Evolution signatures in genome network properties

Ricardo M. Ferreira[2],José Luiz Rybarczyk-Filho[2], Rodrigo J. S. Dalmolin[3], Mauro A. A. Castro[1,2], José C. F. Moreira[3], Leonardo G. Brunnet[2]& Rita M. C. de Almeida[1,2]

National Institute of Science and Technology for Complex Systems[1], Instituto de Física[2], and Departamento de Bioquímica[3], Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves, 9500, 91051-970 C.P. 15051, Porto Alegre, Brazil.

Correspondence to: Rita M. C. de Almeida (rita@if.ufrgs.br).

Keywords: Gene network, Genome evolution, Protein-protein interaction matrix, Monte Carlo dynamics.

ABSTRACT

Genomes maybe organized as networks where protein-protein association plays the role of network links. The resulting networks are far from being random and their topological properties are a consequence of the underlying mechanisms for genome evolution. Considering data on protein-protein association networks from STRING database, we present experimental evidence that degree distribution is not scale free, presenting an increased probability for high degree nodes. We also show that the degree distribution approaches a scale invariant state as the number of genes in the network increases, although real genomes still present finite size effects.

Based on the experimental evidence unveiled by these data analyses, we propose a simulation model for genome evolution, where genes in a network are either acquired *de novo* using a preferential attachment rule, or duplicated, with a duplication probability that linearly grows with gene degree and decreases with its clustering coefficient. The results show that topological distributions are better described than in previous genome evolution models. This model correctly predicts that, in order to produce protein-protein association networks with number of links and number of nodes in the observed range, it is necessary 90% of gene duplication and 10% of *de novo* gene acquisition.

If this scenario is true, it implies a universal mechanism for genome evolution.

Introduction. Genome evolution is determined first by the processes that modify DNA and then by those mechanisms that either neutrally keep or naturally select these mutations by their phenotypic effects. The connection between DNA variations and the consequent phenotypic alterations is far from being simple and is elusive to determine. However, it is reasonable to assume that, after evolutionary time spans, these DNA variation mechanisms have left their mark on the genome.

Phenotypic effects are consequence of the existing associations between proteins which rule cellular metabolism. As proteins are expressed from genes, protein-protein associations will express eventual changes in genotypes and are prone to natural selection. Consequently we may speculate that natural selection, by defining genome evolution mechanisms, has left its mark on organisms protein-protein association matrices. This is not a novel idea. Barabási and collaborators[1,2] have described genomes of different organisms as networks where nodes are either genes or proteins, and links correspond to associations between the nodes. They proposed an evolution dynamics for the genome considering that genes are sequentially added to a network following a preferential attachment rule: each newly incorporated gene interacts with a gene already on the network with a probability that is proportional to its degree, that is, to the number of other genes with which it already interacts. The resulting artificial network is scale free and described well the available experimental data at that date.

However, the properties of a gene already in the network are not the only drive for a novel gene attachment. There are different molecular mechanisms acting as novelty source in gene formation, such as exon shuffling, retroposition, mobile elements, horizontal gene transfer, gene duplication, etc., and the connections of a new gene certainly reflect its origin together with the nature of the genes it connects to[3]. Among the mechanisms involved in new genes creation, gene duplication is recognizably the most important and there is plenty of evidence that it plays an essential role on genome evolution[4]. One major feature of a duplicated gene consists of inheriting its parent connections and this property is determinant to the whole network design.

Vázquez and collaborators[5,6] proposed a model for genome evolution where genes are incorporated by duplication followed by mutations which are translated as adding and/or deleting links on a protein-protein association matrix. In this model, genes are randomly chosen to duplicate and parameters are set to produce gene networks where the probability that a gene product is associated to k other proteins decays as a power law as k increases. A drawback for this approach, using randomly chosen genes, lays on the experimental fact that the probability to fix a given duplication episode greatly varies according to the properties of the duplicating gene[7-9].

Since the contributions by Barabási and collaborators, the amount and quality of data regarding both genomes and protein-protein association have greatly increased. For example, STRING database increased from few organisms at 2001 to 1133 organisms in 2011[10-12]. Also, databases regarding protein-protein association for some organisms have been largely enhanced. Here we analyze data considering 268 core organisms, which strongly suggest that highly connected genes stem from duplication mechanisms acting preferentially on genes that are highly connected, but not excessively clustered. This conclusions are made evident here by

presenting the quantities as functions of $\frac{k}{k_{\max}}$, where $k_{\max}$ is the maximum degree in the network. We also propose an adequate ordering for genes to evince topological properties of the protein-protein association matrix.

Considering these experimentally based conclusions we propose a genome evolution dynamics where, besides a Barabási mechanism of acquiring genes *de novo* based on preferential attachment, we also consider gene duplication, where the probability that a gene duplicates grows with its degree and decreases depending on how clustered it is. The results of these simulations are capable of describing different aspects of the network topology, besides predicting the ratio of duplicated and *de novo* acquired genes.

Building protein-protein association matrices. We considered all 268 core organisms in STRING database, version 8.3[10-12], with confidence scores 0.700, 0.800, and 0.900 using "experimental" and "database" (95% of these interactions) added with "neighborhood", "fusion", "co-expression", and "co-occurrence" evidence. This information renders possible to build a network, where each node corresponds to a protein with at least one known protein-protein association, and links correspond to these associations. To each network node i we assign a degree k_i, which is the number of links arriving at that node. For each organism and score we produce a network and calculate the probability $P'(k)$ that a protein has k links, defined as

$$P^{'(k)} = \frac{N(k)}{N} \quad , \tag{1}$$

where N is the number of nodes and $N(k)$ is the number of nodes with degree k. To compare different organisms, with different genome sizes, we considered a rescaled probability of finding a protein with a given degree k, as follows

$$P\left(\frac{k}{k_{\max}}\right) = k_{\max} P^{'(k)} \quad , \tag{2}$$

where $k_{\max}$ is the maximum degree presented by the proteins of an organism.

Figure 1a presents the average, taken in intervals of $\frac{\Delta k}{k_{\max}} = 0.01$, of the network degree distribution, $P\left(\frac{k}{k_{\max}}\right)$ versus $\frac{k}{k_{\max}}$, for three different confidence scores: 0.700, 0.800 and 0.900. The inset presents the degree distributions of all 268 core organisms, with different colors for different scores. The blue line in Fig. 1a is a power law fit, $F\left(\frac{k}{k_{\max}}\right) = 0.02\left(\frac{k}{k_{\max}}\right)^{-2.4}$, which describes $P\left(\frac{k}{k_{\max}}\right)$ for only a limited interval of $\left(\frac{k}{k_{\max}}\right)$. At values of $\frac{k}{k_{\max}}$ near 0.9, this degree distribution presents a local maximum, associated to the cloud of points with higher values of probability presented in the inset. The probability of proteins with degree near $k_{\max}$ increases and indicates a genome evolution dynamics where high degree genes are probable to appear. As the main mechanism of genome evolution is gene duplication[3,4], it is reasonable to assume that the local maximum in $P\left(\frac{k}{k_{\max}}\right)$ for large $\frac{k}{k_{\max}}$ is due to high duplication probability for more connected genes. Figure 1b presents the same data in a linear plot, where the standard deviations for each average value of $P\left(\frac{k}{k_{\max}}\right)$ are shown, to evince that deviations from the power law fit is significant. Each point is an average over 268 organisms, justifying a Z test for significance. The difference between the power law fit and the average $P\left(\frac{k}{k_{\max}}\right)$ for confidence score 0.800 is shown in the inset for Fig.1b, in units of standard deviations for $P\left(\frac{k}{k_{\max}}\right)$, calculated in intervals of $\frac{\Delta k}{k_{\max}} = 0.01$. The maximum in degree

distribution is significantly different from the power law.

Figure 1c plots as a function of $\frac{k}{k_{\max}}$ the average clustering coefficient $\langle C \rangle$, defined as the fraction of existing connections between the neighbors of a gene with k neighbors in relation to the maximum number of such connections $\left(\frac{k(k-1)}{2}\right)$. The inset in Fig. 1c individually shows the corresponding data for all core organisms. For all three scores this curve is initially constant, presenting local minimum and maximum for, roughly, $\frac{k}{k_{\max}} \approx 0.2$ and $\frac{k}{k_{\max}} \approx 0.8$, respectively, decreasing after that: the most connected genes are not the maximally clustered. Observe that, while the maximum in $P\left(\frac{k}{k_{\max}}\right)$ occurs for $\frac{k}{k_{\max}} \approx 0.9$, the maximum for the clustering coefficient occurs before that.

