Supplementary Material
for
Mass balanced randomization of metabolic networks

Georg Basler[1], Oliver Ebenhöh[2], Joachim Selbig[1,3], Zoran Nikoloski[1,3]

[1]University of Potsdam, Institute for Biochemistry and Biology, D-14476 Potsdam, Germany
[2]University of Aberdeen, Institute of Medical Sciences, AB25 2ZD Aberdeen, UK
[3]Max Planck Institute for Molecular Plant Physiology, D-14476 Potsdam, Germany

Contents

1 Overview of randomization methods

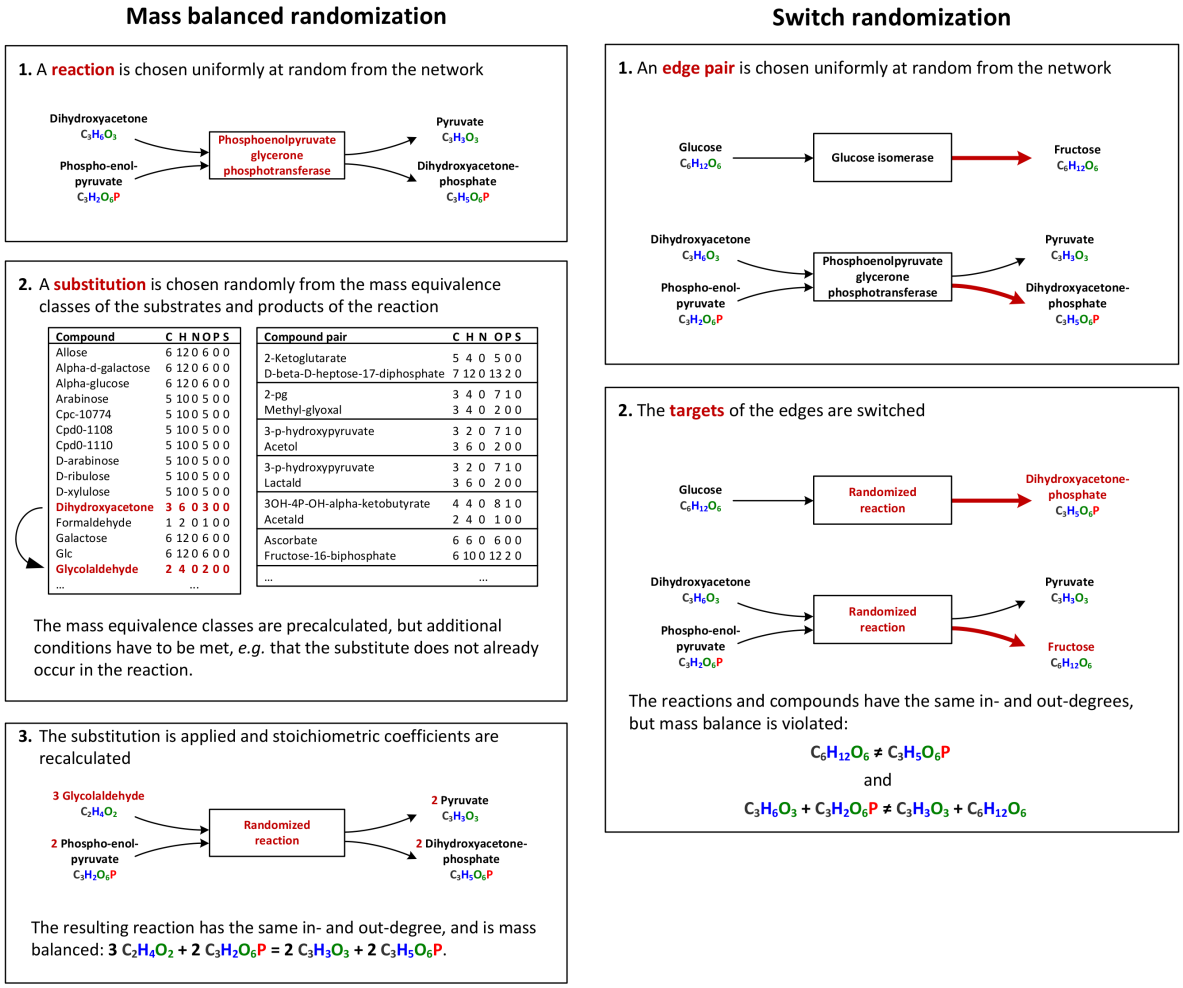


Figure 1: Workflow schemes depicting the mass balanced (left) and switch (right) randomization methods. The procedures are repeated a large number of times in order to obtain fully randomized networks.

2 Algorithms

Algorithm 1: Mass equivalence class calculation

Input:
Set of compounds, V_c
Output:
Mass equivalence classes, $\sigma = \{\sigma(c), \sigma(c,k)\},\ (c,k) \in V_c \times V_c,\ c \neq k$
σ := $\{\}$
$\forall c, k \in V_c$: $\sigma(c)$:= $\{\}$, $\sigma(c,k)$:= $\{\}$
foreach $c \in V_c$ **do**
1 **if** $\sigma(c) \notin \sigma$ **then**
2 add $\sigma(c)$ to σ
3 add c to $\sigma(c)$
4 **foreach** $\sigma(x) \in \sigma$ **do**
5 **foreach** $k \in \sigma(x), k \neq c$ **do**
6 **if** $\sigma(c,k) \notin \sigma$ **then**
7 add $\sigma(c,k)$ to σ
8 add (c,k) to $\sigma(c,k)$

Algorithm for calculating the mass equivalence classes for all individual compounds and pairs of compounds. Lines 1 and 6 involve testing whether a mass equivalent compound, respectively pair of compounds, is already in σ; likewise, lines 3 and 8 require retrieving the corresponding mass equivalence class from σ. Both can be done in constant time when using a hash map for σ, with the basis of the mass vector(s) as hash key, as the basis uniquely identifies a mass equivalence class (see Tables 1 and 2 in the main manuscript). Thus, the time complexity of the algorithm is in $O(|V_c|^2)$, as we iterate over each pair of compounds exactly once in line 5.

Algorithm 2: Mass balanced randomization of metabolic networks

Input:
