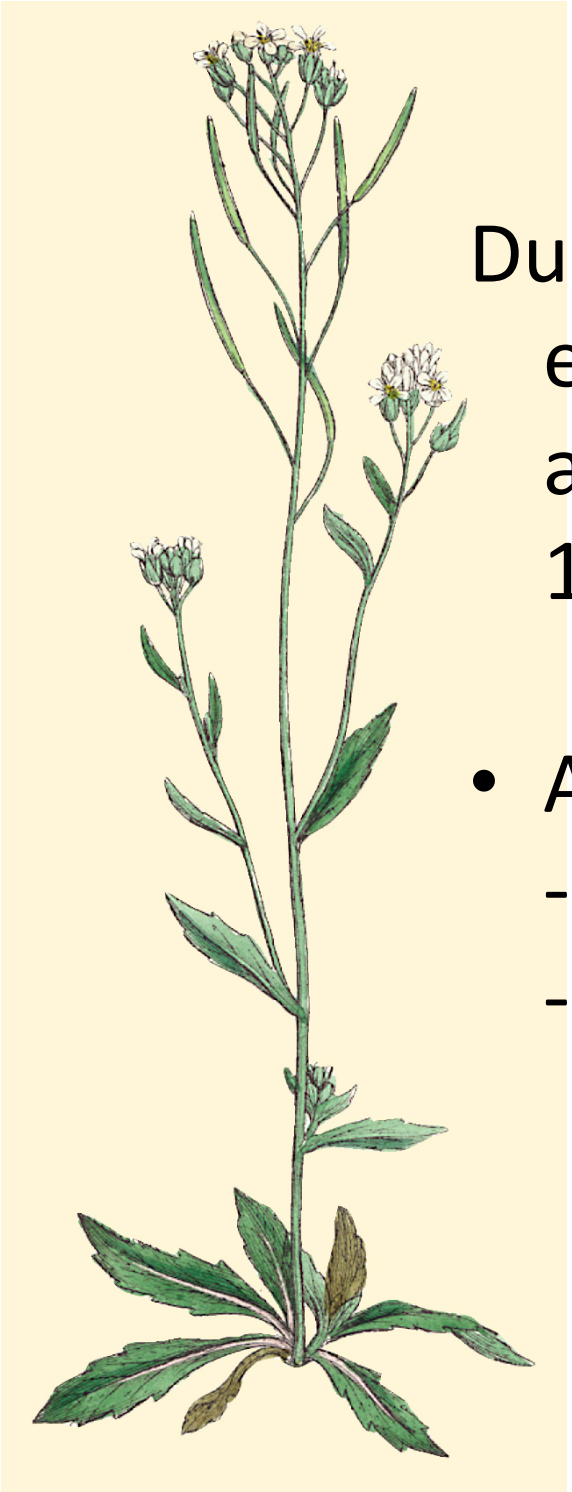
What can biological networks tell us about the fate of duplicate genes in *Arabidopsis thaliana*?

Justin WHALLEY, Julien CHIQUET, Etienne BIRMELÉ, Kousuke HANADA and Carène RIZZON

LABORATOIRE Statistique et Genome, UMR8071 CNRS, 23 bvd de France, 91037, France

justin.whalley@genopole.cnrs.fr



Duplicated genes are very important in evolution as they are implied in the appearance of new functions (Ohno, 1970).

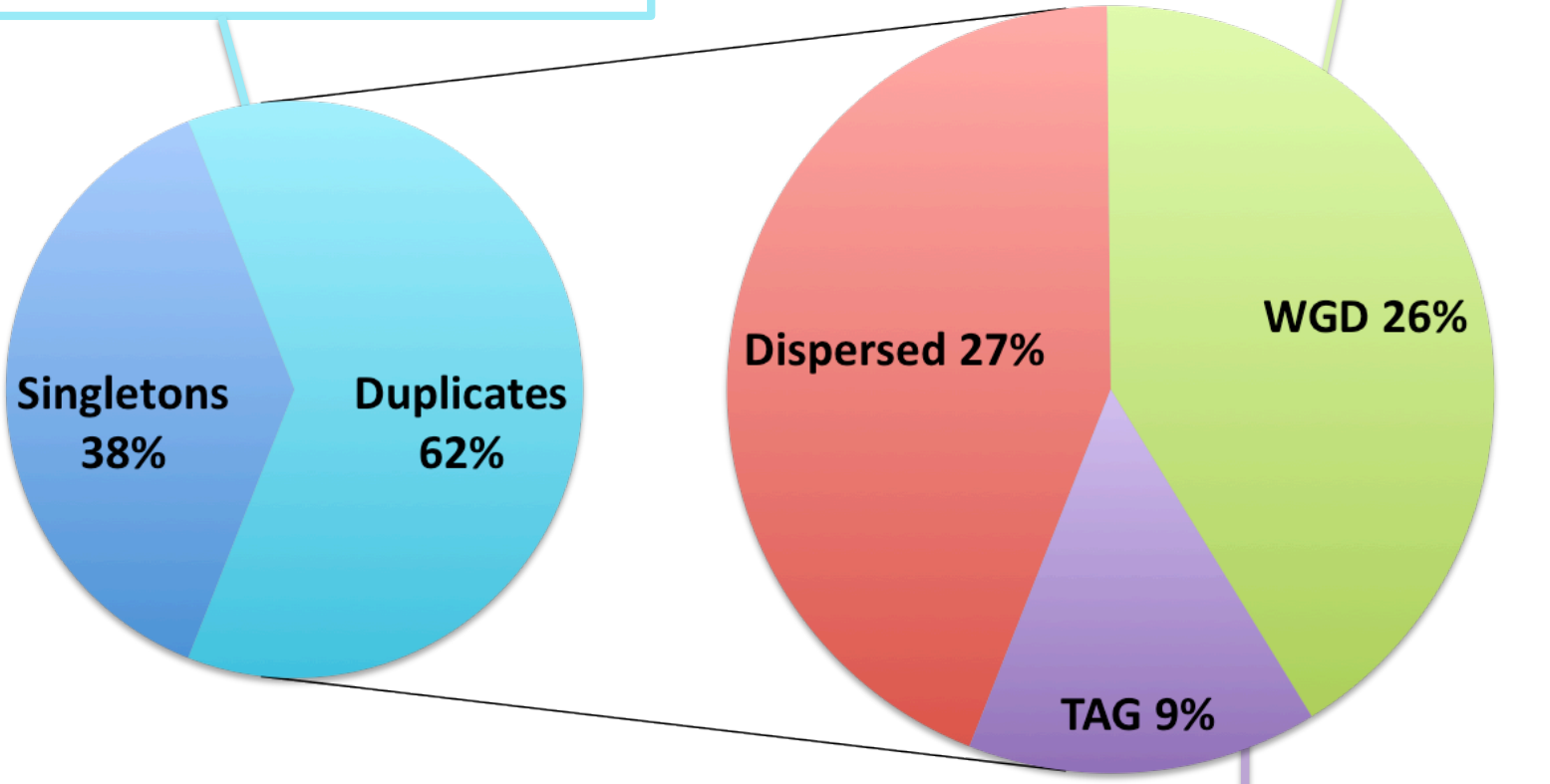
- Analysis of expression data:
  - using standard methods
  - using inferred networks

Arabidopsis thaliana

Types of Duplication

The 16977 duplicate genes are clustered into 5253 families.

Whole Genome Duplicates are defined from the literature (Bowers 2003).



Tandemly Arrayed Genes

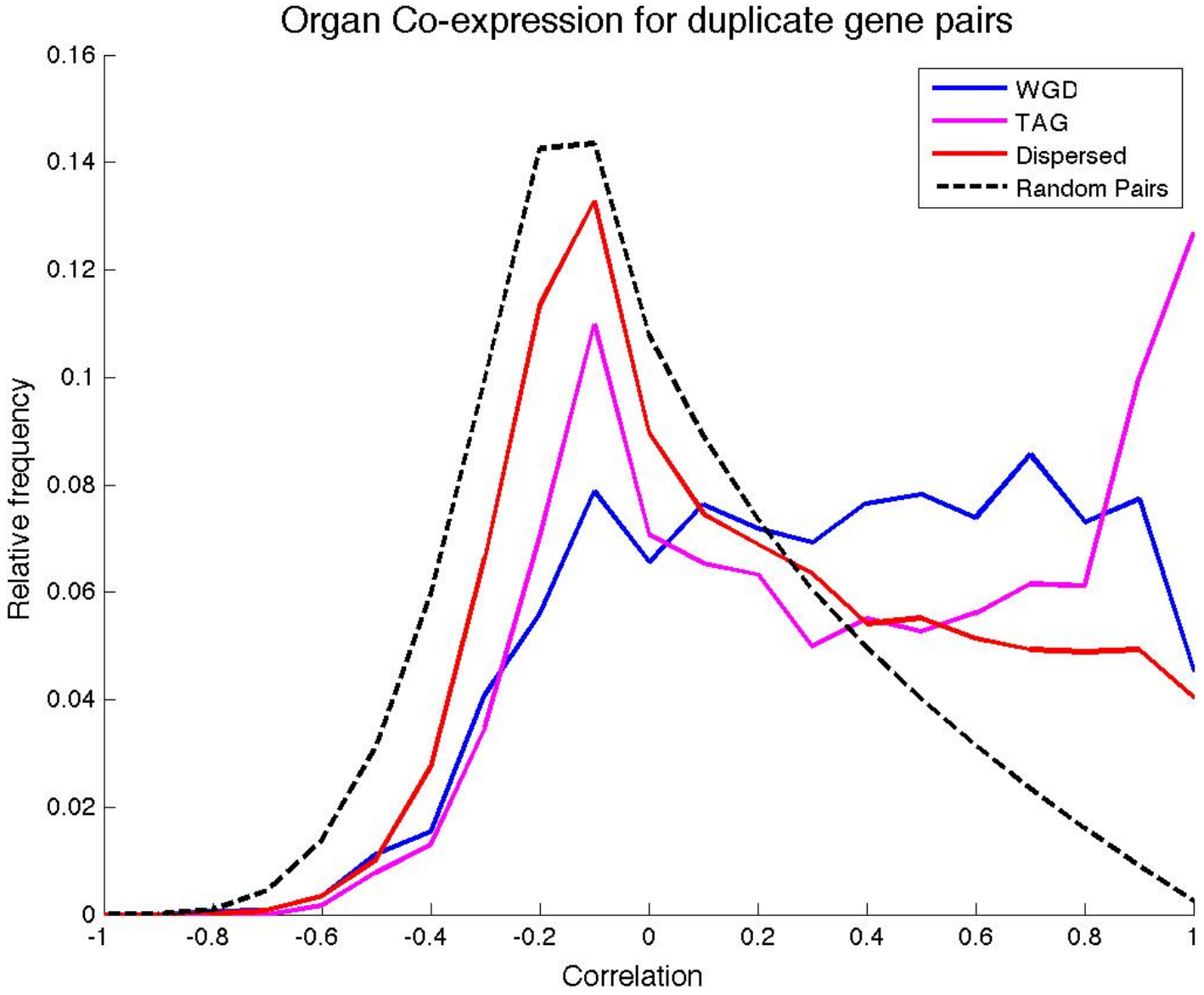
Expression Data

Using an Affymetrix array
Left with 26 601 genes with gene expression under 33 different conditions, grouped in 3 categories: **Organ, Stress and Light**.

