PhD Thesis

Evolutionary Fingerprints in Genome-scale Networks

Moritz Schütte
2011

Max Planck Institute
of Molecular Plant Physiology

Kumulative Dissertation von

Moritz Schütte

geb.: 13.08.1980
Matrikelnummer: 742127

ausgeführt in der Arbeitsgruppe
”Systems Biology and Mathematical Modelling”
am Max-Planck-Institut für
molekulare Pflanzenphysiologie

zum Thema

Evolutionary Fingerprints in Genome-scale Networks

zur Erlangung des akademischen Grades
”doctor rerum naturalium”
(Dr. rer. nat.)
in der Wissenschaftsdisziplin ”Systembiologie/Bioinformatik”

an der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Potsdam

Published online at the
Institutional Repository of the University of Potsdam:
URL http://opus.kobv.de/ubp/volltexte/2012/5748/
URN urn:nbn:de:kobv:517-opus-57483
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus-57483

Abstract

Mathematical modeling of biological phenomena has experienced increasing interest since new high-throughput technologies give access to growing amounts of molecular data. These modeling approaches are especially able to test hypotheses which are not yet experimentally accessible or guide an experimental setup. One particular attempt investigates the evolutionary dynamics responsible for today's composition of organisms. Computer simulations either propose an evolutionary mechanism and thus reproduce a recent finding or rebuild an evolutionary process in order to learn about its mechanism. The quest for evolutionary fingerprints in metabolic and gene-coexpression networks is the central topic of this cumulative thesis based on four published articles.

An understanding of the actual origin of life will probably remain an insoluble problem. However, one can argue that after a first simple metabolism has evolved, the further evolution of metabolism occurred in parallel with the evolution of the sequences of the catalyzing enzymes (manuscripts of Chapter 2 and 3). Indications of such a coevolution can be found when correlating the change in sequence between two enzymes with their distance on the metabolic network which is obtained from the KEGG database. We observe that there exists a small but significant correlation primarily on nearest neighbors. This indicates that enzymes catalyzing subsequent reactions tend to be descended from the same precursor. Since this correlation is relatively small one can at least assume that, if new enzymes are no "genetic children" of the previous enzymes, they certainly be descended from any of the already existing ones. Following this hypothesis, we introduce a model of enzyme-pathway coevolution. By iteratively adding enzymes, this model explores the metabolic network in a manner similar to diffusion. With implementation of an Gillespie-like algorithm we are able to introduce a tunable parameter that controls the weight of sequence similarity when choosing a new enzyme. Furthermore, this method also defines a time difference between successive evolutionary innovations in terms of a new enzyme. Overall, these simulations generate putative time-courses of the evolutionary walk on the metabolic network. By a time-series analysis, we find that the acquisition of new enzymes appears in bursts which are pronounced when the influence of the sequence similarity is higher. This behavior strongly resembles punctuated equilibrium which denotes the observation that new species tend to appear in bursts as well rather than in a gradual manner. Thus, our model helps to establish a better understanding of punctuated equilibrium giving a potential description at molecular level. From the time-courses we also extract a tentative order of new enzymes, metabolites, and even organisms. The consistence of this order with previous findings provides evidence for the validity of our approach.

While the sequence of a gene is actually subject to mutations, its expression profile might also indirectly change through the evolutionary events in the cellular interplay. Gene coexpression data is investigated in the manuscripts of Chapter 4 and 5. This data is simply accessible by microarray experiments and commonly illustrated using coexpression networks where genes are nodes and get linked once they show a significant coexpression. Since the large number of genes makes an illustration of the entire coexpression network difficult, clustering helps to show the network on a metalevel. Various clustering techniques already exist. However, we introduce a novel one which maintains control of the cluster sizes and thus assures proper visual inspection. An application of the method on *Arabidopsis thaliana* reveals that genes causing a severe phenotype often show a functional uniqueness in their network vicinity. This leads to 20 genes of so far unknown phenotype which are however suggested to be essential for plant growth. Of these, six indeed provoke such a severe phenotype, shown by mutant analysis. By an inspection of the degree distribution of the *A. thaliana* coexpression network, we identified two characteristics. The distribution deviates from the frequently observed power-law by a sharp truncation which follows after an over-representation of highly connected nodes. For a better understanding, we developed an evolutionary model which mimics the growth of a coexpression network by gene duplication which underlies a strong selection criterion, and slight mutational changes in the expression profile. Despite the simplicity of our assumption, we can reproduce the observed properties in *A. thaliana* as well as in *E. coli* and *S. cerevisiae*. The over-representation of high-degree nodes could be identified with mutually well connected genes of similar functional families: zinc fingers (PF00096),

flagella, and ribosomes respectively.

In conclusion, these four manuscripts demonstrate the usefulness of mathematical models and statistical tools as a source of new biological insight. While the clustering approach of gene coexpression data leads to the phenotypic characterization of so far unknown genes and thus supports genome annotation, our model approaches offer explanations for observed properties of the coexpression network and furthermore substantiate punctuated equilibrium as an evolutionary process by a deeper understanding of an underlying molecular mechanism.

Allgemeinverständliche Zusammenfassung

Die biologische Zelle ist ein sehr kompliziertes Gebilde. Bei ihrer Betrachtung gilt es, das Zusammenspiel von Tausenden bis Millionen von Genen, Regulatoren, Proteinen oder Molekülen zu beschreiben und zu verstehen. Durch enorme Verbesserungen experimenteller Messgeräte gelingt es mittlerweile allerdings in geringer Zeit enorme Datenmengen zu messen, seien dies z.B. die Entschlüsselung eines Genoms oder die Konzentrationen der Moleküle in einer Zelle. Die Systembiologie nimmt sich dem Problem an, aus diesem Datenmeer ein quantitatives Verständnis für die Gesamtheit der Wechselwirkungen in der Zelle zu entwickeln. Dabei stellt die mathematische Modellierung und computergestützte Analyse ein eminent wichtiges Werkzeug dar, lassen sich doch am Computer in kurzer Zeit eine Vielzahl von Fällen testen und daraus Hypothesen generieren, die experimentell verifiziert werden können.

Diese Doktorarbeit beschäftigt sich damit, wie durch mathematische Modellierung Rückschlüsse auf die Evolution und deren Mechanismen geschlossen werden können. Dabei besteht die Arbeit aus zwei Teilen. Zum Einen wurde ein Modell entwickelt, dass die Evolution des Stoffwechsels nachbaut. Der zweite Teil beschäftigt sich mit der Analyse von Genexpressionsdaten, d.h. der Stärke mit der ein bestimmtes Gen in ein Protein umgewandelt, “exprimiert”, wird.

Der Stoffwechsel bezeichnet die Gesamtheit der chemischen Vorgänge in einem Organismus; zum Einen werden Nahrungsstoffe für den Organismus verwertbar zerlegt, zum Anderen aber auch neue Stoffe aufgebaut. Da für nahezu jede chemische Reaktion ein katalysierendes Enzym benötigt wird, ist davon auszugehen, dass sich der Stoffwechsel parallel zu den Enzymen entwickelt hat. Auf dieser Annahme basiert das entwickelte Modell zur Enzyme-Stoffwechsel-Koevolution. Von einer Anfangsmenge von Enzymen und Molekülen ausgehend, die etwa in einer primitiven Atmosphäre vorgekommen sind, werden sukzessive Enzyme und die nun katalysierbaren Reaktionen hinzugefügt, wodurch die Stoffwechselkapazität anwächst. Die Auswahl eines neuen Enzyms geschieht dabei in Abhängigkeit von der Ähnlichkeit mit bereits vorhandenen und ist so an den evolutionären Vorgang der Mutation angelehnt: je ähnlicher ein neues Enzym zu den vorhandenen ist, desto schneller kann es hinzugefügt werden. Dieser Vorgang wird wiederholt, bis der Stoffwechsel die heutige Form angenommen hat. Interessant ist vor allem der zeitliche Verlauf dieser Evolution, der mittels einer Zeitreihenanalyse untersucht wird. Dabei zeigt sich, dass neue Enzyme gebündelt in Gruppen kurzer Zeitfolge auftreten, gefolgt von Intervallen relativer Stille. Dasselbe Phänomen kennt man von der Evolution neuer Arten, die ebenfalls gebündelt auftreten, und wird Punktualismus genannt. Diese Arbeit liefert somit ein besseres Verständnis dieses Phänomens durch eine Beschreibung auf molekularer Ebene.

Im zweiten Projekt werden Genexpressionsdaten von Pflanzen analysiert. Einerseits geschieht dies mit einem eigens entwickelten Cluster-Algorithmus. Hier läßt sich beobachten, dass Gene mit einer ähnlichen Funktion oft auch ein ähnliches Expressionsmuster aufweisen. Das Clustering liefert einige Genkandidaten, deren Funktion bisher unbekannt war, von denen aber nun vermutet werden konnte, dass sie enorm wichtig für das Wachstum der Pflanze sind. Durch Experimente von Pflanzen mit und ohne diese Gene zeigte sich, dass sechs neuen Genen dieses essentielle Erscheinungsbild zugeordnet werden kann.

Weiterhin wurden Netzwerke der Genexpressionsdaten einer Pflanze, eines Pilzes und eines Bakteriums untersucht. In diesen Netzwerken werden zwei Gene verbunden, falls sie ein sehr ähnliches Expressionsprofil aufweisen. Nun zeigten diese Netzwerke sehr ähnliche und charakteristische Eigenschaften auf. Im Rahmen dieser Arbeit wurde daher ein weiteres evolutionäres Modell entwickelt, das die Expressionsprofile anhand von Duplikation, Mutation und Selektion beschreibt. Obwohl das Modell auf sehr simplen Eigenschaften beruht, spiegelt es die beobachteten Eigenschaften sehr gut wider, und es läßt sich der Schluss ziehen, dass diese als Resultat der Evolution betrachtet werden können.

Die Ergebnisse dieser Arbeiten sind als Doktorarbeit in kumulativer Form bestehend aus vier veröffentlichten Artikeln vereinigt.

Acknowledgments

I sincerely want to thank Oliver Ebenhöh for his advice and introduction into Systems Biology. I very much enjoyed working in his lab and am grateful not only for his sound scientific supervision and experience but also for the very personal and pleasant work environment he created. Also, I would like to thank Joachim Selbig for his mentoring during my teaching experience and for being my official supervisor. I always appreciated his advice in statistics questions and his patience with all kinds of questions on formalities around a PhD.

I am very grateful to Zoran Nikoloski and all members of AG Nikoloski/Ebenhöh, in particular Alexander Skupin, for stimulating discussions, rewarding collaborations, and fun even at work. Moreover, I like to thank Marco Ende for his patience with endless IT questions.

Many thanks to Daniel Segrè for having given me the opportunity to spend five months in his lab at Boston University and especially for his excellent advice during our common research projects. I believe that this experience has had a strong impact on developing my scientific skills and broadening my horizon. I also thank Niels Klitgord, Caroline Lyman, and all members of the Segrè lab for having created such a kind and welcoming atmosphere during this research visit.

I also would like to thank Marek Mutwil, Staffan Persson, and Björn Usadel for fruitful collaborations in the projects of the last two manuscripts.

Further, I thank the International Research and Training Group "Genomics and Systems Biology of Molecular Networks" and all its members for profitable feedback during the retreats, seminars, and IBSB conferences. Further, I am grateful for funding by the DFG, in particular for the time I was able to spend in Boston.