Mass balanced metabolic network, $G = (V_c \cup V_r, E)$,
Mass equivalence classes, $\sigma = \sigma(c) \cup \sigma(c,k),\ (c,k) \in V_c \times V_c,\ c \neq k$,
Set of preserved compounds, $D \subset V_c$,
Number of iterations, $t \in \mathbb{N}^+$
Output:
Randomized mass balanced network
Repeat t times:
1 Choose a reaction $r \in V_r$ uniformly at random
2 **foreach** $c \in r \setminus D$ **do**
3 **foreach** $c' \in \sigma(c), c' \notin r \cup D$ **do**
4 add (c, c') to $\Psi_s(r)$
5 **foreach** $(c,k) \in (r_{in} \times r_{in}) \cup (r_{out} \times r_{out}), c, k \notin D$ **do**
6 **foreach** $(c', k') \in \sigma(c,k), c', k' \notin r \cup D$ **do**
7 Let $A = (m_{c'}, m_{k'})$ be the $(n \times 2)$ matrix of mass vectors of length n
8 Solve $As = b$ with $b = s_{c,r} \cdot m_c + s_{k,r} \cdot m_k$
9 **if** *there is a solution* $s_1, s_2 \in \mathbb{N}^+$ **then**
10 add $(c, k, c', k', s_1, s_2, 1)$ to $\Psi_p(r)$
11 **else if** *there is a solution* $s_1, s_2 \in \mathbb{Q}^+$ **then**
12 Let $f > 0$ be the smallest integer, such that $fs_1, fs_2 \in \mathbb{N}^+$
13 add $(c, k, c', k', s_1, s_2, f)$ to $\Psi_p(r)$
14 $\Psi(r) := \Psi_s(r) \cup \Psi_p(r)$
15 Choose a number $u \in \mathbb{N}^+$ uniformly at random from $[1, |\Psi(r)|]$
16 Let d_u be the u-th substitution in $\Psi(r)$
17 **if** d_u *is an individual substitution* (c, c') **then**
18 **if** c *is a substrate of* r **then**
19 replace the edge (c, r) by (c', r)
20 **else**
21 replace the edge (r, c) by (r, c')
22 Let $f > 0$ be the smallest integer, such that $\frac{f}{m_{c'}} \cdot s_{c,r} m_c \in \mathbb{N}^+$
23 $s_{c',r} := \frac{1}{m_{c'}} \cdot s_{c,r} m_c$
24 Multiply the stoichiometric coefficients of r by f
25 **else if** d_u *is a pair substitution* $(c, k, c', k', s_1, s_2, f)$ **then**
26 **if** c,k *are substrates of* r **then**
27 replace the edges (c, r) and (k, r) by (c', r) and (k', r)
28 **else**
29 replace the edges (r, c) and (r, k) by (r, c') and (r, k')
30 $s_{c',r} := s_1$
31 $s_{k',r} := s_2$
32 Multiply the stoichiometric coefficients of r by f