From Kousuke Hanada, Riken Plant Centre, Japan

Gene Co-expression

using Spearman's rank correlation coefficients



Inferring Networks from Gene Expression data

BIOINFORMATICS APPLICATIONS NOTE

Vol. 25 no. 3 2009, pages 417–418
doi:10.1093/bioinformatics/btn637

Systems biology

SIMoNe: Statistical Inference for MOdular NEtworks

Julien Chiquet*, Alexander Smith, Gilles Grasseau, Catherine Matias and Christophe Ambroise

UMR CNRS 8071 Statistique et Génome, 523, place des Terrasses, F-91000 Évry, France

Received on October 23, 2008; revised on December 4, 2008; accepted on December 6, 2008

Advance Access publication December 10, 2008

Associate Editor: John Quackenbush

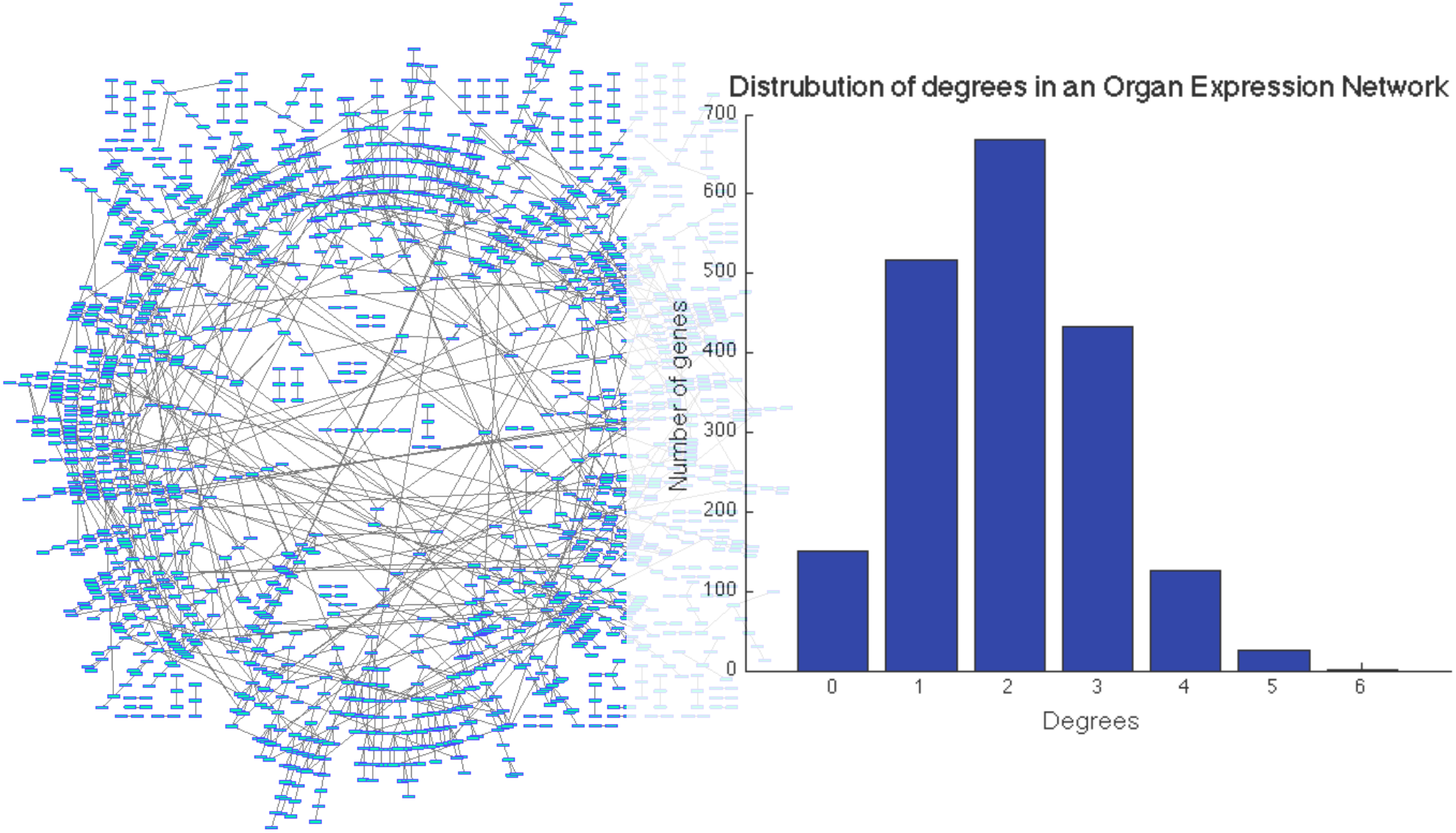
ABSTRACT

Summary: The R package SIMoNe (Statistical Inference for MOdular NEtworks) enables inference of gene-regulatory networks based on partial correlation coefficients from microarray experiments. Modelling gene expression data with a Gaussian graphical model (hereafter GGM), the algorithm estimates non-zero entries of the concentration matrix, in a sparse and possibly high-dimensional setting. Its originality lies in the fact that it searches for a latent modular structure to drive the inference procedure through adaptive penalization of the concentration matrix.

may be computed using the inverse covariance matrix (hereafter concentration matrix). Detecting non-zero entries in this matrix is in fact equivalent to reconstructing the graph of conditional dependencies.

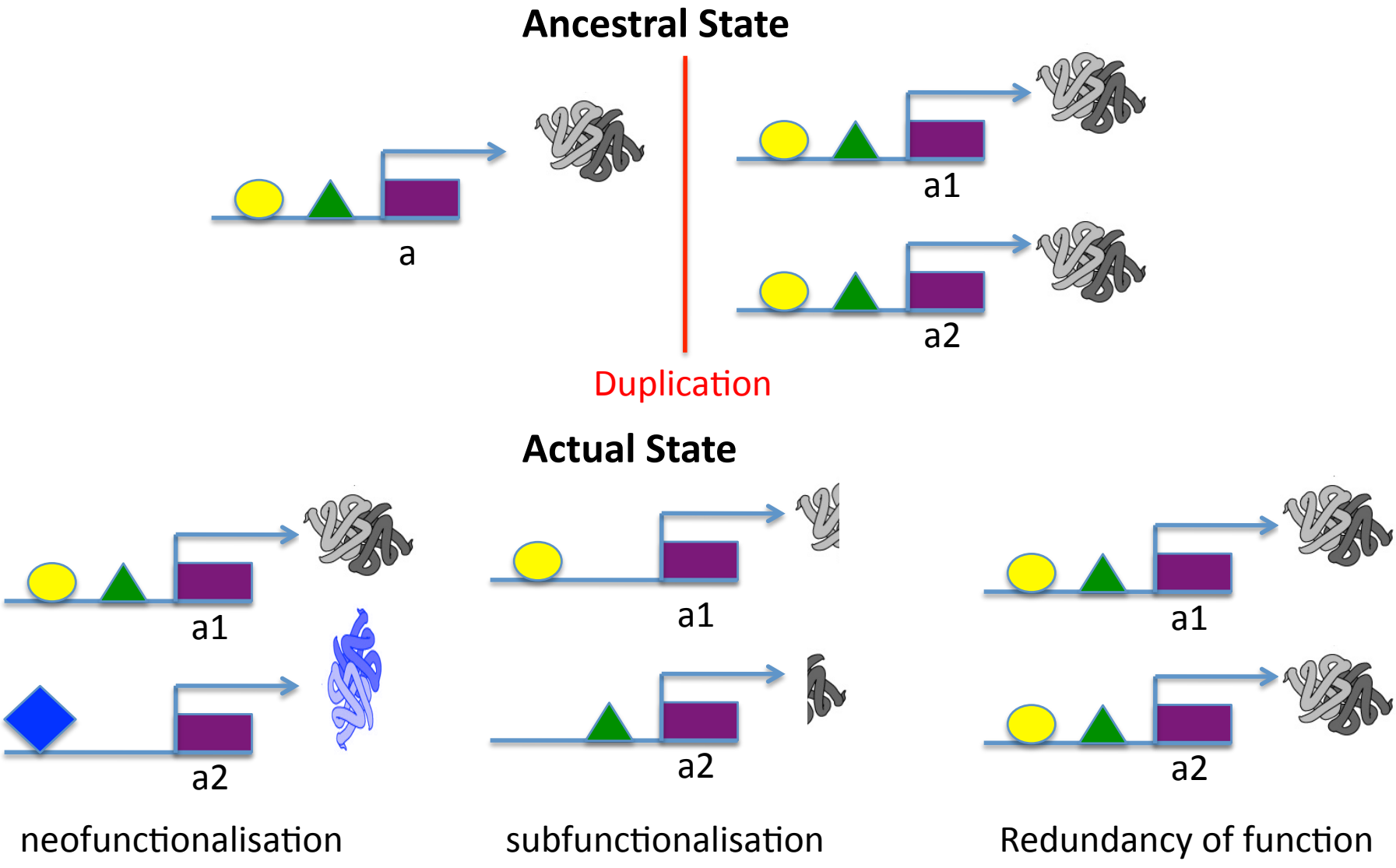
Ideally, the concentration matrix can be estimated by inverting the empirical covariance matrix. However, in the high-dimensional setting the latter is not invertible. Moreover, such a procedure does not lead to a sparse estimate, whereas biological evidence advocates for sparse networks. Sparsity means that the concentration matrix has a large number of zero entries. In this context, several estimation