Figure 1d plots the average degree k_{nn} of the neighbors of a gene as a function of $\frac{k}{k_{\max}}$. The inset individually shows the corresponding data for all core organisms. For all scores this curve is initially increasing, presenting a local maximum at roughly $\frac{k}{k_{\max}} \approx 0.9$, decreasing after that. It means that the most connected genes are not connected to the highest k genes. Observe also that the maxima in both $P\left(\frac{k}{k_{\max}}\right)$ and k_{nn} occur for $\frac{k}{k_{\max}} \approx 0.9$.

Summarizing, these plots indicate that *i)* $P\left(\frac{k}{k_{\max}}\right)$ is not power law; *ii)* $P\left(\frac{k}{k_{\max}}\right)$ presents a local maximum for $\frac{k}{k_{\max}} \approx 0.9$; *iii)* the clustering coefficient is not uniform, presenting local minimum and maximum; and *iv)* the network is assortative up to $\frac{k}{k_{\max}} \approx 0.9$, with k_{nn} decreasing after that. These observations suggest node modules of high average degree which are highly clustered. This behavior is evinced by the superposition of data from a large number of organisms, plotted against a normalized degree $\frac{k}{k_{\max}}$. For comparison, Fig. S1 presents plots normalized by the total number of genes of each organism, where this behavior is not as clearly unveiled.

Ordering algorithm. To investigate further the consequences of genome evolution dynamics we chose those organisms for which there is more information regarding protein-protein association. Figure 2a shows the number of links versus the number of genes for the 268 core organisms for 0.800 confidence score. Observe that data for very well studied organisms as *Homo sapiens* or *Arabidopsis thaliana*, present larger numbers of genes and links, that is, more information is available. In what follows we considered 6 organisms, marked with orange dots in Fig. 2a (*Homo sapiens*, *Mus musculus*, *Arabidopsis thaliana, Drosophila melanogaster, Saccharomyces cerevisiae,* and *Escherichia coli).*

Protein-protein association data may be organized on a matrix M where each axis represents the protein list in a given order. The matrix elements M_{ij} are assigned with value 1 (0) if there is (not) an association between the genes at positions i and j of the list. For illustrational purposes, these association matrices may be represented by plots where a black dot at position (i, j) indicates that $M_{ij} = 1$.

We obtain the sets of genes of each organism from STRING database and dispose them in randomly ordered lists. Each possible order for a gene list implies a different configuration for matrix M, for which a cost function E may be defined as

$$E = \sum_{i=1}^{N}\sum_{j\neq i}^{N} d_{ij}^{\alpha}\left(\left|M_{i,j} - M_{i+1,j}\right| + \left|M_{i,j} - M_{i-1,j}\right| + \left|M_{i,j} - M_{i,j+1}\right| + \left|M_{i,j} - M_{i,j-1}\right|\right), \qquad (3)$$

where $d_{ij} = \sqrt{|i^2 - j^2|}$ is proportional to the distance on the matrix from the point (i,j) to the diagonal (when $i = j$), and α is a parameter, here taken $\alpha = 8$. Minimization of this function, by changing the genes localization on the list, implies approximating mutually interacting genes, as discussed in Ref. 13.

The ordering algorithm starts from a randomly ordered matrix configuration and proceeds by randomly choosing a pair of genes whose positions are tentatively swapped. The cost function for this changed configuration is calculated and, in case the cost decreases, the change is accepted. If the cost function increases by ΔE, the change is accepted with probability $e^{-\Delta E/T}$, where T is a parameter. This procedure is intended to avoid metastable states in the optimization of Eq.(3). Finally, when $\Delta E = 0$, the change is accepted with 50% probability. The algorithm proceeds by randomly choosing another pair of genes and the procedure is repeated until the value of the cost function is stabilized.

Randomly ordered lists yield association matrix configurations with black dots spread over the whole plot (see Figs. S2-S7 of Supplementary Materials). Ordering the gene list by minimizing the cost function evinces topological properties of protein-protein association networks. Figure 3a-f presents the ordered matrices for the six organisms listed above. Observe that points concentrate near the diagonal, implying that there may be an association ($M_{ij} = 1$) between the products of genes localized at not far apart positions i and j. Not all networks may be put in formats like those shown by Figs. 3a-f. See Fig. 3-g which represents a network built using Barabási-Albert algorithm, or an Erdös-Rényi network, presented on Fig. S8 of Supplementary Materials. In fact, this format reveals that genomes (Figs.3a-f) do not present one central hub linked to the whole network (which could indicate scale free networks) but, contrarily, present many hubs with neighborhoods that do not span the entire system.

Figures 1 and 3 evince different aspects of real genomes. First, degree distribution is not a power law. Second, there is an accumulation of high degree nodes, which may be explained by an enhanced duplication probability for highly connected gene products. Finally, hub genes are not central to the whole network, which presents hierarchical clusters.

Another experimental aspect is relevant for genome evolution. Duplication events can be assessed by analyzing gene families, *i.e.,* genes sharing the same ancestral gene. Some gene families have mainly orthologs, while others are composed by a great number of paralogs, indicating many duplication episodes[7,14]. The reason why some genes are prone to duplicate while others avoid duplication is controversial. However, duplication is clearly not randomly fixed and functional characteristics of the parent gene certainly influence new born genes fates. It has been discussed that genes presenting substrate promiscuity are prone to fix duplication while other genes avoid duplication because it probably leads to deleterious effects[15].

Huang and collaborators[16] demonstrated that highly connected proteins with low clustering coefficient (intermodular hubs) possess a higher proportion of duplicated genes as compared with proteins that are highly connected and highly clustered (intramodular hubs). According to those authors, intramodular hubs represent the network most stable and conservative part, while intermodular hubs represent the evolutionary dynamic network regions with a high duplication rate.

Genome evolution model. These experimentally determined characteristics of genomes may

be explained by an evolution dynamics with two different gene acquisition mechanisms: *de novo* formation and duplication. The first mechanism follows Barabási preferential attachment rule, which simulates an enhanced attachment probability shown by genes with more active domains[2]. The second mechanism describes the experimental facts discussed above: genes are chosen with higher probability when they are more connected, but less clustered. Clustering coefficient C_i for the i^{th} gene is defined as[17,18]

$$C_i = \frac{2}{k_i(k_i-1)} \sum_{j=1}^{N} \sum_{l=1}^{N} M_{ij} M_{jl} M_{li} \quad , \tag{4}$$

which gives the ratio of existing links between the neighbors of the i^{th} gene to the maximum possible number of such links. The duplication probability for the i^{th} gene is defined as

$$p_i^D = \frac{k_i(1-C_i)}{\sum_j k_j(1-C_j)}, \tag{5}$$

where the denominator guarantees a normalized probability. This assumption reproduces the experimental facts that *i)* degree distributions have a local maximum for $\frac{k}{k_{\max}}$ near 1 (Fig.1) and *ii)* more clustered genes are less prone to duplicate[7,16,19].

Simulations start with 5 nodes, each linked to two others, forming a ring. To acquire a new gene we first choose either *de novo* mechanism, with probability $(1-q)$, or duplication, with probability q. If the *de novo* mechanism is chosen, each existing node i is linked to the new one with probability $\frac{k_i}{\sum_j k_j}$, and the procedure is repeated until the new node presents at least one link. In case of duplication, the node to be duplicated is chosen using the probability defined in Eq.(5). Duplication implies creating a new node linked to its parent and with the same neighbors.

After duplication, mutations are implemented by deleting links between either the parent or the child with a common neighbor with probability r. In fact, a hallmark of gene duplication is the subsequent speciation of at least one gene copy [20].

This simulation model has two parameters, duplication probability q and mutation probability r. For the numbers of links and genes of simulated networks to fall in the same intervals as more extensively investigated organisms (Fig. 2.a), q must be of the order of 0.90, which is experimentally verified: Zhou *et al*.[4] have studied *Drosophila melanogaster* genome and compared it to other organisms in *D. melanogaster* subgroup. They have found that duplication is responsible for 80% of new genes, and 10% is generated by retroposition, here taken as an additional form of gene duplication. We are left with one single parameter r, set to 0.05 to match the observed relation between number of links and nodes presented by protein-protein association matrices of real organisms (Fig.2a).