Finally, I am very grateful to the Max Planck Institute of Molecular Plant Physiology for the inspiring atmosphere and the possibility to use the great facilities. Also, I like to acknowledge the PotsdamGraduateSchool for the various offered courses, especially for my participation and gain of teaching experience in the JuniorTeachingProfessionals program.

Contents

1. Introduction

The biological cell is a highly complex system. A complete description of the cellular machinery requires an incorporation of various interacting players like proteins, different types of RNA, and DNA. The field of *Systems Biology* accepts this challenge and can be seen as ”an approach to biology that seeks to understand and predict the quantitative features of a multicomponent biological system” [209]. Due to the complexity of the underlying biology and the variety of potential research goals, *Systems Biology* attracts researchers from manifold fields such as biology, informatics, physics, or medical sciences. Since a quantitative description is desired, experimental data is indispensable. Luckily, advancements in measurement techniques have been accelerated in recent years. From the beginning of sequencing of single genes, bacteriophage or bacterial genomes [86, 53, 55], the rapid developments in high-throughput-methods continuously allow for measuring massive amounts of data every day. New types of ”next-generation sequencing techniques” even bring the idea of a personal genome for medical treatment to life [166, 120, 157, 23, 31, 77]. Together with sequence data, huge data bases of biochemical and experimental data have been collected, such as KEGG, Pfam, or BioCyc [91, 54, 93]. It is evident that a deeper understanding of such complex systems can not be reached without a sound theoretical and modeling framework.

A very prominent approach is the description by the mathematical tool of graph theory. Therein, a graph is given by an ensemble of nodes that represent the data points and edges between the nodes that depict a certain interaction of the nodes [3, 21]. The number of edges of one node is called its degree and its distribution often follows a power-law type [15, 113]. Biological examples of graph-theoretical applications include protein-protein interaction, signaling, gene coexpression, or metabolic networks [2, 41, 128, 94, 83] of which the last two are subject of this thesis. Since the architecture and mathematical description of networks remain similar between various scientific areas, one can quickly adapt and apply methods borrowed from other fields. Multiple measures on networks exist, such as the degree distribution, the centrality and essentiality of a node [83], the clustering coefficient and modularity [135], or the path length between two nodes. In this way, one can gain insight into the data shaping the current status of the network and identify biological interactions which can be understood as a top-down approach. Contrary, bottom-up, it is also of great interest to investigate putative growth models that lead to this network status and which, in biological terms, may point at evolutionary mechanisms.

With improved sequencing techniques, it became possible to quickly sequence entire organisms and functionally annotate their genome using homology measures with respect to known sequences. Thus, genome-scale metabolic networks could be reconstructed from the annotated genomes [48, 59, 42]. Since these comprise thousands of reactions with equally many and mostly unknown kinetic parameters [72, 181], genome-scale dynamic modeling using differential equations for the rates remains almost impossible. Simplified methods that aim either at a quantitative or at a qualitative description of the network characteristics were developed.

Flux-Balance analysis (FBA) successfully predicts quantitative steady-state flux distributions assuming that an organism’s metabolism has evolved in order to maximize a certain objective such as biomass production [161, 159, 162, 160, 145]. For unicellular organisms, as *Escherichia coli* [47], the simple growth maximization assumption might hold, whereas it is truly not the case for higher organisms and different tissues [42, 172]. Reversely, instead of maximizing the biological objective, FBA variants engineer metabolic fluxes in order to maximize the production of certain target chemicals by proper knock-out strategies [25, 26, 144]. An investigation of the effect of gene knock-outs on the metabolic performance is crucial for the characterization of a mutant’s response on the gene knock-out [170, 171], to analyze the cellular interplay by epistatic interactions [169], or understand diseases and find potential

drug targets [172].

While FBA predicts a biologically meaningful flux state, it does not give a systematic investigation of the metabolic network capacities. Elementary Flux modes (EFM) are possible minimal sets of enzymes (minimal in the sense that no enzyme can be removed) that provide a steady-state flux through the network [165]. Hence, alternative routes can be estimated, weaknesses in the network as putative drug targets can be identified, or complementary, by minimal cut-sets knock-out targets proposed [102]. Although this method works well for small networks, the number of EFMs increases dramatically with growing network complexity. For an *E. coli* core model with about 100 reactions, the number of possible EFMs is of the order of $10^4 - 10^5$ [103]. Therefore, the calculation of the modes leads to the new problem of their interpretation.

An alternative approach that, likewise as EFM but without the problem of computational explosion, characterizes metabolic network properties is the so-called ”scope” or ”network expansion” method [44, 45, 70, 69]. This method follows the idea of a forward evolution [66] which assumes that new metabolic pathways build up from the available substrate to a new product; the contrary perspective is the retrograde evolution where a certain metabolite gets depleted in the surroundings and thus the organism needs to find a reverse pathway to synthesize this particular metabolite [75]. The network expansion works iteratively, starting from a ”seed” of metabolites, and those reactions are added to the network, for which all substrates are currently contained in the seed. Then in the next step, the products of the identified reactions are added to the seed and the iteration starts again until no further reaction can be added. The starting seed could either represent an environmental pool of metabolites or different growth media. It has been shown how the metabolism depends on environmental conditions [68, 22] such as the rise of atmospheric free oxygen as a major evolutionary landmark [150]. Further, the method provides a framework to extend organism-specific metabolic models by additional reactions which connect the present model to so far excluded but experimentally measured metabolites [30].

Natural entities have grown according to the laws of evolution, namely mutation and selection (and optionally ”natural cooperation” [138]). Mathematical models can produce putative scenarios of a certain step in evolutionary history. Likewise, they can explain observed properties or designs of the current biological status by assumed evolutionary principles and thus depict them as evolutionary fingerprints. Surely, evolutionary models will never reflect the actual evolution but have proven to be successful in a variety of cases. There exist numerous scientific articles of which we can only list some important representatives. The manifold of gene families, species, and protein folds could thus be explained from an *ab initio* perspective [216], while the protein world might have evolved from a ”protein Big Bang” [41]. In order to show that very general principles emerge, artificial chemistry models can be utilized to find evidence for an observed property. Pfeiffer et al. [143] show that the connectivity in metabolic models is closely connected to group transfer reactions. One assumed outcome of evolution is the formation of functional modules as design principles [127] which are reused by the organism and thus form a kind of toolbox [178, 118]. By an investigation of the modularity and robustness of artificial metabolic networks it was found that redundant genes often form synthetic lethal gene pairs and tend to appear in the same modules [73]. Design principles as auto-catalytic circles or recurring modules emerge as a consequence of an optimized path length in artificial chemistry pathways [151] while the optimization for metabolic paths by minimal enzymatic steps has recently experienced empirical evidence in central carbon metabolism [136, 121].

The concept of evolutionary modeling will be central in this cumulative thesis. In the first two chapters we will introduce a model for a putative scenario on the evolution of metabolism from the first small atmospheric molecules as a parallel evolution of metabolic pathways and their corresponding enzymes. In the last two chapters gene coexpression will be analyzed from two perspectives: first by a novel clustering technique and later by an evolutionary model.

Metabolic Evolution

Several evolutionary theories about the origin of life are widely discussed [95, 43], none of which known to be fully true or giving rise to clear rejection. The quest for early prebiotic structures that display the ability of self-reproduction has lead to multiple evidence. RNA molecules may carry information and simultaneously act catalytically as ribozymes which has also experienced experimental support when RNA sequences were found which are capable of elongating single stranded RNA [84]. In further experiments, self-replicating liposomes and micelles have been created [7, 206] and it has been argued that lipids [168] and possibly other polymers [96] may have formed early self-replicating entities. Regardless of the exact nature of the first self-replicating molecules, it is evident that early auto-catalytic reaction networks must fulfill the properties of a closed collectively auto-catalytic set [96], which basically states that each involved molecule, precursor and catalyst, which requires catalysis can be produced in excess by the reaction network itself. For a metabolic network, this entails that strictly speaking, also synthesis of enzymes from amino acids must be included. Smith and Morowitz [174] provided an example which circumvents this problem. In their metabolism first scenario, they demonstrate that the reductive tricarboxylic acid (rTCA) cycle is auto-catalytic and that its reactions were, under early earth conditions, likely to occur with a higher probability than competing chemical reactions even without the presence of catalysts, thus making the rTCA cycle a good candidate for a primordial metabolic core. However, recent theoretical considerations [201] have questioned the evolvability of such self-sustaining auto-catalytic metabolism-first scenarios by pointing out that the propagation of the compositional information from one generation to the next is too inaccurate for efficient selective pressures to work.

While the actual origin of life remains uncertain, after the formation of an early metabolism the further evolution of the metabolic network must have happened in parallel with the development of new catalyzing enzymes. An investigation of this hypothesis has been carried out using a mathematical model of this coevolution, the manuscripts of Chapter 2 and 3.

In a first approach, we investigated whether there exists a sequence similarity between enzymes that catalyze neighboring metabolic reactions. While similar studies have been carried out for single organisms [112, 152, 79], we focused on the metabolic network of a pseudo-organism comprised of all reactions in the KEGG database [90, 92, 89, 91] to see whether this assumption holds as a general principle. Since the set of protein sequences catalyzing chemical reactions from all organisms is huge, approximately one million in KEGG release 53, a complete investigation using all sequences is not feasible. Therefore, we first constructed a consensus sequence set which contains representatives of all classes of functionally equivalent enzymes (i.e. classified by Enzyme Commission (EC) numbers). Even within the set of enzymes with the same EC number sequences vary from identical to completely different which in this case is measured by sequence alignments using BLAST [76]. The Clusters of Orthologous Groups of proteins (COG) database provides the required classification of conserved sequence domains but has not been updated in the recent past [187, 186, 188, 185]. Therefore, we used the COG database as a benchmark to construct a consensus set which reduces the amount of sequences by about four orders of magnitude but still conserves a similar distribution of EC number frequency. The resulting sequences are compared using different symmetric and asymmetric measures of pairwise sequence distance. Then, we compared the calculated sequence distances with the distance of enzymes on the metabolic network. We define two enzymes to be neighbors, i.e. distance one, if they share a metabolite in any catalyzed reaction. Because some metabolites as cofactors like ATP/ADP, NADP/NADPH, or NAD/NADH or small molecules like water, O_2, or CO_2 participate in a variety of reactions, this leads to bypasses which connect most metabolites on a very short path. But, since these paths do not reflect the actual biochemical relations or the route of a metabolic flux, these metabolites were excluded from the calculation of the shortest paths. Finally, we obtained evidence for a small but significant correlation between the sequence distances and the distances defined on the metabolic network which is most visible for nearest neighbors.

Punctuated equilibrium is a concept in the evolution of species and states that evolutionary change does not appear in more or less equal temporal steps, called phyletic gradualism, but rather happens

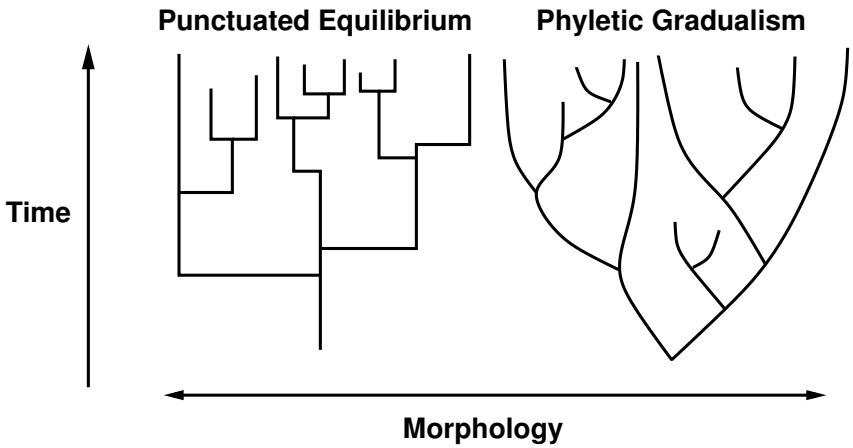


Figure 1.1: Scheme of the competing evolutionary theories. In the scenario of punctuated equilibrium the evolutionary change appears in bursts of rapid change followed by silent intervals. The contrary idea of phyletic gradualism suggests a continuous divergence from the ancestor.

in rapid bursts of morphological variation followed by silent intervals [49, 38] (see Fig. 1.1). While the concept has been observed experimentally in *E. coli* populations [50] and theoretically in the field of self-organized criticality [10], a description explaining potential molecular reasons for the macroscopic phenomenon has not been presented before.

