Complexity from Gene Duplication: A Network-based approach Project Proposal

Zoran Nikoloski[1,*], Joachim Selbig[1,2]

[1] Institute for Biochemistry and Biology, University of Potsdam
[2] Max-Planck Institute for Molecular Plant Physiology
Potsdam, Brandenburg, Germany
[*]E-mail: nikoloski@mpimp-golm.mpg.de

1 Introduction

Since the seminal work of Ohno [19], gene duplication—the process of duplicating a DNA segment that contains a gene—has been recognized as the impetus for creating new genes and as an important source of evolutionary innovation and adaptation. Gene duplication may arise primarily by two distinct mechanisms [12]. In the first, *tandem* or *segmental duplication*, unequal recombination occurs between homologous sequences at two places in the genome. In the second, *retroposition*, a transcribed mRNA sequence is reverse-transcribed, and the resultant DNA is inserted into a chromosome. In contrast to retroposition, tandem duplications may be of sequences containing many genes and intergenic regions. Finally, errors in segregation during meiosis can also result in extra copies of entire chromosomes or duplication of the entire genome. The impact of these processes seems to be extraordinary; for instance, it has been shown that 50% of prokaryotic genes [5, 24] and over 30-65% of eukaryotic genes [29] are products of gene duplication.

There are already several computational approaches to reveal the number and type of duplications which are based on identifying segments with a significant number of homologous genes or gene pairs [7]. Once a library of such segments has been obtained, one may investigate the possibility for existence of simple mechanisms by which homologous genes had been created.

The fate of the paralogs with respect to their functional characterization includes four possible scenarios: (1) retention of the ancestral function, (2) migration to assume distinct functions through (2.a) neofunctionalization (NEO-F), (2.b) subfunctionalization (SUB-F), also known as biological division of labor, (2.c) escape from adaptive conflict (EAC), also known as

accumulation of labor and (3) non-functionalization (NON-F), contributing to loss of genes.

It has already been observed that coordinated migration of duplicated genes is a rare event, as paralogs diversify more frequently at the level of regulation, less frequently through changes in their cellular component, biological process and molecular interactions, and rarely in biochemical function [26]. Thus, regarding the gene duplication process and the functional characterization of genes, three issues emerge as prominent: gene fission/fusion, alternative splicing, transcriptional regulation, and protein interaction.

Gene fission and fusion, the process by which a single gene is split into two separate genes and two adjacent genes are fused into a single gene, respectively, are among the primary processes that generate new genes [22]. Recently, gene fission has been directly related to gene duplication with subsequent partial NON-F [25].

Besides gene duplication, alternative splicing may serve as a variant for enhancing protein diversity in eukaryotes [14]. The origins of alternative splicing and its advantages as opposed to gene duplication as a mechanism of generating protein diversity remain elusive [2]. Recent studies reveal that there is a trade-off between alternative splicing and gene duplication [16, 11]. After a gene duplication event, the duplicates either lose the splice variants or the singletons acquire them. Moreover, alternative splicing may occur through exon duplication, resulting in more exons in the singletons that may serve as internal paralogs. These findings allude to a possible interplay between alternative splicing and gene duplication that may reveal some intersting evolutionary mechanisms.

To address the biological complexity from gene duplication, one may also focus on the dependence of transcription factor regulation and gene duplication. It is already known that the complexity of an organism correlates with an increase in both the ratio and absolute number of transcription factors [23]. After duplication, the gene may undergo changes of its coding sequence, which results in different proteins (and, therefore, different functions), or its upstream region, which renders possible the recognition by different transcription factors. Several evolutionary scenarios are possible if the ancestor and its copy both remain in the transcriptional gene class [8].

Project objectives Combining the insights about gene duplication events with gene functional characterization could enable the investigation of evolutionary processes in unprecedented detail. Mathematical studies of growing networks offer a possibility for developing models that could readily lend

themselves as means for testing evolutionary hypotheses. The goal of this project is three-fold:

1. Devise mathematical models of gene duplications at various levels of detail which accurately match the properties of empirical networks for gene duplication,

2. Provide a unifying framework for ranking of models in terms of their ability to mimic the empirical findings,

3. Develop and implement a method for parameter ranking which can be used for establishing the importance of each mechanism involved in the duplication process.

