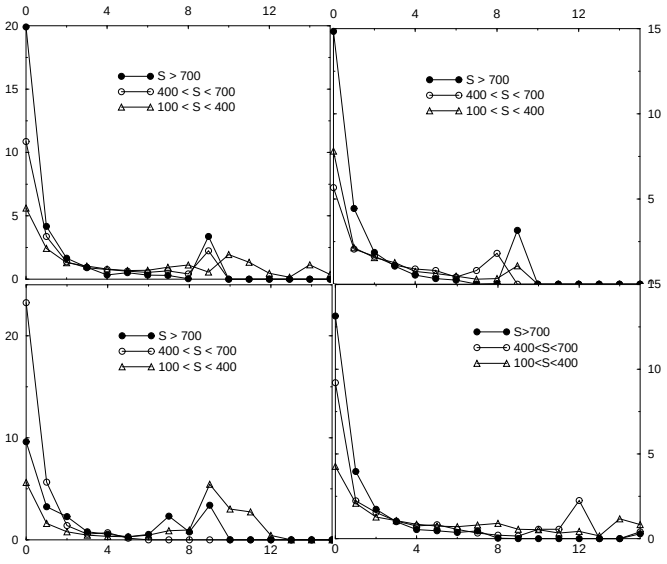

FIG. 7. Correlation between distance and functional association score — ratios of the numbers or observed reactions on the real and randomly rewired graphs with removed functional edges between members of the same operon. Top to bottom and left to right — functional association score based on neighborhood in the genome, domain fusion, philogenetic cooccurrence, and combined score.

B. Mesoscale analysis

DxC graph clusters

The edges on this graph are the product of metabolic and functional association connection — an edge exists when there is both a metabolic link and a functional link with a score greater than a certain cutoff. This graph splits into several connected components, depending on the cutoff on the functional association score. We search for these components using depth-first search algorithm(5).

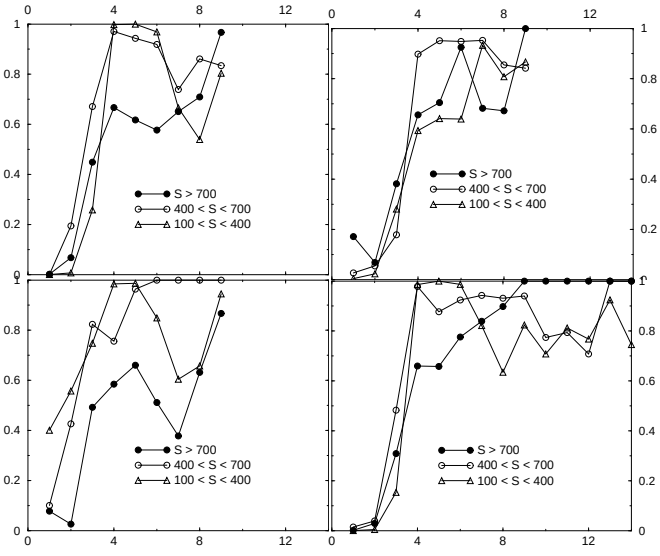

FIG. 8. Correlation between distance and functional association score — probability to observe at least the same number of reactions on a randomly rewired graph as on the real graph. Functional edges between same-operon enzymes are removed. Top to bottom and left to right — functional association score based on neighborhood in the genome, domain fusion, philogenetic cooccurrence, and combined score.

DUC graph clusters

This is the graph where edges of both types are present, and clusters on this graph are defined such that every node in the cluster is connected to every other node by a path through both metabolic and regulatory links, and every node on this path belongs to this cluster.

To search for clusters on this graph, we start with searching for connected components on the metabolic graph (call it blue-edge graph). For every connected "blue" component found, we look for parts of this component which are connected components on the functional association graph (red-edge graph). But these "blue-red" components, in turn, may now be disconnected on the metabolic graph, since some of the "blue" paths between their nodes may have been passing through nodes which were not connected with "red" paths. We therefore repeat the search for connected components on the metabolic "blue" graph within "blue-red" connected components. Then the search on the "red" graph is performed on these "blue-red-blue" clusters. This repetitive

search continues until no more partitions of the clusters can be obtained (see Fig. 9 for an illustration.) For most clusters we found the search converged after four or five iterations.

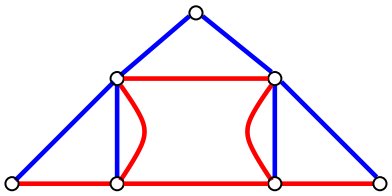

FIG. 9. An illustration of a search for clusters on DUC graph. The first approximation will find that all seven vertices belong to a connected conponent on the blue graph. The second approximation will exclude the top point from this set, but the remaining six points, although a connected component on the red graph, are not a cluster on the blue graph. The third approximation will finally identify the two clusters such that every node in the cluster is connected to every other node by a path through both blue and red links, and every node on this path belongs to this cluster.

Monte Carlo simulation

This method was described in(6). Here we will briefly recapitulate the main ideas for the graph of one type of edges, and then describe the generalization for the *E.coli* metabolic graph.