The Inferred Network

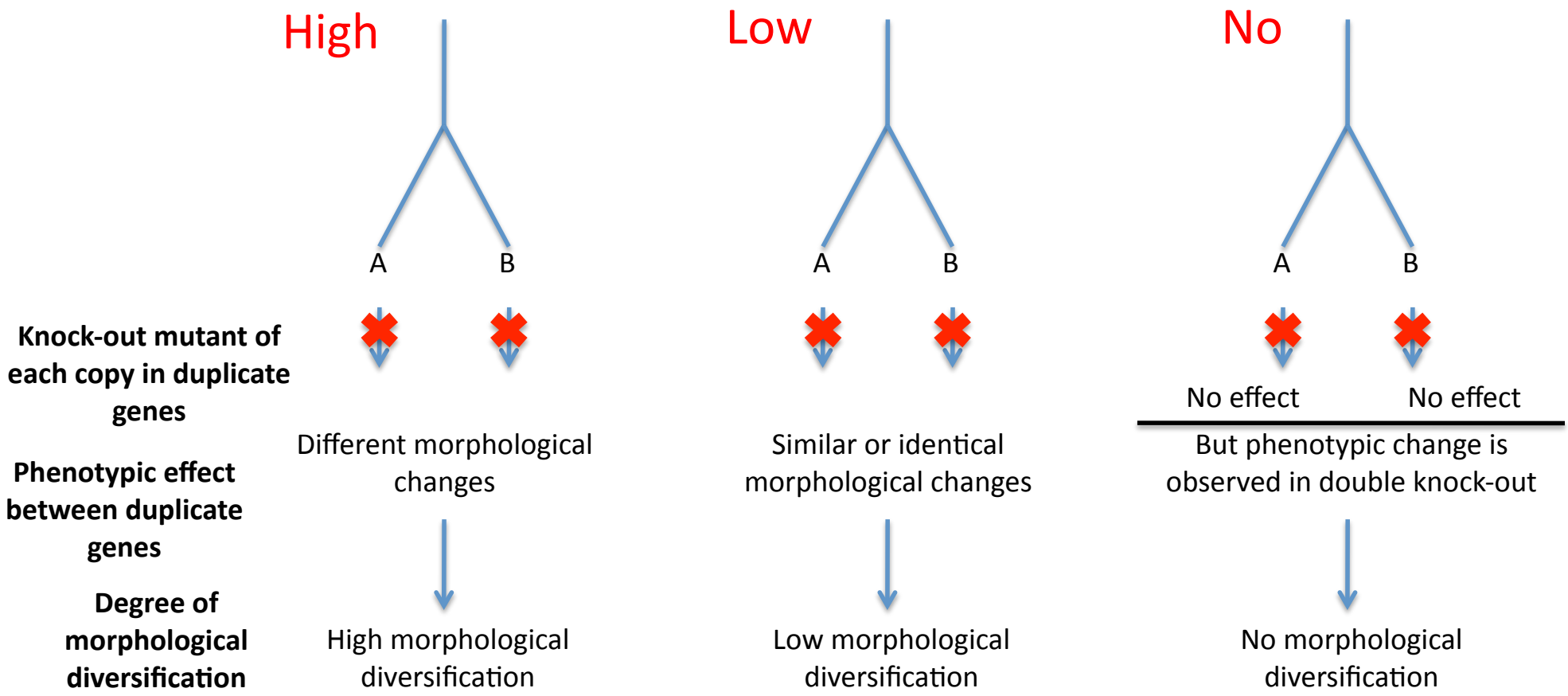


A Sparse network
Metric for measurement: shortest path between genes.

How duplicated genes are maintained



Morphological Knockout Data

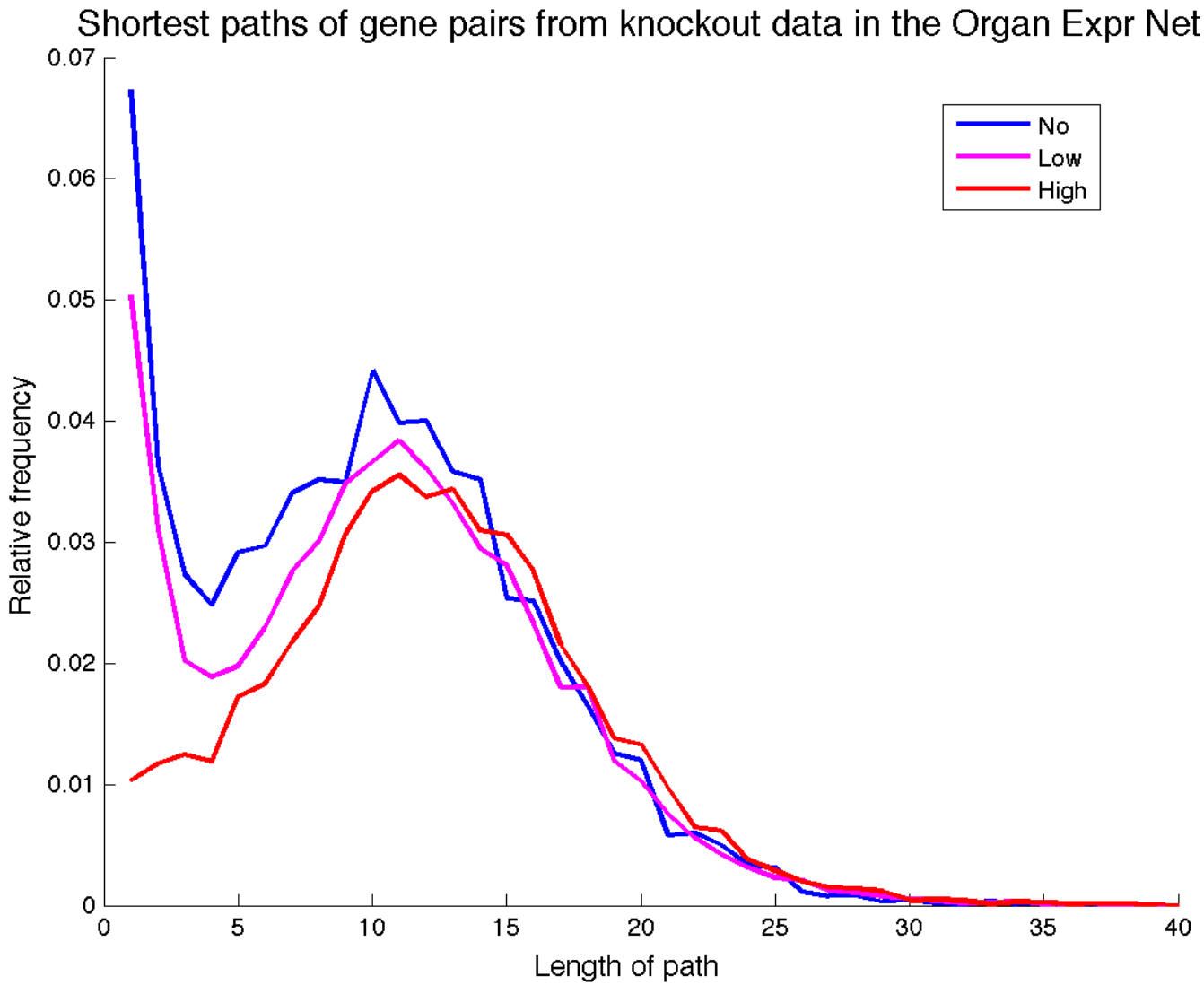


Increased expression and protein divergence in duplicate genes is associated with morphological diversification