We also simulated two other well described models for genome evolution: Barabási and Albert[1] model, based on a preferential attachment rule, and Vazquez *et al.*[5,6] model, where genomes are built by duplicating randomly chosen genes. For both models, parameters are set to ensure that the number of links and nodes are roughly the same as in the protein-protein association networks obtained from STRING database for confidence score 0.800. In Barabási-Albert model, each new node is connected with 15 neighbors, and in the duplication-divergence model each node is linked with its parent, and has 0.4 of mutation probability. For brevity, we considered the most cited models in the literature although other interesting models also address genome growth[21-25]. Anyway, as far as we know, no previous model simultaneously describes all properties shown by protein-protein association networks of 268 core organisms as presented in Figs. 1-3.

Figs. 2a, 2c, and 2e present, as a function of N, the plots of number of links N_L, average degree $\langle k \rangle$, and maximum degree, $k_{\max}$, for experimental results (dots) and simulated models (solid lines). As discussed, the chosen model parameters ensure that the simulated number of links crosses the region with best investigated organisms. The experimental points indicate that the number of links is proportional to the number of nodes, that is, $N_L \sim N^1$. This behavior is clearly shown by both Barabási-Albert and our model, and is further evinced by Fig. 2c, that shows a constant average degree for experimental dots and these two models. Finally, Fig. 2e shows that, for the simulations, $k_{\max}$ increases with, roughly, $\sqrt{N}$. The experimental results are not in contradiction, although they are not conclusive. Anyway, this behavior explains why using $k_{\max}$ instead of N as the normalization constant in Eq. 2 yield different results.

Figs. 2b, 2d, and 2f present $P\left(\frac{k}{k_{\max}}\right)$ versus $\frac{k}{k_{\max}}$ for the three simulated models, measured at different instants. Observe that clearly Barabási-Albert and our model converge to a scaling invariant distributions that superpose as $N \to \infty$, while for the Vasquez (D-D) model this convergence is either not true or too slow. This is a relevant point: although real genomes are finite, we may speculate that when large enough they present a scale invariant degree distribution. If this is true, the data collapse predicted by scaling invariance, together with a significant fit of the collapsed degree distribution of all core organisms, is as a strong evidence of a common mechanism universally ruling genome growth.

On the other hand, experimental degree distributions may present finite size effects. This is clear in Fig. S9, where we show $P\left(\frac{k}{k_{\max}}\right)$ versus $\frac{k}{k_{\max}}$ for the experimental data (score 0.800) averaged over genomes whose protein-protein association networks present N in the ranges $N < 1000$, $1000 < N < 2000$, ..., $6000 < N$. The degree distributions seem to converge to a scale invariant state, but for the smaller networks the finite size effects are visible. Both experimental data and D-A model results show that smaller networks present a higher local maximum in $P\left(\frac{k}{k_{\max}}\right)$ for large $\frac{k}{k_{\max}}$. To properly compare the simulations results with experimental networks with variable sizes, we considered a weighted average of the degree distribution, as follows.

For each model, we produced 10 samples in each size range listed above, and obtained the distributions of degree, clustering coefficient, and average degree of the neighbors as functions of $\frac{k}{k_{\max}}$. To compare with the set of all 268 core organisms, presenting, respectively, 32, 110, 74, 39, 10, 1 and 2 organisms in each size range, we produced weighted averages over the size ranges for the topological distributions, using the weights 32/268, 110/268, 74/268, 39/268, 10/268, 1/268 and 2/268. These results are shown in Fig. 4. Other parameters values in each model yield different results, shown in Figs. S10-S13 in Supplementary Materials: The description of topological quantities are worse in these cases. Similar averages for the six, best investigated organisms are shown in Fig. S14 of Supplementary Materials.

Figures 3g-i present ordered association matrices for simulated networks. Barabási-Albert (B-A) model (Fig. 3g) clearly shows only one module, with a central hub connected to all network. Duplication-Divergence (D-D) model, on the other hand, shows a slimmer structure around the diagonal, and Duplication-Acquisition (D-A) model presents a central hub not connected to the whole network. Figure S15 of Supplementary Materials presents the same panels, zooming at the central regions: the hierarchical structure of clusters, evinced by small solid squares, is clearly present in organisms and Duplication-Acquisition model. Figure S16 of Supplementary Materials present the orderings obtained with $\alpha = 1$, which stress further the clustered structures.

Duplication-Acquisition model reproduces the topology of protein-protein association networks. For each network, we calculated the weighted average for probability $P\left(\frac{k}{k_{\max}}\right)$, the clustering coefficient $\langle C\rangle_{k/k_{\max}}$, and the relative degree $\langle k_{nn}\rangle_{k/k_{\max}}$ of the neighbors of a node with degree $\frac{k}{k_{\max}}$, defined as

$$\langle k_{nn}\rangle_{k/k_{\max}} = \frac{1}{N(k)}\sum_{i=1}^{N}\frac{\delta(k_i-k)}{k_i}\sum_{\langle j\rangle_i}^{k_i}\frac{k_j}{k_{\max}}, \quad (6)$$

where $\langle j\rangle_i$ stands for a sum over the nodes j that are neighbors to node i.

The black dots in Figs. 4 refer to protein-protein association networks of the 268 core organisms, which present large clustering coefficients for all degrees, decreasing as $\frac{k}{k_{\max}}$ approaches 1: very high degree nodes are less clustered than less connected nodes. In organisms, the average number of connections of the neighbors, $\langle k_{nn}\rangle$, first increases with the node degree and then decreases, reinforcing the fact of very high degree nodes not presenting the largest clustering coefficient. Figure 4 presents three columns, one for each model, where we show the *i)* the experimental data as black points, weighted averages for *ii)* experimental points as green lines and for *iii)* simulation as red lines. The first column shows that B-A model produces a degree distribution $P\left(k/k_{\max}\right)$ that follows a power law, a clustering coefficient that is roughly constant at a value that is much less than those shown by experimental data. Furthermore, $\langle k_{nn}\rangle$ does not depend on $k/k_{\max}$. The deviation from the experimental dots reflects that Barabási-Albert model yield scale free networks with a global central hub.

The second column presents the results for the Duplication-Divergence (D-D) model. Here, this distribution clearly does not follow a power law, due to the chosen parameters (link deleting probability of 0.4), that fixed the ratio of number of links to number of nodes to the desired values (see Fig. 2a). The average clustering coefficient decreases too abruptly, as compared to experimental data: as degree increases, the clustering decreases as $\left(\sim\left(\frac{k}{k_{\max}}\right)^{-0.7}\right)$. However, the average degree of the neighbors presents a mild increase, meaning that genes connect to groups of genes with slightly larger degrees. For comparison see Figs. S17-S18 in Supplementary Materials.

The third column in Fig. 4 refers to the results of our model. In Fig. 4c, $P\left(k/k_{\max}\right)$ describes very well the experimental data. For high values of $\frac{k}{k_{\max}}$, degree distribution reproduces the local maximum as shown by real organisms, although for smaller degrees. The clustering coefficient, shown in Fig. 4f, describes the major part of the interval, presenting a more intense decrease as $\frac{k}{k_{\max}} \to 1$. The varying character of assortativeness as $\frac{k}{k_{\max}}$ increases is also evident in Fig. 4i: $\langle k_{\mathrm{nn}}\rangle$ first increases to a maximum up to $0.45\ k_{\max}$.

Comparing the three columns we conclude that D-A model better catches the topological properties of protein-protein association networks, according to the currently available experimental data, although the description is not perfect.

Conclusions. In this paper we have presented experimental evidence that degree distribution is not scale free, presenting an increased probability for high degree nodes, and that there are a few hub nodes in these networks, probably organized in a hierarchical way. Furthermore,

when scaled by the maximum degree in each network, $k_{\max}$, the degree distribution seems to approach a scale invariant state as the number of genes in the network increases. However, real genomes still present finite size effects. If this scenario is true, it indicates a universal mechanism for genome evolution.

We propose a simulation model for genome evolution, Duplication-Acquisition model, where genes in a network are either acquired *de novo* using a preferential attachment rule, or duplicated, with a duplication probability that linearly grows with gene degree and decreases with its clustering coefficient. With this simple rule, topological distributions are well described. This model correctly predicts that, to produce protein-protein association networks with number of links and number of nodes in the observed range, it is necessary 90% of gene duplication and 10% of *de novo* gene acquisition.

To compare the networks we ordered gene lists for each organism and model to produce protein-protein association matrices yielding images of the network association structure. These images suggest that there is a system scale that is less than its size (see Fig.3), with, possibly, a hierarchical modular organization, as predicted by the Duplication-Acquisition model (see Fig. S16).

