Electronic Supplementary Material (ESI) for Molecular BioSystems

Supplementary material for

Inferring gene functions through dissection of relevance networks: Interleaving the intra- and inter-species views

Sebastian Klie,[a] Marek Mutwil,[a] Staffan Persson[a] and Zoran Nikoloski*[,a]

[a] *Max-Planck Institute of Molecular Plant Physiology, Potsdam, Germany*
[*] *Corresponding author, E-mail: nikoloski@mpimp-golm.mpg.de*

Yeast microarray gene-expression data-set

The compendium of microarray experiments used to derive proximity and relevance networks includes publicly available data obtained from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/)[1]. Initially, a total of 1340 microarray experiments (Affymetrix Yeast Genome S, 98, 9.335 probe sets) were downloaded. In addition, the quality was ensured for each individual microarray experiment through a two-stage procedure: First, visual inspection of box plots of raw positive match data and RMA residuals of RMA-normalized data by using the RMA express program was performed. CEL files exhibiting either artefacts on the RMA residual plots or visible deviations from the majority on the positive match box plots were removed from further analysis. Second, automated outlier detection, available from the R Bioconductor package arrayQualityMetrics[2], was performed by conducting (1) between-array comparisons based on distance between arrays and Principal Component Analysis, (2) inspection of array-wide probe intensity distributions by boxplots and density plots, (3) variance-mean dependence of each array, and (4) individual array quality assessment by MA plots. From these analyses, 1176 microarrays were retained, which were finally normalized using quantile normalization via the `simpleAffy` R package.

Gene function prediction using network-based majority voting

In order to analyse whether the degree of incompleteness of biological annotation influences the performance of *guilt-by-association* (GBA) based network-driven gene function prediction, we employ the aforementioned yeast transcriptomics dataset after quality control. Further, one proximity network and one relevance network were constructed under the constraint that they are of same density, *i.e.*, the networks contain the same number of edges. The edges were obtained by using the Pearson's correlation coefficient, quantifying the similarity between gene profiles from the 1176 experiments. For both networks, the density is 0.006; it corresponds to a threshold $\tau = 0.62$ for the correlation coefficient in case of the relevance network and a cut-off of 55 for the highest reciprocal rank in case of the proximity network. Both networks are comprised of the same set of 4161 nodes, representing genes annotated with both GO-MF (molecular function, 4301 genes) or GO-BP terms (biological process, 4929) of the gene-ontology GO[3].

The influence of successively sparser annotation is simulated by artificially removing different fractions of those 4161 annotated genes, so that the percentage of unannotated genes varies between 10% and 90%. For each percentage of missing annotation, we randomly sampled with uniform probability a given fraction of genes and treated them as unannotated. For the set of unannotated genes, we employed a majority voting network-based algorithm for automated annotation via the GBA principle. For each of the two sub-ontologies – GO-MF and GO-BP, we derived predictions separately. Furthermore, for every node, we considered all of its neighbours in the majority voting.

For the validation of the obtained predictions, we compared the highest scoring prediction (*i.e.*, the one with the most votes) and the third highest scoring prediction with the original annotation. This validation strategy is based on the observation that 50% of the currently annotated genes are annotated with 3 or more terms. A validation of only the first term is consequently not optimally reflecting the structure of Yeast gene annotation. Additionally, correctly inferred predictions up to the third term allow for a more differential characterization of an unannotated gene.

Furthermore, to reduce issues related to deriving a trivial or less specific annotation which corresponds to the root nodes within the hierarchy of GO terms, we removed terms the 10 terms with the lowest information content[4] (*cf.* Material and Methods in the main text). An example of terms exhibiting a low information content are within the GO-BP sub-ontology 'biological_process' (GO:0008150), *i.e.* the root term or 'transport' (GO:0006810). For GO-

MF, examples of removed terms include 'binding' (GO:0005488) and, again, the root node 'molecular_function' (GO:0003674). Note, that more specific terms corresponding to higher information content values are retained. These include, for instance, the children of the term 'binding', 'secretion' (GO:0046903) and 'ion transport' (GO:0006811).