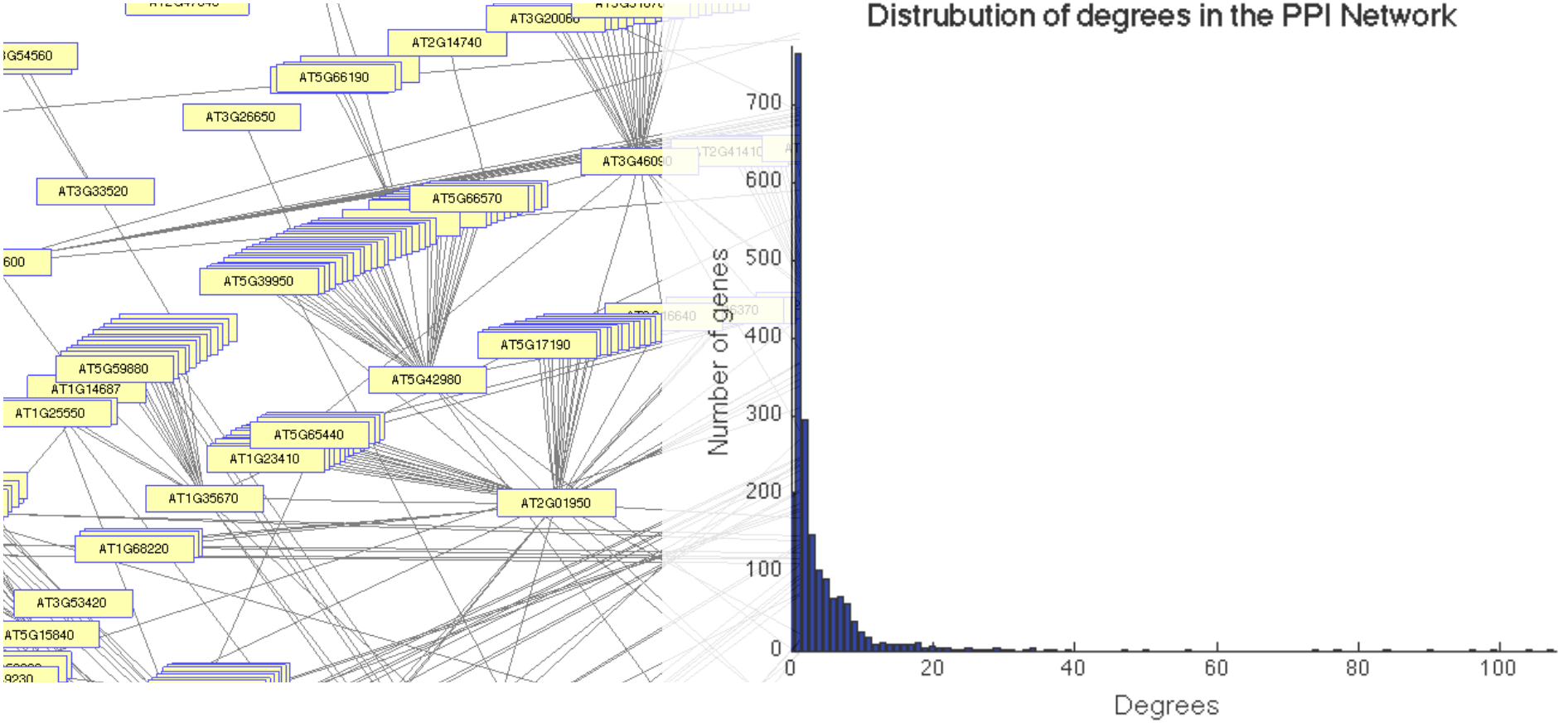
Hanada K. et al (2009) PLOS

497 gene pairs	Of which we have expression data for: 445

Network inferred for morphological knockout data



Protein-protein interaction networks



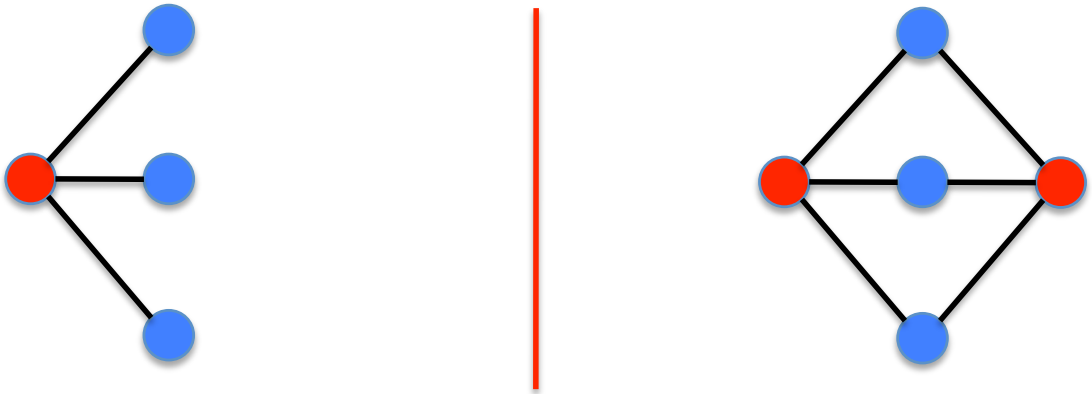
PPI Network:
Scale Free
Metric used: Jaccard index = shared_neighbours/all_neighbours

