Genome Biology

Genome Biology
Biology for the post-genomic era

Metabolism evolution by gene duplication: a network perspective

Genome Biology 2007, **8**:R26 doi:10.1186/gb-2007-8-2-r26

Juan Javier Diaz-Mejia (jdime@ibt.unam.mx)
Ernesto Perez-Rueda (erueda@ibt.unam.mx)
Lorenzo Segovia (lorenzo@ibt.unam.mx)

ISSN 1465-6906

Article type Research

Submission date 4 July 2006

Acceptance date 27 February 2007

Publication date 27 February 2007

Article URL http://genomebiology.com/2007/8/2/R26

Metabolism evolution by gene duplication: a network perspective

Juan Javier Díaz-Mejía [1], Ernesto Pérez-Rueda [1], Lorenzo Segovia [1*]

1 Departamento de Ingeniería Celular y Biocatálisis, Instituto de Biotecnología, Universidad Nacional Autónoma de México. Av. Universidad 2001, Col. Chamilpa, Cuernavaca, Morelos, México. CP 62210.

E-mail addresses:
JJDM: jdime@ibt.unam.mx
EPR: erueda@ibt.unam.mx
LS: lorenzo@ibt.unam.mx
***** Corresponding author

Abstract

Background

Gene duplication, followed by divergence, is known as a main source of metabolic versatility. The patchwork and stepwise models help to understand these processes, but their assumptions are relatively simplistic. In this work, we used a network based approach to determine the influence of metabolic constraints on the retention of duplicated genes.

Results

We detected duplicated genes looking for enzymes sharing homolog domains. Our results uncover an increased retention of duplicates between reactions catalyzing consecutive reactions, as illustrated by the ligases acting in the biosynthesis of peptidoglycan. As a consequence, metabolic networks show a high retention of duplicates within functional modules. We found a preferential coupling of reactions that partially explain this bias. A similar behavior was found in enzyme-enzyme interaction networks, but we failed to detect it in gene transcriptional regulatory and non-enzymatic protein-protein interaction networks. Thus, we suggest that this bias results from laws governing substrate-enzyme-product relationships. Additionally, our findings confirm a high retention of duplicates between chemically similar reactions, as illustrated by the origin of fatty acids metabolism. However, the retention of duplicates between chemically dissimilar reactions is also greater than expected by chance. Finally we detected a significant retention of duplicates as groups, instead of single entities.

Conclusions

Collectively, our findings provide evidence that *in silico* models trying to explain the origin and evolution of metabolism are improved with the inclusion of specific functional constraints, such as the preferential coupling of reactions. Our findings suggest that the stepwise and patchwork models are not independent of each other. In fact, the network perspective used herein permits us to reconcile and combine these models.

Background

The classical view of metabolism is that relatively isolated sets of reactions or pathways allow the synthesis and degradation of compounds. The new perspective views metabolic components (substrates, products, cofactors, and enzymes) as part of a single network. Defining metabolism as pathways is not always straightforward because some functional properties, such as the closeness between reactions from different pathways, are visible only when metabolism is analyzed from a network perspective [1]. A way to do this is representing metabolism with a compound-centric network, wherein nodes (substrates and products) participating in the same reaction are connected. Otherwise, in an enzyme-centric network, nodes (enzymes) producing a compound are connected with other nodes consuming the same compound. These tools have shown that metabolism has a scale-free topology [2, 3], meaning that the majority of nodes show a low degree of connectivity, but a small fraction of highly connected nodes dominates the topology of the network. Another property of metabolic networks is their hierarchical modularity [4, 5], showing groups of highly clustered, functionally related nodes.

Recent models have successfully simulated the origin of scale-free networks by gene duplication [6], while their modular organization has been explained by the preferential attachment of new nodes to the most connected preexisting ones [5]. Nevertheless, these models do not take into account the functional constraints of metabolism [6]. For instance, carbon-nitrogen ligases (EC:6.3.-.-) tend to act consecutively, reducing their chance to associate with enzymes catalyzing other reaction types (Figure 1). We call this property "preferential coupling of reactions", and suggest that it reflects a biochemical necessity, for example to synthesize the peptidoglycan of bacterial cell wall. Our results evidence the importance of including functional constraints to improve the models of the origin and evolution of metabolic networks. In fact, a recent model simulating the origin of highly connected compounds [7] is significantly improved when reactions are considered as coupled reaction pairs, instead of single entities.

The first hypotheses on the origin and evolution of enzyme-driven metabolism were formulated based on the idea that gene duplication, followed by divergence, can lead the origin of new metabolic reactions. The two pioneering models about this paradigm, *stepwise* [8] (or retrograde) and *patchwork* [3] exhibit two main differences: i) The *stepwise* model posits that, as consequence of the depletion of a substrate, gene duplication can provide enzymes capable of supplying the exhausted substrate, giving rise to homolog enzymes catalyzing consecutive

reactions. Whereas the *patchwork* model postulates that duplication of genes encoding promiscuous enzymes (capable of catalyzing various reactions) allows each descendent enzyme to specialize in one of the ancestral reactions. In this regard, enzymes generated by *patchwork* can catalyze reactions separated by distances greater than those originated by *stepwise*. ii) The *stepwise* model invokes consecutive reactions, hence it can originate enzymes catalyzing chemically dissimilar reactions (CDR) but preserving specificity for the type of substrate [9, 10]. In contrast, the *patchwork* model considers that promiscuous enzymes tend to catalyze chemically similar reactions (CSR) even while acting on different types of substrates [9, 10]. A simple way to determine whether similar reactions are, is comparing the first two digits of their EC numbers (EC:a.b.-.-) [10-12].

Some authors have used the differences between the *stepwise* and *patchwork* models attempting to clarify their possible contribution to specific cases in the evolution of metabolism. Collectively, these analyses suggest the *patchwork* model as the most common mechanism generating metabolic versatility [9-12]. A major difficulty with these analyses is the significant fraction of consecutive CSR catalyzed by homolog enzymes [10, 11]. The *stepwise* model could explain the origin of such reactions because they are consecutive, but also the *patchwork* model could do because they are CSR. For example, amidophosphoribosyl transferase and xanthine phosphoribosyltransferase, are homolog enzymes catalyzing consecutive reactions. Thus, their origin could be attributed to the *stepwise* model. Nonetheless they catalyze CSR and hence their origin could also be explained by the *patchwork* model (Figure 1A). Similarly, the origin of four homolog carbon-nitrogen ligases of the peptidoglycan biosynthesis agrees with the *stepwise* model because they catalyze consecutive reactions, but also with the *patchwork* model because they catalyze CSR [10] (Figure 1B). Here, we determined that the fraction of consecutive CSR is significantly greater than the one expected by chance (see below), implying that the origin of such reactions can be explained complementary by *stepwise* and *patchwork*. We suggest that using a network based approach these two models can be reconciled.