Within the automated process of gene-function prediction by majority voting, we repeated the predictions for each fraction of un-annotated genes 1000 times, and, in every iteration, a different set of randomly unannotated genes was sampled.

Characterizing degree distributions observed in proximity and relevance networks

As shown in the main material (section 'Relevance and proximity networks'), relevance and proximity networks of the same density strongly differ in their structure. The different approaches of establishing edges in both networks ultimately yield different distribution of the degrees of nodes. In the case of the scale-free relevance network, this means that the fraction $P(x)$ of nodes in the network having x connections to other nodes follows for large values of x the power-distribution

$$P(x) \sim c x^{-\alpha},$$

which is invariant to scaling implying that the constant c simply multiplies the original power-law relation. By using maximum likelihood methods available in the R package `igraph` (function `power.law.fit`) we determined the exponent alpha=1.89 (see Fig. 2, Materials and Methods in the main text).

In contrast to the relevance network, visual inspection of the degree distribution from the proximity network indicated that it does not follow a power-law distribution. While relevance networks resemble a scale-free network[5-7], *i.e.*, a network whose degree distribution follows a power law, we claim that proximity networks resemble a gamma-distribution. Again, by fitting various typical distributions (*e.g.* exponential, (log-) normal, Poisson and *t*-distributions) and determining the maximum likelihood, we determined the gamma-distribution to be the best representative for the observed degree distribution. All distributions were fitted using the function `fitdistr` from the R package `MASS`. A gamma distribution can be expressed in terms of the gamma function Γ as

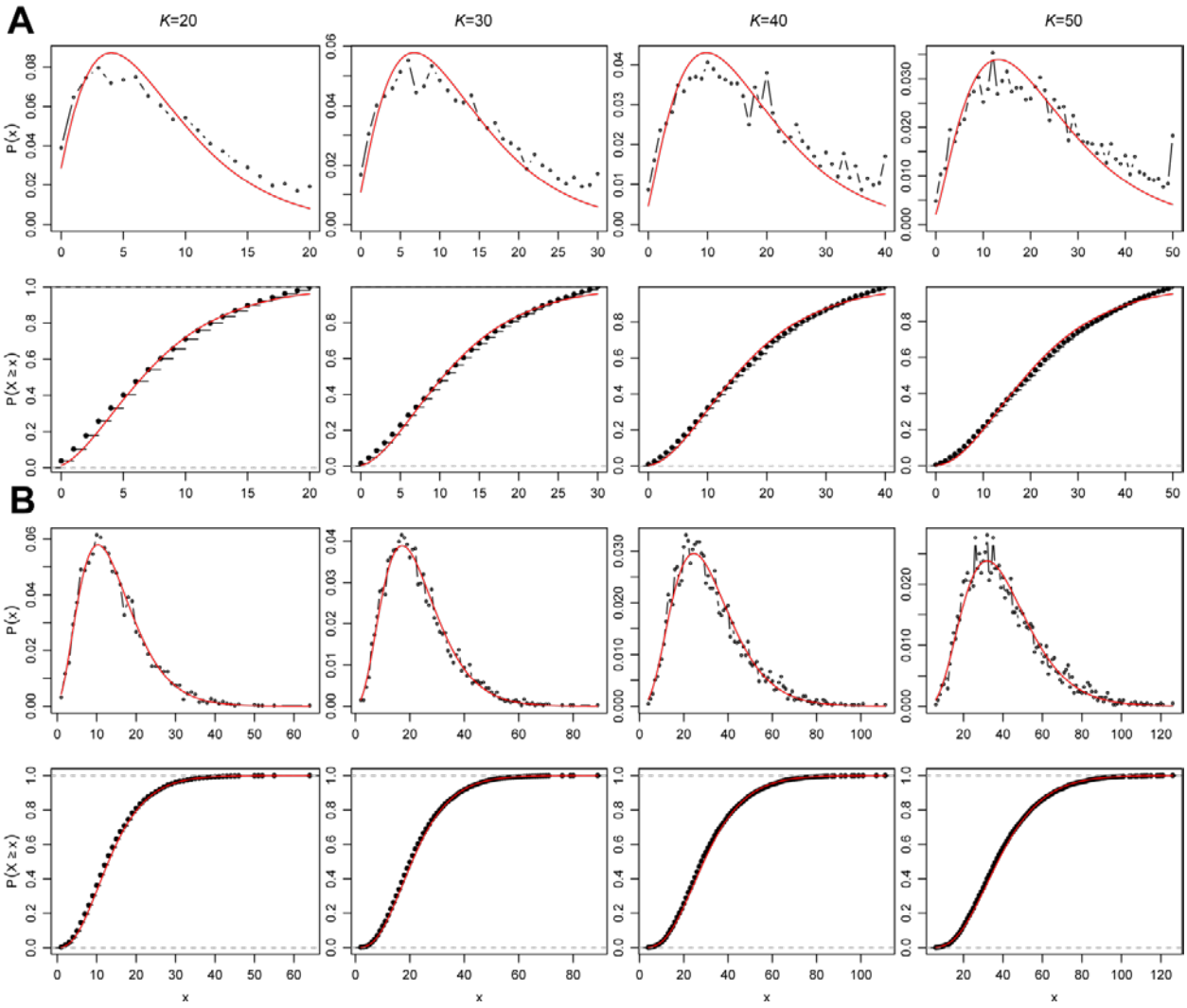
$$g(x;\, k, \theta) = \frac{1}{\theta^k} \frac{1}{\Gamma(k)} x^{k-1} e^{-\frac{x}{\theta}},$$