2946 genes

Of which we have expression data for: 2749

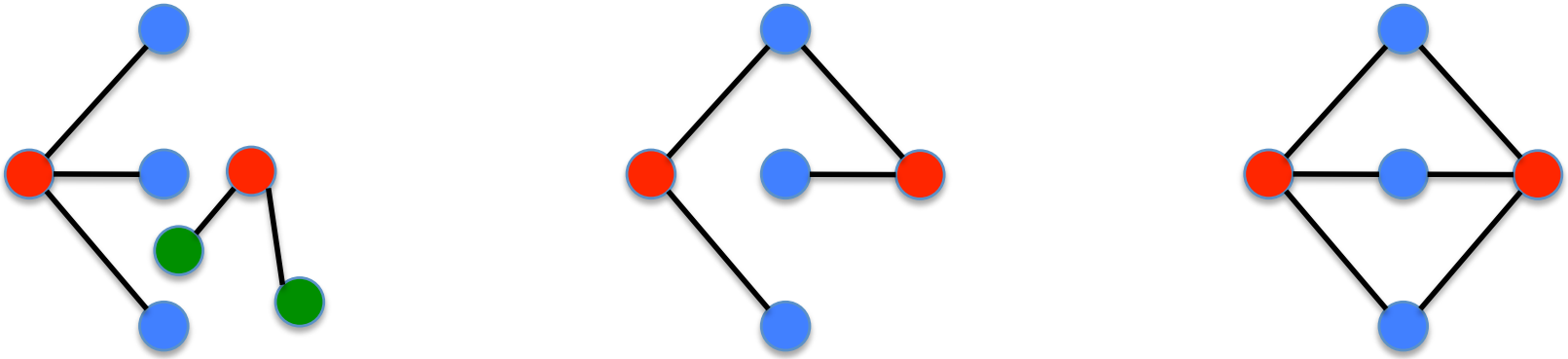
Duplication in PPI networks

Ancestral State



Duplication event

Actual State



neofunctionalisation

Jaccard = 0

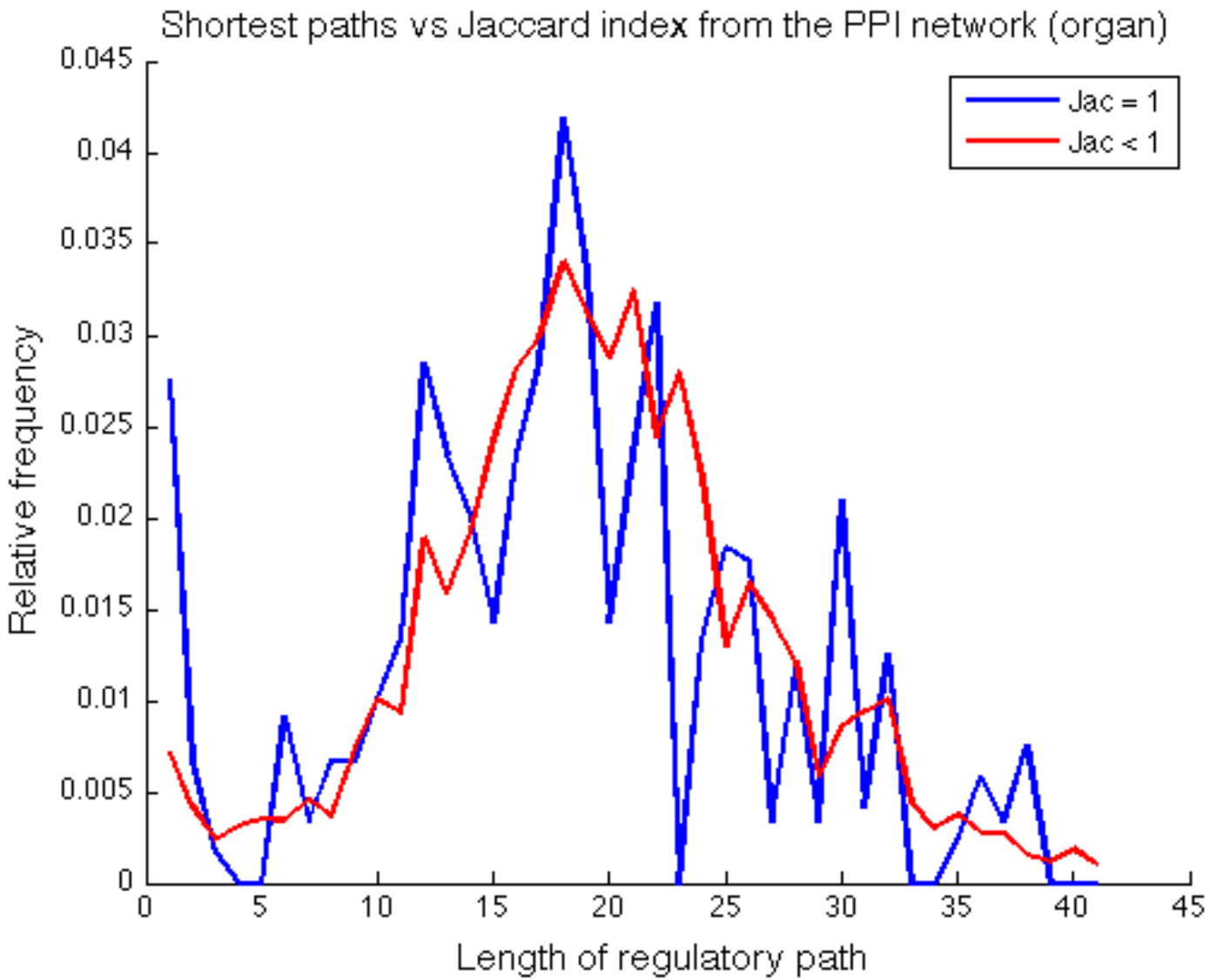
subfunctionalisation

Jaccard = 1/3

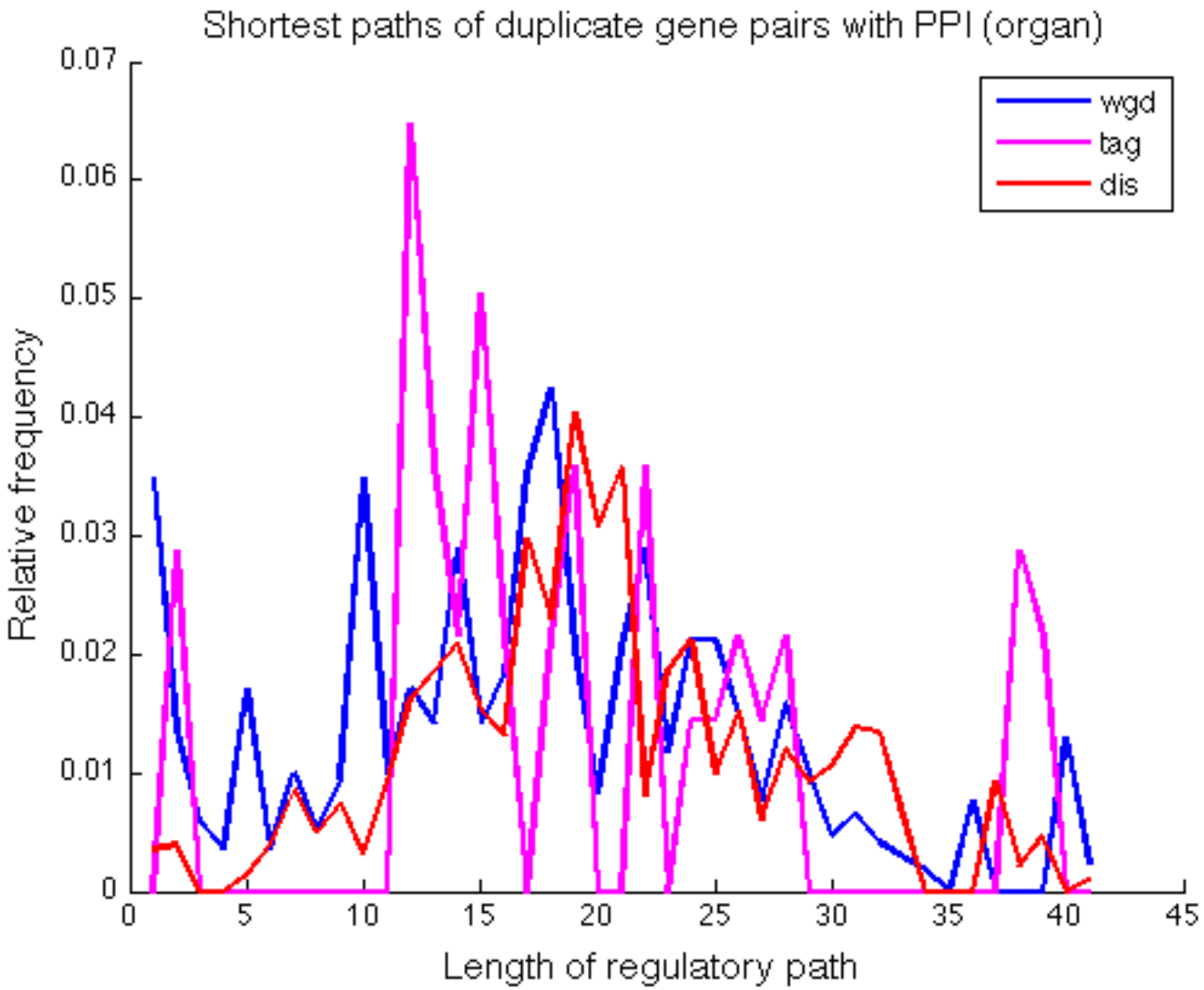
Redundancy of function

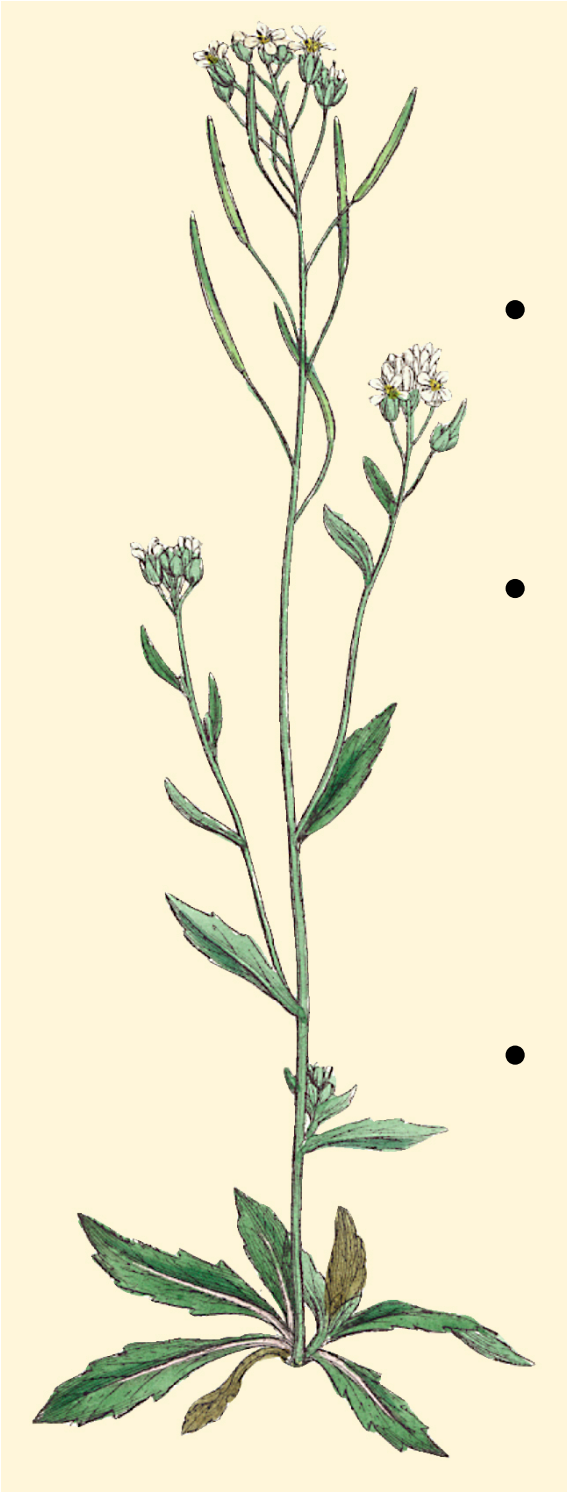
Jaccard = 1

Organ expression network analysed for Jaccard index in the PPI



Organ expression network analysed for type of duplication





Discussion

- Biological Results – Type of Duplication matters.
- Inferred networks provide us more detailed analysis then standard methods.

Further research

- How do genes within families relate to each other on the inferred network?

Arabidopsis thaliana

Thank you!

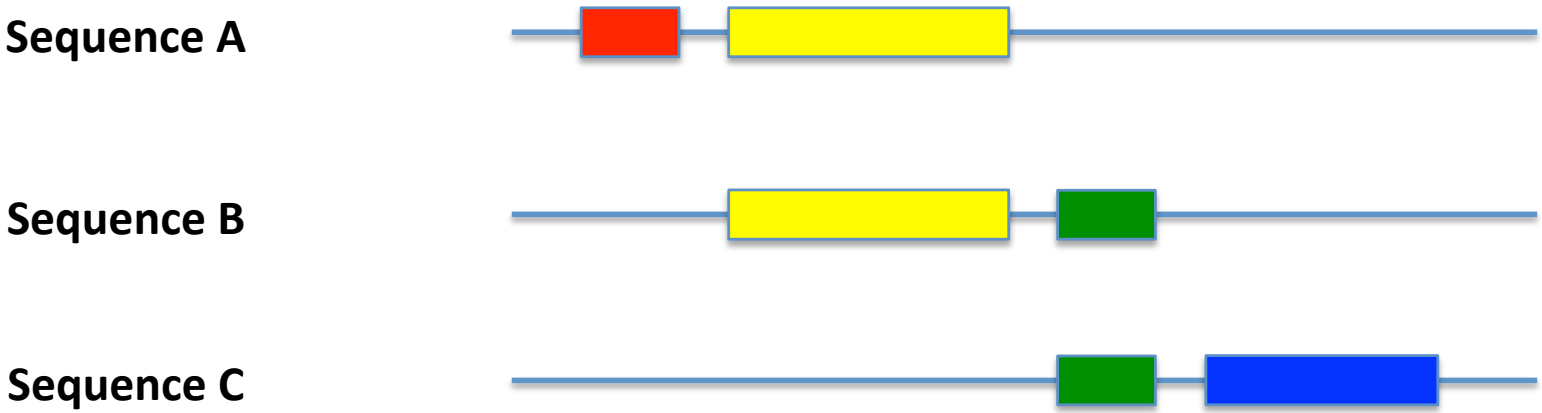
Statistique et Génome

stat.genopole.cnrs.fr

riken.jp

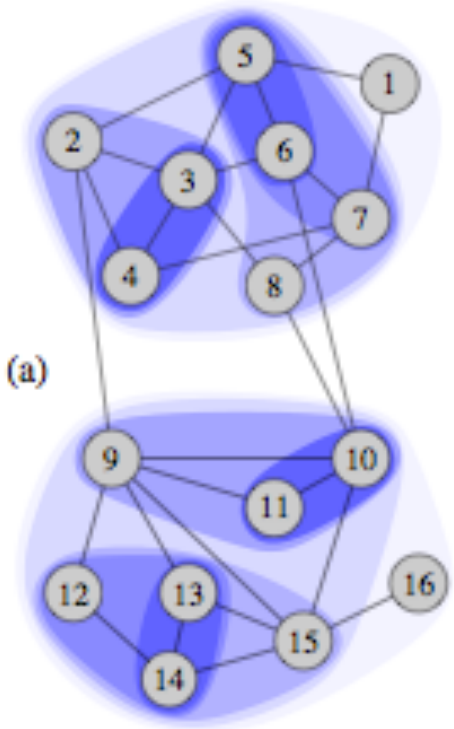
Defining Gene Families

Finding homologous genes:
BLASTP search found 978 880 gene pairs (23 386 genes)



Defining Gene Families

Clustering with the Walktrap Algorithm

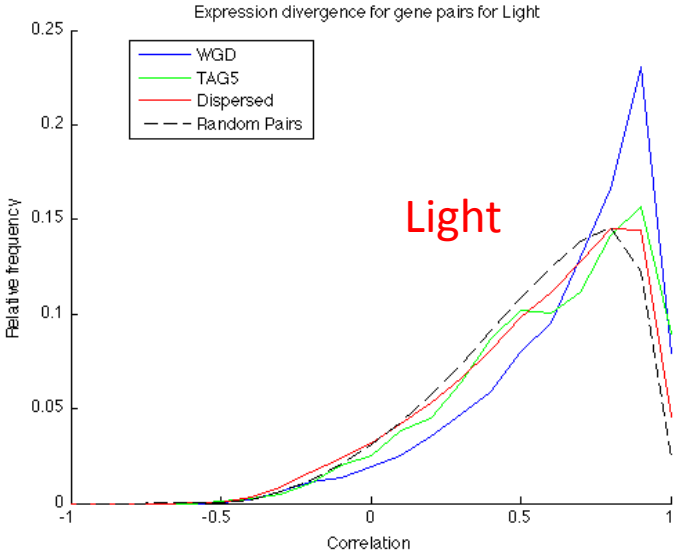


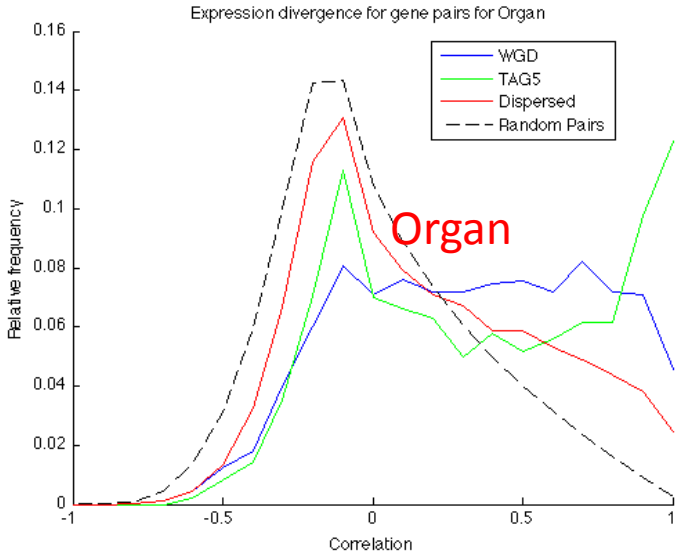
From:
Fig 1a in Computing Communities in Large Networks Using Random Walks
Pascal Pons and Matthieu Latapy
Computer and Information Sciences - ISCIS 2005
Lecture Notes in Computer Science, 2005, Volume 3733/2005, 284-293