This method is used to find a tight subgraph of a predetermined number of nodes M. At time $t = 0$ a random set of M nodes is selected. For each pair of nodes i, j from this set the shortest path L_{ij} between i and j on the graph is calculated. Denote the sum of all shortest paths $\sum_{ij} L_{ij}$ for this set as L_0. At every time step one of M nodes is picked at random, and one node is picked at random out of all its neighbors. The new sum of all shortest paths L_1 is calculated if the original node were to be replaced by this neighbor. If $L_1 < L_0$, the replacement takes place with probability 1, if $L_1 > L_0$ — with probability $\exp(-(L_0 - L_1)/T)$, where T is the effective temperature. Every tenth time step an attempt is made to replace one of the nodes from the current set with a node which has no edges to the current set to avoid getting caught in an isolated disconnected subgraph. This process is repeated until the original set converges to a complete subgraph, or for a predetermined number of steps, after which the tightest subgraph (the subgraph corresponding to the smallest L_0) is recorded.

On the *E.coli* metabolic graph two types of edges are present. We calculate the shortes path for every pair of nodes on the metabolic graph. For regulatory graph we use the empirical rule to transform decreasing functional association confidence into increasing edge lengths. High-confidence scores are assigned a edge length $l = 1$, medium-confidence — edge length $l = 2$, low confidence — edge length $l = 4$, and score below 100 — edge length $l = 16$, which is chosen to be larger than the longest path length between any two connected reactions on the graph of metabolic edges.

The value to optimize is the generalized "potential energy"

$$E_{gen} = d^2 + Rl^2$$

where d is the length of the shortest metabolic path

The Monte Carlo simulation is run starting with a connected set of nodes, meaning every node is a neighbor of at least one of the other nodes. At every step when a node is picked, an attempt is made to replace it with a neighbor of any of the remaining nodes rather than its neighbor or an arbitrary node on the graph. The replacement is made by the same rules as in the previous paragraph. The recorded clusters are merged and redundant clusters are removed.

Examples of modules mapped on metabolic pathways

The cysteine pathway breaks into two modules (*cysDN, cysC, cysH, cysIJ*) and (*cysE, cysK, cysM, metA, metB*), the latter containing two genes of the methionine pathway. This way the cysteine and methionine pathways are re-distributed between the modules that look reasonable from the biochemical point of view (Fig. 10).

Another unexpected mode of genomic association is observed in the pathways of purine and pyrimidine biosynthesis. These pathways are linked together by a single module (Fig. 11). Such fusion of purine and pyrimidine pathways may be reflected in co-regulation of their genes by *E. coli* by the PurR transcription factor. Purine biosynthesis is also split at the IMP junction, revealing IMP to GMP production line as a single module (*guaA, guaB, guaC*). This separation seems surprising since *guaA* and *guaB* are also regulated by PurR. However, weak genomic associations with other genes in the pathway bring *guaA-guaB-guaC* into a separate module.

Central metabolism. Few modules are present in the large pathways of the central metabolism (glycolysis, pentosephosphate pathway, the Krebs (TCA) cycle, respiration) (Fig. 12).

Although strict thresholds yield only small clusters of associated reactions (e.g. a module of nonoxidative branch of the pentose phosphate pathway), large super-pathway modules containing representatives from several pathways are obtained at low thresholds. For example, a part of the EMP pathway, degradation of several carbon sources, and the nonoxidative branch of the pentose phosphate pathway form a single super-pathway module.

Lack of modules mapping to traditional pathway in the central metabolism suggests high diversity in its structure and evolution in bacteria, as well as complexity of its regulation (e.g. a cascade of 11 transcription factors regulating 3 genes, *aslL, zwf*, and *gnd*, in the pentose phosphate pathway(7). This agrees with observations of Glazko and Mushegian(8) and earlier analyses

of Dandekar *et al.*(9) and Huynen *et al.*(10) who demonstrated high diversity of the Krebs cycle and the glycolysis pathway. Examples of super-pathway modules (obtained mostly by the Monte Carlo search) include cell wall and membrane biosynthesis, biosynthesis of certain amino acid whose genes demonstrate strong linkage, central metabolism (see above), enterochelin, and tetrapyrrole pathways, thus corresponding to large functional systems.