In this article we reconstructed the *E. coli* K12 and multiorganismal enzyme-centric metabolic networks according to the BioCyc [13, 14] and the KEGG [15] databases. The protein sequences of their enzymes were compared to detect duplicated genes (hereafter called duplicates). We evaluated the influence of both the chemical similarity and the distance between reactions on the rate of

retained duplicates. We also estimated whether the preferential coupling of reactions and the modularity of networks affect this rate. Finally, we detected cases in which duplicates have been retained as groups and determined how general this mechanism is.

Results and Discussion

The preferential coupling of reactions in metabolic networks reflects a functional constraint

Metabolism follows logical rules implying that specific reactions and fluxes are temporally and spatially compartmentalized [16]. We searched for some of these rules in our data, determining whether the combination of reactions types (EC:a.b.-.-) is constrained by biochemical necessities, or it simply results from random processes. To do this, we determined the frequency for each reaction pair type (EC:a.b.-.- → EC:w.x.-.-) in metabolic networks, and compared it against the values expected by chance. To calculate expected values a set of null "Maslov-Sneppen" models [17] was generated. The models are randomly rewired versions of the original network, preserving the degree of connectivity for each node (see Methods). The results evidence that certain reaction types tend to act consecutively (Figure 1D). To illustrate the biological relevance of this finding consider the case of carbon-nitrogen ligases (EC:6.3.-.-), which tend to be followed by other EC:6.3.-.-, for example to synthesize peptidoglycan (Figure 1B). In fact, a recent study uncovers that metabolites also show a preferential coupling [18]. We consider that these biases reflect underlying biochemical mechanisms and substrate stoichiometric necessities. In the following sections we discuss the relevance of this finding to the retention of duplicates.

Influence of chemical similarity on the retention of duplicates

We computed the retention of duplicates originating both CSR and CDR. The resulting frequencies were compared against the values expected by chance, using "Maslov-Sneppen" models, to determine whether they can be attributable to a biological pressure. Figure 2A shows that retention of duplicates between CSR is 6-fold greater than the observed between CDR. This agrees with previous reports [10-12]. However, note that duplicates leading both CSR and CDR at distances < 3 are more frequent than expected by chance (Z-score > 3, $P < 0.001$). The main implication of this finding is that the retention of duplicates generating both CSR and CDR is not a

random process, but reflects underlying biological phenomena. Thus, gene duplication is an important source of metabolic variability, but also of biochemical innovations.

Influence of distance between reactions on the retention of duplicates

In addition to the retention of duplicates generating CSR and CDR, Figure 2A shows an increased retention of duplicates between closer reactions. This is between consecutive reactions, or between reactions separated by one or two metabolic steps. The explanation of this phenomenon is non-trivial because there is not a biological trait clearly associable to shorter distance between reactions. Thus, we compared the results from metabolic networks with those from other biological networks to determine whether our observation is general. We identified duplicates within a gene transcription regulatory network [19] and within a validated protein-protein interaction network [20], both from *E. coli*. The regulatory network did not show a significant influence of the distance between transcription factors and target genes on the retention of duplicates (Figure 2C). In contrast, the protein-protein interaction network (Figure 2D) shows an increased retention of duplicates between closer proteins. A more detailed analysis uncovers that this increase is mainly influenced by enzyme-enzyme interactions. In fact, the fraction of non-enzymatic duplicates, mainly composed by replication, transcription, translation and protein folding complexes, is not significantly different from random (Z-score < 3, $P > 0.001$). Thus, it seems that the increased retention of duplicates between closer proteins is characteristic of metabolic networks and enzyme-enzyme complexes. From this observation, we propose that laws governing substrate-enzyme-product relationships in metabolic networks are different from those acting on protein-DNA and non-enzymatic protein-protein interactions. A possible reason for this is that metabolic interactions have a small molecule as intermediate, the substrate/product compound, whereas the protein-protein and protein-DNA interactions require larger surfaces, and their retention could be more difficult. In fact, some authors have shown that regulatory interactions are quickly lost [21]. In contrast, protein-protein interactions are preserved in a higher degree, specifically those involved in metabolic processes [22].

What distinguishes metabolic from other biological network types that could increase the retention of duplicates between closer nodes? We found that the preferential coupling of reactions is an important constraint characterizing metabolic networks. Thus, we simulated the retention of duplicates in a set of null "functionally" similar models including this constraint. These models are

randomly rewired versions from the original network, preserving both the degree of connectivity and the preferential coupling of reactions (see Methods). The retention of duplicates simulated using "Maslov-Sneppen" models (red circles in Figure 2A) show a behavior independent of the distance between proteins. In contrast, using "functionally" similar models (red circles in Figure 2B) an increased retention of duplicates between closer nodes was detected, approximating better what happens in real metabolic networks. This implies that the preferential coupling of reactions partially explains the increased retention of duplicates between closer reactions. Because this coupling of reactions is exclusive for metabolism, this finding also helps to understand why in transcriptional regulatory and non-enzymatic protein-protein interaction networks this behavior was not detected.

Finally, we controlled for various network and enzyme properties on the retention of duplicates. First, we considered whether the increased retention of duplicates is restricted to a region of the network. To evaluate this we randomly sampled the network and computed the retention of duplicates within samples. The main finding (blue bars in Figure 1A and 1B) is that the increased retention of duplicates between closer reactions remains statistically significant (Z-Score > 3, $P < 0.001$), and is not restricted to a region of the network. Second, we evaluated the influence of highly promiscuous compounds (hubs) on the retention of duplicates, gradually excluding hubs from network reconstructions and computing the retention of duplicates each time. The increased retention of duplicates between closer enzymes remains statistically significant (Z-Score > 3, $P < 0.001$) (Additional data file 4). Similar results were found analyzing different metabolic networks (Additional data file 4). Third, because there is a significant number of enzymes consisting of two or more domains, having only one EC number assigned, and vice versa [23], their direct comparison can cause false positives. To avoid this, we manually split enzyme sequences by functional domains. Additionally, in one control (Additional data file 5), we extracted the subset of monodomain enzymes and repeated the analyses of retention of duplicates. In a second control (Additional data file 5), we required that all domains between duplicates are homologs. The results from these two controls support the ones discussed in previous sections. Fourth, we redefined our criterion of chemical similarity, using both the first digit of EC numbers (EC:a.-.-.-) and the first 3-digits (EC:a.b.c-). As expected, these new criteria modify the relative rates of CSR and CDR retained duplicates (Additional data file 5), but the increased retention of duplicates at closer distances remains significant, supporting our previous conclusions. Finally, because we used a

method to detect remote homology, based on hidden Markov models, we controlled for this method conducting a search for homologs using Blast (closer homologs) and Psi-Blast (remote homologs) (Additional data file 5). As expected the rate of retained duplicates changes when considering only closer homologues, but the increased retention of duplicates between closer reactions remains statistically significant (Z-Score > 3, P < 0.001). Collectively, these controls evidence that the increased retention of duplicates at shorter distances is independent of the way in which metabolic databases are constructed, their size, and the hub prevalence. The manual validation of enzyme domains and network databases could precise our findings, but the main conclusions are robust.

