Effects of metabolic network structure on genetic interactions and genome organization

Ph. D. thesis summary

Károly Kovács

Supervisor: Balázs Papp, Ph. D.

Biological Research Center of the Hungarian Academy of Sciences

Ph. D. School of Biology

University of Szeged, Faculty of Science and Informatics

Szeged, 2012

Introduction

With the recent availability of large-scale genomic and phenotypic datasets it has become possible, for the first time, to study the mapping from genotype to visible phenotypic traits in a systematic way and on an unprecedented scale. Bioinformatics analysis and integration of genome-scale datasets into large-scale mathematical models emerged as important methods to accomplish this goal. Metabolism is arguably the best characterized cellular subsystem which renders it as an excellent candidate to examine the link between genotype and phenotype. My thesis consists of two separate studies, both of which examine how the structure of the metabolic network influences the relations of the involved gene pairs. In the first part, the metabolic network is used as a tool to better understand interactions between mutations. In the second part, we investigated how natural selection acting on the performance of metabolic pathways might shape genome structure. My thesis is connected to the field of genomics by examining genome anatomy, and to systems biology by the system-level investigation of metabolic networks.

Modularity and predictability of genetic interactions in the *Saccharomyces cerevisiae* metabolic network

Our work is the first systematic, large-scale analysis of genetic interactions in a metabolic network (Szappanos et al., 2011). Genetic interactions, the non-independence of mutation effects, underlie various biological phenomena and illuminate gene functions. Despite efforts to globally map epistasis in model organisms, it remains poorly understood how genetic interactions arise from the operation of biomolecular networks.

Our work is based on the first large scale empirical dataset of genetic interactions among genes encoding metabolic enzymes, including ~185 000 double gene deletions in *Saccharomyces cerevisiae* with quantitative genetic interaction data (data produced in collaborator Charles Boone's lab). Among these gene pairs we defined double deletions that result in higher or lower fitness than expected (based on a multiplicative model) meaning positive or negative genetic interactions. Fitness was estimated based on colony size of haploid yeast strains.

Aims

The two main aims of our work were to: i) better understand how genetic interactions are related to the functional modularity of the metabolic network ii) estimate genetic interaction predictability based on genomic and metabolic network data.

More specifically, we set out to test two earlier predictions about the distribution of interactions within and between metabolic functional modules. A prior computational study based on FBA suggested that i) genetic interactions are enriched within metabolic annotation groups, and ii) interactions between different functional groups tend to be either exclusively negative or exclusively positive, a property termed 'monochromaticity' (Segre et al. 2005).

We also asked how well we can predict genetic interactions based on our knowledge of metabolic genes using a genome-scale biochemical model of the metabolic network (FBA) and a statistical/data mining method.

Methods

- Randomisation tests using Perl programming language
- Building and evaluating logistic regression and random forest models using R statistical environment (R Development Core Team, 2009), and random forest R package (Liaw and Wiener, 2002)

Results

1. Most of the genetic interactions are between functional modules. Using our large-scale genetic interaction map we found partial support for the above theoretical expectations. We report a modest but significant enrichment of both negative (1.6-fold) and positive (2.5-fold) genetic interactions within traditionally defined functional modules. However, the majority of genetic interactions occur between genes assigned to different metabolic functions (93% of negative and 90% of positive).

As an alternative to functional groups defined based on classical biochemical pathways, flux coupling provides a biochemically sound, unbiased definition of functional relatedness and has strong physiological and evolutionary relevance (Papin et al., 2004; Price et al., 2004). We used computationally identified flux-coupled gene pairs, that is, pairs of reactions where the activity of one reaction implies the activity of the other, either reciprocally or in one direction. In agreement with results obtained using annotation groups, although we find that both negative (1.4-fold) and positive (2.3-fold) interactions are significantly enriched in flux-coupled pairs, the overwhelming majority (> 97%) of both forms of interactions occur between uncoupled genes. In conclusion, both definitions of functional relatedness reveal that most genetic interactions connect across distinct functional modules.

2. Genetic interactions between modules are monochromatic, but only to a limited degree. We asked whether interactions between different functional groups tend to be either exclusively negative or positive. In agreement with earlier theoretical predictions, we found a statistically significant excess of monochromaticity among pairs of functional groups in the real data compared to randomized interaction maps. Nevertheless, monochromaticity in our genetic interaction map is modest: only ~24−34% more monochromatic pairs were found than expected by chance.