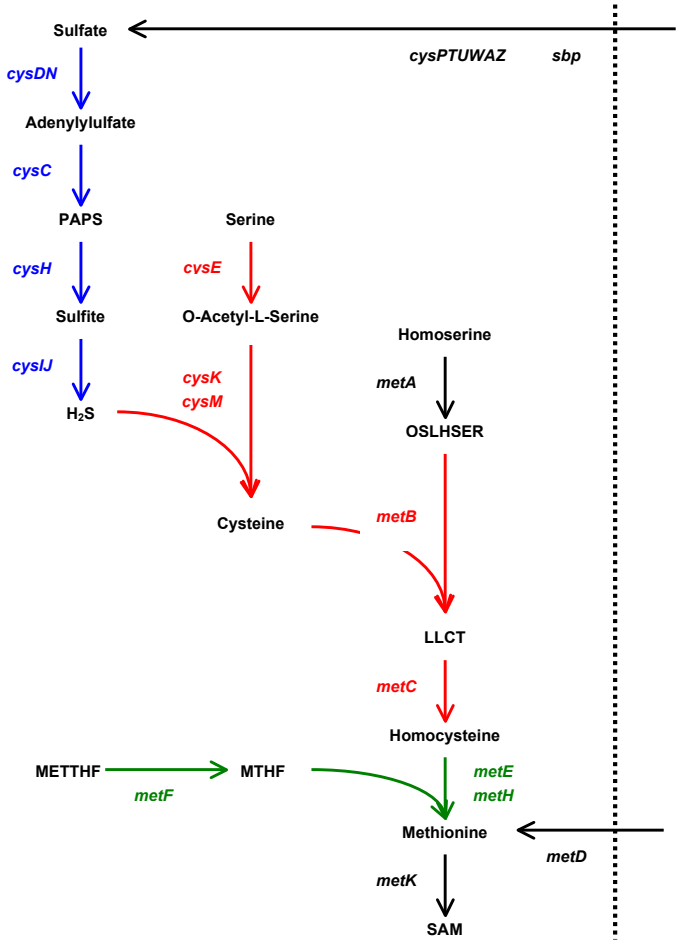

FIG. 10. Cystein and methionine biosynthesis. Left and middle: systein; right: methionine; horizontal: one-carbon metabolism (partial), broken line: membrane with transporters. Colored (blue, red, green): clusters.

Subunits *hisP, hisQ, hisM* have strong association between themselves ($S > 900$), while week ($100 < S < 300$) with *hisJ* subunit of the *hisPQMJ* ATP-dependent histidine transport system. The origin of this weak link becomes apparent if one recalls that *hisPQM* subunits can also function as lysine/arginine/ornithine transporter when working with *argT* protein. The *hisPQM* system achieves its flexibility by using either *ArgT* or *HisJ* as the periplasmic component. This example suggests that weak association between subunits can indicate functional diversity of the complex.

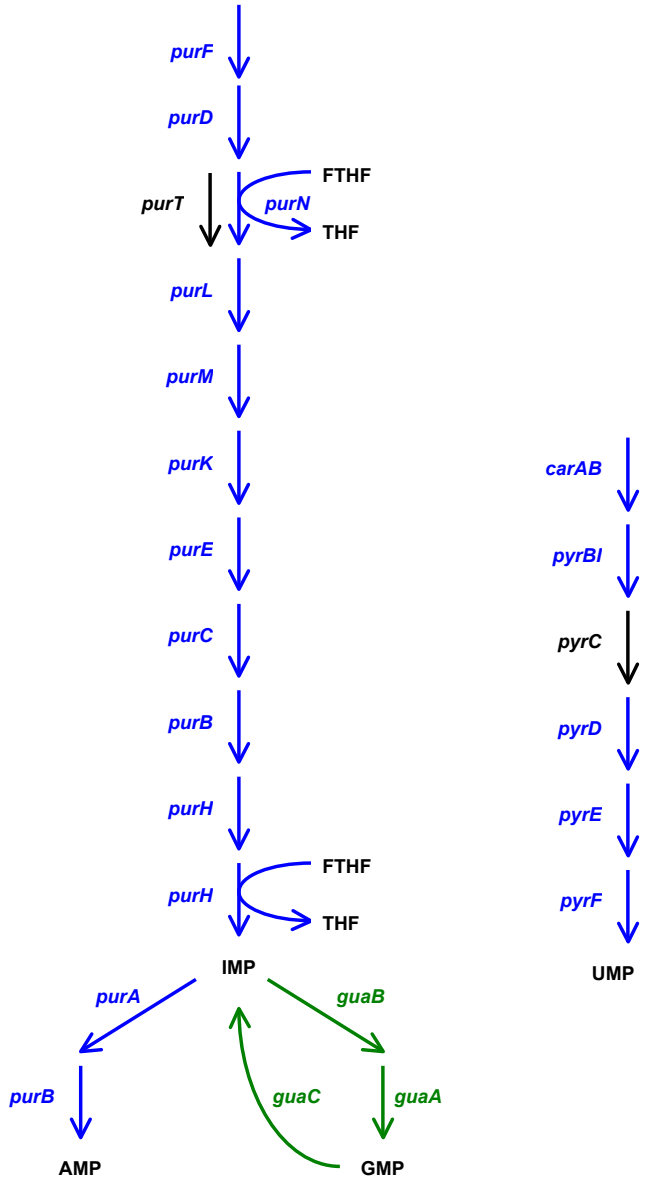

FIG. 11. Purine (left) and pyrimidine (right) pathways. Blue: hybrid purine-pyrimidine module. Green: GMP module.

Figs. 14 and 15 show two more examples of modules: fucose and rhamnose pathways and clusters and aromatic amino acids and folate pathways.

Effect of operon-related edges.

As we argued in *Macroscale* section, the effect of strong functional edges due to the presence of the corresponding enzymes in the same operon is expected to be minimal. We marked the operon-related functional edges in the cluster tables on the project web site in the *Mesoscale* section. As one can see, while the density of functional edges in some of the modules will be lower, these modules will remain statistically significant.