Providing an accurate model of gene duplication, as an underlying mechanism for protein diversification, may in fact provide new insights in the evolution of (superimposed) processes, such as: gene fission/fusion, alternative splicing, transcriptional regulation, and protein interaction.

2 Proposed Approach

As more genomes become available and their annotations are improved, the global view of the gene duplication process and the insights into its impact on gene fission/fusion, regulation, and splicing become hampered by the complexity and size of the possible relations that may arise on the system level representation. The abstraction of all of these processes into networks offers a fruitful approach, the main goal of which is to relate the structure of the network to the biological function. Our approach to the study of gene duplication comprises three steps.

Step I. For a set of given genomes, one may infer the evolutionary relationships (*i.e.*, "is ancestor of" relation) of orthology and paralogy by using two existing algorithms, EvolMAP [20] and SYNERGY [26]. According to Fitch [9], orthologs are genes that share a common ancestor at a speciation event, while paralogs are related through duplication events. These are not simple one-to-one relationships, as two paralogs are both orthologous to the same gene in another species or can result in a one- or many-to-none orthologous relationships when genes are lost in a particular species or lineage. Moreover, since gene fission/fusion and unresolved ambiguities in the computational approaches, yielding the ancestor relation, may produce more than

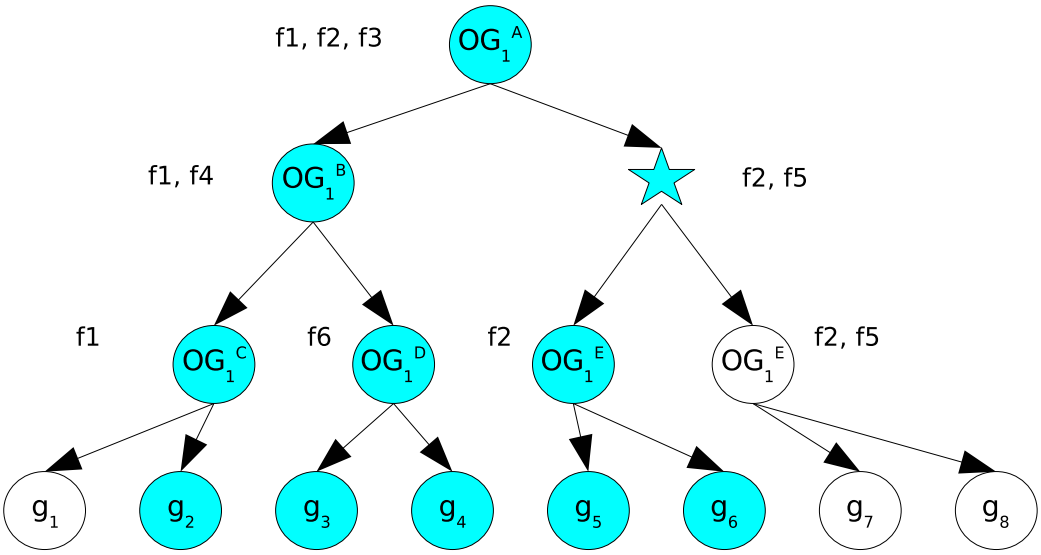

Figure 1: **Genome representation** DAG representation of genomes from five species, A, B, C, D, and E. The subgraph corresponding to genes of species A (blue leaves) is marked in blue. Labels indicate the putative functions of orthogroups and genes.

one ancestor per gene, the genome of an organism can be represented by a *directed acyclic graph* (*DAG*). The nodes of the DAG representation are divided into leaves (nodes with out-degree 0) and internal nodes. A leaf node is a gene, while an internal node is an orthogroup (*i.e.*, a set of genes that descended from a single common ancestor gene). To our knowledge, no such network-based approach for analysis of genomes has been suggested.

The DAG representation of a genome can be coupled with assignment of labels to each gene representing its functional characterization and possibility for alternative splicing. In such a way, one arrives at a *labeled DAG*, as shown in Figure 1.