The simulation model is not perfect. Most probably phenotypic effects caused by gene acquisition, duplication, or mutation cannot be fully grasped by network gene properties only and, consequently, this model is an over-simplification. However it does point towards a positive correlation between duplication probability and degree, while indicating a negative correlation between duplication probability and clustering coefficient.

ACKNOWLEDGEMENTS

We acknowledge fruitful discussions with Prof. Diego Bonnato, Centro de Biotecnologia, and support from the Centro de Física Computacional, Universidade Federal do Rio Grande do Sul.

FUNDING

This work has been partially funded by Brazilian agencies Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS).

REFERENCES

1. Barabási, A. L. & Albert, R. Emergence of Scaling in Random Networks. S*cience* **286**, 509-512 (1999).
2. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. & Barabasi,A. –L. The large scale organization of metabolic networks; *Nature* **407,**651-654 (2000).
3. Long, M., Thornton, K. & Wang, W. The origin of new genes: glimpses from the young and old. *Nature Reviews***4**, 865-875 (2003).
4. Zhou, Q., Zhang, G.-j., Zhang, Y. & Spring, C. On the origin of new genes in Drosophila. *Genome Research***18**, 1446-1455 (2008).
5. Vázquez, A. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physical Review E***15**, 1-15 (2007).

6. Vázquez, A., Flammini, A., Maritan, A. & Vespignani, A. Modeling of Protein Interaction Networks. *Complexus* **65**, 38-44 (2003).
7. Dalmolin, R.J.S., Castro, M.A.A., Rybarczyk-Filho, J.L., Souza, L.H.T., de Almeida, R.M.C., and Moreira, J.C.F. Evolutionary plasticity determination by orthologous groups distribution. *Biology Direct* **6**, 22 (2011).
8. Koonin, E.V., Wolf, Y.I. Constraints and plasticity in genome and molecular-phenome evolution. *Nat Rev Genet***11**,487-498 (2010).
9. Wall, D.P., Hirsh, A.E., Fraser, H.B., Kumm, J., Giaever, G., Eisen, M.B., Feldman, M.W. Functional genomic analysis of the rates of protein evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 5483-5488 (2005).
10. Mering, C. v. *et al*. STRING: Known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic acids research* **33**, D433-D437 (2005).
11. Mering, C. v. *et al*. STRING 7- Recent developments in the integration and prediction of protein interactions. *Nucleic acids research* **35**, D358-D362 (2007).
12. Jensen, L. *et al*. STRING 8 - A global view on proteins and their functional interactions in 630 organisms. *Nucleic acids research* **37**, D412-D416 (2009).
13. Rybarczyk-Filho, J.L., *et al*.Towards a genome-wide transcriptogram:the *Saccharomyces cerevisiae* case. *Nucleic acids research* **39**, 3005-3016 (2011).
14. Koonin, E.V. *et al.* A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biology* **5**, R7(2004).
15. Conant, G.C. and Wolfe, K.H. Turning a hobby into a job: How duplicated genes find new functions. *Nature Reviews Genetics* **9**, 938-950 (2008).
16. Li, L., Huang, Y., Xia, X., and Sun, Z. Preferential duplication in the sparse part of yeast protein interaction network. *Molecular Biology Evolution* **23**, 2467-2473 (2006).
17. Colizza, V., Flammini, A., Maritan, A. & Vespignani, A. Characterization and modeling of protein–protein interaction networks. *Physica A***352**, 1-27 (2005).
18. Costa, L. F., Rodrigues, F. A., Travieso, G. & Boas, P. R. V. Characterization of complex networks: A survey of measurements. *Advances in Physics* **56**, 167-242 (2007).
19. Castro, M.A.A., *et. al.* Evolutionary origins of human apoptosis and genome stability gene networks. *Nucleic acids research* **36** ,6269-93 (2008).
20. Innan, H. and Kondrashov, F. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics* **11**, 97-108 (2010).
21. Mithani, A., Preston, G. M. & Hein, J. A stochastic model for the evolution of metabolic networks with neighbor dependence. *Bioinformatics* **12**, 1528-1535 (2009).
22. Evlampiev, K. & Isambert, H. Modeling protein network evolution under genome duplication and domain shuffling. *BMC systems biology* **1**, 49 (2007).
23. Kim, W. K. & Marcotte, E. M. Age-dependent evolution of the yeast protein interaction network suggests a limited role of gene duplication and divergence. *PLoS computational biology* **11**, e1000232 (2008).
24. Berg, J., LÄssig, M. & Wagner, A. Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC evolutionary biology* **1**, 51 (2004).
25. Takemoto, K. & Oosawa, C. Modeling for evolving biological networks with scale-free connectivity, hierarchical modularity, and disassortativity. *Mathematical biosciences* **2**, 454-468 (2007).