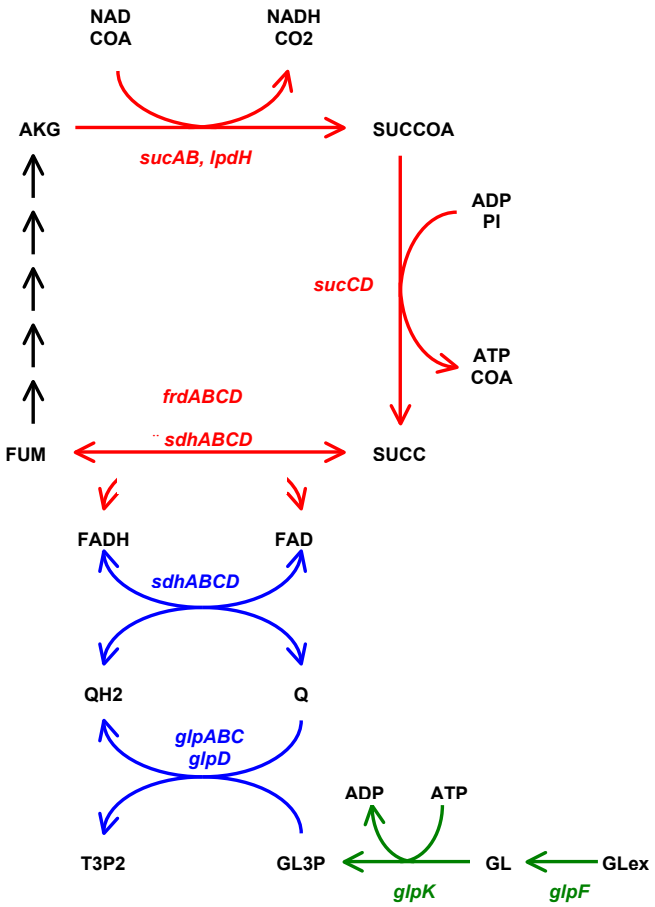

FIG. 12. A hybrid cluster. Red — the Krebs (TCA) cycle. Blue — respiration (partial). Green — alternative carbon sources (glucerol).

III. STATISTICAL SIGNIFICANCE

To estimate statistical significance of the identified modules, we introduced generalized Q value as a measure of the density of connections within the module. Our Monte Carlo technique that minimizes the sums of metabolic and functional distances equivalently maximizes this Q value.

The generalized Q value is defined as follows

$$Q_{gen} = \frac{\Sigma_{i=1}^{n}\Sigma_{j=i+1}^{n} d_{ij} + \Sigma_{i=1}^{n}\Sigma_{j=i+1}^{n} f_{ij}}{n(n+1)},$$

where d_{ij} is metabolic edge between vertices i and j, f_{ij} — weighted functional edge, n — the cluster size. The denominator reflects the fact that there can be at most $n(n+1)/2$ metabolic edges and $n(n+1)/2$ functional edges. The weight of metabolic edge is 1, the weights of functional edges are as follows: $f = 1$ for association score $S > 700$, $f = 1/2$ for $400 < S < 700$, $f = 1/4$ for $100 < S < 400$, and $f = 1/16$ for $S < 100$. This is consistent with our choise of functional edge length for Monte Carlo optimization.

The probability to observe a cluster with connections density no less than Q_{gen} approximately follows the Fisher-Tippett extreme value distribution (EVD)

$$P_{evd}(m) = \exp\left(-\exp\left(-\alpha(Q_{gen} - u)\right)\right) \qquad (1)$$

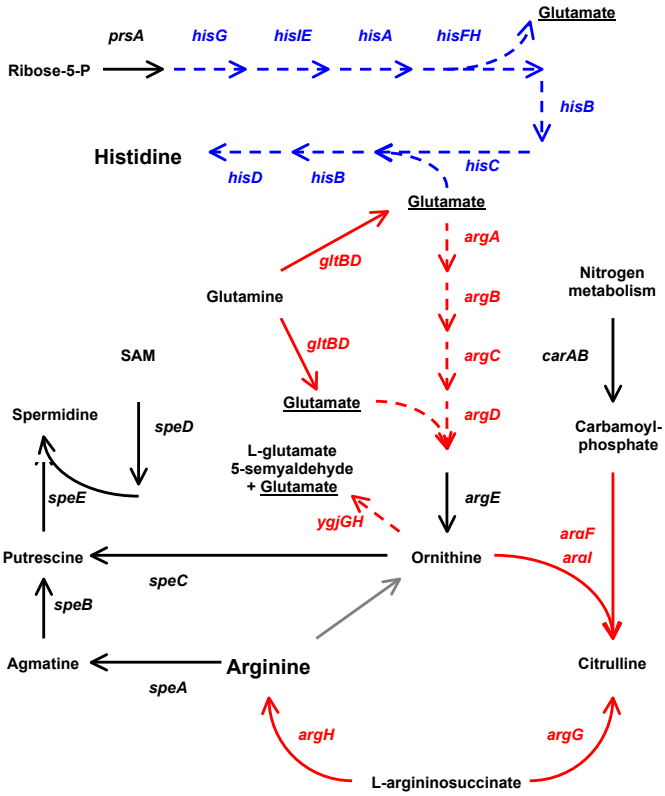

FIG. 13. Arginine and histidine pathways. Left (*spe* genes): spermidine/putrescine biosynthesis, not in any cluster. Red: arginine biosynthesis (X4). Blue: histidine biosynthesis (X4). Broken: arginine-histidine biosynthesis (U7).

where α and u are parameters of the distribution.