Detailed algorithm for mass balanced randomization of a metabolic network. For a randomly chosen reaction, the set of individual compound substitutions (lines 2-4) and the set of pair substitutions (lines 5-13) are determined from the mass equivalence classes. Optionally, a set of preserved compounds D may be specified, *e.g.* cofactors, which remain unmodified. For pair substitutions, it is necessary to determine whether there are stoichiometric coefficients satisfying mass balance (lines 7-8). If there is no rational solution, the pair substitution is neglected. In lines 14-16, a substitution is chosen uniformly at random from the set of all possible substitutions. In lines 18-21 and 26-29, the edges corresponding to the chosen substituted compounds are replaced by new edges connecting the substitutes. For an individual compound substitution, this involves determining the new stoichiometric coefficients of the reaction, which can always be found due to the linear dependence of mass vectors (lines 22-24). For a pair a substitution, the previously determined solution is used (lines 30-32). Note that the stoichiometric coefficients of compounds other than the substitutes are modified only if $f > 1$, which is the case if the substituted (sum of) mass vector(s) is no integer multiple of the new (sum of) mass vector(s) (see Table 3 in the main manuscript). For a full randomization, the number of iterations, t, should be chosen as the number of compounds and pairs of compounds available for substitutions. We use $t = \lceil |V_r| \cdot \overline{d}(V_r)^2 \rceil$ as an upper approximation, where $\overline{d}(V_r)$ is the average (undirected) reaction degree.

3 Metabolic networks

Table 1: Summary statistics for the seven analyzed genome-scale metabolic networks. $|V_r|$: number of reactions; $|V_c|$: number of compounds; CHNOPS: number of compounds consisting only of carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur; $T_{classes}$: runtime for calculating the mass equivalence classes; T_{rand}: runtime for generating 1000 mass balanced randomized networks; T_{switch}: runtime for generating 1000 switch randomized networks; $|\sigma_s|$, $|\sigma_p|$: number of mass equivalence classes for individual and pairs of compounds, respectively. All runtime calculations were performed on a single core of an Intel Xeon Processor E5345 with 2.33GHz and 16GB RAM running Fedora 13 Linux 32-bit and Sun Java SE 1.6.0 update 6 in JVM server mode.