The topology of the DAG (without its labels) can be studied in two possible ways: (1) by defining an optimal decomposition of a DAG, G, into rooted trees, T_i, $1 \leq i \leq k$ whose union is the G itself, one may arrive at processes which generate the rooted trees and classify them into several categories: random, plane-oriented, or distance-based trees, (2) definition of random growth processes for DAG generation which closely match the empirical properties of G, namely: in- and out-degree distribution, average

height, or average width, to name a few. While the first approach has been explored in depth, the second is still in its infancy. Therefore, addressing these questions would lead to novel insights in both biology and mathematics.

For the labeled DAG and a set of gene functions $\mathcal{F}$, one may investigate several biologically relevant hypotheses: Is it more likely that a gene of function F_1 is an ancestor of a gene of function F_2?, How likely is that two functions F_1 and F_2 arise at a critical point in the evolution of G)? (this question is related to the notion of critical point in the "evolution" of random graph processes for which a rich theory already exists), or What is the structural characteristic of G that enables emergence or enrichment in genes of function F_1?

Furthermore, we point out that such representation, although very simple, (a) enables investigation of functional emergence and distribution in terms of dynamic competing processes on G, (b) provides means for quantifying the coupling of NEO-F, SUB-F, EAC, and NON-F.

The outcome of this step will include:

1. Set of rules for generating random DAG topologies representative of a realistic genome,

2. Set of rules that describe label distribution and emergence in terms of the topology,

3. Procedure for randomizing a DAG and its labels to test biological hypotheses related to gene function, and, finally,

4. Unifying theory that ties structural and biological properties to address and assess the relationship between the fate of paralogs.

The DAG can be overlaid with two types of networks: gene regulatory and protein interaction networks (GRN and PIN, respectively). To arrive at these overlay networks, one has to add node representations for the gene products (proteins) to the proposed DAG representation.

Step II. GRNs are directed networks modeling the transcriptional regulation of genes. By overlaying a GRN on the DAG representing a genome, one can investigate the effects of gene duplication on the growth of GRN [23]. Synthetic models of GRN growth by gene duplication have already been investigated in [8] and [10]. It would be interesting to create a model that mimics real-world data, including upstream and downstream gene regions

for specific organisms (*e.g.*, *Arabidopsis Thaliana*, for which necessary data are available).

Step III. PINs are undirected networks modeling the interaction of proteins encoded by genes. By overlaying a PIN on the DAG one can deduce the evolutionary laws of PINs and their dependence on the labeled DAG structure. Similarly to the proposed study of GRNs, we plan to arrive at a model that mimics protein interaction data for an investigated species.

We plan to provide ranking for the set of proposed models and analyze the robustness of the network models.

3 Proposed Methods

In this section we described the statistical and graph-theoretic methods for achieving the goals of our three-step approach.

3.1 Statistical Tools for the Analysis of Global Network Properties

Here, we describe the methods aimed at studying a graph property $\mathcal{P}$ defined for the nodes or edges of a given graph, *i.e.* a graph property that can be expressed in terms of a probability distribution. For instance, the degree of a node is the number its neighbors. The degrees of all nodes can be summarized by the degree distribution. Given a probability model (*i.e.*, exponential, Possion, power-law) our goal is to determine the parameters which best describe the property $\mathcal{P}$. Moreover, once the parameters for a set of probability models have been determined, we should be able to distinguish which model provides the best description for $\mathcal{P}$.

3.1.1 Maximum Likelihood Inference

For a network G, the property $\mathcal{P}$ defined for the nodes/edges of G can be treated as independent observations in the limit of an infinite size network. We take a composite likelihood approach to inference [6]: For a given functional model $Pr(k, \theta)$ of the distribution for a property $\mathcal{P}$ we can use maximum likelihood estimation applied to the composite likelihood in order to estimate the parameter which best characterizes the distribution of the

data. The composite likelihood of the model given by the observed data $K = \{k_1, \ldots, k_n\}$ is defined by

$$L(\theta) = \prod_{i=1}^{n} Pr(k_i; \theta). \tag{1}$$

Taking the logarithm, one may obtain the log-likelihood

$$lk(M) = lk(\theta) = \sum_{i=1}^{n} \log Pr(k_i; \theta). \tag{2}$$

The maximum likelihood estimate (MLE), $\hat{\theta}$, of θ is the value of θ for which Equations (1) and (2) are maximized. For this value of θ the observed data is more probable to occur than for any other value of the parameter.