3. Most genetic interactions cannot be predicted either based on gene pair characteristics or biochemical modelling. We estimatedhow well we can predict genetic interactions based on our knowledge of metabolic genes. We assessed the predictive power of two computational approaches: a genome-scale biochemical model of the metabolic network (FBA) which computes the growth of single and double deletant strains and a statistical/data mining method. In the second approach we compiled a dataset of gene-pair characteristics (e.g. coexpression), following earlier studies (Wong et al., 2004; Ulitsky et al., 2009) and metabolic network features (e.g. shortest path of reactions) but omitting any information on genetic interactions. We used a classical statistical method (logistic regression) and a new data-mining method (random forest (Breiman, 2001)) to classify genetic interactions based on these features.

Using the biochemical model we can predict negative and positive interactions up to 50% and 11% rates of true predicted interactions (precision), respectively. Although this confirms that the highest predicted interaction scores have high physiological

relevance, we find that only a minority of empirical interactions are captured by the model (2.8% and 12.9% for negative and positive interactions, respectively).

The statistical approach using genomic and metabolic network data gives better predictions. Although an increased fraction of *in vivo* interactions could be retrieved, ~70% of negative and ~75% of positive interactions were still predicted with very low (<10%) precision. Notably, incorporating fitness and genetic interaction scores derived from the biochemical model into statistical models boosts the precision of negative interaction predictions, indicating that biochemical modeling provides unique information that is not captured by purely statistical data integration. We conclude that the majority of genetic interactions are not well understood either in terms of biochemical processes or statistical associations.

Colinearity of gene order in *Escherichia coli* metabolic operons

It is well established that gene order in prokaryotic genomes is not random. This is most evident when looking at operons, these often encode enzymes involved in the same metabolic pathway or proteins from the same complex. However, it is almost completely unexplored whether gene order within operons is governed by chance or could have any functional significance.

Aims

1. To empirically test whether gene order within operons reflects the functional order of the encoded enzymes in a metabolic pathway (colinearity) in *E. coli*.

2. To build general mathematical models of operon expression coupled to a linear metabolic pathway with four enzymes, in order to gain insight into the potential interplays between gene oder and the flux of a metabolic pathway.

3. To test and make predictions for three different adaptive scenarios of colinearity by model simulations.

4. To empirically test the predictions of three competing hypotheses for the origin of colinearity.

Methods

- Dataset compilation using EcoCyc database (Keseler et al., 2009) and published microarray screens
- Randomisation tests using Perl programming language

Results

1. There is a significant trend for colinearity in *E.coli* metabolic operons. Approximately 60% of the intra-operonic gene pairs show this pattern, compared to 50% expected if gene order was random.

There is no known mutational bias which can result in colinearity thus we looked for adaptive scenarios. We argued that colinearity might have a fitness advantage (i.e. increased growth rate) by increasing pathway productivity. At first sight, colinearity is unexpected as gene order should not affect the steady-state pathway productivity. Indeed, this is confirmed by our mathematical model. We considered three extensions of the steady-state model that could potentially account for colinearity.

2. Polarity cannot explain colinearity.

Colinearity in polar operons can increase steady-state pathway flux, where polarity refers to a decreasing mRNA abundance profile along the operon. The hypothesis is based on the theoretical finding that decreasing enzyme concentrations along the path can increase the flux along the pathway when the total enzyme concentration is fixed (Heinrich & Klipp, 1996). Thus, in a polar operon colinear arrangement can increase steady-state flux. This theoretical prediction is corroborated by our simulations and it predicts that we should observe colinearity in polar operons only. However, in contrast to this prediction, we failed to find an enrichment of colinearity in polar operons in *E.coli*.

3. There is no link between colinearity and expression variability.

The second hypothesis explains colinearity by faster metabolic processing immediately after up-regulation upon environmental change. According to experimental measurements there is a time delay between the expression of two consecutive genes in an operon (Alpers & Tomkins, 1965, 1966). Thus right after activating the operon, the end-product can appear faster if the gene order is colinear. This verbal argument is confirmed by our model. This hypothesis predicts that operons showing high expression variation across conditions should more often display colinearity compared to constitutively expressed operons. However, using publicly available gene expression data we failed to find support for this hypothesis.