where $k > 0$ and $\theta > 0$ are the shape and scale parameter, respectively. After determination of both the shape and scale parameter, we verified the accordance of both the observed and the fitted distribution by visual inspection of both the obtained probability density function and empirical cumulative distribution function (Supplemental Fig. 1) illustrating the correspondence of the obtained degree distribution and the theoretical gamma distributions. Additionally, we employed two statistical test to quantify the goodness-of-fit using the Chi-Square Goodness-of-Fit test[8] and Kolmogorov-Smirnoff (KS) test[9]. In the case of the KS test, the observed degree distribution is tested against a theoretical distribution obtained using the fitted gamma distribution. Here, H_0 assumes that both samples originate from the same distribution. A 'discrete' gamma distribution is obtained by generation and subsequent binning of random values derived from gamma distribution using the fitted parameters. The employed KS test can be found in the `dgof` R package and is based on considerations of Conover [10] for discrete distributions.

Furthermore, we applied the Chi-square Goodness-of-Fit test, as it is a favourable alternative to the Kolmogorov-Smirnoff test[9], particularly in the case of discrete distributions combined with a large sample size, such as the obtained node-degree distributions. Here, the observed frequencies of each node-degree are tested against theoretically obtained probabilities in the respective integer intervals for each degree using the fitted reference gamma distribution (*i.e.* the probability density function). Finally, exact *p*-values can be obtained using Monte Carlo simulation[11] available in the function `chisq.test` in the R stats package. Again, as in the case of the KS test, H_0 in the Chi-Square Goodness-of-Fit test assumes that both distributions are the same.

Supplementary Table 1 summarizes the *p*-values obtained using both statistical test, separately for proximity networks using different thresholds for *highest reciprocal ranks* (Supplemental Table 1A), as well as *mutual ranks* (Supplemental Table 1B, for the definitions please refer to the main text). In the case of proximity networks based on mutual ranks, for most tested thresholds (20 to 80) the goodness-of-fit of a gamma distribution can be confirmed by failing to reject of H_0 at a significance level of 1% and 5%. In case of the proximity networks based on the highest reciprocal ranks, *p*-values indicate that H_0 can be rejected. However, the visual inspection of the obtained degree distribution and theoretically obtained gamma distributions suggests that the gamma distribution closely approximates the degree distribution.