In Chapter 3, we present a model that generates putative scenarios of the metabolic evolution. For this, we extend the method of network expansion, introduced earlier [44, 45, 70, 69], by not only taking into account the addition of possible new reactions to the network but by a parallel evolution of the reaction network and the catalyzing enzymes based on their mutual enzyme similarity. We start from a set of metabolites that putatively were present in a prebiotic atmosphere. This set is comprised of small molecules built of C,N,O,P,S, and H from which most necessary metabolites, such as amino acids, lipids, and carbohydrates, can be formed. Recent studies also suggest the possibility of bacteria living of arsenic instead of phosphorus [213] although these results have been questioned [153]. Candidates for the first enzymes are identified if they contain certain conserved sequence fragments that are common in a large number of proteomes. The choice of the next emerging enzyme and the specific time point of its appearance is implemented using the Gillespie algorithm [64] and thus allows for a definition of a time coordinate of the coevolution. Surely, this time does not reflect a geological time scale at all and thus does not depict an actual time of evolutionary events. However, it provides a tool to estimate the time intervals between two events which here depict the appearance of a new enzyme. Further, through the Gillespie method we introduce a parameter that triggers the influence of sequence information on the choice of the next enzyme. After the generation of various simulations, we performed a time-series analysis of the results that clearly shows that, if enzymes evolve by small mutational changes in the sequences rather than by randomly picking new sequences, new enzymes tend to appear in bursts reflecting the principle of punctuated equilibrium at a molecular level. Specifically, we find a high coefficient of variation of the inter-enzyme time intervals and substantiated this by calculation of the autocorrelation function, which exhibits higher correlations for emphasized sequence relations.

Since a model is a theoretical construct and needs to produce realistic evolutionary walks on the network, we analyzed the temporal order of the appearance of metabolites, enzymes, and organisms. The amino acids appear, on average over 200 simulations, in good correlation both with an order from robustness against reaction removal [30] and with an order obtained by the frequency of appearance of an amino acid in the consensus set of sequences. Furthermore, the enzymes were mapped on specific pathway functions given through the KEGG database and we identified all pairs of enzymes that appear in the same temporal order in 200 simulations. In order to differ between effects from stoichiometry (i.e. biochemistry) and from the sequence relations (i.e. evolution), we subtracted all temporal pairs of sequences that were also found in simulations where new enzymes were chosen randomly. This

indicates that these pairs may be results of purely stoichiometric constraints. We observe that during the evolutionary scenario, central-carbon metabolism appears first whereas synthesis of secondary metabolites emerges late. To generate a tree of life, we defined an enzyme repertoire for every organism and identified the time point when 80% of this repertoire has evolved. This cut-off is chosen to define the appearance of an organism. Of course, this is rough and error-prone but due to missing enzymes in the data and the fact that not every enzyme was necessary to bring an organism to life or enzymes might have been invented later, this cut-off seems a good compromise between precision and quality. Despite this simplification, we find that organisms appear in good correlation with their complexity: organisms with small networks first and with larger ones later. While this holds for eukaryotes, the situation for bacteria looks different. Here, the correlation is much smaller and we observe a big cloud of bacteria of all sizes that appear roughly together. Altogether, these analyses substantiate our results and show the applicability of this modeling approach.

Analysis of Gene Coexpression Networks

Besides the genetic sequence information, the gene expression profile determines the cellular status. Thus, in a second project, we investigated the networks constructed through the coexpression profile of genes. In this network type, two genes are connected if they are significantly strongly coexpressed. These gene coexpression networks were investigated by two approaches: first, by a novel clustering mechanism and secondly, since their degree distribution somehow deviates from the often observed power-law behavior, by an evolutionary model that reproduces this observation, Chapter 4 and 5.

There already is a variety of clustering methods. These may vary either by a concrete goal for which the clustering is designed, such as a parameter that determines the number of clusters [71], or by a goal such as finding local densities [8] or high flow [197] in the network to determine clusters. While such criteria may lead to clustering solutions which are comprised of very many or unequally sized clusters, they are hard to interpret and to visualize. Therefore, we designed a clustering method, called Heuristic Cluster Chiseling Algorithm (HCCA), which controls the cluster size, Chapter 4. In order to check the implemented clustering algorithm, we compared it to the existing methods Markov clustering (MCL), MCODE, and k-means [197, 8, 71]. For this, we used the graph measures modularity, Davies-Bouldin score, and the adjusted Rand index as well as enrichment of biological functions given by MapMan terms within the clusters. The novel method could outperform the existing algorithms for most parameter combinations. As biological validation, the method was applied on the gene coexpression network constructed of microarray data of *Arabidopsis thaliana*. By a comparison between essentiality and functional uniqueness in a network vicinity, we identified twenty genes which are putatively essential for plant growth but for which no phenotype has been annotated yet. Through a mutant analysis the usefulness of the approach could be shown, because six of these twenty genes indeed show a severe phenotype: two gametophyte, two embryo, one seedling lethal, and one pale green dwarf.

The degree distribution of these coexpression networks deviates by two properties from the widely accepted form of a power-law [14]. It shows characteristic humps and a sharp truncation for high degrees. We analyzed these properties for the *Arabidopsis* network used before and for the networks of *E. coli* and *Saccharomyces cerevisiae*. Since these properties might have the same origin, we generated an evolutionary model aiming at reproducing those. The coexpression profile of genes is potentially influenced by their evolutionary history. Therefore, our model mimics the evolution of a genome given by its expression profile. Then, by gene duplication subject to a special selectivity procedure, inspired by the Fermi-Dirac distribution, this genome grows until it reaches a predefined size similar to those of the data of the organisms. Following this method, the final degree distributions show the same characteristics that were observed in the real data and we thus conclude that the selective pressure during evolution might have shaped the expression profile.

In summary, this cumulative thesis comprises four articles published in international scientific journals and each presented as a separate chapter. The format of the papers has been slightly changed to combine it into the thesis. The first two deal with a model of chemical evolution which as a main

finding seems to underlie dynamics of a bursting enzyme appearance, shown by time-series analysis, that gives a first potential explanation for the concept of punctuated equilibrium at a molecular level. The last two chapters investigate gene coexpression data. We could show that a novel clustering approach has predictive power as it suggests phenotypes to previously unknown genes which is experimentally confirmed by mutant analysis. Further, we investigated that network characteristics in the gene coexpression network of three organisms can be explained by an evolutionary model.

2. Coevolution of Metabolism and Protein Sequences[†]

2.1. Abstract

The set of chemicals producible and usable by metabolic pathways must have evolved in parallel with the enzymes that catalyze them. One implication of this common historical path should be a correspondence between the innovation steps that gradually added new metabolic reactions to the biosphere-level biochemical toolkit, and the gradual sequence changes that must have slowly shaped the corresponding enzyme structures. However, global signatures of a long-term coevolution have not been identified. Here we search for such signatures by computing correlations between inter-reaction distances on a metabolic network, and sequence distances of the corresponding enzyme proteins. We perform our calculations using the set of all known metabolic reactions, available from the KEGG database. Reaction-reaction distance on the metabolic network is computed as the length of the shortest path on a projection of the metabolic network, in which nodes are reactions and edges indicate whether two reactions share a common metabolite, after removal of cofactors. Estimating the distance between enzyme sequences in a meaningful way requires some special care: for each enzyme commission (EC) number, we select from KEGG a consensus set of protein sequences using the cluster of orthologous groups of proteins (COG) database. We define the evolutionary "distance" between protein sequences as an asymmetric transition probability between two enzymes, derived from the corresponding pairwise BLAST scores. By comparing the distances between sequences to the minimal distances on the metabolic reaction graph, we find a small but statistically significant correlation between the two measures. This suggests that the evolutionary walk in enzyme sequence space has locally mirrored, to some extent, the gradual expansion of metabolism.

2.2. Introduction

The evolutionary walk from an early proto-metabolism to the current biochemical pathways must have been shaped by innovations concurrently involving enzymes and chemical compounds [107]. While it is generally assumed that today's enzymes have evolved from a few ancestors that were able to catalyze the first reactions, a clear correspondence between the evolution of metabolic functions and their catalyzing enzymes remains to be established. The evolution of metabolic pathways has been addressed by several competing models, including the patchwork model [214, 82], and models of forward [66] and retrograde [75] evolution. In addition, several studies have addressed the relation between sequence homology and protein function. These studies have been widely used for the prediction of protein functions [192, 154, 211] associated with newly sequenced genes and for the analysis of relations between sequences and functions in Gene Ontology terms [85].

To date, there exist only few studies of evolutionary relation between enzyme sequence homology and distance on the metabolic reaction network. Most of these studies are restricted to networks of single organisms. These works show a possible link between homology and metabolic network distance. For example in *E.coli* [112, 152], it was found that homologous protein sequences are more likely to be found in close vicinity on the metabolic network than what expected by chance and the same trend has

[†]Published as:
M. Schütte, N. Klitgord, D. Segrè, O. Ebenhöh, *Co-evolution of metabolism and protein sequences*, Genome Informatics 22, 156–166 (2010)

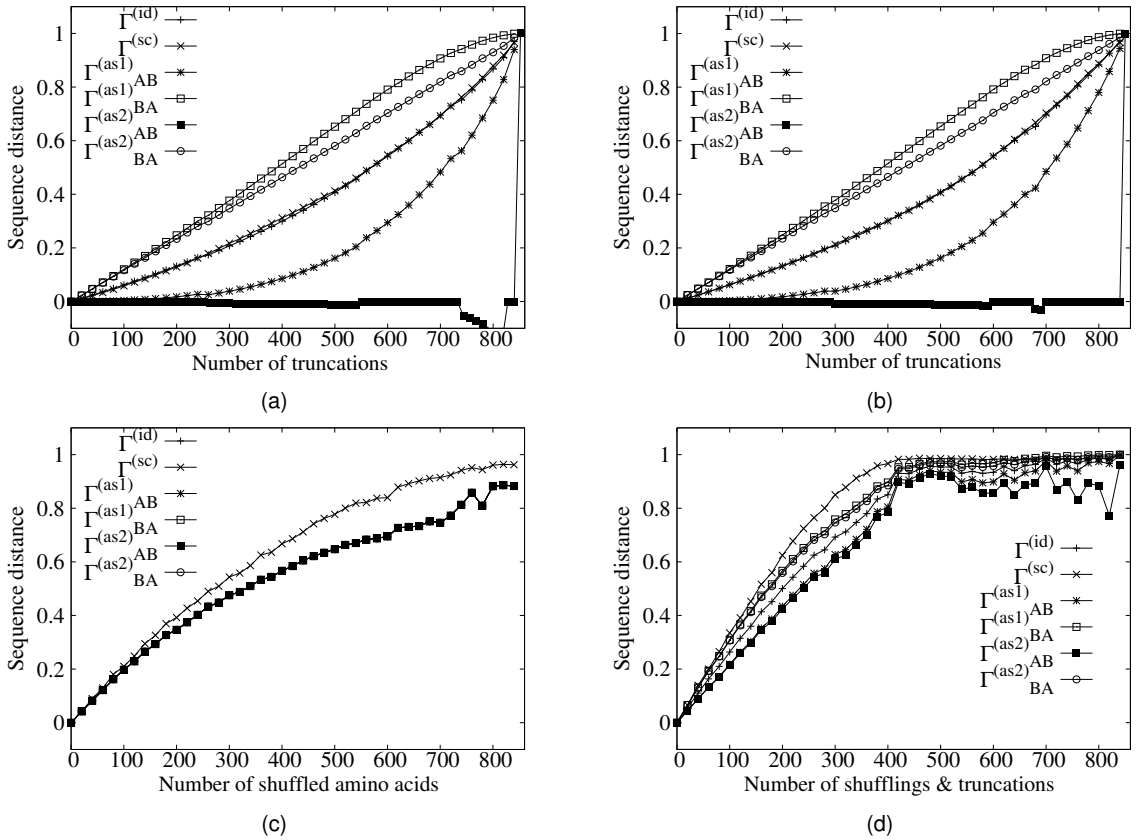


Figure 2.1: Simulation of four evolutionary scenarios. We take a test sequence, change it and blast it against an original copy. In (a) and (b) we iteratively delete amino acids from the start (a) or end (b) of the test sequence. (c): The amino acids are increasingly shuffled. (d) combines (a) and (c): we shuffle and truncate the sequence.