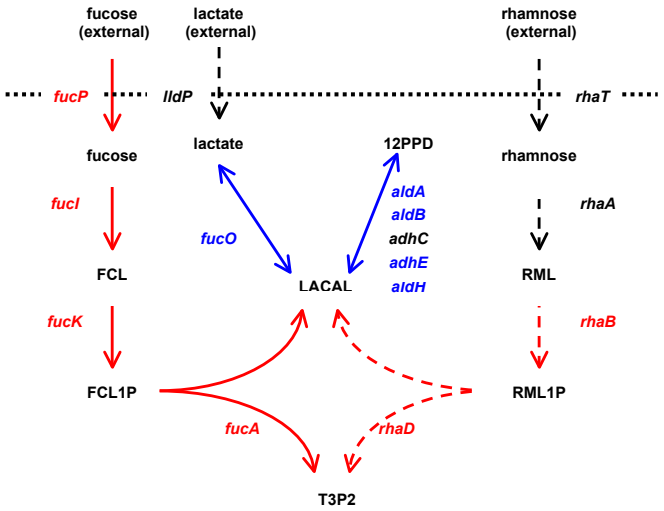

FIG. 14. Fucose and rhamnose pathways and clusters. Fucose pathway: solid lines. Colored (red and blue): two clusters.

We used this property to estimate statistical significance of Monte Carlo clusters. First, we generate 1000 randomly rewired networks and run MC search on each of them. This way we obtain clusters with Q_{gen} obeying EVD and derived parameters $\alpha(n)$ and $u(n)$ as a function of cluster size n. Second, we analyzed clusters discovered in the real network and computed P_{evd} for each of them using equation 1. We noticed that $\alpha(n)$ and $u(n)$ scale approximately linearly with the cluster size n.

$$\alpha(n) = \frac{1}{a_1 n + a_2}; \quad u(n) = u_1 n + u_2 \qquad (2)$$

allowing computation of P_{evd} for a cluster of any size n.

To discard the statistically insignificant clusters, we then calculate the expectation value E_{evd} to encounter cluster with given size and density of connections out of N network vertices. However, the question arises as to the number of nodes to choose from. One cannot simply use N, since the network may contain many nodes with very low connectivities which are smaller than the lowest connectivity of any node in a densely connected cluster. We therefore approximate the expectation value by $E_{evd} = P_{evd}\binom{N^\star}{n}$, where we choose $N^\star$ to be the number of vertices on the network with connectivities larger than the average connectivity of the cluster. Clusters with $E_{evd} < E_{cutoff} = 0.1$ are said to be statistically significant. Fig. 16 presents distribution of Q_{gen} and their EVD approximations obtained using randomly rewired networks together with the clusters discovered in the real network.

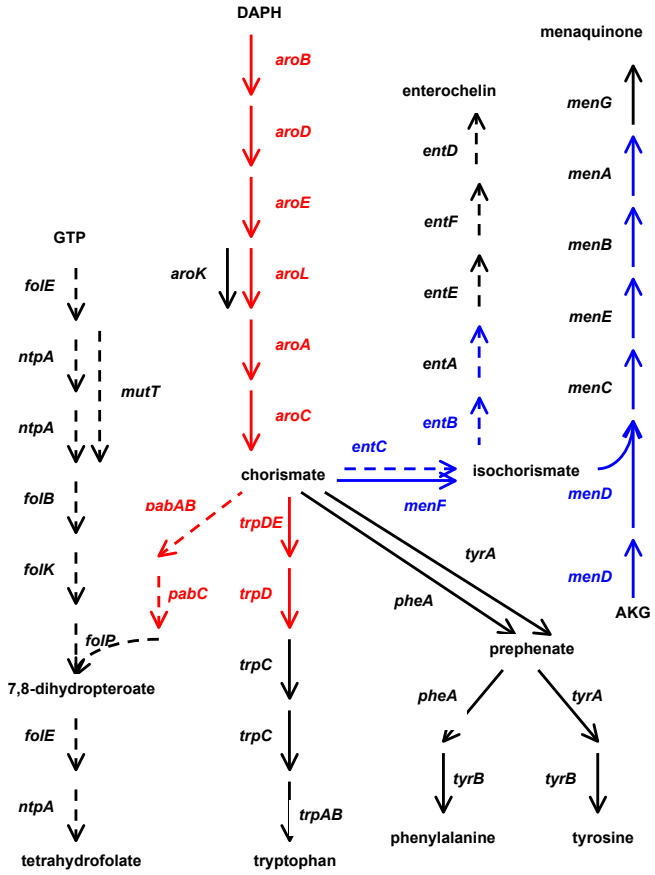

FIG. 15. Aromatic amino acids and folate pathways. Pathways: folate — broken, left; aromatic a-a — solid, left; enterochelin — broken, right; menaquinone — solid, right. Clusters: aromatic/folate — red; enterochelin/menaquinone — blue.