Supplemental figures



Supupplemetal Fig. 1. Visualization of the obtained degree distribution of proximity networks using the *highest reciprocal rank* approach (A) and the *mutual rank* approach (B) for different selected threshold (*K*=20, 30, 40, 50; left-to-right). The theoretical degree distribution following a gamma distribution is shown for comparison (red line) both for the probability density function (first and third row) and compared to the empirical cumulative distribution function (ECDF, third and fourth row).

Supplemental tables

A) *highest reciprocal rank*

K	log-likelihood	AIC	*p*-value (KS-test)	*p*-value (chi-square *gof*)
10	-9702.28	19513.44	<0.001	<0.001
15	-11311.51	22678.36	<0.001	<0.001
20	-12475.12	24981.94	<0.001	<0.001
25	-13398.30	26811.69	<0.001	<0.001
30	-14172.90	28355.32	<0.001	<0.001
35	-14813.58	29637.71	<0.001	<0.001
40	-15395.24	30809.66	<0.001	<0.001
45	-15904.93	31828.56	<0.001	<0.001
50	-16362.49	32754.32	<0.001	<0.001
55	-16769.20	33577.05	<0.001	<0.001
60	-17127.02	34298.27	<0.001	<0.001
65	-17469.21	34994.99	<0.001	<0.001
70	-17785.97	35619.26	<0.001	<0.001
75	-18080.16	36207.20	<0.001	<0.001
80	-18356.43	36773.44	<0.001	<0.001

B) *mutual ranks*

K	log-likelihood	AIC	*p*-value (KS-test)	*p*-value (chi-square *gof*)
10	-11300.44	22605.43	<0.001	<0.001
15	-12930.71	25867.07	0.002	<0.001
20	-14059.73	28128.52	**0.285**	0.001
25	-14963.76	29934.40	**0.098**	**0.064**
30	-15693.95	31394.96	**0.727**	**0.364**
35	-16301.66	32612.11	**0.367**	**0.273**
40	-16830.51	33669.18	**0.164**	**0.106**
45	-17294.38	34596.84	**0.538**	**0.179**
50	-17717.87	35443.25	**0.394**	**0.046**
55	-18102.56	36212.96	**0.887**	0.026
60	-18458.32	36925.70	**0.826**	0.020
65	-18781.24	37569.19	**0.321**	**0.341**
70	-19074.10	38154.03	**0.701**	0.014
75	-19347.33	38700.53	**0.673**	<0.001
80	-19589.58	39184.43	**0.533**	0.006

Supplemental Table 1. Parameters describing the goodness-of-fit for the gamma distribution and the observed degree-distribution for proximity networks, using highest reciprocal ranks (A) and mutual ranks (B) using different thresholds for the number of nearest neighbours considered (column 1).

Supplemental references

1. R. Edgar, M. Domrachev and A. E. Lash, *Nucleic Acids Research*, 2002, **30**, 207-210.
2. A. Kauffmann, R. Gentleman and W. Huber, *Bioinformatics*, 2009, **25**, 415-416.
3. M. Ashburner, *Nature Genetics*, 2000, **25**, 25-29.
4. P. Resnik, *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1995, 448-453.
5. A. L. Barabasi and R. Albert, *Physical Review Letters*, 2000, **85**, 5234-5237.
6. A.-L. Barabasi and Z. N. Oltvai, *Nat Rev Genet*, 2004, **5**, 101-113.
7. J. Ruan, A. Dean and W. Zhang, *BMC Systems Biology*, 2010, **4**, 8.
8. R. R. Sokal and F. J. Rohlf, *Biometry*, W.H. Freeman and Company, New York, 2003.
9. G. W. Snedecor and W. G. Cochran, *Statistical methods*, Iowa State University Press, 1989.
10. W. J. Conover, *Journal of the American Statistical Association*, 1972, **67**, 591-596.
11. A. C. A. Hope, *Journal of the Royal Statistical Society. Series B (Methodological)*, 1968, **30**, 582-598.