Influence of network modularity on retention of duplicates

It has been reported that metabolic networks possess modular architecture [4, 5]. Enzymes constituting a module are highly clustered neighbors, and consequently one could expect a higher retention of duplicates within modules than between them. To test this hypothesis we used a hierarchical clustering algorithm to detect modules in metabolic networks (Figure 3A. See Methods). Then we calculated a paired measure of evolutionary distance (ED) for all-against-all metabolic pathways. This measure reflects the retention of duplicates between pathways within and between modules. Our definition of (ED) is similar to the one used to determine the relatedness between genomes based on protein domain content [24] (see Methods). It is important to emphasize that (ED) is not the distance between nodes referred in previous sections. The results show that metabolic pathways of the same module tend to have a lower (ED) (Figure 3B). This implies a greater retention of duplicates within modules than between them. For instance, considering the *E. coli* metabolic network as a whole, the total retention of duplicates among CSR is ~15 %. In contrast, if one module is extracted, such as the metabolism of amino acids (blue portion in Figures 3A and 3B, indicated by pink arrows), and the retention of duplicates within it is calculated, the resulting fraction is ~50%. To assess the significance of (ED) values we compared them against those expected by chance. To do this, we simulated a null scenario preserving both the connectivity and interaction partners of the original network, but the domain content across proteins was randomly shuffled (see Methods). This analysis demonstrates that the retention of duplicates within modules is significantly greater than between them (Z-Score > 3, P < 0.001) (Fig 3C). Thus, we propose that the capability of metabolic networks to grow modularly by gene duplication is closely related with two factors: the closeness between reactions and the kind of substrate(s) participating

within each module. Further studies evaluating the influence of metabolites similarity on the retention of duplicates could help to understand this phenomenon.

Retention of duplicates as groups and single entities

Finally we determined the frequency of duplicates retained as groups (pairs of consecutive reactions), instead of single entities. To illustrate this idea, let us show the case of fatty acid degradative (β-oxidation) and biosynthetic routes (Figure 4A). These pathways are chemically similar, but act in opposite directions and differ in their acyl-carrier groups. We determined that enzymes catalyzing CSR in these pathways were originated by gene duplication. Thus, we suggest that an ancestral pathway catalyzed both the fatty acid degradation and their biosynthesis. The direction of such ancestral pathway could be dependent on the acyl-carriers and fatty acids available. To have a first approximation of how general this observation is, we performed an all-against-all comparison of the enzymes catalyzing consecutive CSR (EC:a.b.-.- → EC:w.x.-.-). Our results indicate that ~15% of enzymes have at least one homolog in metabolism. From this fraction, two thirds are retained as isolated duplicates (scenario III in Figure 4B) and one third is retained as groups (scenario II in Figure 4B). Interestingly, the retention of both groups and isolated duplicates is greater than expected by chance (Z-scores > 50). In contrast, the case where retention of duplicates was not detected is lower than expected (Z-scores < -20). Thus, we suggest that models trying to explain the growing of metabolism by gene duplication should include the retention of both groups and isolated duplicates.

Conclusions

We used an enzyme-centric network approach to estimate the retention of duplicates in metabolism from various sources (species and databases). The observed frequencies were compared against null models to determine their significance. Collectively, our results highlight the influence of both distance and chemical similarity between reactions on the retention of duplicates. Specifically, we found an increased retention of duplicates between consecutive reactions (Figure 2A and 2B), and demonstrate that this bias can be partially attributed to the preferential coupling of reactions (Figure 2B). A similar analysis using a gene transcriptional regulatory and protein-protein interaction networks shows that this behavior is characteristic of enzymatic relationships. Thus, we

propose that the laws governing substrate-enzyme-product interactions are different from those acting on protein-DNA and non-enzymatic protein-protein interactions (Figure 2C and 2D). This is reflected as a higher retention of duplicates within network modules than between them (Figure 3). Additionally, our results evidence a significant retention of duplicates acting on both CSR and CDR (Figure 2), supporting the idea that gene duplication is not only important to generate metabolic variants but also innovations [9-12]. A synergic influence of closeness and chemical similarity between reactions explains the high retention of duplicates between consecutive CSR (Figure 2A). Our hypothesis that duplicates are significantly retained as groups can be extended to several series of reactions (Figure 4).

We consider that gene duplication should be studied as a single process, instead of distinguishing the retention of duplicates by the *stepwise* or *patchwork* models. The difficulties derived from the traditional conception of these models are avoided with the network based approach used herein, reconciling the *stepwise* and *patchwork* models.

Biological networks share general topological properties, such as their scale-free behavior and hierarchical modularity. In fact, some of these properties have been found in social and technological networks [2, 5, 19, 25, 26]. Our findings coincide with previous studies suggesting that the next step in modeling the origin and evolution of networks must consider not only the properties they share but also those differentiating them [7, 25, 27]. In particular, we improved the modeling of metabolic networks including the preferential coupling of reactions. A more detailed analysis contemplating other functional constrains, such as metabolite similarity and binding versus catalytic enzyme properties, as well as massive gene duplications and horizontal gene transfer, could enhance our understanding of the influence of metabolic versatility in the evolution of species.

Materials and Methods

Networks reconstruction

Enzyme-centric metabolic networks according to two databases BioCyc v8.0 (EcoCyc and MetaCyc) and KEGG v0.4 (EcoKegg and the full KEGG, refered RefKegg) were reconstructed as follow: if the reaction R1 produces the compound A, and A is the substrate of R2, a directed link

between the EC numbers of R1 and R2 was established. In reversible reactions, a second link, from the EC number of R2 to the EC number of R1, was added. To obtain information of reactions from BioCyc the following files were used: "reactions.dat" (substrate/product), "enzrxns.dat" (reversibility) and "reaction-links.dat" (EC numbers). The “xml” files from KEGG provide similar information in their sections: "reaction" (substrate/product and reversibility) and "entries id" (EC numbers). For each network hubs were detected, and the links established solely by hubs were gradually eliminated. The reconstructed networks, eliminating the top 20 hubs, posses the following number of nodes/edges: EcoCyc (976/4473), EcoKegg (804/2410), MetaCyc (964/4230), RefKegg (2575/11499).

Detection of retained duplicates

Enzyme sequences were retrieved, according to the desired EC number, from the following databases: EcoCyc, UNIPROT [28], BRENDA [29], and KEGG. A manual split of sequences by functional domains, according to UNIPROT, was carried out to avoid false positives caused by multifunctional enzyme comparisons. The final set has 4534 domain sequences, representing 1527 EC numbers completely annotated and 348 partial annotations. To detect duplicates, sequences were compared against the hidden Markov models of homolog domains of SUPERFAMILY v1.65 [30] and PFAM v16 [31] databases. The HMMER v2.3.1 suite of programs [32] was used for this comparison, with an E-value = 0.001 as threshold. We assumed chemically similar those reactions catalyzed by enzymes whose EC numbers share the first two digits (EC:a.b.-.-). A network adjacency matrix containing every pair of nodes (i,j) was imputed to the Floyd-Warshall algorithm [33] to determine the distance (minimal path length) between each pair (i,j). The adjacency matrix contained all reactions with known substrate/products, including those without an assigned enzyme (gene). This strategy permits to determine the retention of duplicates as a function of both the distance and the chemical similarity between reactions.