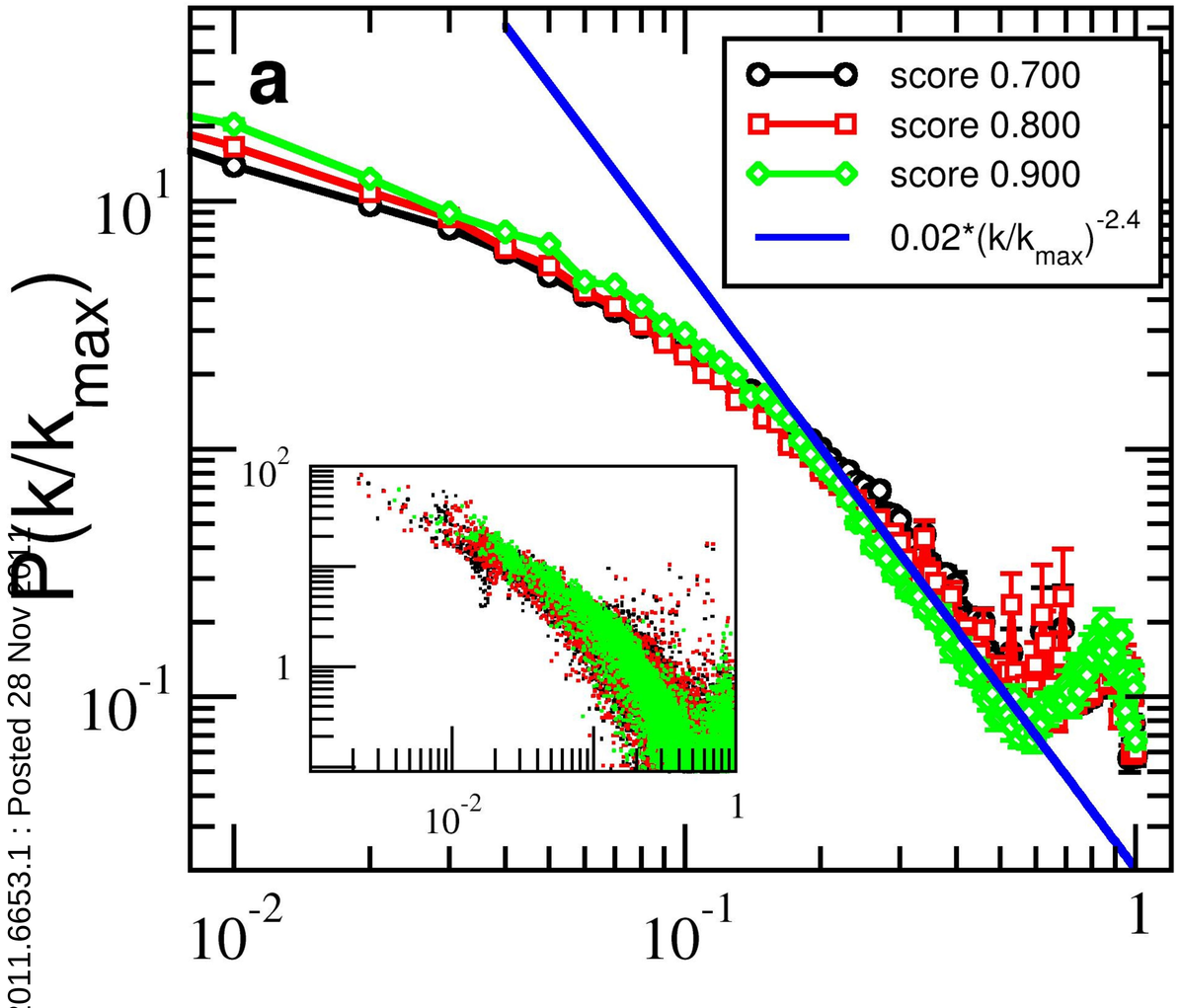
LEGENDS

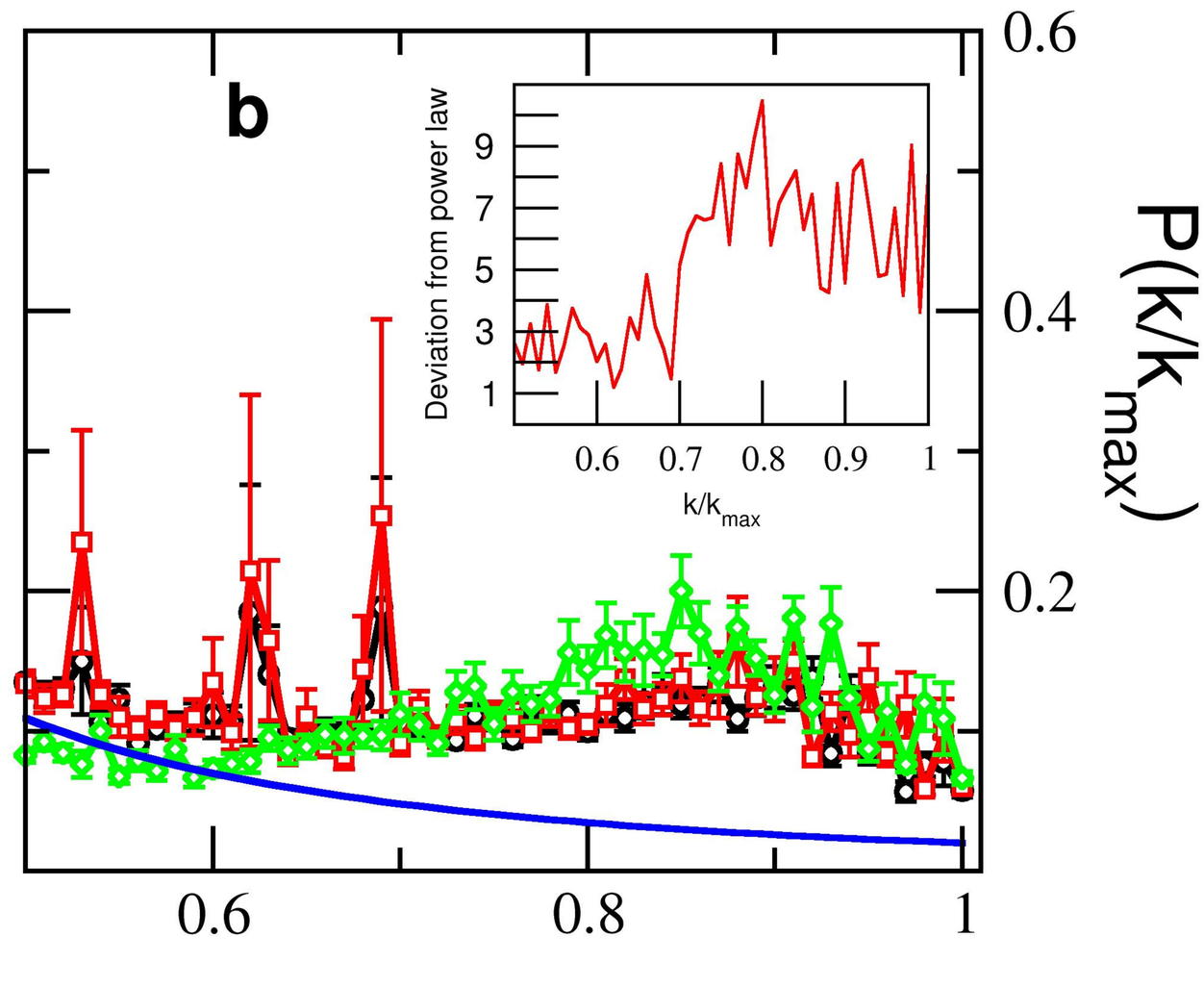
Fig. 1 – Topological quantities for all 268 core organisms from STRING database for three different confidence scores: 0.700, 0.800 and 0.900 (black, red and green lines in all graphs, respectively). All measurements are taken as functions of node degree, rescaled by the maximum degree of the corresponding network. All averages were taken over intervals of $k/k_{\max} = 0.01$. (**a**) Average degree distribution compared with a tentative power law fit (blue line). (**b**) Average degree distribution in linear scale, showing the increase in the degree distribution for higher degree. The inset presents the distance between the power law fit and the average of networks with score 0.800 measured in number of standard deviations. (**c)** Clustering coefficient and (**d)** mean nearest neighbor degree averaged over all core organisms. The insets in panels (**a**), (**c**) and (**d**) show individual results for all core organisms for each score.

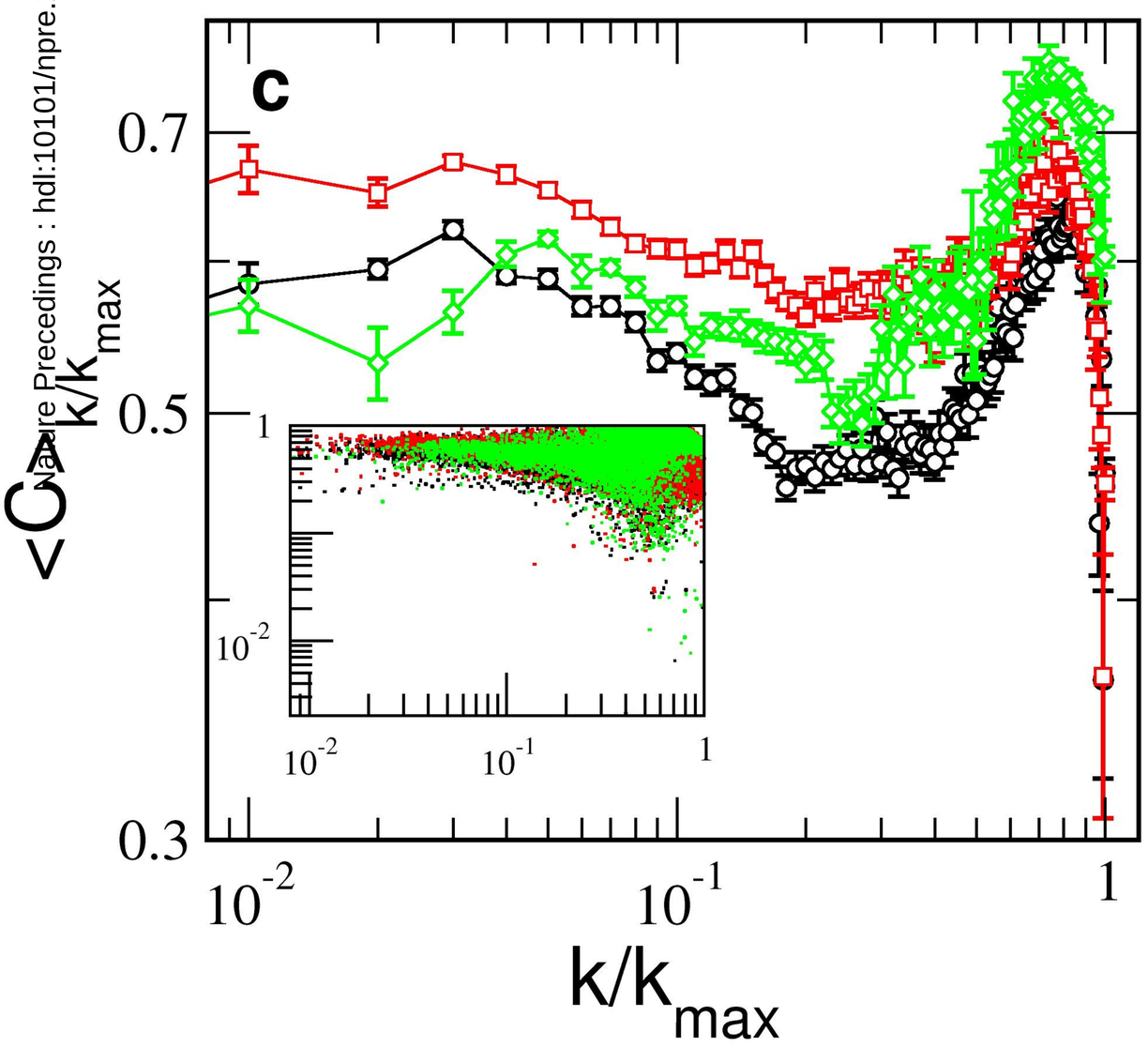
Fig 2. - Evolution of simulated Barabási-Albert, duplication-divergence and duplication-acquisiton networks (red, blue and green lines, respectively). The black dots represent all core organisms from STRING database, where six well studied organisms are highlighted in orange. (**a**) Number of links, (**c**) mean degree and (**e**) maximum degree are shown as functions of the total number of nodes in the network. The degree distribution was calculated in five snapshots of the evolution of (**b**) Barabási-Albert, (**d**) duplication-divergence, and (**f**) duplication-acquisition models, in intervals of 2000 nodes.