3.1.2 Model Selection with Akaike Weights

Selecting a model which best describes a given data set can be obtained by using the Akaike information criterion (AIC) [1]. The AIC for a model $Pr(k; \theta)$ is defined by

$$AIC = -2lk(\hat{\theta}) - 2d, \tag{3}$$

where $\hat{\theta}$ is the MLE of θ and d is the number of parameters required to define the model, *i.e.*, the dimension of θ. The model with the minimum AIC is chosen as the best. Note that, due to the parametrization in terms of d, a more complicated model is only accepted as better if it contains more information about the data than a simpler model. In order to compare r different models, we define the relative difference:

$$\Delta_j^{AIC} = AIC_j - \min_{j=1}^{r} AIC_j, \ 1 \leq j \leq r. \tag{4}$$

The relative likelihoods of the different models are given by

$$e^{-\frac{\Delta_j^{AIC}}{2}}. \tag{5}$$

The so-called Akaike weights, w_j, $1 \leq j \leq r$ can then be obtained by normalizing the relative likelihoods:

$$w_j = \frac{e^{-\frac{\Delta_j^{AIC}}{2}}}{\sum_{j=1}^{r} e^{-\frac{\Delta_j^{AIC}}{2}}}. \quad (6)$$

Given a data set, the Akaike weight w_j can be interpreted as the probability that model j, $1 \leq j \leq r$ is the best model for the observed data.

3.1.3 Goodness-of-Fit

Besides the described Akaike weight, the performance of a model can be estimated by two other goodness-of-fit statistics: Kolmogorov-Smirnov (KS) and Anderson-Darling (AD). The KS statistic is defined as

$$D_1 = \max_{i=1}^{n} |\hat{C}(i) - C(i)|, \quad (7)$$

while the AD statistic is given by

$$D_2 = \max_{i=1}^{n} \frac{|\hat{C}(i) - C(i)|}{\sqrt{C(i)\,(1 - C(i))}}, \quad (8)$$

where $\hat{C(i)}$ and $C(i)$ are the empirical and the theoretical cumulative distribution functions, respectively. In other words $C(i) = \sum_{i=1}^{n} Pr(i)$ and $\hat{C}(i) = \sum_{i=1}^{n} \hat{Pr}(i)$. If $\hat{C}(i)$ depends on a parameter θ then $\hat{C}(i) = \sum_{i=1}^{n} \hat{Pr}(i;\theta)$ from the estimated distribution.

We can then calculate p-values for the values of D_1 and D_2 by the following procedure: if there are n observed data, draw n numbers from $Pr(k;\theta)$ and calculate D_1 and D_2; repeat this step N times in order to get the null distributions for D_1 and D_2. The approximations of the p-values allow to test how close are the estimated and the empirical distributions for a property $\mathcal{P}$. This is in fact a parametric bootstrap procedure which uses the estimated model for a property $\mathcal{P}$.

3.2 Statistical Tools for Selection of Network Models

Recent work on fitting network models comes from social sciences, where the so-called exponential random models (ERMs also known as $p*$ models) are introduced [27]. Recently, ERMs were applied to the analysis of biological network structure [21]. The $p*$ models focus on "local" structural features

of networks (*e.g.*, characteristics of nodes that determine a presence of an edge). Contrary to this approach, we aim at estimating the likelihood of a graph model that may approximate the structure of a given biological network without relying on specific global properties.

First, we need a specification of the model, and to this end we employ probabilistic inductive classes of graphs [13]. A *probabilistic inductive class of graphs* (*PICG*), I, is given by:

1. class B of initial graphs, the basis of PICG,
2. class R of generating rules, each with distinguished left element to which the rule is applied to obtain the right element,
3. probability distribution specifying how the initial graph is chosen from class B,
4. probability distribution specifying how the rules from class R are applied, and, finally,
5. probability distribution specifying how the left elements for every rule in class R are chosen.

Let the graph model be described by a vector-valued parameter θ (item 3 in the definition of the PICG) and let G_t be an observation from the model—a graph after applying the rules (from item 2 of the PICG definition) t times. We are intersted in finding the MLE of the graph G_t, which is tantamount to calculating the likelihood of G_t as a function of θ.