The function ($1/distance_{ij}^2$) was used to construct a matrix of normalized associations for all pairs (i,j). This matrix was used to perform a hierarchical clustering to detect network modules. To do this, we used the Kendall's τ algorithm implemented in the program CLUSTER 3.0 [34]. Similar results were obtained using the Spearman rank correlation. To determine the retention of duplicates within and between modules we calculated the evolutionary distance (ED) for each pair of pathways as follow:

$ED = A' / (A' + AB)$

where A' is the number of enzymes of the smaller pathway (pA) without homologs in the second pathway (pB). AB is the number of enzymes of pA with homologs in pB. In the limit, when all the enzymes of pA have homologs in pB, the value of (ED) tends to zero. In contrast, when the two pathways share no homologs the value of (ED) tends to one.

Significance tests

To determine whether the higher retention of duplicates at closer distances could be restricted to a portion of the network we conducted 10,000 half random samplings of the real network and calculated the frequency of retained duplicates within each sample. Additionally, we determined the significance of these frequencies, comparing them against the values expected by chance using two sets of null models. The first set, comprising 10,000 "Maslov-Sneppen" models, preserve the degree of connectivity for each node of the original network, but edges were randomly rewired. To construct these models, two edges of the original network were randomly chosen and their inputs were switched. This was repeated until the original network was completely rewired (see lower panel of Figure 2A). The second set, comprising 10,000 "functionally" similar models, preserve both the degree of connectivity and the preferential coupling of reactions of the original network. To construct these models, two edges of the original network were randomly chosen, but their inputs were switched only if both the inputting and outputting nodes represent chemically similar reactions (see lower panel of Figure 2B). Otherwise, other two edges were chosen, and the former ones were returned for further choices. This was repeated until the network was completely rewired. Some edges, from chemically similar groups with even number of pairs, remain unpaired after rewiring their group. They were added to models in their original form. These pairs represent less than 5% of the models.

We used the Z-score (Z_i) to determine the significance of real frequencies as follow:

$Z_i = (Nreal_i - \langle Nrand_i \rangle)/std(Nrand_i)$

where $Nreal_i$ is the frequency of an attribute (i) in the real network. For example, the frequency for each reaction pair type, the number of retained duplicates at a given distance, and so on. $\langle Nrand_i \rangle$ and $std(Nrand_i)$ are the average frequency and standard deviation of (i) in null models. A Z-score $\geq$ 3 implies that the frequency of (i) in the real network is significantly greater than expected by chance ($P < 0.001$). In contrast a Z-score $\leq$ -3 indicates that (i) is significantly underrepresented in the real network.

To determine the significance of (ED) values within and between modules, we compared the actual values against the ones expected using 1000 null models. These models preserve the networks intact (connectivity and wiring), but the domain content was shuffled across proteins. A Z-score ≤ -3 implies that retention of duplicates between two pathways is greater than expected by chance ($P < 0.001$).

Additional data files

The following additional data are available with the online version of this paper. Additional file 1 shows the reconstructed metabolic networks from various databases (EcoKegg, EcoCyc, RefKegg and MetaCyc) eliminating hubs gradually in each database. Additional file 2 shows the amino acid sequences of enzymes analyzed in this work. Additional file 3 shows the domains detected in such sequences, grouped by EC numbers. Additional file 4 shows the results of retention of duplicates in various databases, gradually removing hubs Additional file 5 shows the controls for the multidomain enzymes, the criteria of chemical similarity, and the method used to detect duplicates.

Acknowledgments

We thank Gerardo May for helping us to implement the Floyd-Warshall algorithm. Virginia Walbot, Sergio Encarnación, Cei Abreu, Ricardo Rodriguez de la Vega, Cesar Hidalgo and the two anonymous referees for their helpful comments in the preparation of the manuscript. This work was partially supported by grant 43502 of Mexican Science and Technology Research Council (CONACYT). JJDM was the recipient of a graduate studies scholarship from CONACYT and DGEP-UNAM.

References

1. Schuster S, Fell DA, Dandekar T: **A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks.** *Nat Biotechnol* 2000, **18:**326-332.

2. Wagner A, Fell DA: **The small world inside large metabolic networks.** *Proc Biol Sci* 2001, **268:**1803-1810.

3. Jensen RA: **Enzyme recruitment in the evolution of new function.** *Annu Rev Microbiol* 1976, **30:**409-425.

4. von Mering C, Zdobnov EM, Tsoka S, Ciccarelli FD, Pereira-Leal JB, Ouzounis CA, Bork P: **Genome evolution reveals biochemical networks and functional modules.** *Proc Natl Acad Sci U S A* 2003, **100:**15428-15433.

5. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL: **Hierarchical organization of modularity in metabolic networks.** *Science* 2002, **297:**1551-1555.

6. Pastor-Satorras R, Smith E, Sole RV: **Evolving protein interaction networks through gene duplication.** *J Theor Biol* 2003, **222:**199-210.

7. Pfeiffer T, Soyer OS, Bonhoeffer S: **The evolution of connectivity in metabolic networks.** *PLoS Biol* 2005, **3:**e228.

8. Horowitz NH: **On the evolution of biochemical synthesis.** *Proc Natl Acad Sci USA* 1945, **31:**153-157.

9. Gerlt JA, Babbitt PC: **Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies.** *Annu Rev Biochem* 2001, **70:**209-246.

10. Light S, Kraulis P: **Network analysis of metabolic enzyme evolution in Escherichia coli.** *BMC Bioinformatics* 2004, **5:**15.

11. Alves R, Chaleil RA, Sternberg MJ: **Evolution of enzymes in metabolism: a network perspective.** *J Mol Biol* 2002, **320:**751-770.

12. Teichmann SA, Rison SC, Thornton JM, Riley M, Gough J, Chothia C: **The evolution and structural anatomy of the small molecule metabolic pathways in Escherichia coli.** *J Mol Biol* 2001, **311:**693-708.

13. Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, Paley SM, Pellegrini-Toole A, Bonavides C, Gama-Castro S: **The EcoCyc Database.** *Nucleic Acids Res* 2002, **30:**56-58.

14. Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, Arnaud M, Pick J, Rhee SY, Karp PD: **MetaCyc: a multiorganism database of metabolic pathways and enzymes.** *Nucleic Acids Res* 2004, **32:**D438-442.

15. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28:**27-30.

16. Tu BP, Kudlicki A, Rowicka M, McKnight SL: **Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes.** *Science* 2005, **310:**1152-1158.