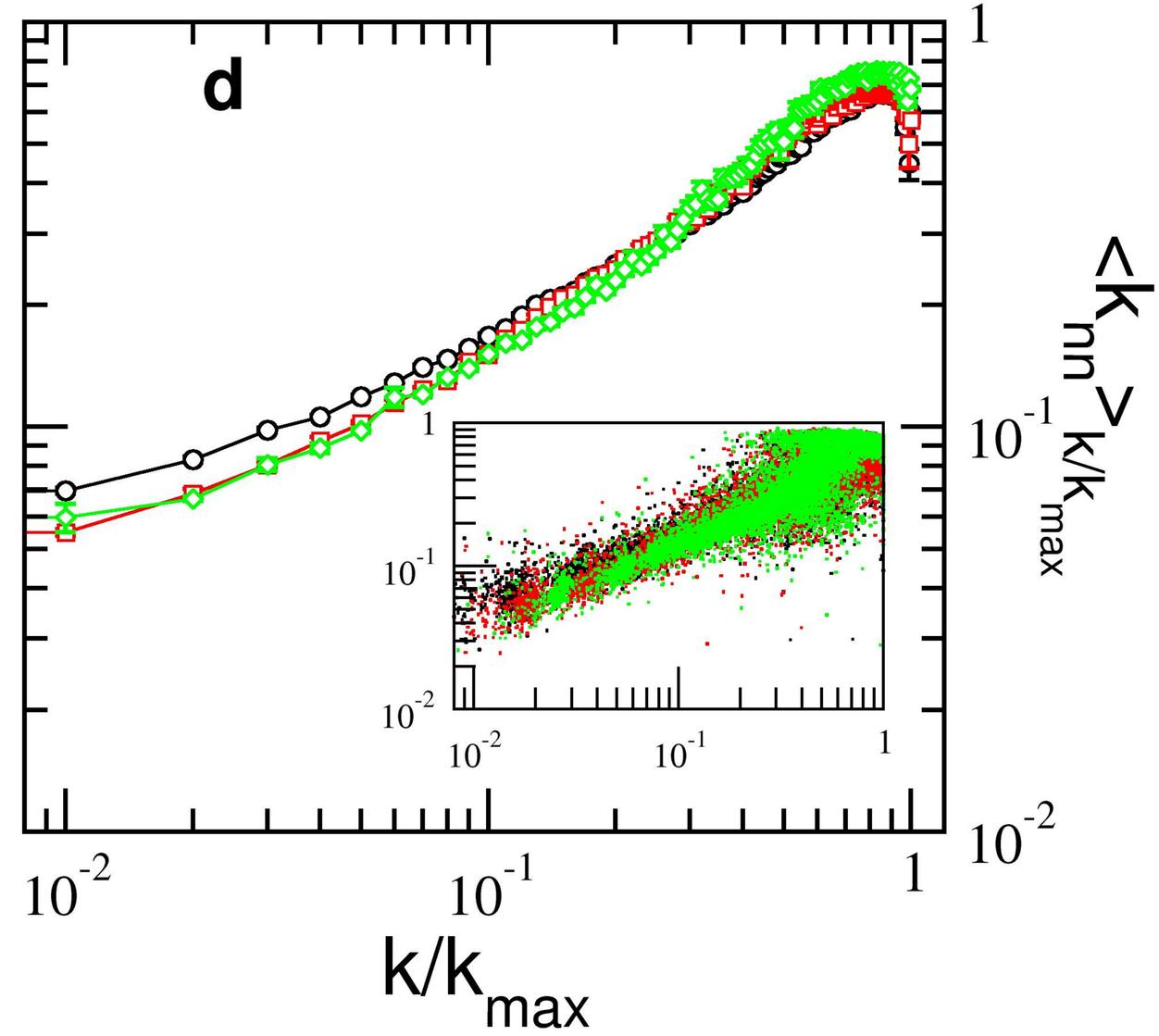
Fig 3. - Ordered association matrices. This figure presents the association matrices for *Homo sapiens*, *Mus musculus*, *Arabidopsis thaliana, Drosophila melanogaster, Saccharomyces cerevisiae, Escherichia coli*, Barbási-Albert model, duplication-divergence model and duplication-acquisition model after running the ordering algorithm. The black dots represent interactions between two nodes.

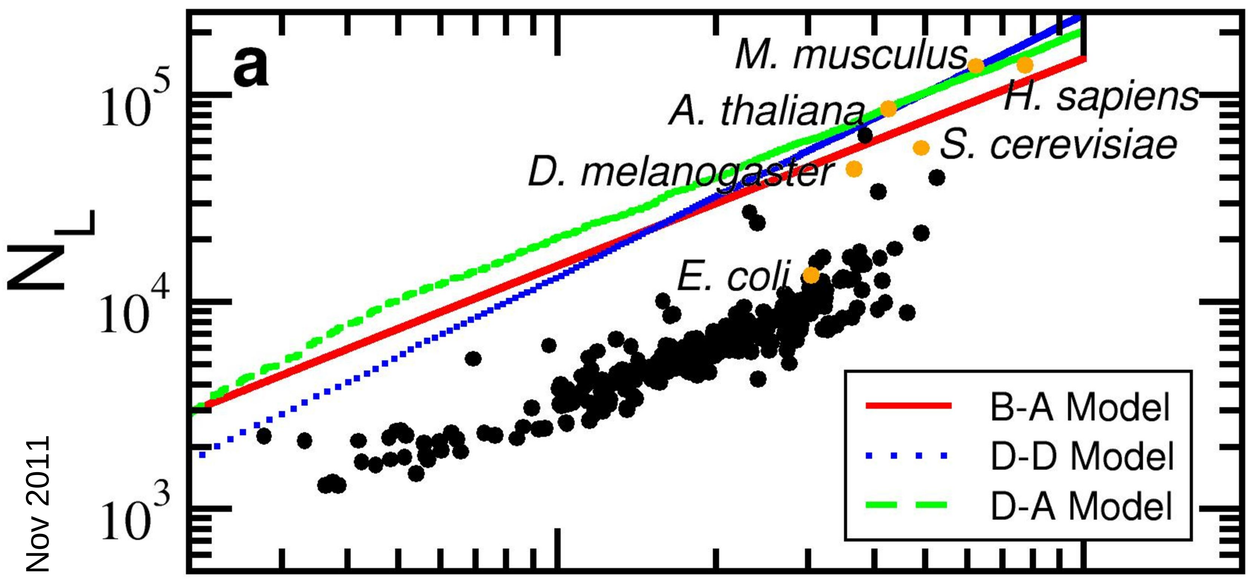
Fig 4. - Comparison of topological measures for simulated networks. The black dots represent the superposed networks for all core organisms from string database with confidence score 0.800, the green lines are averages taken in intervals of $k/k_{\max} = 0.01$, and the red lines are weighted averages of simulated networks. The upper, central, and lower rows show, respectively, degree distribution, clustering coefficient, and nearest neighbor mean degree. Each column refers to a simulated model: Barabási-Albert on the left, duplication-divergence on the center and duplication-acquisition on the right.