3. Empirical data supports „stochastic stalling" hypothesis.

Small numbers of molecules are frequently involved in the process of gene expression and could lead to significant stochasticity in protein abundance (Elowitz et al., 2002). Whereas enzymes encoded in a highly expressed operon are likely to be always present in the cell whenever the operon is induced, stochasticity might play an important role in weakly expressed operons as enzymes could either decay or be diluted by cell division between two expression episodes (Cai et al. 2006), hence recurrently stalling metabolism. Colinearity could minimize the effect of such stochastic enzyme losses by speeding up the reinitiation of stalled metabolic transformations, in a similar manner as it provides a transient advantage after up-regulation of an inactive pathway. This argument is also supported by our model. This hypothesis specifically predicts that colinearity should be restricted to lowly expressed operons. Indeed, we found that only genes with low mRNA abundance show colinearity, hence supporting the „stochastic stalling" hypothesis.

To sum up, our work is the first reporting a non-random pattern of operonic gene order: in lowly expressed metabolic operons gene order reflects the functional order of the encoded enzymes (Kovács et al., 2009). Empirical tests of different adaptive scenarios for colinearity in *E. coli* supports the hypothesis that the advantage of colinearity is to minimize metabolic stalling owing to stochastic protein loss.

References

Alpers, D.H., and Tomkins, G.M. (1965). The order of induction and deinduction of the enzymes of the lactose operon in E. coli. Proc. Natl. Acad. Sci. U.S.A *53*, 797–802.

Breiman, L. (2001). Random forests. Machine Learning *45*, 5–32.

Cai, L., Friedman, N., and Xie, X.S. (2006). Stochastic protein expression in individual cells at the single molecule level. Nature *440*, 358–362.

Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic Gene Expression in a Single Cell. Science *297*, 1183–1186.

Heinrich, R., and Klipp, E. (1996). Control Analysis of Unbranched Enzymatic Chains in States of Maximal Activity. Journal of Theoretical Biology *182*, 243–252.

Kovács, K., Hurst, L.D., and Papp, B. (2009). Stochasticity in Protein Levels Drives Colinearity of Gene Order in Metabolic Operons of Escherichia coli. PLoS Biol *7*, e1000115.

Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest. Resampling Methods in R: The Boot Package 18.

Papin, J.A., Stelling, J., Price, N.D., Klamt, S., Schuster, S., and Palsson, B.O. (2004). Comparison of network-based pathway analysis methods. Trends in Biotechnology *22*, 400–405.

Price, N.D., Reed, J.L., and Palsson, B.Ø. (2004). Genome-scale models of microbial cells: evaluating the consequences of constraints. Nature Reviews Microbiology *2*, 886–897.

R Development Core Team (2009). R: A language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing).

Segre, D., DeLuna, A., Church, G.M., and Kishony, R. (2005). Modular epistasis in yeast metabolism. Nat Genet *37*, 77–83.

Szappanos, B., Kovács, K., Szamecz, B., Honti, F., Costanzo, M., Baryshnikova, A., Gelius-Dietrich, G., Lercher, M.J., Jelasity, M., Myers, C.L., et al. (2011). An integrated approach to characterize genetic interaction networks in yeast metabolism. Nature Genetics *43*, 656–662.

Ulitsky, I., Krogan, N.J., and Shamir, R. (2009). Towards accurate imputation of quantitative genetic interactions. Genome Biol *10*, R140.

Wong, S.L., Zhang, L.V., Tong, A.H.Y., Li, Z., Goldberg, D.S., King, O.D., Lesage, G., Vidal, M., Andrews, B., Bussey, H., et al. (2004). Combining biological networks to predict genetic interactions. Proceedings of the National Academy of Sciences of the United States of America *101*, 15682–15687.

List of publications

Fehér, T. *, Bogos, B. *, Méhi, O. *, Fekete, G., Csörgő, B., **Kovács, K**., Pósfai, G., Papp, B., Hurst, L.D., Pál, C. (2012) Competition between Transposable Elements and Mutator Genes in Bacteria *Mol Biol Evol 29: 3153*

IF: 5,550

Szappanos, B. *, **Kovács, K. ***, Szamecz, B., Honti, F., Costanzo, F., Baryshnikova, A., Gelius-Dietrich, G., Lercher, M.J., Jelasity, M., Myers, C.L., Andrews, B.J., Boone, C., Oliver, S.G., Pál, C., Papp, B. (2011) An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nature Genetics 43: 656*

IF: 35,532

Kovács, K.*, Hurst, L.D., Papp, B. (2009) Stochasticity in Protein Levels Drives Colinearity of Gene Order in Metabolic Operons of Escherichia coli. *PloS Biol. 7: e1000115.*

IF: 12,916

* first author