17. Maslov S, Sneppen K: **Specificity and stability in topology of protein networks.** *Science* 2002, **296:**910-913.

18. Becker SA, Price ND, Palsson BO: **Metabolite coupling in genome-scale metabolic networks.** *BMC Bioinformatics* 2006, **7:**111.

19. Shen-Orr SS, Milo R, Mangan S, Alon U: **Network motifs in the transcriptional regulation network of Escherichia coli.** *Nat Genet* 2002, **31:**64-68.

20. Butland G, Peregrin-Alvarez JM, Li J, Yang W, Yang X, Canadien V, Starostine A, Richards D, Beattie B, Krogan N, et al: **Interaction network containing conserved and essential protein complexes in Escherichia coli.** *Nature* 2005, **433:**531-537.

21. Madan Babu M, Teichmann SA, Aravind L: **Evolutionary dynamics of prokaryotic transcriptional regulatory networks.** *J Mol Biol* 2006, **358:**614-633.

22. Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T: **Conserved patterns of protein interaction in multiple species.** *Proc Natl Acad Sci U S A* 2005, **102:**1974-1979.

23. Todd AE, Orengo CA, Thornton JM: **Evolution of function in protein superfamilies, from a structural perspective.** *J Mol Biol* 2001, **307:**1113-1143.

24. Yang S, Doolittle RF, Bourne PE: **Phylogeny determined by protein domain content.** *Proc Natl Acad Sci U S A* 2005, **102:**373-378.

25. Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S, Ayzenshtat I, Sheffer M, Alon U: **Superfamilies of evolved and designed networks.** *Science* 2004, **303:**1538-1542.

26. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: **The large-scale organization of metabolic networks.** *Nature* 2000, **407:**651-654.

27. Artzy-Randrup Y, Fleishman SJ, Ben-Tal N, Stone L: **Comment on "Network motifs: simple building blocks of complex networks" and "Superfamilies of evolved and designed networks".** *Science* 2004, **305:**1107; author reply 1107.

28. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, et al: **UniProt: the Universal Protein knowledgebase.** *Nucleic Acids Res* 2004, **32:**D115-119.

29. Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D: **BRENDA, the enzyme database: updates and major new developments.** *Nucleic Acids Res* 2004, **32:**D431-433.

30. Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure.** *J Mol Biol* 2001, **313:**903-919.

31. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, et al: **The Pfam protein families database.** *Nucleic Acids Res* 2004, **32:**D138-141.

32. Eddy SR: **Hidden Markov models.** *Curr Opin Struct Biol* 1996, **6:**361-365.

33. Lipschutz S: *Data Structures.* McGraw-Hill; 1987.

34. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci U S A* 1998, **95:**14863-14868.

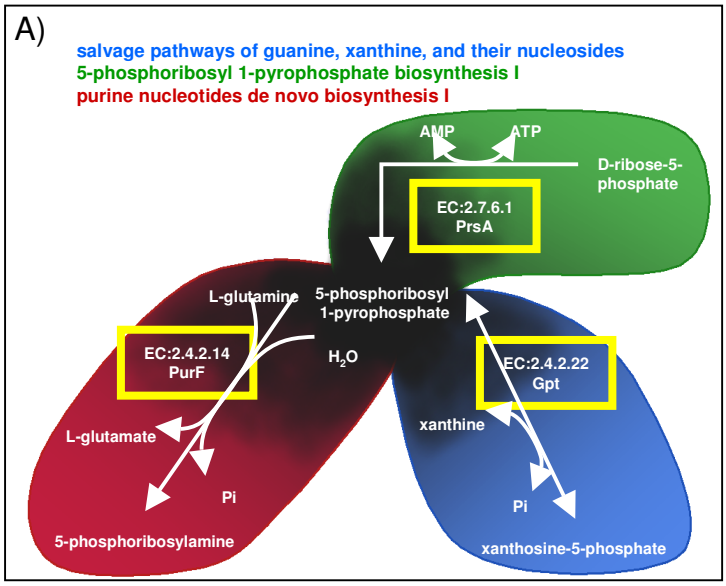
Figure legends

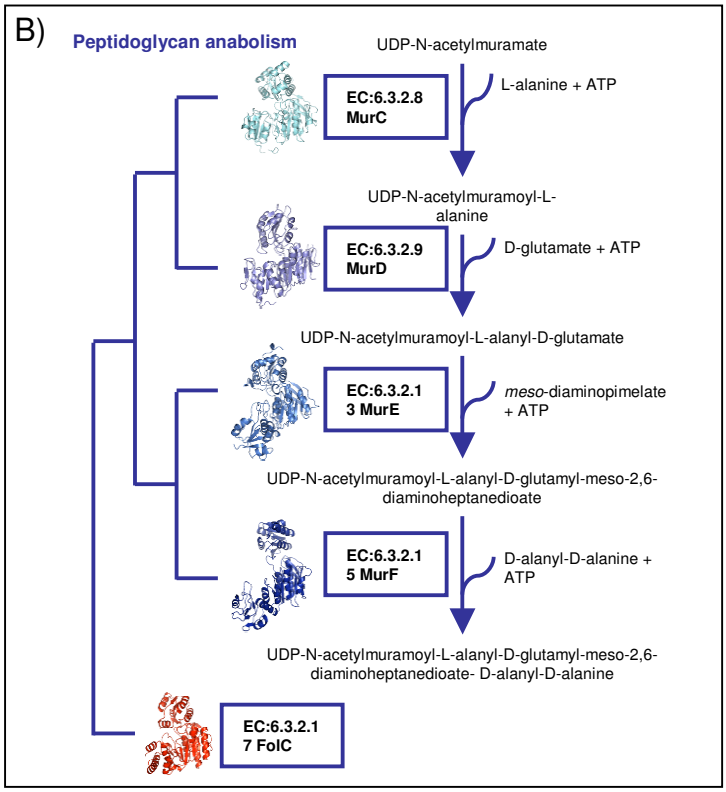
Figure 1. Preferential coupling of reactions in metabolic networks. **A)** Homolog transferases catalyzing consecutive CSR. Their origin can be explained by both the *stepwise* and the *patchwork* models. **B)** Homolog ligases whose origin can be explained by both the *stepwise* and the *patchwork* models. A distant homologue (FolC) acts in the folate metabolism. **C)** Frequencies of reaction types (EC:a.b.-.-) in the *E. coli* K12 metabolic network, according to KEGG (hereafter called EcoKegg). **D)** Frequencies of consecutive reaction types (EC:a.b.-.- → EC:w.x.-.-) in EcoKegg were compared against the expected values using a set of null "Maslov-Sneppen" models (see Methods). The Z-score (color scale bar at top) indicates the number of standard deviations between the real and the average expected frequencies. Consecutive reaction types overrepresented in real networks are shown in green-to-yellow, underrepresented ones are shown in red. The diagonal (pink box) highlights consecutive CSR, including the ligases synthesizing peptidoglycan (pink arrow). Reaction types were sorted vertically using a hierarchical clustering to detect highly related reactions types, such as EC:1.5.-.-, EC:1.7.-.- and EC:2.1.-.-. (center of plot).