Clustering methods compared

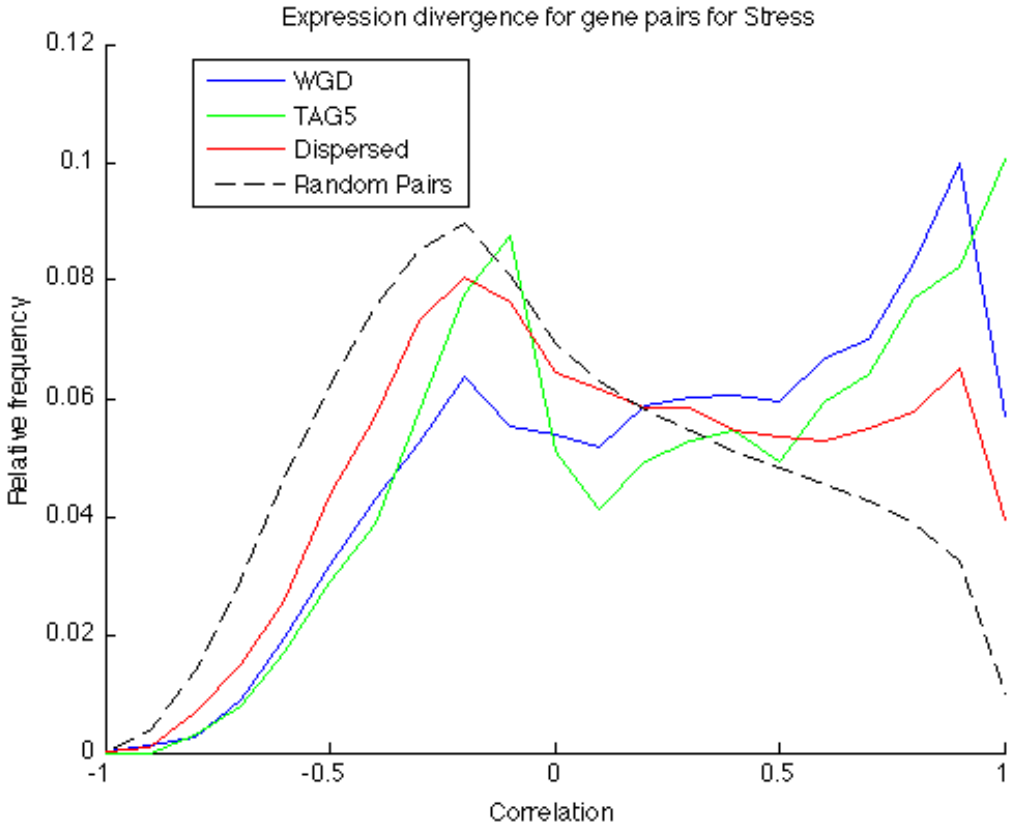
	Walktrap 30	Walktrap 50	Walktrap 70	MCL	Ensembl
No. of duplicate genes	17085	16977	16619	17693	9438
No. of families	5092	5253	5690	6598	4116
Maximum family size	61	61	48	194	12
Families that are the same as a Walktrap 30		4949	4574	798	0
Familes that contain a Walktrap 30 family		0	0	4908	1852
Families that are contained by Walktrap 30 family		302	1109	1587	277
Families that have no match with a Walktrap 30 family		2	7	1145	3271