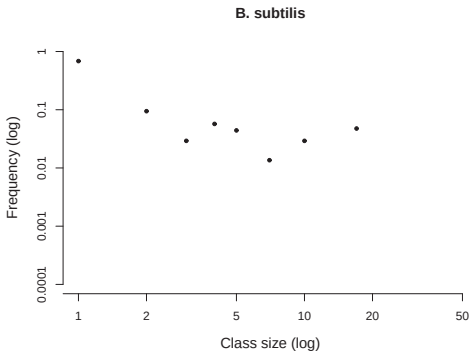
Network	$\|V_r\|$	$\|V_c\|$	CHNOPS	$T_{classes}$	T_{rand}	T_{switch}	$\|\sigma_s\|$	$\|\sigma_p\|$
B. subtilis [1]	855	777	736	15.4s	75.6m	490.7m	567	114116
S. cerevisiae [2]	1203	995	596	9.5s	176.1m	2228.6m	474	76076
E. coli (iAF1260) [3]	1481	1039	887	28.4s	124.9m	1809.5m	702	190230
E. coli (EcoCyc) [4]	1622	1088	785	23.5s	124.1m	1887.2m	611	133933
C. reinhardtii [5]	1541	1377	1054	43.5s	120.1m	1735.8m	780	192586
A. thaliana [6]	2508	2190	1790	191.9s	308.2m	5927.0m	1280	482925
H. sapiens [7]	2819	2690	1953	165.2s	537.9m	10981.4m	1214	386315

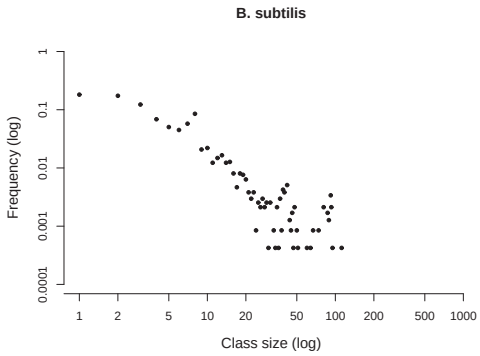
4 Uniformity properties

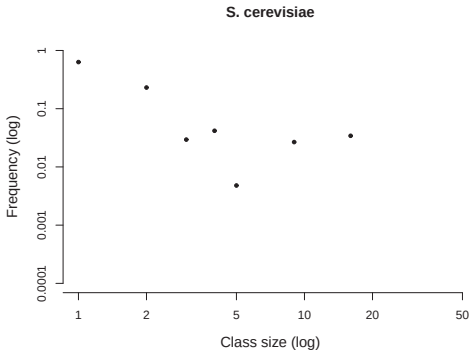
Table 2: Scaling coefficient γ and expected degree difference of adjacent nodes δ were obtained from a random walk, average degree of the transition graph $\overline{d}(\Sigma_G)$ was obtained from sampling 10^3 random walks on each network. P_{uni}: lower bound for the worst-case probability, that nodes from Σ_G are sampled almost uniformly at random after $t = 10^6$ steps; Ω_s: size of the sample space for individual substitutions.

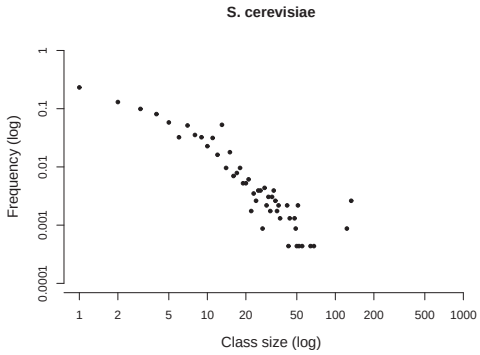
Network	γ	δ	$\overline{d}(\Sigma_G)$	P_{uni}	Ω_s
B. subtilis	1.85	6.91	15641	0.73	$4.4 \cdot 10^{645}$
S. cerevisiae	2.04	6.15	16032	0.82	$3.8 \cdot 10^{610}$
E. coli (iAF1260)	1.89	6.52	21769	0.81	$4.3 \cdot 10^{1089}$
E. coli (EcoCyc)	1.87	7.14	19490	0.80	$3.0 \cdot 10^{957}$
C. reinhardtii	1.77	7.38	25047	0.78	$6.3 \cdot 10^{972}$
A. thaliana	1.63	12.50	54201	0.80	$1.2 \cdot 10^{2033}$
H. sapiens	1.24	22.55	158138	0.53	$2.8 \cdot 10^{2210}$