We say that a node (edge) is *removable* if it can be obtained by one of the rules in item (2) of the PICG definition, above. By focusing on the removable nodes, we first define $\delta(G_t, v)$ as the graph obtained by deleting node v from G_t. We can then obtain the following definition of the likelihood:

$$L(\theta, G_t) = \frac{1}{t} \sum_{v \text{ is removable}} \omega(\theta, G_t, v) L(\theta, \delta(G_t, v)), \tag{9}$$

where

$$\omega(\theta, G_t, v) = P_\theta(G_t | \delta(G_t, v)), \tag{10}$$

and the factor $1/t$ denotes the probability that v is the last added node (by assuming that the rules in R add one node per time step).

Note that Equation (10) implies checking all of the possible node (edge) orderings, and, therefore, although $L(\theta, G_t)$ can be evaluated recursively, in practice is intractable due to the super-exponentially many permutations. However, the problem can be solved by using Importance Sampling [18], which allows writing the likelihood as an expectation over a Markov chain [4, 28, 17].

3.3 Biological Robustness from Topology

Robustness is a property that allows a system to maintain its functions despite external and internal perturbations. Kitano [15] specifies that biological robustness is tightly related to the existence of bow-tie structures and their hierarchical ordering.

A directed graph with bow-tie structure consists of four parts: giant strong component (GSC), in-subset (IS), out-subset (OS) and isolated subset (IS). The GSC is the largest strongly connected components, IS consists of nodes that can reach the GSC but cannot be reached from it, while OS consists of nodes that are accessible from the GSC, but do not link back to it. The IS contains nodes that can neither reach nor be reached from the GSC. By designing an algorithm for creation of bow-tie hierarchy and defining a set of indices on such hierarchy one can quantify the robustness of a directed biological network. The analysis of metabolic networks presented in [30] is a similar first step, yielding a visualization tool rather than a method for quantifying the robustness. We plan to analyze the gene regulatory network overlaid on the DAG representation of a genome to reveal some biological properties from the inherent topology of the network. Centrality indices and measures of congestion are some possibilities we would like to investigate.

For a DAG G representing an investigated genome, first we define an *st*-DAG as DAG with a special node s, denoting a source, and a special node t, denoting a target. It is trivial to show that any DAG can be turned into an *st*-DAG which then can be analyzed in terms of how far it is from a series parallel graph.

The class of series-parallel (*sp*-) graphs is defined as follows: A graph is an *sp*-graph, if it may be turned into an by a sequence of the following operations:

- Replacement of a pair of parallel edges with a single edge that connects their common endpoints and

- Replacement of a pair of edges incident to a vertex of degree 2 other than s or t with a single edge.

An *sp*-index of a *st*-DAG is then defined as the minimum number of node reductions required to reduce a given *st*-DAG into a graph with one directed edge, when used along with series and parallel reductions. A node reduction contracts a node with in-degree (out-degree) of one into its single incoming (outgoing) neighbor. Since the *st*-DAGs associated to different genomes have different number of nodes, we propose to use the normalized *sp*-index as a measure of genome complexity, where normalization is performed with respect to the number of nodes in an *st*-DAG. Simple observations can be used to arrive at the following algorithm for determining the normalized *sp*-index of an *st*-DAG which employs the definitions of a dominator tree T^d, reverse dominator tree T^r, and a complexity graph $G*$. One can then prove that the normalized *sp*-index of the *st*-DAG is given by the size of the maximum matching in $G*$ divided by the order (number of nodes) of G. Reduction of *sp*-graphs with respect to some flow problems is discussed in [3].

References

[1] H. Akaike. Information measures and model selection. *Proceedings of the 44th Session of the International Statistical Institute*, pages 277–291, 1983.

[2] G. Ast. How did alternative splicing evolve? *Nature Reveiws Genetics*, 5:773–782, 2004.

[3] W. Bein and P. Brucker. Greedy concepts for network flow problems. *Discrete Applied Mathematics*, 15:135–144, 1986.

[4] I. Bezaáková, A. Kalai, and R. Santhanam. Graph model selection using maximum likelihood. *Proceedings of International Conference on Machine Learning 2006*, 2006.