For clusters obtained by exact enumeration algorithms ($D \times C$ and $D \cup C$) we estimated the statistical significance by a different approach. The original graph was reshuffled 10,000 times and for each reshuffle all $D \times C$ and $D \cup C$ clusters were identified and the density of metabolic and association links within them recorded. A cluster found on the original network is statistically significant if we find no more than 100 clusters with higher density of connections in 10,000 rewired networks. This corresponds to E value of 0.01.

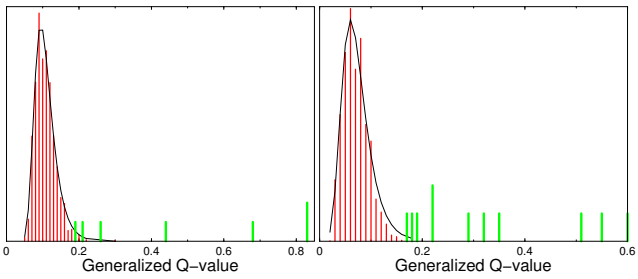

FIG. 16. The distribution of Q values of clusters on the rewired network (red vertical bars) and its approximation by extreme value distribution (solid black line). Green bars correspond to Q-values of the clusters found in the real network. Left — ratio of metabolic and functional edges is one, cluster size = 15. Right — ratio of functional and metabolic edge strengths equals 16, cluster size = 10.

IV. MICROSCALE

A. Two-reaction junctions

In total there are 275 metabolites that participate in exactly two reactions. Reactions catalyzed by isoenzymes are treated as separate, except if a metabolite is participating in two such reactions, these reactions are not considered forming a pair.

Pairs or neighboring reactions can be divided into six categories (Fig. 17).

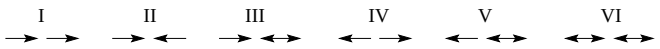

FIG. 17. Pairs of neighboring reactions. Six types if the reversible reactions are viewed as proceding in both directions.

There are only four pairs of type II and three pairs of type IV. All these pairs have functional association score < 100. The histogram of observed versus expected number of reactions as a function of score for the other four types is shown in Fig. 18 (the expected number assumes random rewiring of functional edges). All four types of pairs show similar observed/expected ratios - approximately 2/3 for score > 100 and between 3/1 and 5/1 for score > 700.

But high association scores among reversible junctions may be a consequence of the fact that these junctions are actually the same pattern as type I. To test this, we processed the network viewing reversible reactions as proceeding in one direction only. In this case there are three types of junctions — I, II, and IV. The majority of junctions belong to type I — 236 out of total of 273. Type II has 16 reactions and type IV — 21.

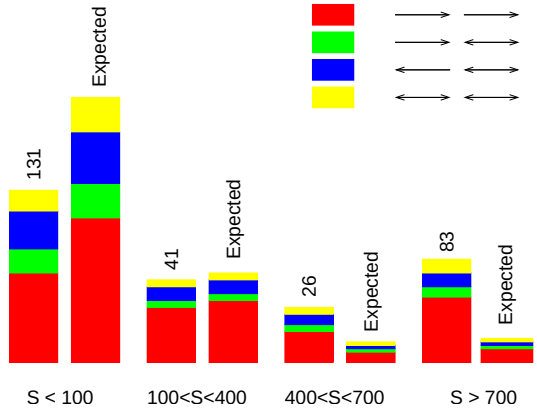

FIG. 18. The histogram of functional association scores for four dominant pair types.

The association among convergent and divergent pairs tends to be similar to that expected on a random association network, while linear flow (type I) two-reaction junctions are several times more often associated than one can expect at random (Fig. 19).

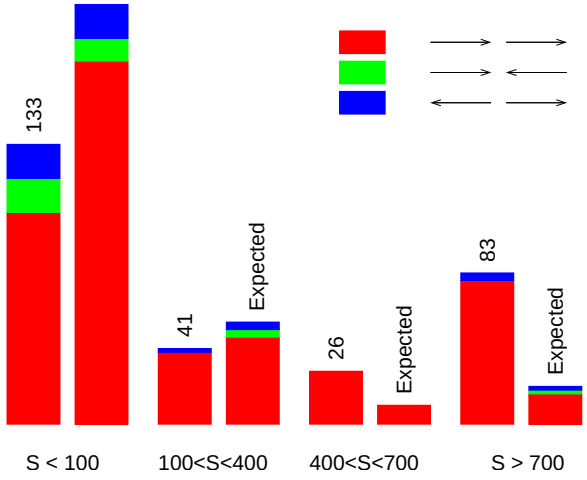

FIG. 19. The histogram of functional association scores for the three pair types when reversible reactions are viewed as proceeding in one direction only.

The full table with numbers of observed versus expected two-reaction junctions is on the web in *Microscale* section.