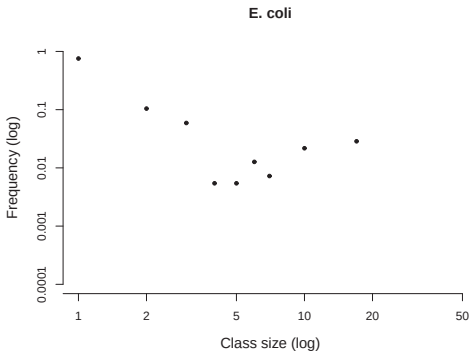
5 Mass equivalence class size distributions

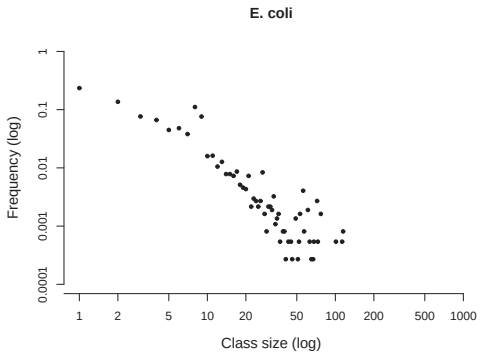












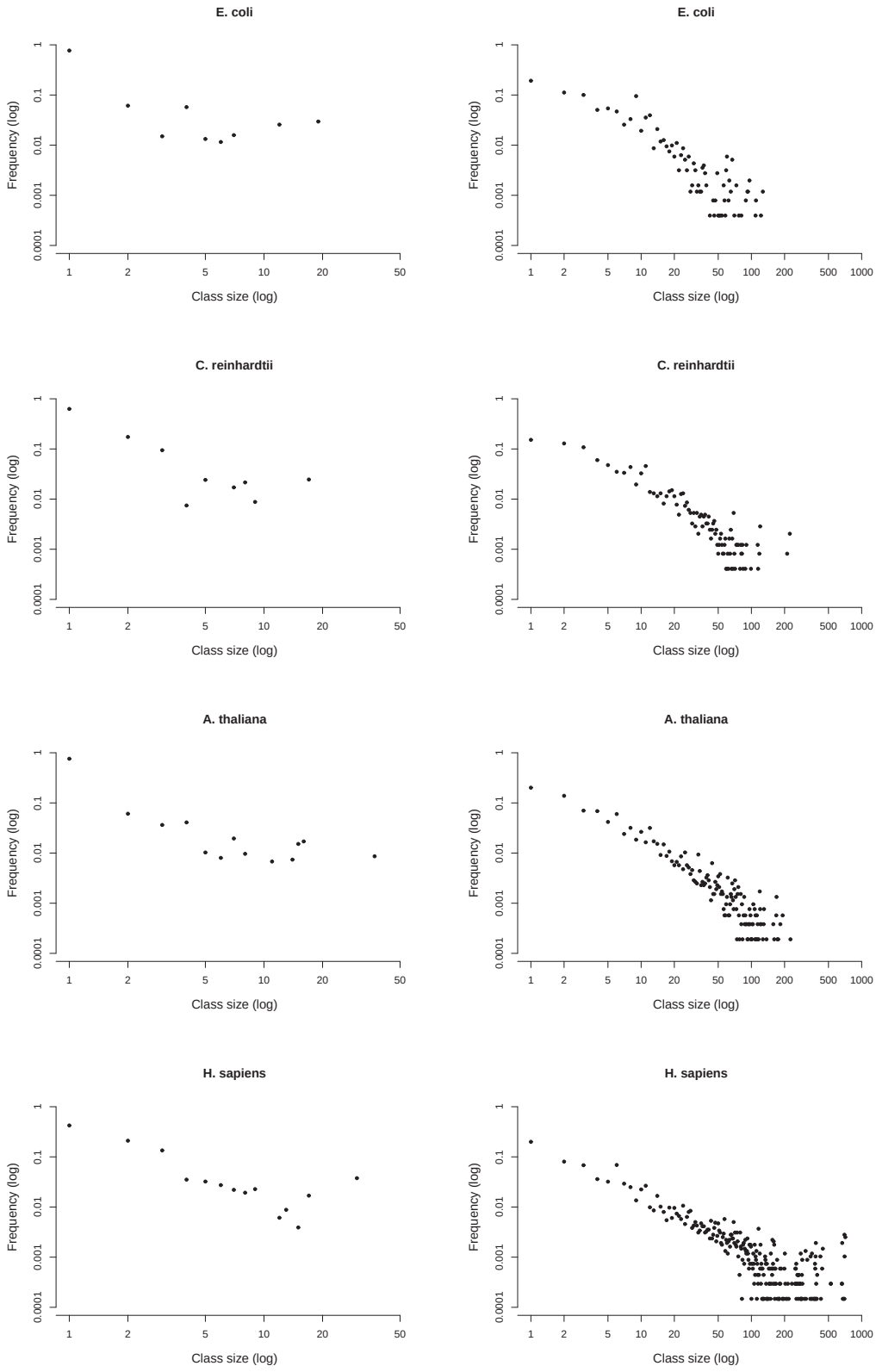


Figure 2: Mass equivalence class size distributions for individual compounds (left column) and pairs of compounds (right column).