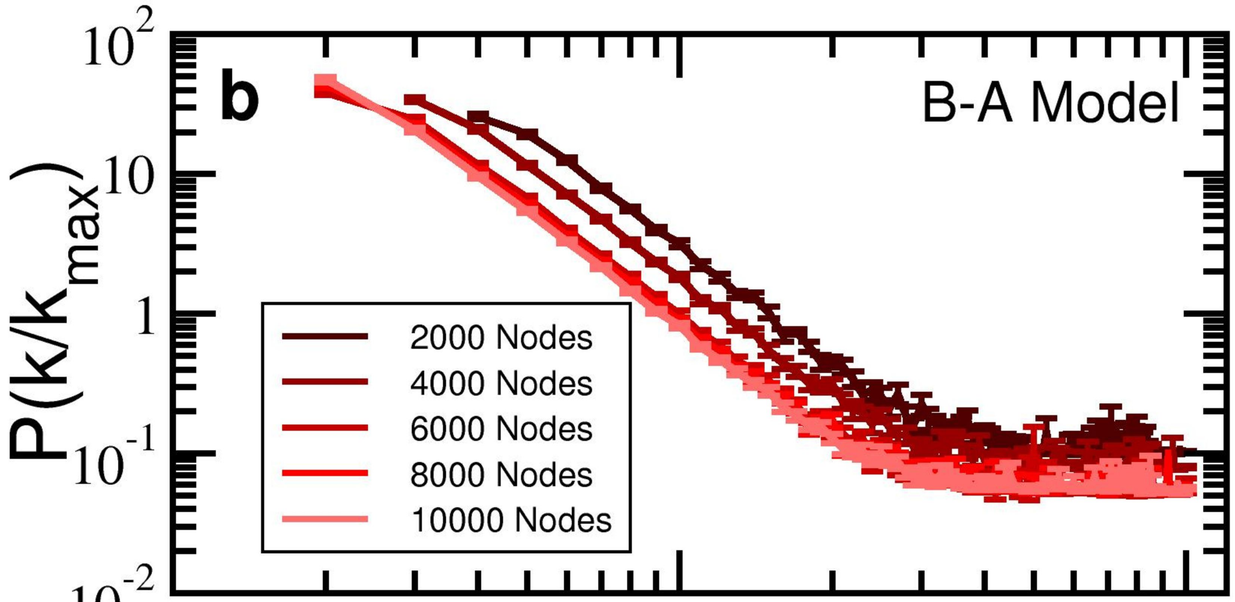


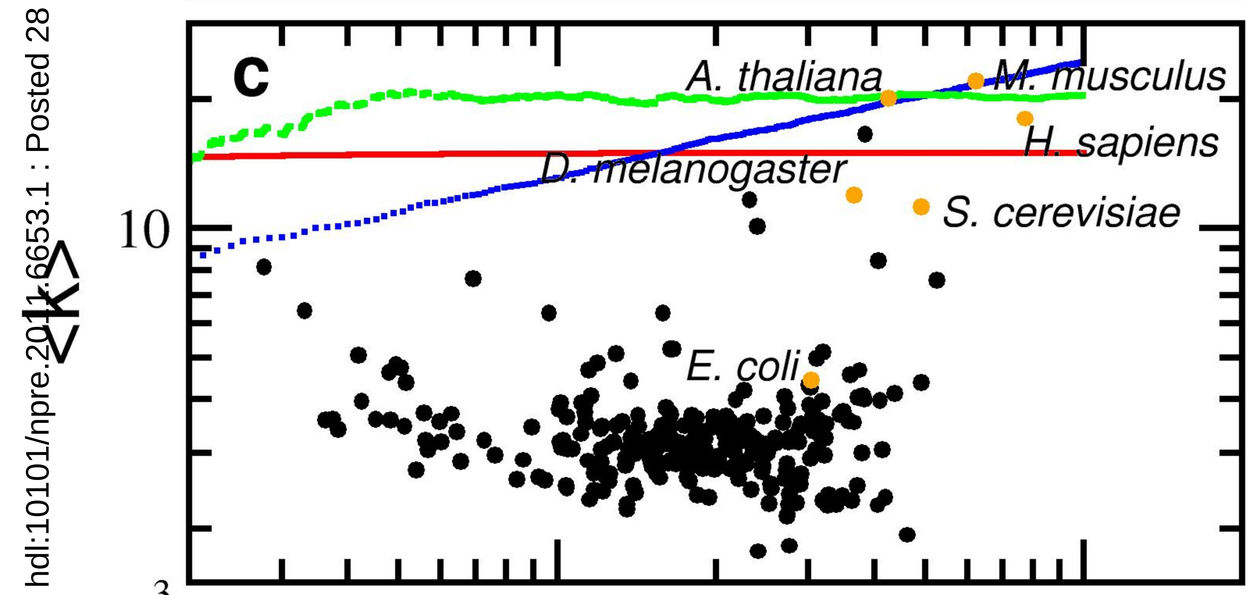


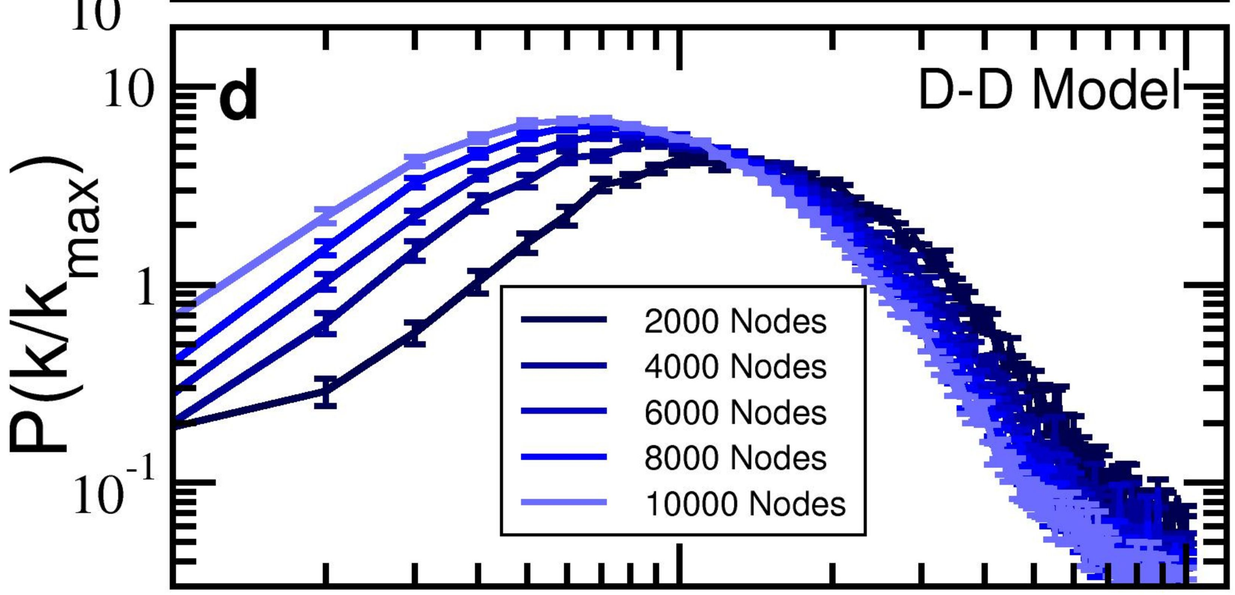


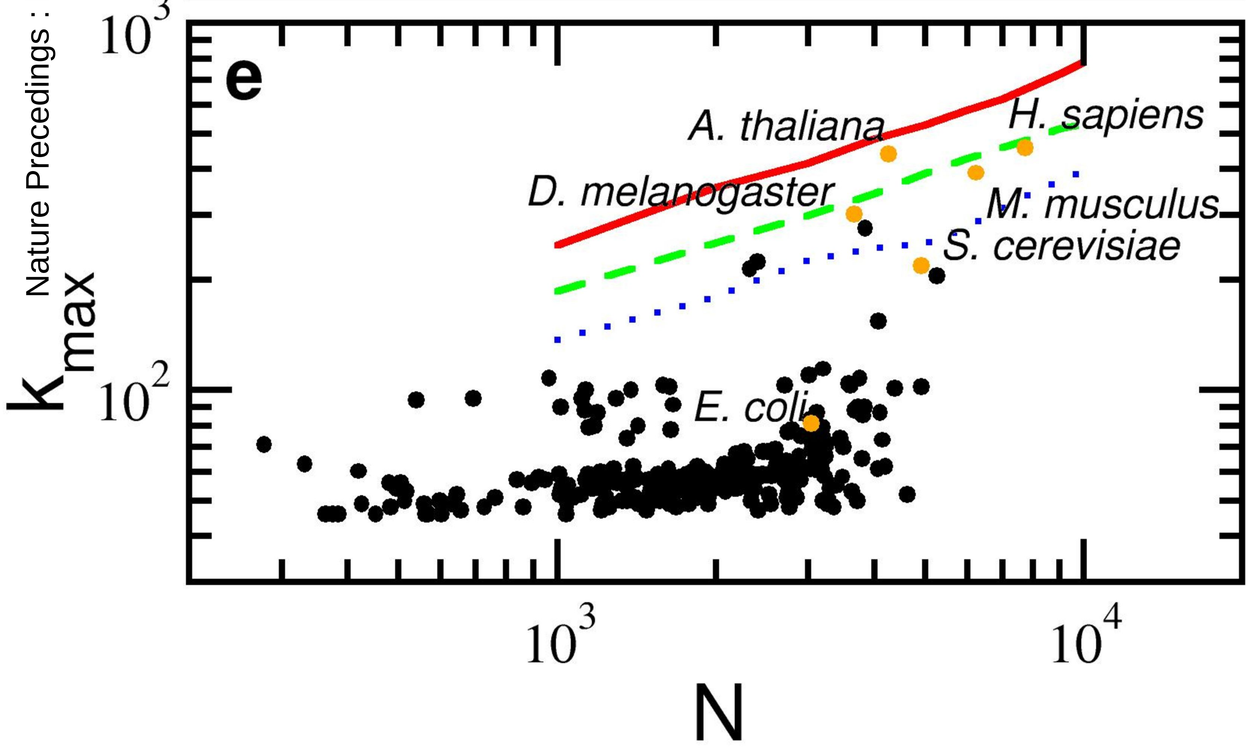