been confirmed for protein-protein interactions [79]. Similarly, in studies of yeast [202] a link has been found between the metabolic network structure and enzymatic evolution. Since single enzymes or even entire operons have been copied and changed to fulfill new functions, high promiscuity in the locality of catalyzing enzymes complicates the search for evolutionary relations [164, 98, 178].

Here, we test the hypothesis that the global scope of metabolism has evolved in parallel with the enzymes that make up the network. To investigate this hypothesis, we have taken a large-scale approach using the entire set of reactions from the KEGG [90, 92, 89] reaction database. If our hypothesis holds true, then we expect to find that a similarity between protein sequences should be reflected by a closeness in the metabolic reaction network. In order to test our hypothesis we define different measures of sequence distances, both symmetric and asymmetric, based on a reciprocal pair-wise BLAST analysis. Asymmetry becomes important if two sequences of different lengths are compared, because it is more likely that a shorter sequence is the evolutionary child of a longer sequence, than the other way around. These distances are then compared with the distances on an enzyme-enzyme network of chemical reactions that we construct from the KEGG database. As KEGG provides multiple protein sequences per reaction, we first choose a consensus sequence set based on the cluster of orthologous groups of proteins (COG) database [187, 185] that greatly reduces the sequence space we must analyze. Our analysis supports our hypothesis, showing that enzymes that are close in the reaction network are enriched for sequence similarity. The observed trend is small but significant against a simulated control.

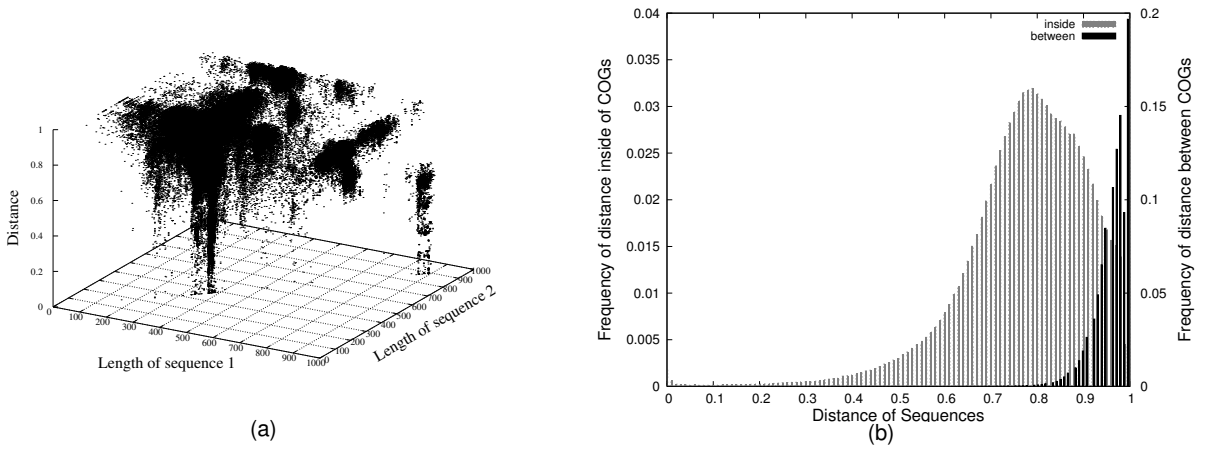


Figure 2.2: (a) All pair-wise blasts of the alcohol dehydrogenase (EC 1.1.1.1) sequences scored by the distance Eq. (2.1). Two apparent clusters of small distance are observable around length 300 and 900. (b) Benchmark of COG inner distance versus the distances between representatives of each COG. The cut-off of 0.9 to differ seems reasonable.

2.3. Protein Sequence Distances and Enzyme Distances

To address our hypothesis on a large scale, we use the entirety of reactions from the KEGG database. We construct a reaction-centric network where reactions are nodes and two reactions are linked if they involve a common metabolite. Some metabolites, such as cofactors, participate in a variety of reactions and thus produce short-cuts that do not carry actual fluxes. To account for this, we extract the cofactor pairs ATP/ADP, NAD^+/NADH, $NADP^+$/NADPH, and CoA/Acetyl-CoA from the reactions where they appear as pairs [79]. Furthermore, we delete highly abundant molecules, H_2O, H^+, and O_2, which appear in more than 500 reactions, from the reaction set. Every reaction is catalyzed by one or more enzymes given in terms of enzyme commission (EC) numbers, exceptions in the reaction set are spontaneous reactions or those for which the catalyst is not known. We link the reactions to the enzyme sequence using the EC numbers. Such, we transform the reaction network to an enzyme-enzyme network. As we only know sequences for roughly half of all EC numbers we still use all reaction links but only calculate shortest paths between enzymes for which we know the sequences of starting and ending enzymes. For this purpose we use the Dijkstra algorithm [40]. Intermediate reactions without catalyzing enzyme, like spontaneous reactions, or with an enzyme without known sequence, are still counted as a step in the distance.

The distance in protein sequence space requires more elaboration. We use the Blastp in the *bl2seq* program to obtain pair-wise comparisons between sequences [76]. A rather intuitive measure obtained from blast is the *identity* measure counting how many residues on a certain aligned fragment are identical between the two sequences. To get a comparable scoring measure we need to normalize the identities by the sequence lengths. It is important to note that we are interested in interpreting such a measure as an estimate of the probability that one sequence has evolved from the other. Since asymmetry can play a crucial role in the transition from one sequence to the other [137], we will take into account not only the total length, but also the difference between the two lengths. We define $\mathcal{I} = \sum_i I_i$ as the sum of the identities of every found alignment. Based on this measure, we define three different estimates of the probability Γ_{AB} that a sequence B (with length L_B) has evolved from a sequence A (with length L_A):

$$\Gamma_{AB}^{(\mathrm{id})} = 1 - \frac{2 \cdot \mathcal{I}}{L_A + L_B}\,, \tag{2.1}$$

$$\Gamma_{AB}^{(\mathrm{as1})} = 1 - \frac{2 \cdot \mathcal{I}}{L_A + L_B}\left[1 + \frac{(L_A - L_B)}{(L_A + L_B)}\right]\,, \tag{2.2}$$

$$\Gamma_{AB}^{(\mathrm{as2})} = 1 - \frac{\mathcal{I}}{L_B}\,. \tag{2.3}$$

The second measure follows from a Taylor expansion of Eq. (2.1). We consider the length difference as a small correction to the mean of the lengths: $(L_A+L_B)/2 \Longrightarrow \bar{L}-\Delta L/2$: $(\bar{L}-\Delta L/2)^{-1} \approx 1/\bar{L}+\Delta L/2\bar{L}^2$. If the directionality of a linear pathway is known, an asymmetric distance like Eqs. (2.2)–(2.3) could in principle be used to test for retrograde or forward evolution [66, 75]. In addition to Eqs. (2.1)–(2.3) we use the *score* values of the best hit obtained from BLAST and normalize it by the score of the sequence with itself:

$$\Gamma_{AB}^{(\mathrm{sc})} = 1 - \frac{2 \cdot \mathrm{score}(A,B)}{\mathrm{score}(A,A) + \mathrm{score}(B,B)}\,. \tag{2.4}$$

To evaluate the utility of these proposed measures, we tested how they perform on simple pairs of sequences, computed so as to simulate a gradual transition from exact identity to large differences, see Fig. 2.1. For the purposes of our simulation, we chose one random but rather long (839 amino acids) test sequence (*glo:Glov_1829 integral membrane sensor signal transduction histidine kinase (EC:2.7.13.3)*). These experiments simulate some of the possible scenarios of sequence changes during enzyme evolution. In the first two experiments, we took a copy of the sequence and iteratively cut off amino acids from either start or end of the copy. This reduced sequence is then blasted against the original one, Figs. 2.1(a) and (b). In experiment three we shuffled increasingly more amino acids in the copy and blasted against the original sequence, (c). The last experiment, (d), is a combination of shuffling and length reduction. Here, we observe an edge at around 0.9 distance. Below this value the behavior seems random.

These experiments show that $\Gamma_{AB}^{(\mathrm{as2})}$, Eq. (2.3), is not an appropriate measure. Specifically, in the first two experiments this measure reaches negative values, resulting from the fact that we sum all identities. Because the aligned fragments become very small, it is likely that the same fragment is found twice or that an overlap between two matches is found in the original copy. Since the measure is only normalized by its own length, this may result in negative values. A similar artifact is also observed in experiment four, where the measure shows a false positive agreement when sequences have been highly shuffled, and greatly truncated. Thus this measure will not be used for further investigations.