B. Three-reaction junctions

There are 45 metabolites participating in exactly three reactions. These reactions can be divided into four types (Fig. 20): convergent, divergent, conflicting directions, and reversible three-reaction junctions. In the latter at least one reaction is reversible, while in the three former all reactions are irreversible.

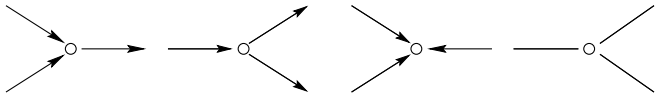

FIG. 20. The four types of three-reaction junctions. Left to right: convergent, divergent, conflicting directions, and reversible junctions.

There is only one conflicting-directions junctions — three irreversible reactions with one of the products being valine. The corresponding enzymes are not functionally associated.

The three-reaction junctions can be divided into five types by the pairwise functional association.

(*i*) no association between any pair
(*ii*) linear — one reaction is associated with only one other reaction along the metabolic flow
(*iii*) linear switch — one of the reactions is catalyzed by isoenzymes, each of which is associated with only one of the other reactions along the metabolic flow
(*iv*) fork — one of the reactions is associated with *both* other reactions along the metabolic flow
(*v*) full association — all three reactions are associated.

There are only seven triplets with one isoenzyme-catalyzed reaction. Only one of those is a linear switch with phenylalanine being the product of one such reaction which then proceeds along two separate paths. The rest of triplets with isoenzyme-catalyzed reactions are linear or unassociated.

Under linear and unassociated categories fall most of three-reaction junctions. At all cutoffs on the functional association score linear junctions occur several times more often than would be expected if the association links were rewired at random (Figs. 21—23).

Among all associated patterns linear association is clearly dominant.

The full table with numbers of observed versus expected three-reaction junctions is on the web in *Microscale* section.

C. Multiple-reaction junctions

There are 131 metabolites participating in more than three reactions. The functional association properties of these multiplets can be characterized by average scores of six types of reation pairs (Fig. 17), or by observed versus expected number of associated pairs of a type within a certain functional association score range.

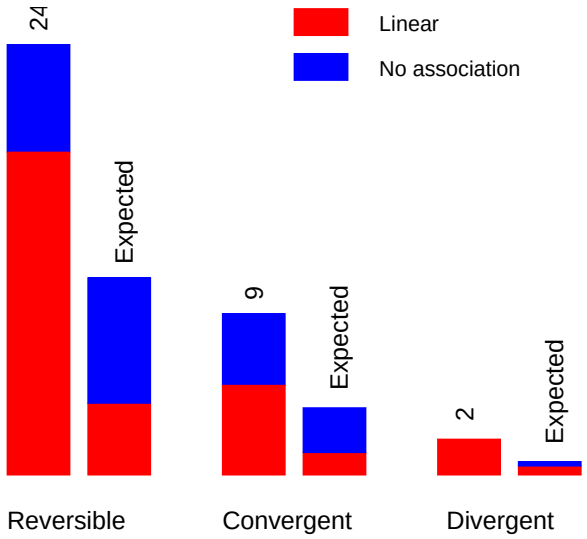

FIG. 21. The histogram of observed versus expected numbers of associated three-reaction junctions with the association score cutoff 100.

With the average network score $S_{\text{network}} = 135$ the highest average score is observed among pairs of type VI — $S_{\text{VI}} = 252$. Pairs of type II and IV have the score similar to the network average ($S_{\text{II}} = 123$ and $S_{\text{IV}} = 144$), while the remaining three types have lower average score — $S_{\text{I}} = 65$, $S_{\text{III}} = 84$, and $S_{\text{V}} = 56$.

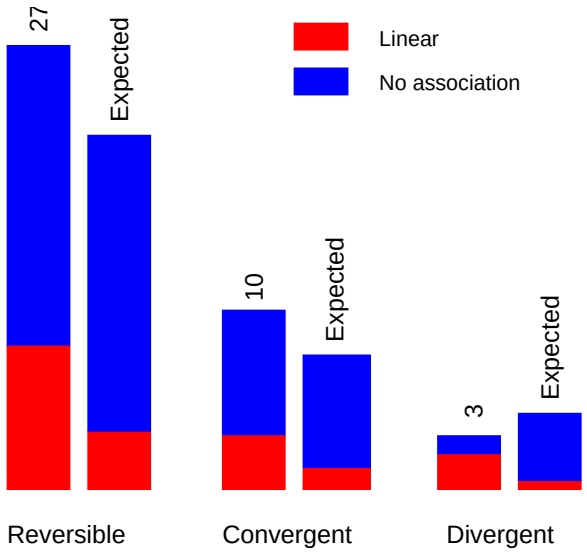

FIG. 22. The histogram of observed versus expected numbers of associated three-reaction junctions with the association score cutoff 400.

Similar results are observed on the histogram of observed versus expected numbers of pair types (Fig. 24).