Gene Co-expression for all duplicates

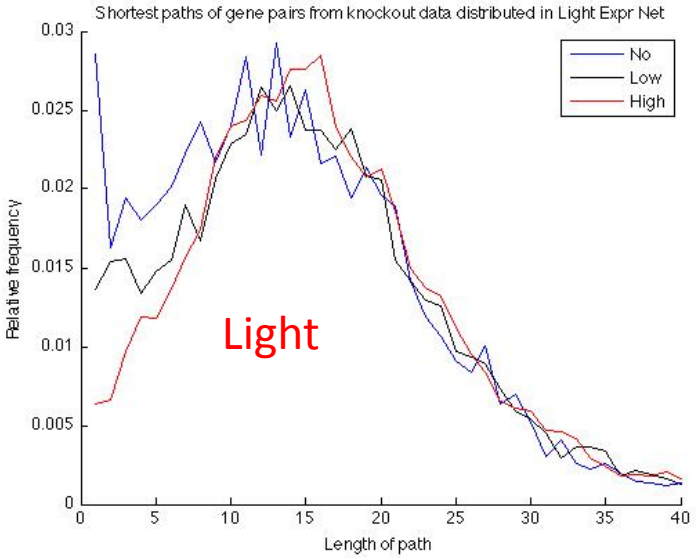


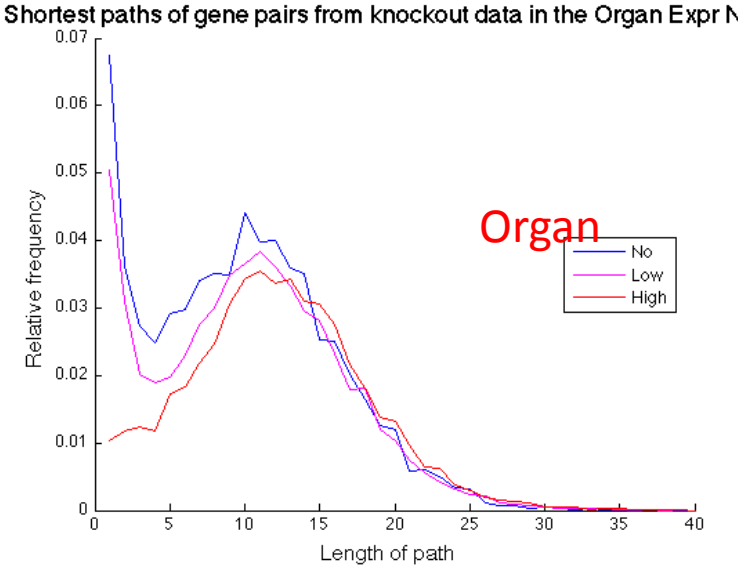


Stress

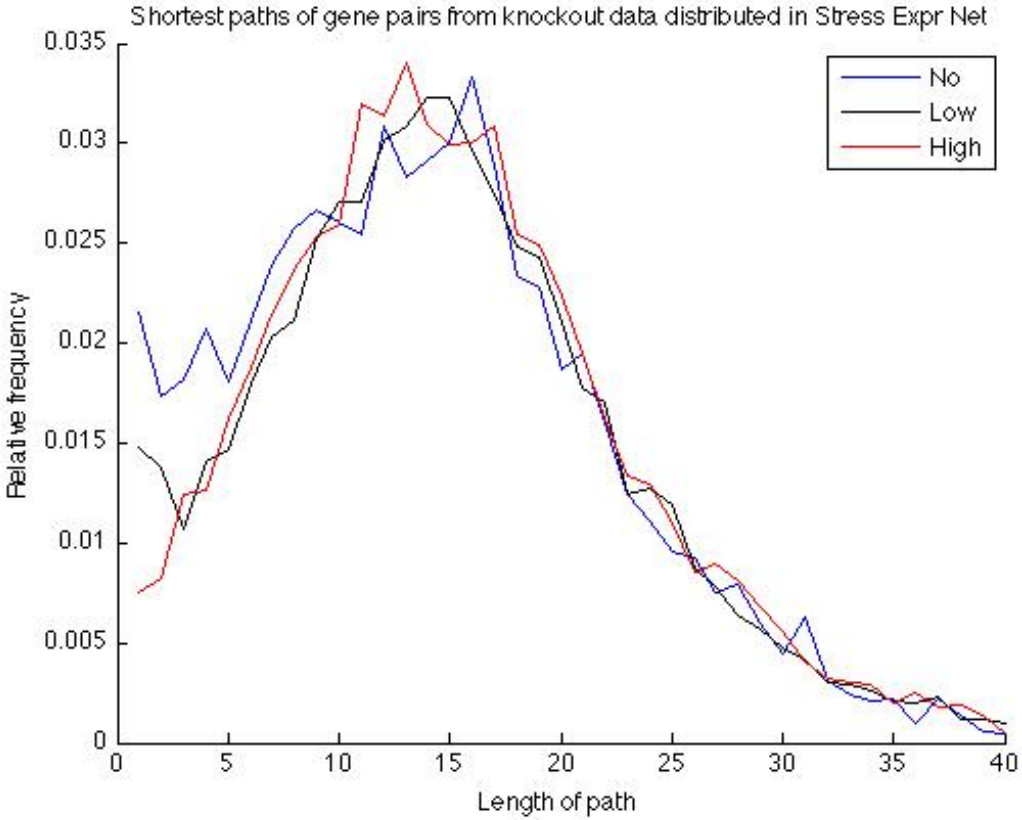


Gene Regulatory networks for Morphological Knockout data

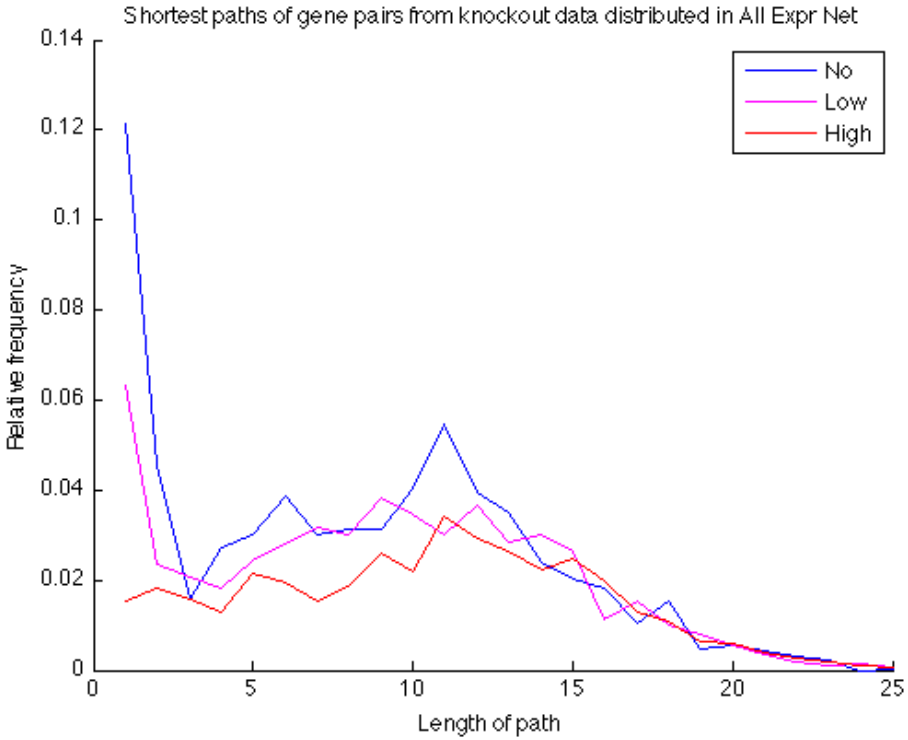


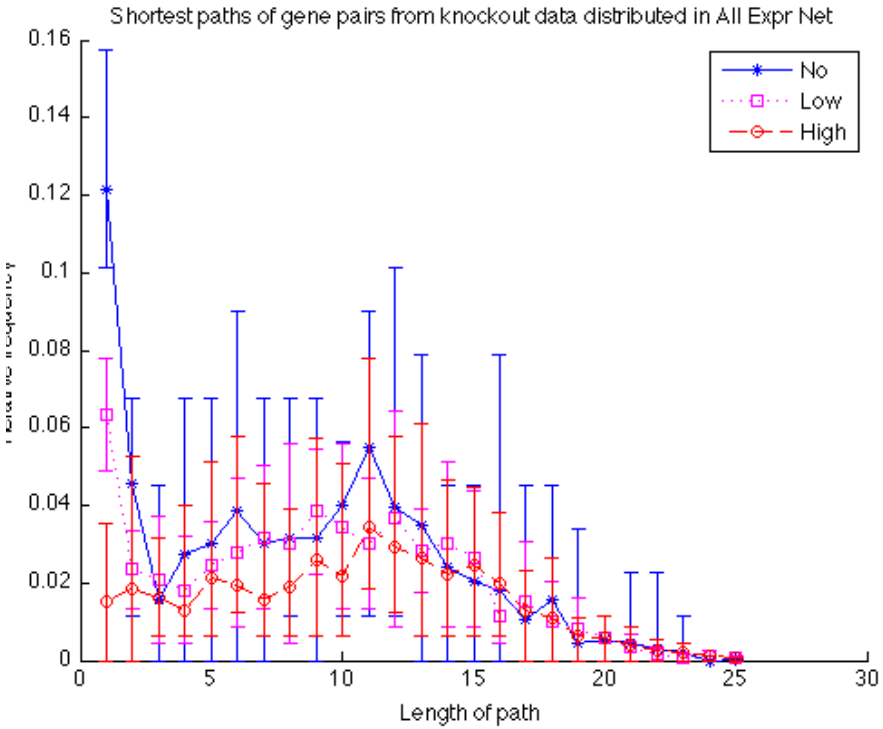


Stress

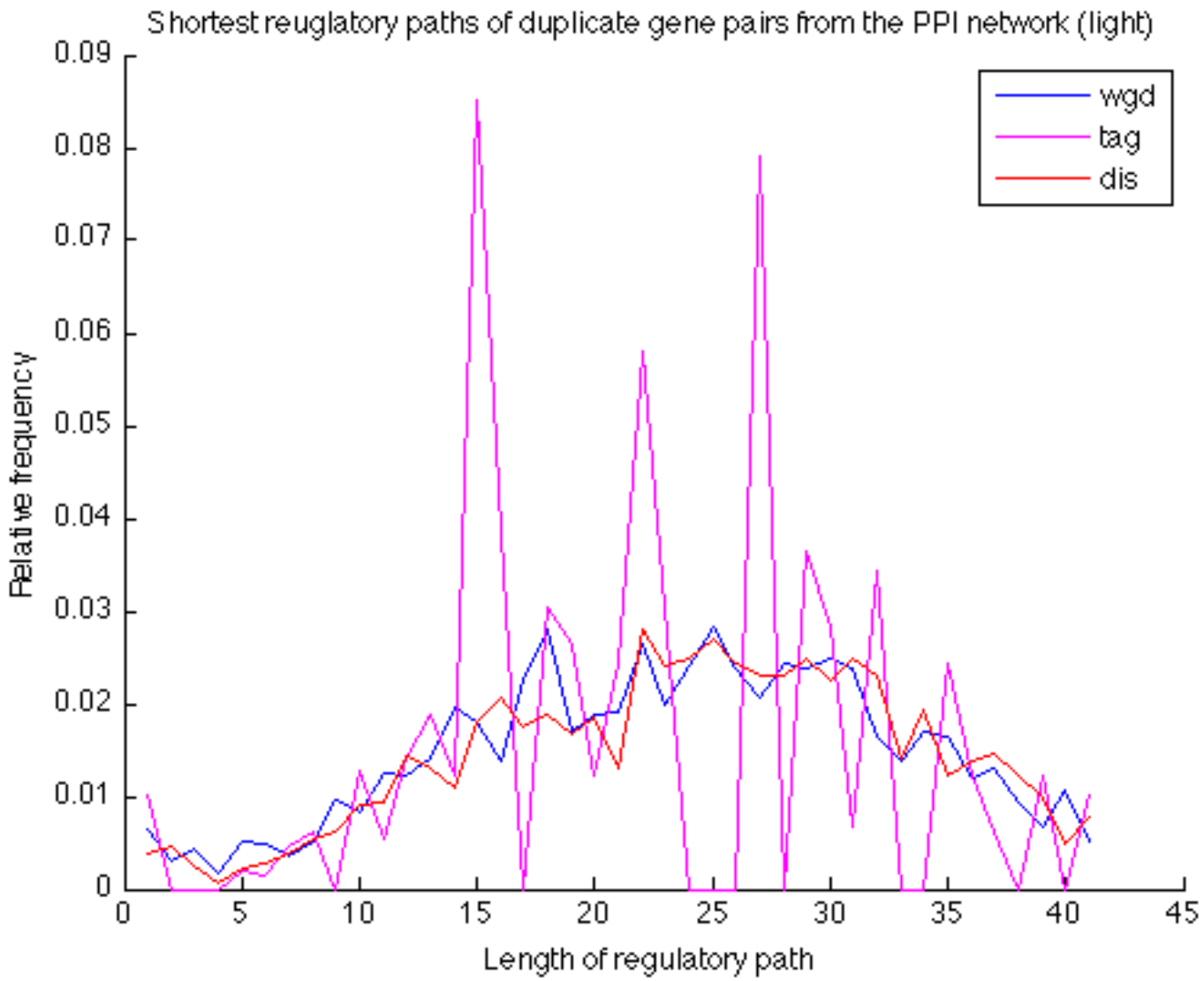


Gene Regulatory networks for Morphological Knockout data

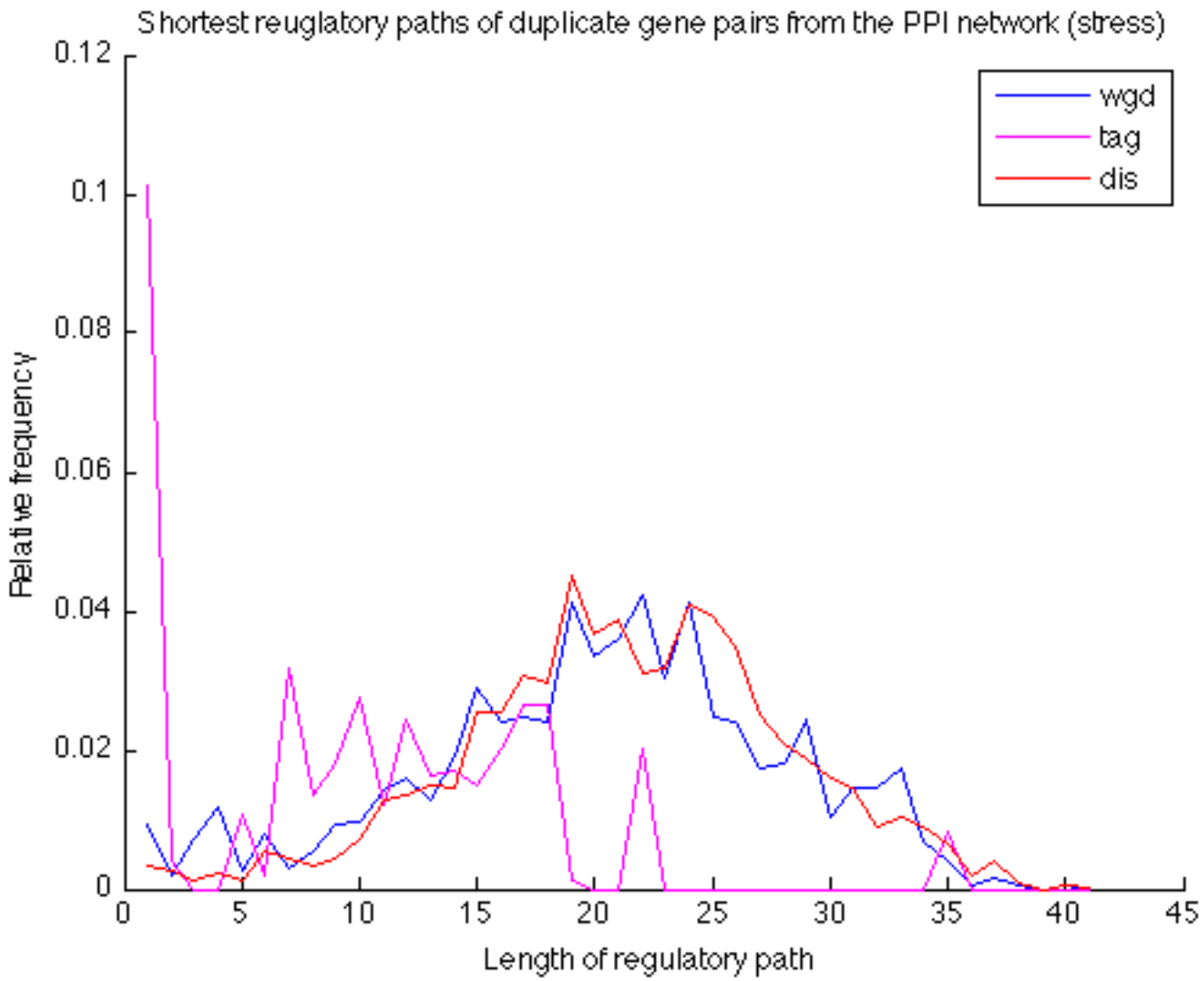