2.4. Consensus Sequence Set

We characterize an enzyme by two features: its sequence and its function. These features can, to some degree, vary independently: One enzyme may have multiple functions, or conversely, one specific function can be performed by multiple sequences [60]. Our goal is to investigate whether the evolutionary distance of sequences relates to their functional distance as determined by the metabolic reaction network where the reactions are defined in terms of EC numbers. Since each EC number can be associated with a rather large number of sequences, we define a consensus set of sequences serving as representatives of the specific function. In total, the current KEGG database contains approximately 750000 sequences that contain one or more EC numbers in their description. As can be seen in Fig. 2.2(a), the distribution of the number of sequences per EC number can vary quite a bit in KEGG. This has been greatly reduced in our consensus sequence set. For example, the sequence by sequence distances for alcohol dehydrogenase, EC 1.1.1.1, are shown in Fig. 2.2(a). The lengths vary by a factor three and even the sequences of similar lengths need not at all be similar. Even the set of 188 alcohol dehydrogenase sequences that are all 350 amino acids in length varies from completely

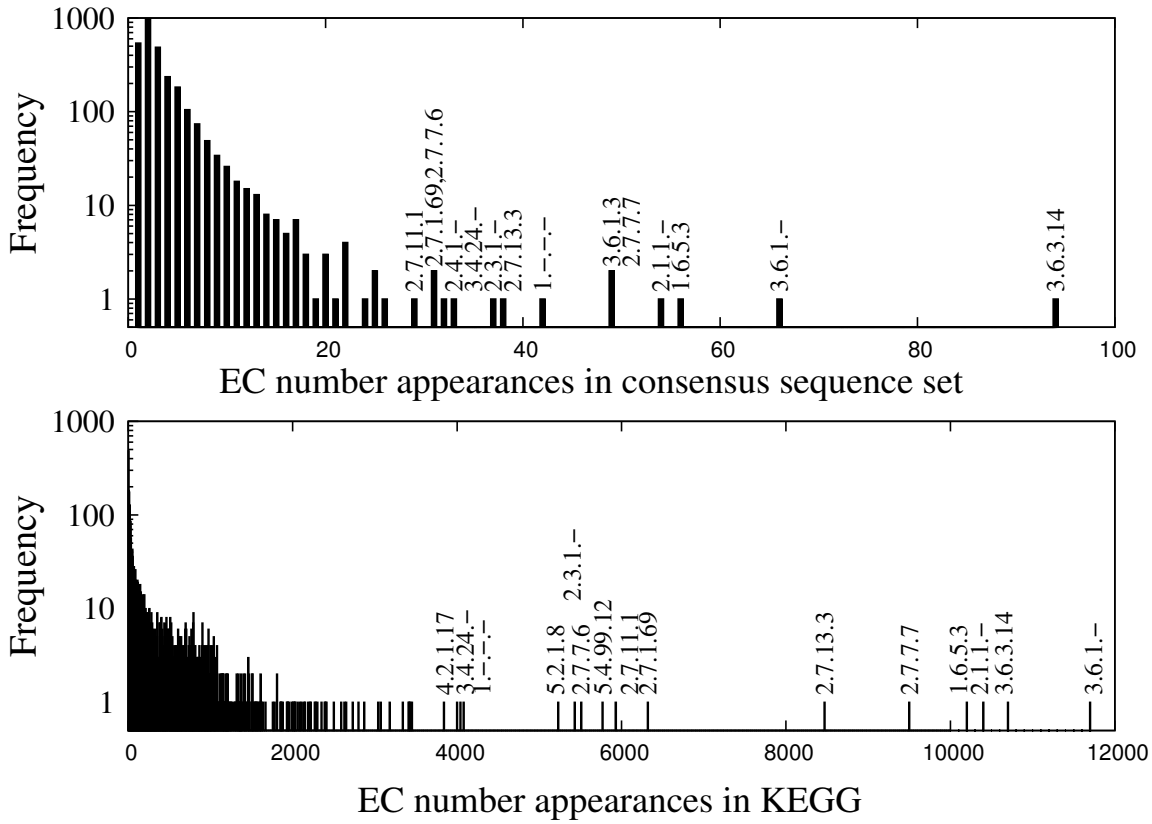


Figure 2.3: Distribution of EC numbers how often they appear in descriptions of different protein sequences. Top in the consensus set, bottom in the KEGG sequences.

identical to completely different with a mean of $\Gamma^{(\mathrm{id})} = 0.72 \pm 0.12$. One can observe several distinct clusters of high similarity in Fig. 2.2(a).

To simplify our task of selecting representative sequences, we utilize the COG database [187, 185] that clusters proteins by their function based on homology. Every COG contains between a few and a few hundred sequences that are related by a duplication or speciation event. We pick the longest sequence of every COG as the representative of this COG [105]. In order to cover as many as possible EC numbers we cluster the remaining KEGG sequences by a very simple procedure. We group the sequences by EC number and perform all pairwise blasts dropping those sequences that have a 0.9 or higher distance according to Eq. (2.1). This loose cut-off is justifiable using the COGs as a benchmark. We calculated all inner-COG distances and compared them with the distances between the representatives of every COG, see Fig. 2.2(b). A second argument for this cutoff comes through the result of Fig. 2.1 (d) where all curves show a somewhat random behavior for distances larger than 0.9.

Following this procedure we obtain a consensus set of 8123 sequences coding for 2821 EC numbers. Fig. 2.3 shows a histogram of the frequency of a certain EC number with that it appears in descriptions of different sequences. The top bar graph shows the distribution in the final consensus set and the bottom one in the starting set from all KEGG sequences. There is a good agreement between the ranking by EC number with Pearson correlation 0.81 and Spearman Rank correlation 0.53.

2.5. Correlation of Network and Sequence Distances

We use the previously defined consensus sequence set to analyze a relationship between distance on the enzyme-enzyme graph and the sequences of the enzymes. We use a sample of 4.8 million shortest paths which all start and end with a sequenced enzyme.

Table 2.1.: Comparison of correlations between enzyme sequence distances and distances obtained form the enzyme-enzyme graph. In the control measurement we shuffle the enzyme distance matrix and repeat the simulation. The distances in the sequence space were calculated with the measures described in section 2 (sample size: 4.8 million shortest paths). Although the correlations are very low, they are highly significant in comparison with the control data.

measure	$\Gamma_{AB}^{(\mathrm{id})}$	control to $\Gamma_{AB}^{(\mathrm{id})}$	$\Gamma_{AB}^{(\mathrm{as1})}$	control to $\Gamma_{AB}^{(\mathrm{as1})}$	$\Gamma_{AB}^{(\mathrm{sc})}$	control to $\Gamma_{AB}^{(\mathrm{sc})}$
correlation	0.0127	0.0002	0.0170	-0.0003	0.0035	0.0004
p-value	10^{-171}	0.7270	10^{-304}	0.4931	10^{-14}	0.4189

Figure 2.4(a) shows a boxplot of the correlation using the measure $\Gamma^{(\mathrm{id})}$, Eq. (2.1). We see a highly significant but small correlation that is mirrored by a trend seen in the outliers where similar enzymes tend to be closer in the network. In order to test the results against the null hypothesis that they appeared by chance, we calculated the p-value which is based on the sample size. We performed a second control calculation utilizing a permutation test where we shuffled the sequence distances to generate a random set. For this control simulation the correlation is completely lost and we observe high p-values, see Tab. 2.1.

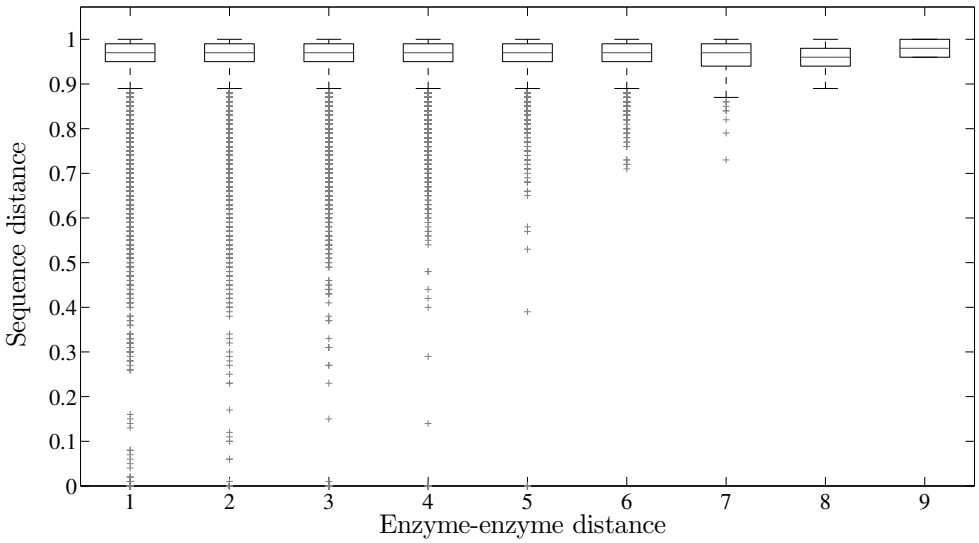
To further quantify the observation, we analyze the results sorting them by particular path-lengths, Fig. 2.4(b). Neighboring enzymes tend to show higher similarity on the sequence level. For enzyme-enzyme distance 1, the relative proportion of distances below 0.8 and 0.7 is enriched. The bar on distance 8 is the highest but it represents only a sample of four similar enzymes of 47.

Table 2.1 compares the results for different measures of sequence similarity. The asymmetric measure Eq. (2.2) yields the highest correlation and significance in calculating the sequence distance. For every enzyme pair we chose the more similar value of the asymmetric sequence distance $\Gamma^{(\mathrm{as1})}$.

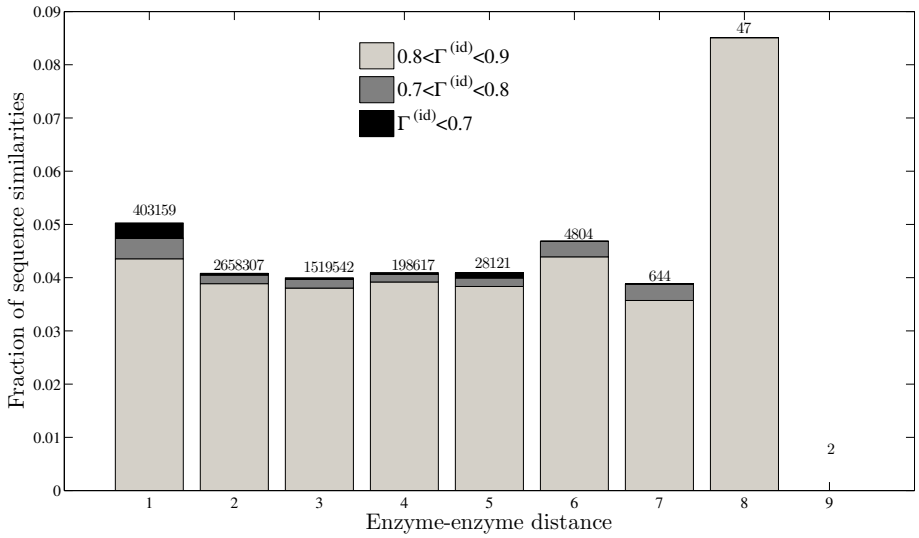
2.6. Conclusion

We have investigated the evolutionary relation of enzyme sequences and their distance in the metabolic network on a large-scale using a consensus sequence set from the entire KEGG database. However, the choice of the consensus set is strongly biased by two aspects of the used database: the choice of sequenced organisms and the accuracy in investigating proteins. A large number of redundant sequences is due to the variety of organisms whose proteomes are sequenced, and whose function was assigned via homology. The sequences in the consensus set come from 27 animals, 6 plants, 21 fungi, 535 bacteria, 51 archea and 17 protists and these result in 1395, 274, 585, 4974, 598, 297 sequences from the particular kingdoms. The variability in plant-specific enzymes might be underestimated as only a few model plants are well investigated. For the carbon-fixating enzyme RuBisCO (EC 4.1.1.39) we obtain only two different sequences from bacteria *Synechocystis* and *Anabaena*. The majority of organisms are bacteria for which lateral gene transfer is an important factor [147, 110]. By the use of the COG database we might neglect this possibility of sequence change. The second bias appears through the way proteins are investigated. As an example we examine ATP synthase, EC 3.6.3.14. This protein is the second most abundant in KEGG and the most abundant in the consensus set, Fig. 2.3. It catalyzes only one reaction, ATP + H_2O + $H^+_{in} \rightleftharpoons$ ADP + phosphate + H^+_{out}. This reaction is essential in most organisms and frequently investigated. We thus capture the variability of sequences for known enzymes but do not grasp it for less known ones.