6 Distributions of differences in degrees

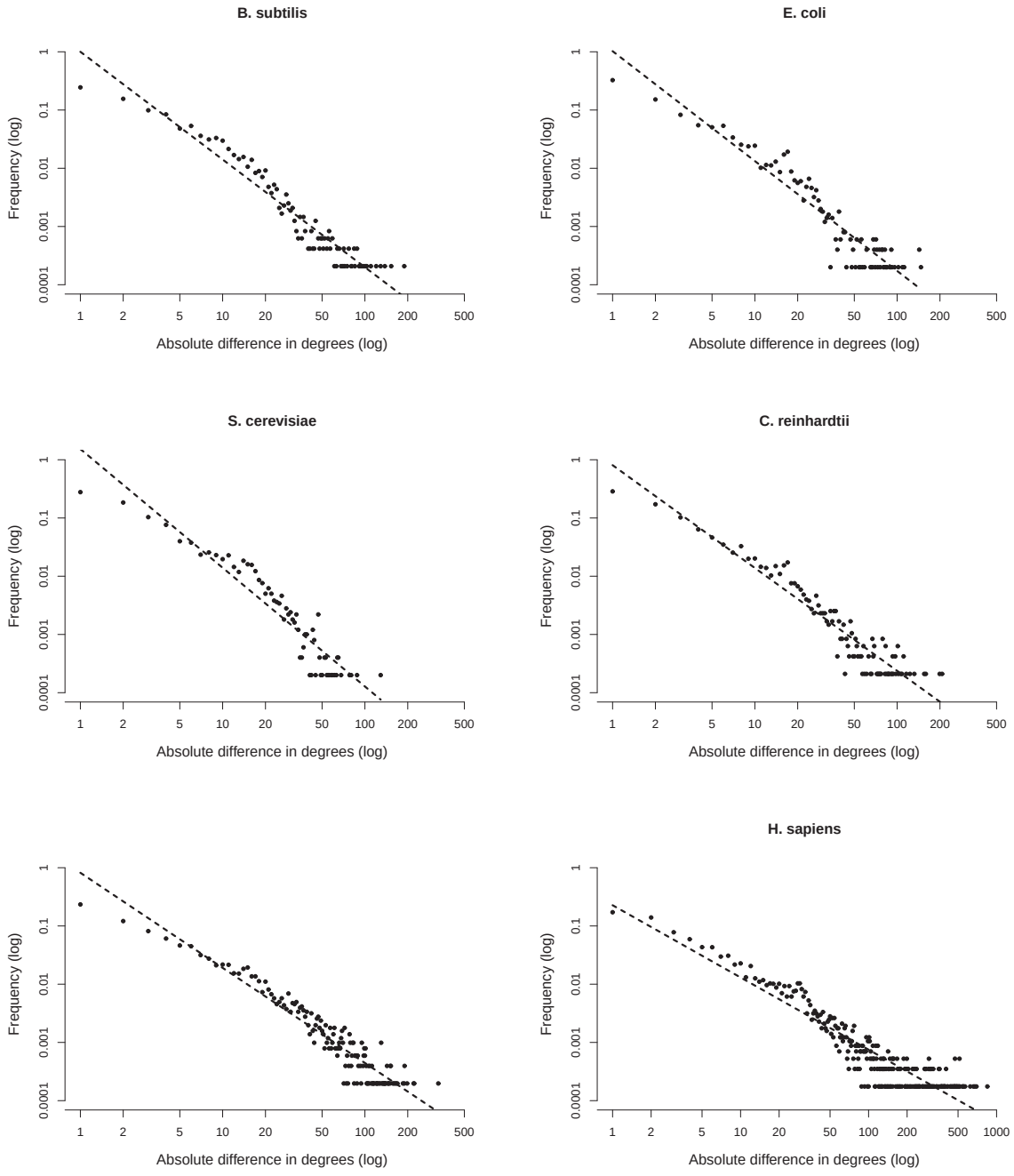


Figure 3: Distributions of absolute differences in degrees between neighbors, sampled by random walks on the transition graphs of *B. subtilis*, *S. cerevisiae*, *E. coli* (iAF1260), *C. reinhardtii*, *A. thaliana*, and *H. sapiens*. The dashed lines show the power-law fit. Scaling coefficients and mean differences are given in Table 2.

References

[1] Oh, Y.-K., Palsson, B. Ø., Park, S. M., Schilling, C. H., and Mahadevan, R. *J Biol Chem* **282**(39), 28791–28799 Sep (2007).

[2] Herrgård, M. J., Swainston, N., Dobson, P., Dunn, W. B., Arga, K. Y., Arvas, M., Blüthgen, N., Borger, S., Costenoble, R., Heinemann, M., Hucka, M., Novre, N. L., Li, P., Liebermeister, W., Mo, M. L., Oliveira, A. P., Petranovic, D., Pettifer, S., Simeonidis, E., Smallbone, K., Spasić, I., Weichart, D., Brent, R., Broomhead, D. S., Westerhoff, H. V., Kirdar, B., Penttil, M., Klipp, E., Palsson, B. Ø., Sauer, U., Oliver, S. G., Mendes, P., Nielsen, J., and Kell, D. B. *Nat Biotechnol* **26**(10), 1155–1160 Oct (2008).

[3] Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V., and Palsson, B. Ø. *Mol Syst Biol* **3**, 121 (2007).

[4] Keseler, I., Bonavides-Martinez, C., Collado-Vides, J., Gama-Castro, S., Gunsalus, R., Johnson, D., Krummenacker, M., Nolan, L., Paley, S., Paulsen, I., Peralta-Gil, M., Santos-Zavaleta, A., Shearer, A., and Karp, P. *Nucleic Acids Research* **37**, D464–D470 Jan (2009).

[5] May, P., Wienkoop, S., Kempa, S., Usadel, B., Christian, N., Rupprecht, J., Weiss, J., Recuenco-Munoz, L., Ebenhöh, O., Weckwerth, W., and Walther, D. *Genetics* **179**(1), 157–166 May (2008).

[6] Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P., and Huala, E. *Nucleic Acids Research* **36**(Database issue), D1009–14 January (2008).

[7] Ma, H., Sorokin, A., Mazein, A., Selkov, A., Selkov, E., Demin, O., and Goryanin, I. *Mol Syst Biol* **3**, 135 (2007).