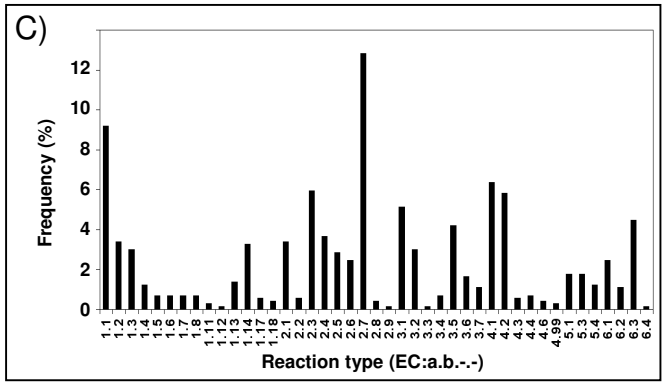
Figure 2. Influence of chemical similarity and distance between reactions on the retention of duplicates. **A)** Frequencies of retained duplicates (histogram bars) in EcoKegg are shown for the whole reaction set (ALL), and the subsets of CSR and CDR, at different distances (metabolic steps). Blue bars indicate three standard deviations (σ) from these frequencies. Deviations were obtained conducting a random sampling. Red circles represent the average expected frequencies, +/- 3 σ, obtained using "Maslov-Sneppen" models. **B)** A similar procedure to A) was conducted, using null "functionally" similar models (red circles), to control the influence of the preferential coupling of reactions. Lower panels in A) and B) illustrate the rewiring constructing null models. In "Maslov-Sneppen" models all nodes are equally eligible. In "functionally" similar models the preferential coupling of reactions restricts the choices. **C)** Retention of duplicates in the gene transcription regulatory network of *E. coli* as function of the distance (number of regulatory interactions) between transcription factors and target genes. **D)** Retention of duplicates in a protein-protein interaction network of *E. coli.* The full set of interactions (ALL), and the subsets of enzyme-enzyme (EC-EC) and the non-enzymatic protein-protein (P-P) interactions are shown. In C) and D) red circles represent "Maslov-Sneppen" models.

Figure 3. Influence of network modularity on the retention of duplicates. A) A hierarchical clustering was carried out to delimitate modules in metabolic networks. Colors denote each module in EcoKegg. **B)** Metabolic pathways (branches in the trees) within and across modules (colors group related branches) were compared using a measure of evolutionary distance (ED). A value of (ED) closer to zero (darker dot) implies a greater retention of duplicates between two pathways. **C)** Observed (ED) values were compared against the ones expected by chance. A Z-score < -3 (green) refers to significant (ED) values ($P < 0.001$).

Figure 4. Retention of duplicates as groups and single entities. A) The fatty acid degradative and biosynthetic routes exemplify the retention of duplicates as group. Same colors in EC number boxes denote duplicates. **B)** Retention of duplicates acting consecutively. Five possible scenarios were analyzed (left panel). Same color boxes denote duplicates. Scenarios (I) and (V) have a common reaction followed or preceded by other two. In (I) gene duplication was detected, in (V) it was not. Scenarios (II), (III) and (IV) imply the existence of two consecutive reaction pairs. In (II) both pairs are duplicates, in (III) only one pair is duplicated, and in (IV) none of the pairs are duplicates. Accordingly, one pair can participate in more than one scenario, looking upstream or downstream the network flux. The right histogram shows the frequency for each scenario. Red circles represent the expected frequencies, +/- 3 σ, obtained using "Maslov-Sneppen" models.


A)
salvage pathways of guanine, xanthine, and their nucleosides
5-phosphoribosyl 1-pyrophosphate biosynthesis I
purine nucleotides de novo biosynthesis I
AMP
ATP
D-ribose-5-phosphate
EC:2.7.6.1
PrsA
L-glutamine
5-phosphoribosyl 1-pyrophosphate
EC:2.4.2.14
PurF
H2O
EC:2.4.2.22
Gpt
xanthine
L-glutamate
Pi
Pi
5-phosphoribosylamine
xanthosine-5-phosphate


B)
Peptidoglycan anabolism
UDP-N-acetylmuramate
EC:6.3.2.8
MurC
L-alanine + ATP
UDP-N-acetylmuramoyl-L-alanine
EC:6.3.2.9
MurD
D-glutamate + ATP
UDP-N-acetylmuramoyl-L-alanyl-D-glutamate
EC:6.3.2.1
3 MurE
meso-diaminopimelate + ATP
UDP-N-acetylmuramoyl-L-alanyl-D-glutamyl-meso-2,6-diaminoheptanedioate
EC:6.3.2.1
5 MurF
D-alanyl-D-alanine + ATP
UDP-N-acetylmuramoyl-L-alanyl-D-glutamyl-meso-2,6-diaminoheptanedioate- D-alanyl-D-alanine
EC:6.3.2.1
7 FolC


C)
Frequency (%)
Reaction type (EC:a.b.-.-)

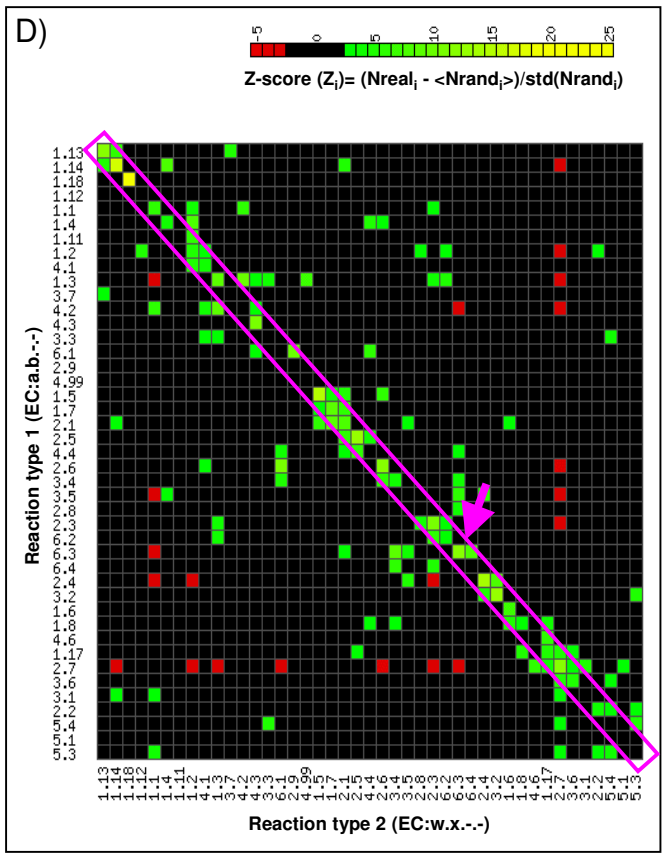

D)
Z-score (Zi)= (Nreali - <Nrandi>)/std(Nrandi)
Reaction type 1 (EC:a.b.-.-)
Reaction type 2 (EC:w.x.-.-)