We have observed a weak correlation between enzymes that are neighbors in the graph representing metabolism, and the corresponding sequences. Our finding extends to an all-organism level results previously obtained for single organisms [112] and using protein-protein interactions [79]. The correlation detected indicates a certain degree of coevolution between the topology of metabolism and its enzyme capabilities. We envisage that future simulations of the evolutionary expansion of metabolism,



(a)



(b)

Figure 2.4: (a) Boxplot for the correlation between distances on the enzyme-enzyme graph and the measure $\Gamma^{(\mathrm{id})}$ for the sequences. The correlation is very low, 0.0127, but significant, see Tab. 2.1. (b) Sequence distances sorted by particular enzyme-enzyme distances. The plot shows the fraction of enzyme sequence pairs within a certain distance of $\Gamma^{(\mathrm{id})}$ compared to all enzymes found in the distance on the network. The small numbers on top of the bars represent the total number of enzymes found in the particular distance. For neighboring enzymes we observe a higher fraction of enzymes with similar sequences.

possibly employing our proposed asymmetric measure of sequence distance, could shed more insight into the nature of this correlation.

2.7. Acknowledgments

We acknowledge financial support from the International Research Training Group *Genomics and Systems Biology of Molecular Networks* IRTG 1360 (MS), the German Federal Ministry of Education and Research, Systems Biology Research Initiative *Go*FORSYS, the Scottish University Life Science Alliance SULSA (OE), the NASA Astrobiology Institute, and the US Department of Energy (NK, DS).

3. Modeling the Complex Dynamics of Enzyme-pathway Coevolution[†]

3.1. Abstract

Metabolic pathways must have coevolved with the corresponding enzyme gene sequences. However, the evolutionary dynamics ensuing from the interplay between metabolic networks and genomes is still poorly understood. Here, we present a computational model that generates putative evolutionary walks on the metabolic network using a parallel evolution of metabolic reactions with their catalyzing enzymes. Starting from an initial set of compounds and enzymes, we expand the metabolic network iteratively by adding new enzymes with a probability that depends on their sequence-based similarity to already present enzymes. Thus, we obtain simulated time courses of chemical evolution in which we can monitor the appearance of new metabolites, enzyme sequences, or even entire organisms. We observe that new enzymes do not appear gradually but rather in clusters which correspond to enzyme classes. A comparison with Brownian motion dynamics indicates that our system displays biased random walks similar to diffusion on the metabolic network with long-range correlations. This suggests that a quantitative molecular principle may underlie the appearance of punctuated equilibrium whereby enzymes occur in bursts rather than by phyletic gradualism. Moreover, the simulated time courses lead to a putative time-order of enzyme and organism appearance. Among the patterns we detect in these evolutionary trends is a significant correlation between the time of appearance and their enzyme repertoire size. Hence, our approach to metabolic evolution may help understand the rise in complexity at the biochemical and genomic levels.

Evolution is a dynamic process in which species become extinct and new species emerge all the time. It is a disputed question whether the emergence of new species proceeds with an approximately constant rate or whether new species rather evolve in short periods with a high speciation rate which are separated by long silent periods in which only few new species evolve. The latter scenario is referred to as 'punctuated equilibrium' and has recently received support from empirical evidence. Here, we present a model of metabolic evolution which suggests that punctuated equilibria can also be observed in the evolution of macromolecules. This finding also supports the hypothesis that underlying molecular mechanisms may be responsible for the phenomenon of punctuated equilibrium in the evolution of new species. Our model uses available amino acid sequences for thousands of enzymes present in several hundred different organisms. By comparing all these sequences, we estimate probabilities that sequences may have evolved from one another. This information allows us to simulate putative scenarios for how today's metabolism might have evolved. By time series analysis we demonstrate that the existing sequence information strongly suggests a punctuated equilibrium behavior, which is considerably less pronounced if sequence information is deliberately neglected.

[†]Published as:
M. Schütte, A. Skupin, D. Segrè, O. Ebenhöh, *Modeling the complex dynamics of enzyme-pathway coevolution*, Chaos 20(4): 045115 (2010)

3.2. Introduction

The evolution of the modern biochemical pathways from an early proto-metabolism must have been shaped by innovations concurrently involving enzymes and chemical compounds [125, 117, 32]. While it is generally assumed that today's enzymes have evolved from a few ancestors that were able to catalyze the first reactions, the details of this evolutionary history are almost as uncertain as the details about the first self-replicating systems themselves [7, 206, 168, 96, 130, 43]. Several scenarios have been proposed, the simplest suggesting a 'forward' evolution in which enzymes evolved that could make use of the end products of existing metabolic pathways [66]. In the reverse assumption of a retrograde evolution, a necessary precursor became depleted and enzymes have evolved that replenish this required resource from other, still abundant, substances [75]. While supporting example pathways may be found for both views, the more complex assumption of a patchwork evolution [214, 82] becomes more relevant when viewing metabolism as a whole. The method of network expansion [44, 70] provides a simple evolutionary model that extends the forward evolution scenario to the metabolic network comprising all biochemical reactions known to date. While this approach was useful to relate structural to functional properties [45] by tracing catalytic properties along the evolutionary tree [46], discovering hints for an early separation of DNA and RNA metabolism [119] and providing insight into the increase of complexity upon the rise of oxygen in the Earth's atmosphere [150], it is clearly too simple to reproduce realistic evolutionary paths. More recent models elaborating on these ideas include the toolbox model of metabolic evolution [118], which assumes that network evolution is driven by the need to explore new resources and can readily explain the apparent quadratic scaling of the numbers of transcription factors with the total number of genes. The view of metabolic evolution as a Markov process in which additions or removal of reactions depend on the numbers of neighboring reactions [129] allows one to estimate parameters for the evolutionary dynamics and to assess possible evolutionary paths between two different network configurations.

The above mentioned examples all provide plausible arguments for a particular evolutionary path, but do not explicitly take into account that, after the appearance of the first catalyzed reaction networks, the discovery of new chemical compounds is strongly linked to the evolution of new enzymes from existing ones. Hints that the evolution of the sequence space defining contemporary enzymes mirrors to some extent the gradual expansion of the chemical space, defined by the variety of metabolites, were recently found by correlating sequence similarities to a distance of the catalyzed reactions on the metabolic network [167].

It was argued [10] that such a coevolution promotes short term avalanches during which a large number of new enzymatic steps could be invented, thus giving rise to a punctuated equilibrium behavior [49, 50]. In this paper, we present a model of metabolic evolution combining genome scale data, tools from bioinformatics, dynamic modeling and time series analysis with the goal of studying the apparent coevolution of small molecules and catalysts in further detail. As a basis for our exploration, we use the KEGG database [90, 91] which provides a comprehensive collection of biochemical reactions from several hundred organisms and information on amino acid sequences of the respective catalyzing enzymes. While previous models [88, 143, 73] investigated the evolution of metabolic networks as idealized artificial processes, our current model explicitly considers available biological data and assumes that those enzymes are more likely to evolve for which a related enzyme has already been discovered. We systematically explore how the evolutionary dynamics depends on the coevolution of metabolites and enzymes by introducing a tunable parameter reflecting the importance of sequence similarity. Thus, we can separate the effects of a sequence-based evolution from one in which the discovery of new enzymes is only restricted by stoichiometric constraints. We find that simulations taking into account existing sequence data display a punctuated equilibrium behavior and thus support the view that evolution, also at the level of metabolic networks, occurs in bursts of rapid sequences of new inventions, rather than in a gradual fashion [49].

3.3. Model Description

The enzymes found in contemporary organisms are highly efficient and usually very specific catalysts for chemical reactions. The amino acid sequences of present-day enzymes are the outcome of a long evolutionary history, in which they were subjected to random mutations and selective pressures favoring only particular sequences which may efficiently perform useful functions. A difficulty in modeling the evolutionary process of enzyme evolution is that neither sequences for early or extinct enzymes nor the precise criteria for the selective pressures are known.

Our proposed simple model for the evolution of metabolism takes these limitations into account. Instead of aiming at describing the evolution of networks of particular organisms, we focus on the network comprising reactions from several hundreds of species. We can thus focus on very general selective principles and ignore the specific pressures that were acting to support the evolution of highly specialized functions. Due to the lack of knowledge of early and now extinct protein sequences, our model is limited to all described biochemical reactions and sequence information available to date.

We mimic the evolution of the network comprising the presently known metabolic reactions by a simple process in which the network grows in size by consecutive addition of single enzymes. The process is initiated by assuming that a certain combination of primitive metabolites are abundant in the environment. We assume that new enzymes may evolve from existing ones through a series of amino acid exchanges. Since the mechanism for such mutations is essentially a random process, we assume that the probability to discover a new functional enzyme from an existing one is higher, the more similar their corresponding sequences are. Thus, at any stage of our simulated evolutionary process in principle every known enzyme may evolve. However, we assume that only those newly discovered enzymes will be positively selected which can perform a useful function. We therefore impose a selective pressure by accepting only those new enzymes which can catalyze a biochemical reaction from reactants that may in principle be produced from reactions already present in the network. A temporal scale is introduced by assuming that evolutionary events with a higher probability tend to occur faster.

The evolutionary simulation is implemented as a Gillespie algorithm [64] for the simulation of stochastic expansion processes and can be summarized in the following 7 steps:

1. A set of primitive compounds and first enzymes is selected. These comprise the initial network.
2. On the basis of the actual network structure, all enzymes that can catalyze a reaction utilizing only substrates present in the network are identified. For each enzyme i, a propensity p_i is calculated based on the sequence similarity to already present enzymes (see below). The propensity describes the probability that the enzyme is discovered per unit time.
3. Depending on the propensities, the time t_{next} of the next evolutionary event is determined by an exponentially distributed random variable with the mean given by $1/\sum p_i$.
4. Which particular enzyme is added at time t_{next} is determined by a uniformly distributed random number. The probability that enzyme j is selected is given by $p_j/\sum p_i$.
5. All reactions catalyzed by the selected enzyme as well as the corresponding products, are added to the network.
6. Due to the incorporation of new substances, new reactions catalyzed by enzymes already present in the network may be executable. These reactions and their products are added as well. The same holds true for any newly occurring spontaneous reactions.
7. The process is repeated with step 2 until no new enzymes can be added to the network.

Iterating this expansion process leads to a series of invented enzymes whose invention times depend on the underlying dynamics. Hence, we use the inter-enzyme intervals (IEIs), which are defined by the sequence of t_{next} and correspond to waiting times, to characterize the evolutionary process.

In contrast to the conceptually similar method of network expansion introduced in [70], in our model enzymes are considered to be the basic units of the networks rather than reactions. As a consequence, the discovery of a new enzyme leads to the addition of all reactions that such enzyme can catalyze. Moreover, whereas in the method of network expansion all reactions which can possibly occur are

simultaneously added to the growing network in each step, we here only add a single enzyme in each expansion event. Defining probabilities for enzyme appearance introduces a stochastic component which is inherent to all evolutionary processes. Further, by assigning characteristic times for the single evolutionary events our model possesses an intrinsic definition of an evolutionary time coordinate.

Like in many applications of the method of network expansion (see e.g. [68, 30]), we also assume that common cofactors do not specifically have to be produced during the expansion process before they can be used (see Methods). The rationale for this is that their metabolic functions can in principle also be carried out by simpler pairs of molecules. For example, the transfer of phosphate groups by ATP/ADP is possible by pyrophosphate and phosphate, as demonstrated in the bacterial phosphotransferase system, the role of NADH/NAD^+ as electron carriers can in principle be performed by metal ions such as Fe^{2+}/Fe^{3+} with different oxidation states.