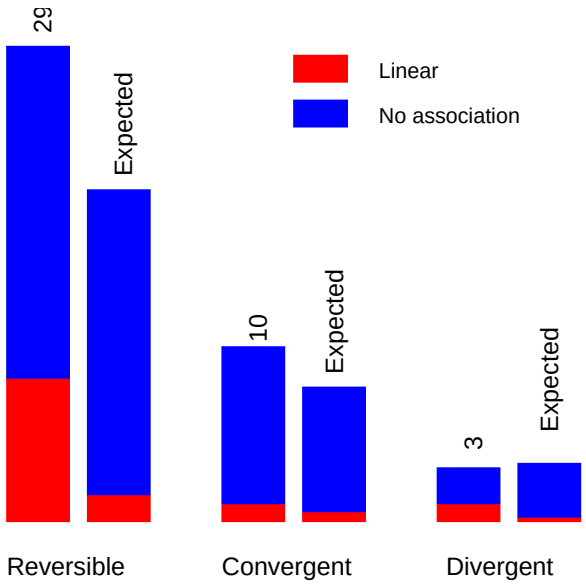

FIG. 23. The histogram of observed versus expected numbers of associated three-reaction junctions with the association score cutoff 700.

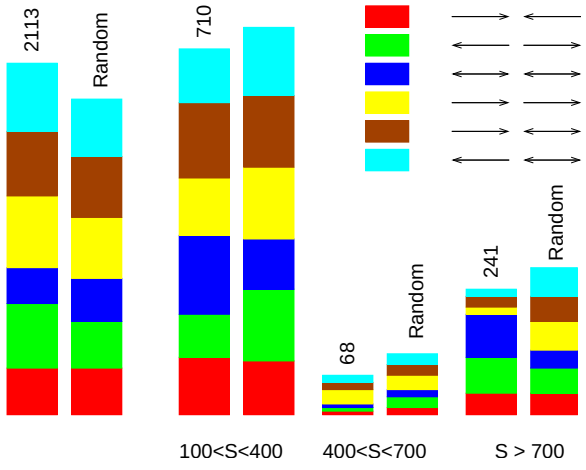

FIG. 24. The histogram of functional association scores for six multiplet types. $S < 100$ columns not to scale.

As in two-reaction junctions, viewing reversible reactions as flowing in one direction only supports the linear metabolic flow. The association of type II and IV pairs is approximately the same as would be expected on a random association network, while tipe I junctions are underassociated. In an n-reaction junction there are $\sim n^2$ possible pairs, and $\sim n$ possible linear flow directions. The fact that type I reactions are underassociated in n-reaction junctions therefore indicated that only a few flow directions are favored, which is consistent with the linearity of metabolic network assumption. The histogram of associated pairs is shown in Fig. 25.

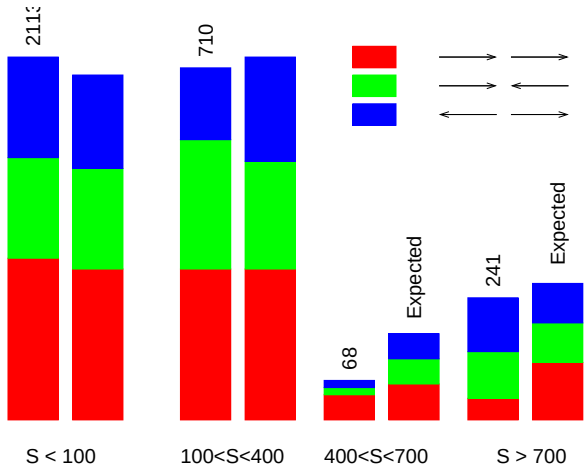

FIG. 25. The histogram of functional association scores for three multiplet types — reversible reactions proceed in one direction only. $S < 100$ columns not to scale.

The full table with numbers of observed versus expected multiple-reaction junctions is on the web in *Microscale* section.

References

1. UCSD Genomics Circuits Research Group. *E.coli* metabolic reactions database Version 1.01. http://gcrg.ucsd.edu

2. Snel, B., Lehmann, G., Bork, P., & Huynen, M.A. (2000) *Nucleic Acids Res.* **28**, 3442-3444.

3. von Mering, C., Huynen, M.A., Jaeggi, D., Schmidt, S., Bork, P., & Snel, B. (2003) *Nucleic Acids Res.* **31**, 258-261.

4. Salgado H, Gama-Castro S, Peralta-Gil M, Diaz-Peredo E, Sanchez-Solano F, Santos-Zavaleta A, Martinez-Flores I, Jimenez-Jacinto V, Bonavides-Martinez C, Segura-Salazar J, Martinez-Antonio A, Collado-Vides J. (2006) *Nucleic Acids Res.* **34**,D394-7.

5. R. Sedgewick, *Algorithms in C++*, (Reading, Mass.: Addison-Wesley, 1992).

6. Spirin, V. & Mirny, L.A. (2003) *Proc. Natl. Acad. Sci. USA* **100**, 12123-12128.

7. Keseler, I.M. *et al* (2003) *Nucleic Acids Res* **31** 258-261.

8. Glazko, G.V. & Mushegian, A.R. (2004) *Genome Biol.* **5**, R32.

9. Dandekar, T., Schuster, S., Snel, B., Huynen, M., & Bork, P. (1999) *Biochem J* **343 Pt1**, 115-124.

10. Huynen, M., Dandekar, T., & Bork, P. (1999) *Trends Microbiol* **7**, 281-291.