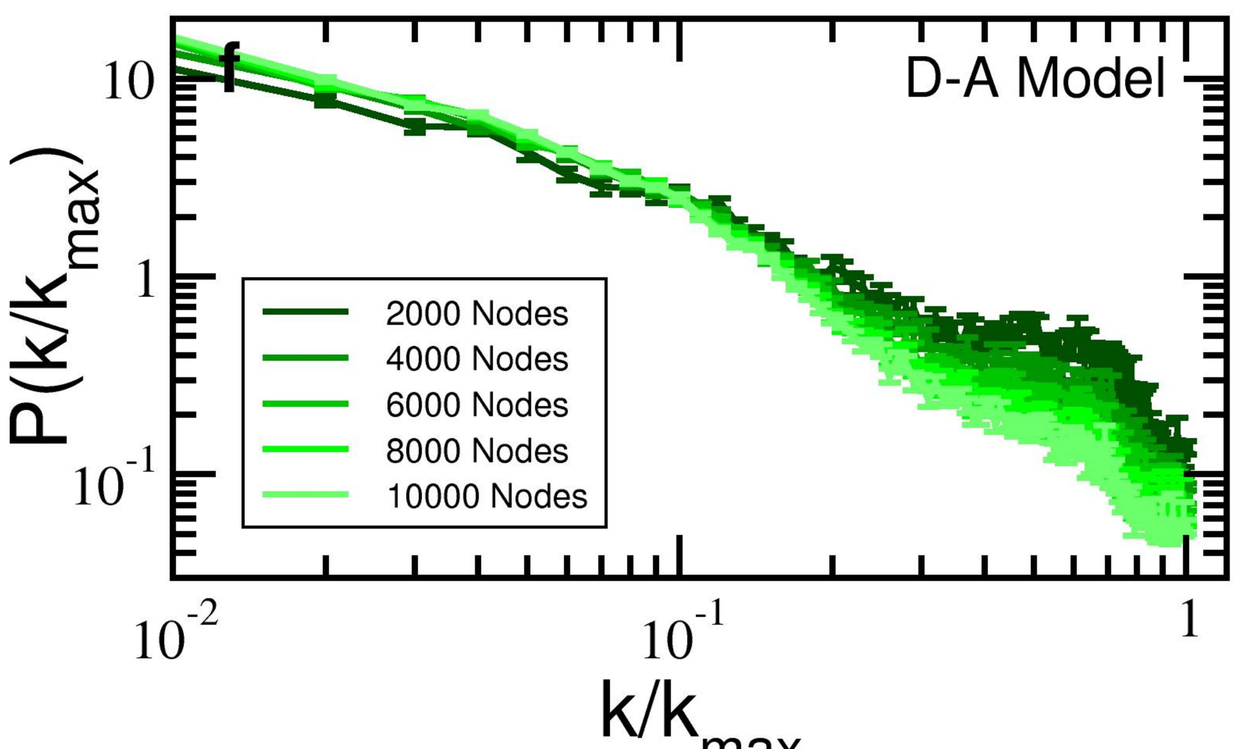


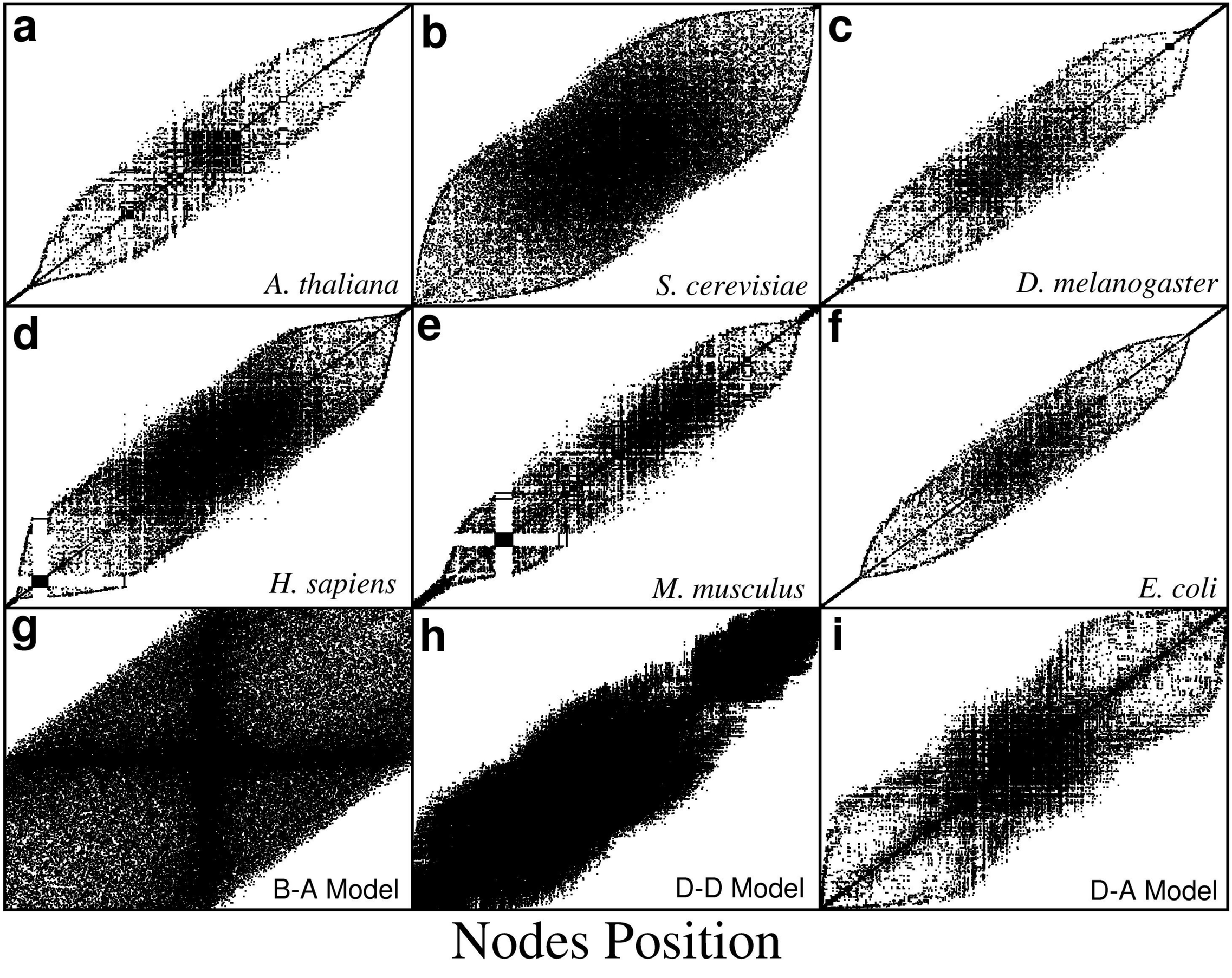












B-A Model

D-D Model

D-A Model

$P(k/k_{max})$

$<k_{nn}>_{k/k_{max}}$

k/k_{max}

a

b

c

d

e

f

g

h

i