3.4. Sequence Distances and Propensities

One particular focus of our model is the investigation of the evolutionary dynamics for different assumptions on how strongly the evolvability of novel enzymes from existing ones depends on the respective sequence similarities.

For roughly half of all functionally different enzymes present in the KEGG database [91], sequence information is available. The amount of information for one particular enzyme commission (EC) number can vary between none and a couple of thousand different sequences. In total KEGG (release 53) provides around one million sequences from various organisms for about 3000 EC numbers. We reduced the space of possible sequences by construction of a consensus set using the clusters of orthologous groups of proteins (COG) database [187, 186, 188, 185] as a benchmark. This enables to drop redundant sequences between functionally equivalent enzymes having the same EC number, see [167] and the Methods section. Herewith, we obtain a set of 11925 sequences that code for 3048 EC numbers for which we calculate all mutual distances using the '*score*' of the best BLAST alignment to assess the probability that sequence A evolves into B. The Blast '*score*' provides putative evolutionary knowledge about single amino acid substitutions, insertions, and deletions from which we define the distance between sequence A and B as

$$D_{AB} = 1 - \frac{2 \cdot \text{score}(A,B)}{\text{score}(A,A) + \text{score}(B,B)} \,. \tag{3.1}$$

The pairwise distance ranges from 0 for identical sequences to 1 for sequences without any significant alignment. For EC numbers for which no sequence is available, we assign distances to all other enzymes randomly from the distribution of all calculated sequence distances. While it is certainly possible that this introduces a bias in our results, we consider this approach the best possibility under the circumstances of incomplete information.

It is plausible to assume that, during evolution, the probability to discover a new enzyme is higher, if a similar enzyme already exists. We denote by $d_i^{\min}$ the minimal distance for enzyme i to all enzymes that have already been found. To have a tunable parameter that weighs the strength of the influence of the protein sequences, we define the propensities for a new enzyme to be discovered by

$$p_i = \frac{1}{d_i^{\min \gamma}}. \tag{3.2}$$

This definition implies that we assume that the expected time to find a new enzyme depends only on the minimal distance to existing enzymes scaled by the exponent γ. The extreme assumption of $\gamma = 0$ leads to equal propensities for all possible new enzymes and thus reflects a hypothetical case in which sequence information has no influence on the selective process, and the evolution of the network is exclusively determined by chemical constraints. A value of $\gamma = 2$ corresponds to the assumption that the possible sequence space is explored in a process analogous to a random walk, for which the average distance covered is proportional to the square root of the elapsed time. The other extreme,

$\gamma \to \infty$, reflects the hypothetical case of the path following least resistance in which the enzyme with the closest distance to an existing enzyme will always be discovered in the next step [10, 212].

3.5. Methods

3.5.1. Data for Network Structure and Sequences

We use the KEGG database, release 53. The "genes.pep" in fasta format is downloaded for protein sequences and the ligand-file for reactions. In order to curate the data erroneous reactions, that are not balanced or contain unspecified parts like a rest group, are rejected. The irreversibility information is obtained by scanning the pathway maps [69, 68]. Reactions that contain the cofactor pairs ATP/ADP, NAD/NADP, NADH/NADPH, or Co-A/Acetyl-CoA are added a second time without the cofactors, assuming that they are possible during expansion without coupling to cofactor usage and production.

3.5.2. Sequence Distance and Consensus Set

In order to calculate the evolutionary distance between any two enzyme we use pairwise sequence alignment using BLAST (version 2.2.22, standalone blast, http://www.ncbi.nlm.nih.gov/blast/blast_overview.shtml): *bl2seq -i SequenceA -j SequenceB -p blastp -F F -o output* with the default substitution matrix BLOSUM62. Then we score the alignment by the best hit of the '*score*' from the output getting the distance D_{AB} given by Eq. (3.1). This is not a distance by mathematical definition as it does not fulfill the triangular inequality. It ranges from 0, identical, to 1, completely different [167]. KEGG release 53 contains around one million sequences containing an EC (enzyme commission) number in their description. We sort these sequences by EC numbers and choose representatives from each set by taking only into account those that have a mutual distance, defined similarly to (3.1), above 0.95 and drop all others. The cutoff has been chosen according to a benchmark using the clusters of orthologous groups of proteins (COG) database, [187, 167]. This reduces the sequence set to 11925 for 3048 EC numbers. For EC numbers that appear in the reaction set but for which we do not have a sequence we randomly pick a distance to any other sequence from the distribution of distances between all known ones.

3.5.3. Seeds

What are the first metabolites and enzymes? We use the following primordial seed of compounds: H_2O, CO_2, H_2SO_4, H_3PO_4, NH_3 and H^+, [125, 29]. In order to identify the putative first enzymes, we use work by Y. Sobolevsky [176, 177, 175] who identified common conserved protein fragments in 131 proteomes. One particularly long fragment *LSGGQQQRVAIARAL* was found in bmn:BMA10247_1739 and tpe:Tpen_0904 and we added the remaining two of the same function 3.6.3.21, eco:b2306 and hpa:HPAG1_0922, to the seed of enzymes.

3.6. Results

3.6.1. The Expansion: Process and Enzyme Sequences

The expansion process starts from a given set of metabolites and enzymes, called the seed [44, 45, 46, 70]. This set represents a putative prebiotic chemical environment. A necessary requirement for the evolution of a substantial reaction network is the presence of all essential chemical elements in the seed. Here, we consider only the atoms H, C, O, N, P and S, because 80% of all metabolites in the KEGG database are composed of these elements. As seed, we choose H_2O, CO_2, H_2SO_4, H_3PO_4, NH_3 and H^+ [125, 29]. The choice of a first enzyme sequence is made by using conserved sequence fragments [176, 177, 175] to be enzymes of the function 3.6.3.21, see Methods.

To study the effect of sequence information, controlled by the parameter γ introduced in Eq. (3.2), we perform simulations with five values $\gamma = 0, 2, 10, 20, 100$. We account for the stochasticity of the

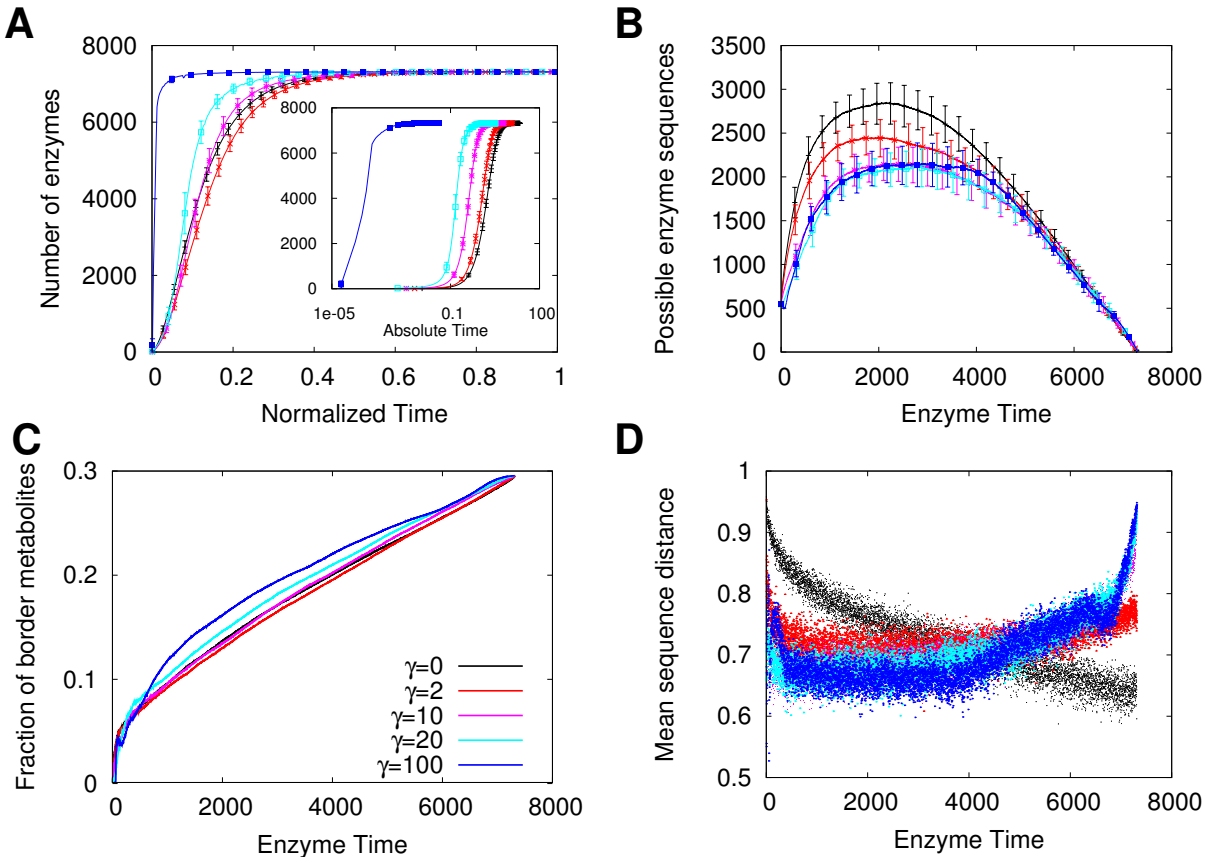


Figure 3.1: Comparison of the expansion process for different strengths of the sequence-information parameter γ. Means of 200 simulations are shown and for all panels the color code given in panel C holds. **A**: The network size measured by the number of enzymes attached to the network is shown over time. The expansion velocity increases with γ in both normalized time and in absolute time (inlet) obtained from the Gillespie algorithm. **B**: The number of attachable enzymes at every step in the expansion process can be understood as the evolvability of the network. Using sequential information leads to a less evolvable but thus denser network. **C**: How quickly do we expand to the border of existing knowledge. At every step in enzyme time we plot the number of detected metabolites which only participate in one reaction in KEGG. Higher γ approach the border faster supporting the assumption of a smarter expansion. **D**: Mean sequence distances between every new enzyme and its duplication partner. The $\gamma = 0$ curve decreases since by chance for any new enzyme on average a similar sequence can be found if more enzymes are present in the current network. For higher γ isolated sequences without any similarity to all others are preferentially found at the end resulting in an increase.

simulated evolutionary walks by performing 200 simulation runs for each selected value of γ, which correspond to scenarios in which sequence information is completely ignored ($\gamma = 0$) to the case in which a strict order of enzyme appearance is imposed by the sequence relatedness ($\gamma = 100$). As a direct consequence of the definition of the propensities of Eq. (3.2), simulations with different γ proceed on very different time scales, with the total time required to explore the entire network (inlet in Fig. 3.1A) being roughly 1000 times longer for the random scenario when compared to the scenario with high γ. In order to compare the velocities of the evolutionary processes between scenarios with different γ's, we normalize for each γ the time by the average final time of the respective 200 simulations and term the resulting temporal measure the *normalized time*, as opposed to the non-normalized *absolute time*. Fig. 3.1A provides a comparison of the expansion processes on both time scales.

The expansion process with maximum sequential order, $\gamma = 100$, leads to the quickest exploration of the network also on the normalized time (Fig. 3.1A). The behavior in terms of the number of metabolites as a function of time looks qualitatively similar Fig. A.1, and in particular obeys the same ranking in dependence on γ.