[5] S. E. Brenner, T. Hubbard, A. Murzin, and C. Chothia. Gene duplication in h. influenzae. *Nature*, 378:140, 1995.

[6] D. R. Cox. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91:729–737, 2004.

[7] Y. V. der Peer. Computational approaches to unveiling ancient genome duplications. *Nature Reviews*, 5:752–763, 2004.

[8] J. Enemark and K. Sneppen. Gene duplication models for directed networks with limits on growth. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(11):P11007, 2007.

[9] W. M. Fitch. Distinguishing homologous from analogous proteins. *Syst. Zool.*, 19:99–113, 1970.

[10] D. V. Foster, S. A. Kauffman, and J. E. S. Socolar. Network growth models and genetic regulatory networks. *Physical Review E*, 73:031912, 2006.

[11] A. L. Hughes and R. Friedman. Alternative splicing, gene duplication and connectivity in the genetic interaction network of the nematode worm caenorhabditis elegans. *Genetica*, 134:181–186, 2008.

[12] M. Hurles. Gene duplication: the genomic trade in spare parts. *PLoS Computational Biology*, 2:900–904, 2004.

[13] N. Kejzar, Z. Nikoloski, and V. Batagelj. Probabilistic inductive classes of graphs. *Mathematical Sociology*, 32:85–109, 2008.

[14] E. Kim, A. Magen, and G. Ast. Different levels of alterantive splicing among eukaryotes. *Nucelic Acids Research*, 35:125–131, 2007.

[15] H. Kitano. Biological robustness. *Nature Reviews*, 5:826–837, 2004.

[16] N. M. Kopelman, D. Lancet, and I. Yanai. Alternative splicing and gene duplication are inversly correlated evolutionary mechanisms. *Nature Genetics*, 37:588–589, 2005.

[17] J. Leskovec and C. Faloutsos. Scalable modeling of real graphs using Kronecker multiplication. *Proceedings of International Conference on Machine Learning 2007*, 2007.

[18] J. S. Liu. *Monte Carlo Strategies in scientific computing*. Springer, New York, 2001.

[19] S. Ohno. *Evolution by Gene Duplication*. Springer, New York, 1970.

[20] O. Sakarya, K. S. Kosik, and T. H. Oakley. Reconstructing ancestral genome content based on symetrical best alignments and dollo parsymony. *Bioinformatics*, 24:606–612, 2008.

[21] Z. M. Saul and V. Filkov. Exploring biological network structure using exponential random graph models. *Bioinformatics*, 23:2604–2611, 2007.

[22] B. Snel, P. Bork, and M. Huynen. Genome evolution gene fusion versus gene fission. *Trends in Genetics*, 16:9–11, 2000.

[23] S. A. Teichmann and M. M. Babu. Gene regulatory network growth by duplication. *Nature Genetics*, 36:492–496, 2004.

[24] S. A. Teichmann, J. Park, and C. Chothia. Structural assignments to the mycoplsma genitalium proteins show extensive gene duplications and domain rearrangements. *Proceedings of the National Academy of Sciences of the USA*, 95:14658–14663, 1998.

[25] W. Wang, H. Yu, and M. Long. Duplication-degeneration as a mechanism of gene fission and the origin of new genes in drosophila species. *Nature Genetics*, 36:523–527, 2004.

[26] I. Wapinski, A. Pfeffer, N. Friedman, and A. Regev. Natural history and evolutionary principles of gene duplication in fungi. *Nature*, 449:54–61, 2007.

[27] S. Wasserman and P. Pattison. Logit models and logistic regression for social networks. *Psychometrika*, 60:401–425, 1996.

[28] C. Wiuf, M. Brameier, O. Hagberg, and M. P. H. Stumpf. A likelihood approach to analysis of network data. *Proceedings of the National Academy of Science of the USA*, 103:7566–7570, 2006.

[29] Z. J. Zhang. Evolution by gene duplication: an update. *Trends in Ecology and Evolution*, 18:292–298, 2003.

[30] J. Zhao, L. Tao, H. Yu, J.-H. Luo, Z. W. Cao, and Y. Li. Bow-tie topological features of metabolic networks and the functional significance. *Chinese Science Bulletin*, 52:1036–1045, 2007.