Figure 1

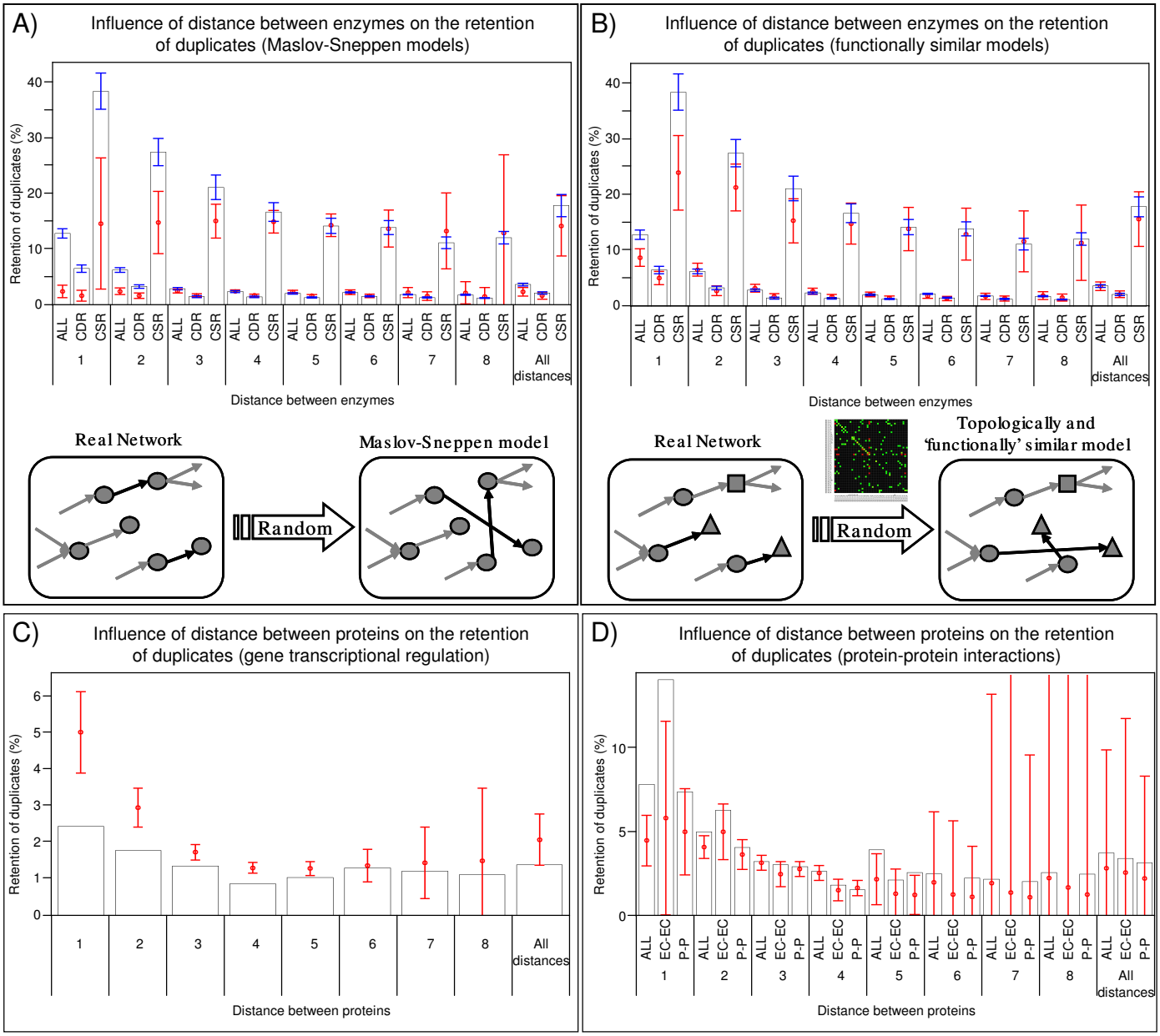
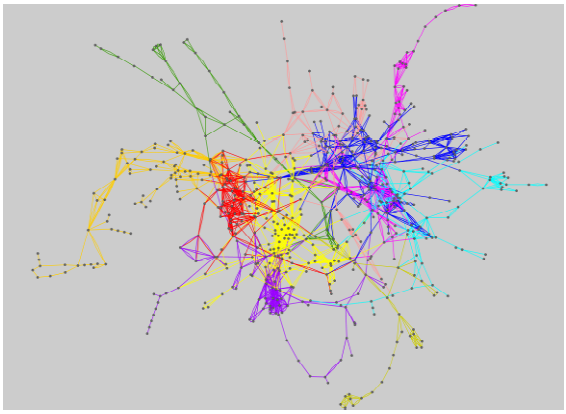
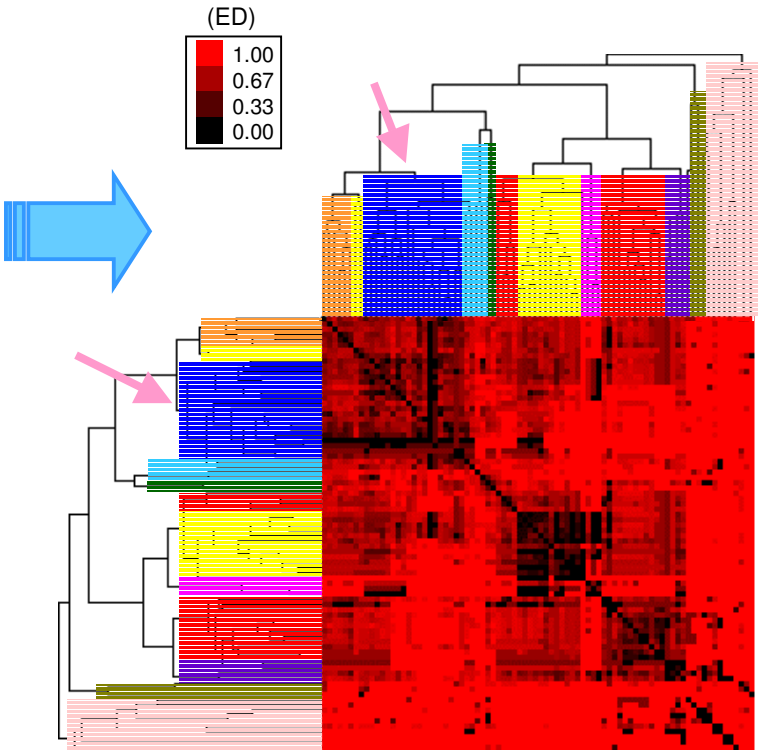