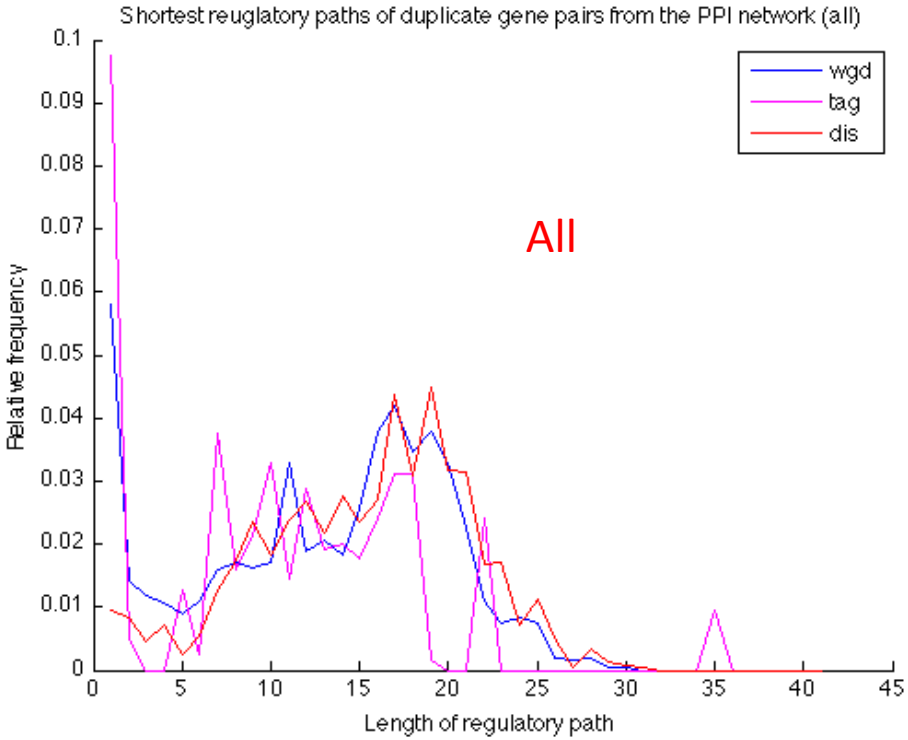
Light co-regulation for PPI genes

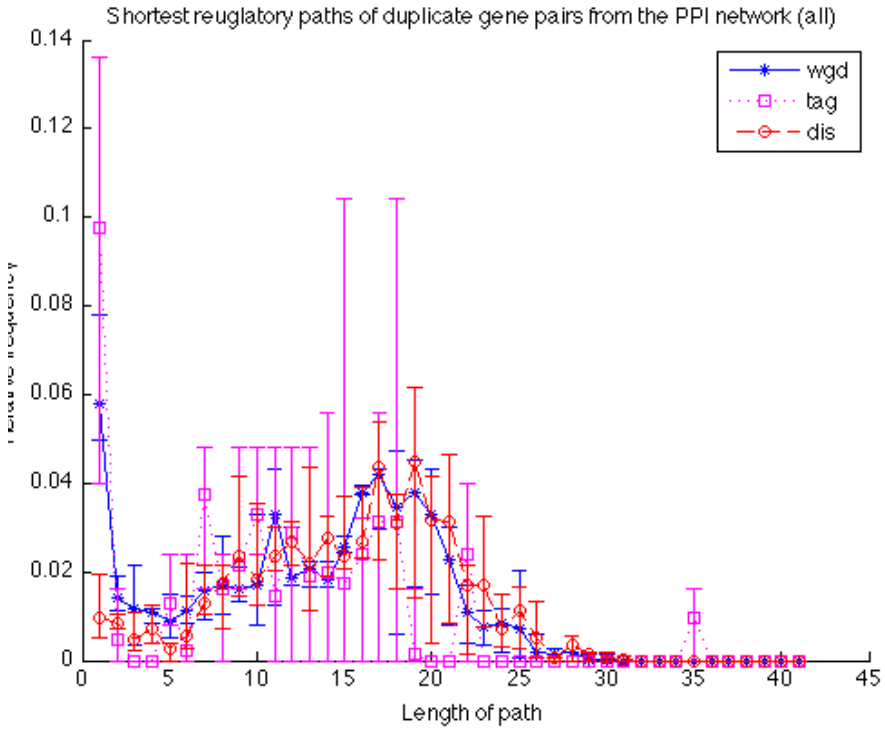


Stress co-regulation for PPI genes

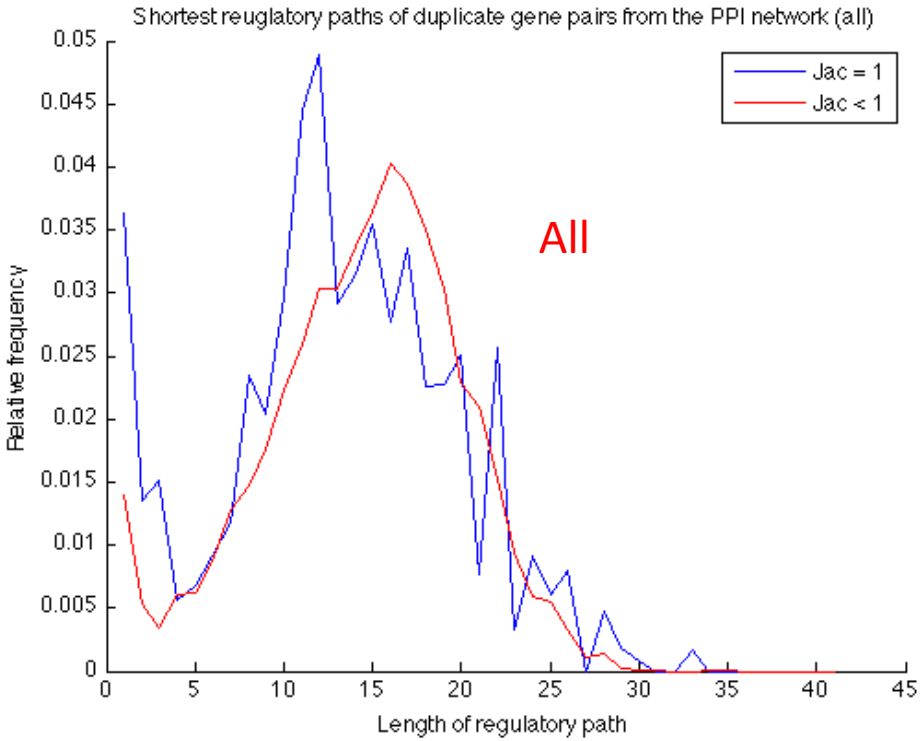


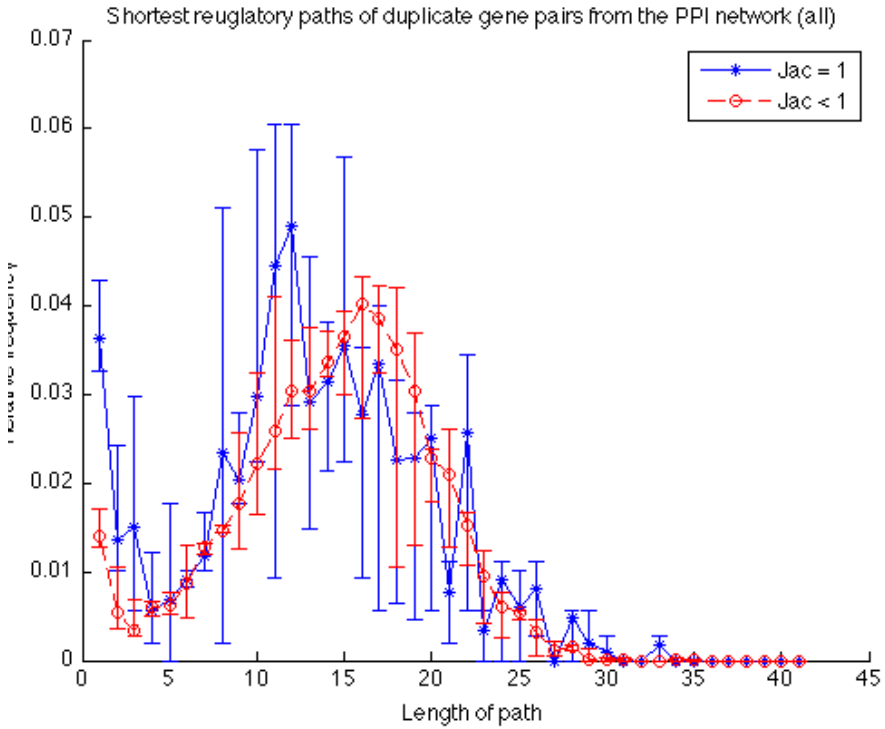
Gene Regulatory networks for duplicates in the PPI





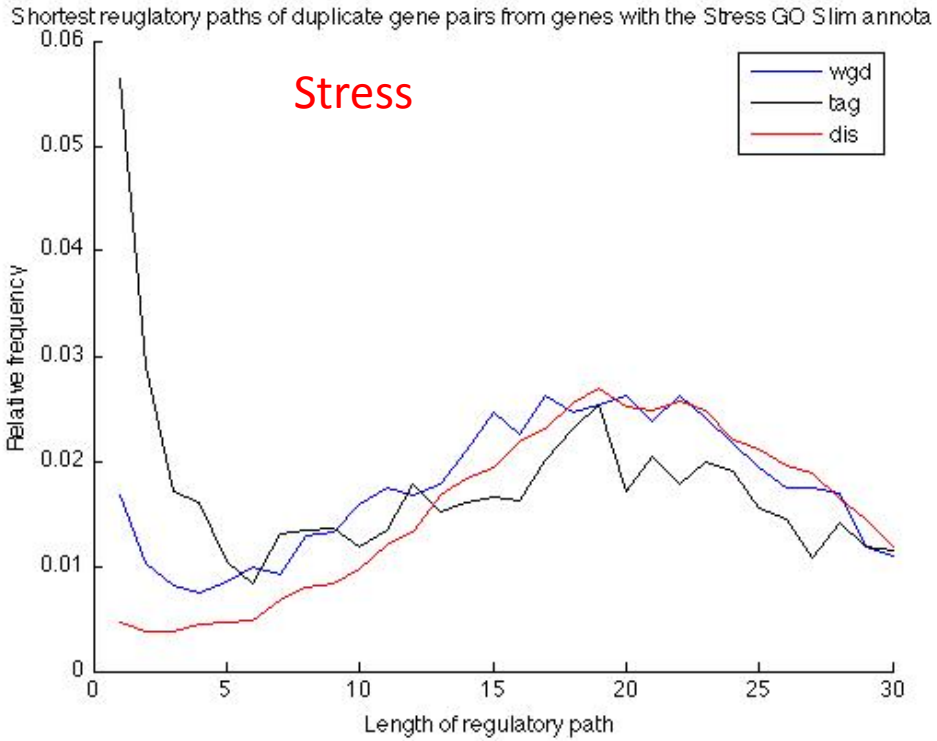
Jaccard index (PPI) vs. Shortest Path (Gene Regulatory Network)





Gene Regulatory networks found using:

Stress GO slim



Response to Light Stimuli GO annotation