A)
Influence of distance between enzymes on the retention of duplicates (Maslov-Sneppen models)
Retention of duplicates (%)
ALL CDR CSR
1 2 3 4 5 6 7 8 All distances
Distance between enzymes
Real Network
Random
Maslov-Sneppen model
B)
Influence of distance between enzymes on the retention of duplicates (functionally similar models)
Retention of duplicates (%)
ALL CDR CSR
1 2 3 4 5 6 7 8 All distances
Distance between enzymes
Real Network
Random
Topologically and 'functionally' similar model
C)
Influence of distance between proteins on the retention of duplicates (gene transcriptional regulation)
Retention of duplicates (%)
1 2 3 4 5 6 7 8 All distances
Distance between proteins
D)
Influence of distance between proteins on the retention of duplicates (protein-protein interactions)
Retention of duplicates (%)
ALL EC-EC P-P
1 2 3 4 5 6 7 8 All distances
Distance between proteins

Figure 2

A) Identification of functional modules



B) A greater frequency of duplicates between pathways implies a lower evolutionary distance (ED)



protein domain content
random shuffling

C) Significance of (ED) values

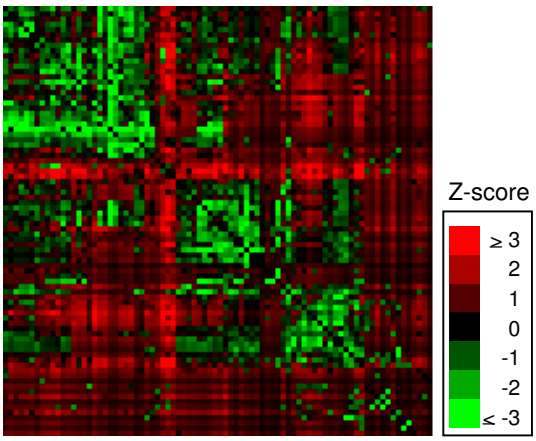


Figure 3

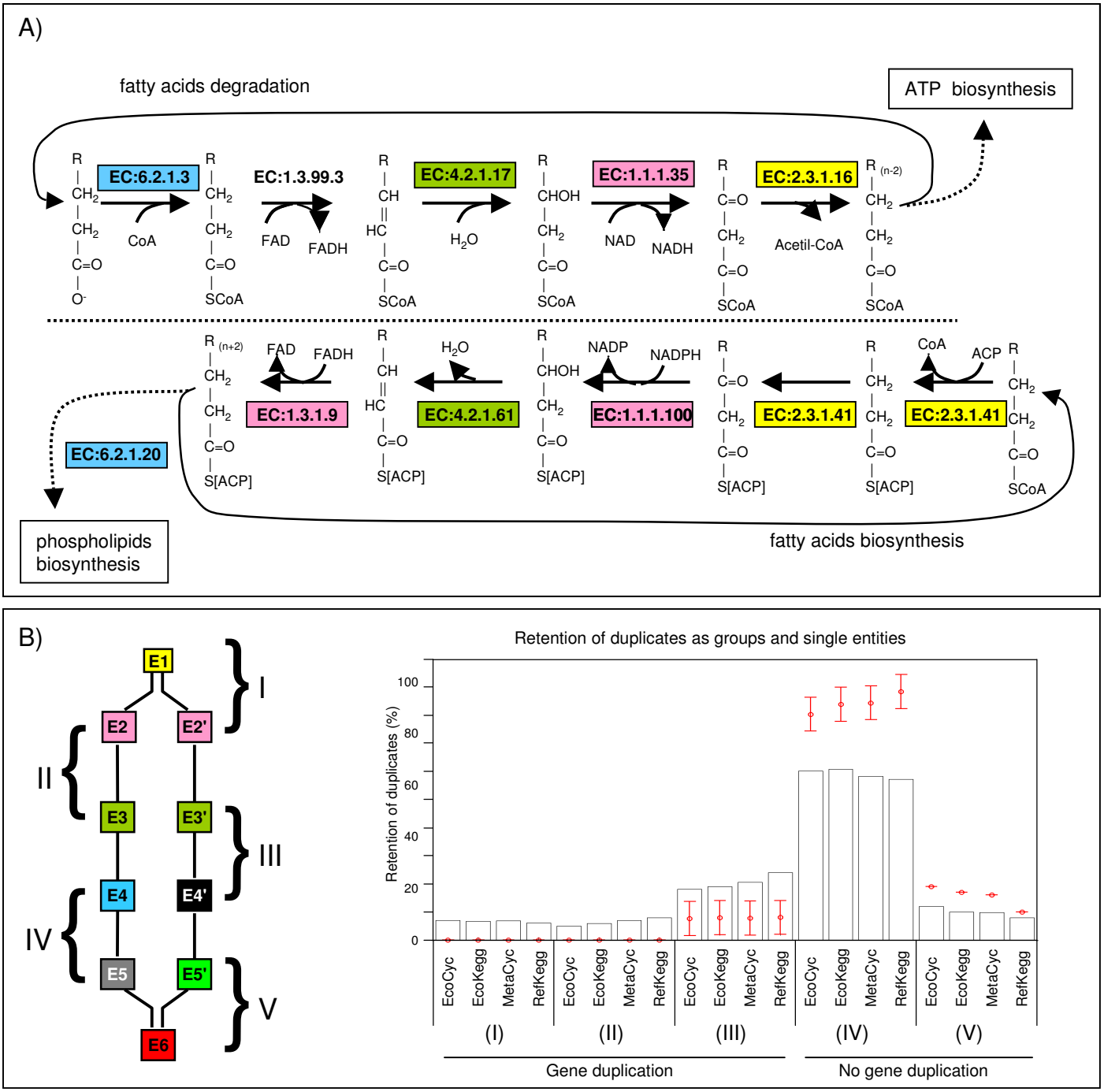

A)
fatty acids degradation
ATP biosynthesis
EC:6.2.1.3
CoA
EC:1.3.99.3
FAD
FADH
EC:4.2.1.17
H2O
EC:1.1.1.35
NAD
NADH
EC:2.3.1.16
Acetil-CoA
EC:1.3.1.9
EC:4.2.1.61
EC:1.1.1.100
EC:2.3.1.41
EC:2.3.1.41
CoA
ACP
EC:6.2.1.20
phospholipids biosynthesis
fatty acids biosynthesis
B)
E1
E2
E2'
E3
E3'
E4
E4'
E5
E5'
E6
I
II
III
IV
V
Retention of duplicates as groups and single entities
Retention of duplicates (%)
EcoCyc
EcoKegg
MetaCyc
RefKegg
(I)
(II)
(III)
(IV)
(V)
Gene duplication
No gene duplication

Figure 4

Additional files provided with this submission:

Additional file 1: Additional_file_1.txt, 6980K
http://genomebiology.com/imedia/3485352910942237/supp1.txt
Additional file 2: Additional_file_2.txt, 2037K
http://genomebiology.com/imedia/1306394283109420/supp2.txt
Additional file 3: Additional_file_3.xls, 234K
http://genomebiology.com/imedia/5784058810942085/supp3.xls
Additional file 4: additional data file 4.ppt, 75K
http://genomebiology.com/imedia/1559842531330154/supp4.ppt
Additional file 5: additional data file 5.ppt, 92K
http://genomebiology.com/imedia/8194581711330154/supp5.ppt