Which novel sequences may actually perform a useful function by catalyzing a biochemical reaction depends on the specific structure of the metabolic network at any given time during the evolutionary process. The number of these potential new enzyme sequences can be understood as a measure of the evolvability of the network [204]. To compare networks of identical sizes for scenarios with different γ, we introduce a third time measure, the *enzyme time*, defined by the current network size determined by the number of contained enzymes. In Fig. 3.1B the evolvability is shown as a function of the enzyme time for different values of γ. For all values of γ, the temporal change of the evolvability can be divided into three phases. Until enzyme time 1500–2000, it increases rapidly before it reaches a plateau which is more pronounced for higher values of γ. In the final phase after enzyme time 5000 the limitation of the enzyme pool results in a rather constant decrease. The off-set on the y-axis for enzyme time 0 in Fig. 3.1B results from a peculiarity of reaction R00086, ATP + $H_2O \Longleftrightarrow$ ADP + P_i. This reaction can be catalyzed by enzymes of 66 EC numbers and is associated with 355 different sequences. Considering that ATP is treated as a cofactor, for which we do not explicitly require that it can be produced by the present network, this reaction can be added even to the initial seed network at enzyme time 0. Addition of this reaction does not expand the chemical functions of the network, but increases the variety of sequences from which new sequences may potentially evolve. Measuring the evolvability in numbers of new executable reactions results in qualitatively similar curves, with the main differences that the offset is not observed and that the curves exhibit a negative skewness instead the positive one, see Fig. A.1.

For both measures, it is remarkable that the evolvability is systematically larger for scenarios with lower γ, in which new sequences are added more randomly. This observation suggests that in this case consecutively added enzymes are rather unrelated in their chemical function, leading to a high metabolic diversification. In contrast, in the scenarios in which sequence information is important, the preferential discovery of enzymes similar to existing ones leads to an evolutionary exploration of local neighborhoods and thus to denser and more functional networks.

To support this hypothesis, we investigate the appearance of metabolites which only occur in a single reaction of the KEGG database and which can thus be seen as the border of the currently known metabolism. Overall, in the KEGG subnetwork that is reached by our simulated evolutionary processes, 29.5% of the metabolites (661 of 2237, see Fig. A.1) belong to this class. In Fig. 3.1C, the appearance of these border metabolites is depicted over enzyme time. Evidently, the influence of sequence information leads to a quicker exploration of the border. This supports the notion that, as a tendency, for large γ pathways are completed in a consecutive order, whereas for small γ pathways tend to be explored in parallel.

In Fig. 3.1D we depict how the actual minimal sequence distance for the selected enzymes changes with enzyme time. For large values of γ, in which a strong preference for sequences with a low minimal distance to existing enzymes prevails, the curve exhibits a characteristic U-shape. The initial drop is explained by the low number of enzymes within the evolving network and the restricted choice of new functional enzymes; the increase late in the process results from the fact that only those sequences

remain unattached which have no noticeable sequence similarity to any other sequence. For smaller γ enzymes are picked at random and the pure increase in network size results in a lower average minimal distance.

3.6.2. Dynamic Bursting in Evolution

The invention of new classes of enzymes often goes along with a completely new sequence structure and may open a new branch in the evolutionary process. Such a novel enzyme can have deep impact on the evolutionary dynamics, since once a new reaction is found, similar reactions may evolve in close temporal neighborhood. Hence, a strong sequence dependency is expected to lead to a bursting like behavior of enzyme attachment. This would reflect the principle of punctuated equilibrium at a molecular level [49, 10, 50]. In the framework of punctuated equilibrium, new species do not appear gradually at equally spaced time points but in rapid successions followed by silent intervals. We exploit the capability of the current model to investigate the evolutionary dynamics to test if the sequence dependency of the network expansion may substantiate this hypothesis.

First, we determine the appearance times for a new enzyme as a function of γ as shown in Fig. 3.2A. Here the invention of 500 enzymes (enzyme time 1500–2000) is plotted for 3 different values of γ by a vertical black line. Again it is obvious that the sequence dependence leads to an acceleration of evolution as can be seen by comparing the normalized time of each panel. A closer view reveals a more homogeneous structure for smaller γ. For $\gamma = 0$, the 500 events are rather homogeneously distributed over time with only few gaps. For $\gamma = 20$, the number and size of visible silent intervals increases because the 500 enzymes are invented more clustered. In case of very strong sequence dependence with $\gamma = 100$, the dynamics exhibit an even stronger clustering of events. The bursting dynamics leads to relatively large inter-cluster distances and subsequently to short intervals within an enzyme-class cluster because the number of enzymes is constant for all three scenarios, see Figs. A.2 and A.3.

These observations are subsumed in Fig. 3.2B where the logarithm of the frequency of inter-enzyme intervals (IEI) in normalized time are plotted. First of all, larger γ lead to shorter IEI in normalized time, corresponding to faster evolutionary dynamics. The bursting like behavior leads to multiple peaks in the distribution for larger γ and a flat plateau for $\gamma = 100$ which has similarly been observed in earlier models [28, 141]. Summarizing, the clustered appearance of new sequences hints at a potential molecular principle associated with punctuated equilibrium dynamics. Interestingly, our distributions deviate from previous studies about self-organized criticality [1, 196]. The differences are probably caused by the different generating processes. While in the former investigations the number of possible events was unlimited, our model has a finite number of events since it is restricted to existing enzymes. This may be seen as a disadvantage of the model but at the same time it may reflect biological constraints such as a limited number of functional protein sequences.

For a further analysis of the evolutionary dynamics, we characterize the process in terms of the IEI by the Coefficient of variation C_v [34]. We use the resulting spike trains shown in Fig. 3.2A to determine the average IEI μ and the corresponding standard deviation σ and calculate the Coefficient of variation $C_v = \sigma/\mu$. For an unbiased evolution ($\gamma = 0$), we expect the characteristics of a Poisson process as a generating process since the time step determined by the Gillespie algorithm is independent of the history and purely random. A Poisson process leads to an exponential distribution of the waiting times [62] implying $C_v = 1$ [33].

Since we are interested in the temporal characteristics of the expansion, we use a sliding window of 100 enzymes to calculate C_v in dependence on the evolutionary steps. Indeed, the C_v for $\gamma = 0$ (black line) fluctuates around 1 as shown in Fig. 3.2C. Increasing the influence of sequence information by increasing γ, leads to systematically increased $C_v > 1$. This is a strong indicator for multiple characteristic time scales [33, 34]. These are given here on the one hand by the typical time to explore a new 'class' of enzymes, a slow process in which a novel sequence, unrelated to existing ones, evolves, and on the other hand by the characteristic time to invent an enzyme with a sequence similar to an already present one. For the shown window size of 100, the C_v for $\gamma = 100$ (blue) exhibits several

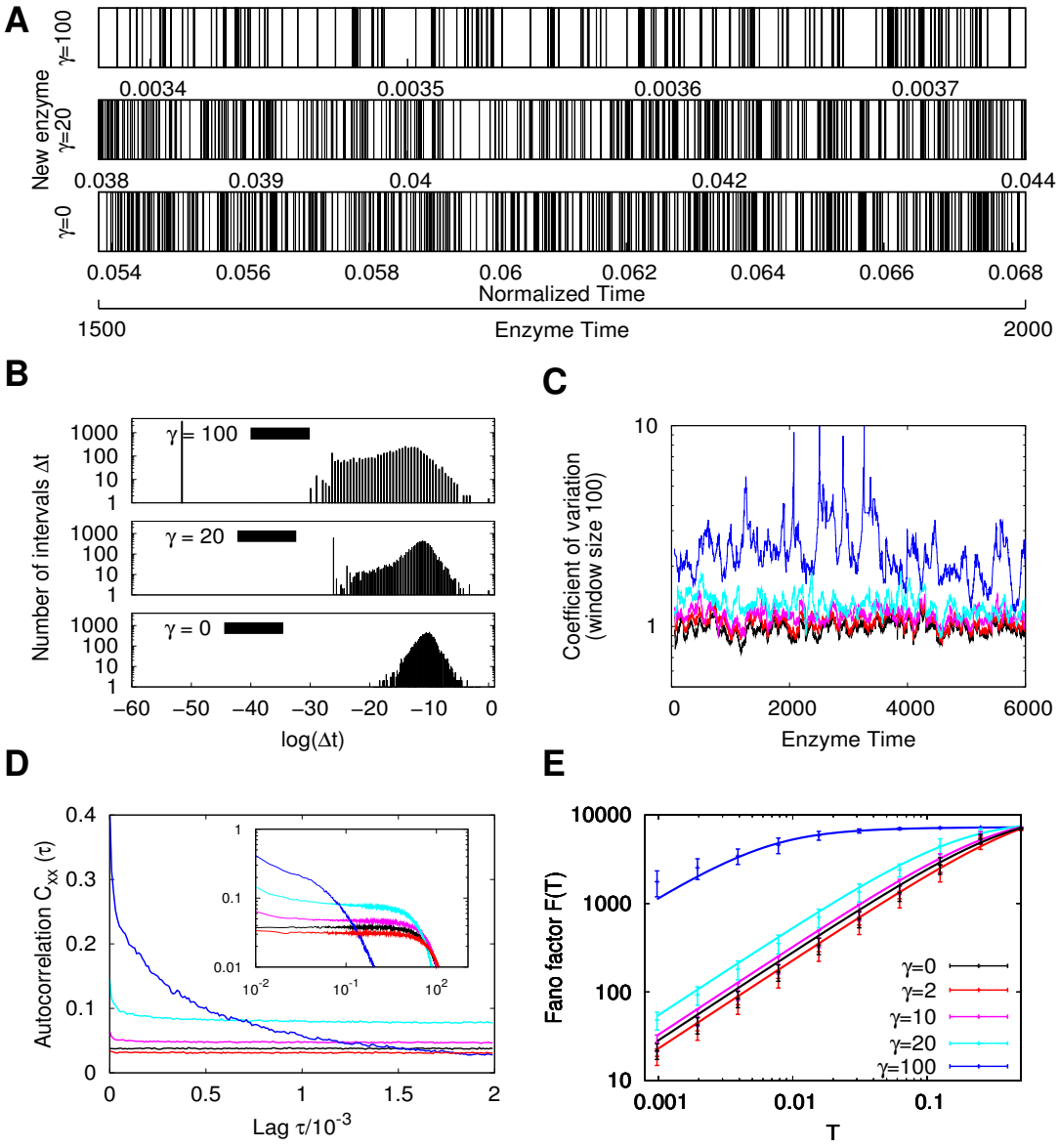


Figure 3.2: The acquisition of new enzymes happens in bursts of increasing strength for larger sequence sensitivity. Here we show one example run in **A**–**C** and means of 200 runs in **D** and **E**. **A**: Spike train with one bar at every incident of a new enzyme. The panel shows a window of 500 new enzymes for each γ on its particular normalized time. While for $\gamma = 0$ the enzymes appear almost equidistantly, larger γ leads to enzyme bursts. **B**: Distribution of time intervals between any two new enzymes (IEI). For higher γ the distributions are shifted to smaller distances and exhibit multiple peaks. **C**: The coefficient of variation, $C_v = \sigma/\mu$, measured in sliding frames of 100 enzymes indicates multiple characteristic time scales. The peaks point to times of evolutionary explosions. **D**: The autocorrelation $C_{xx}(\tau)$ of IEIs supports the bursting behavior further. For large γ IEI are strongly correlated on a short time scale whereas small γ lead to no significant correlation. **E**: The fit of the data to the Fano factor of biased Brownian motion enables to estimate the correlation time τ_{corr}. (For all color panels the legend of panel E holds.)

γ	C_v	$D/10^5$	τ_{corr}	$D \cdot \tau_{\mathrm{corr}}/10^3$
0	1.14 ± 0.03	0.57 ± 0.05	0.20 ± 0.03	11.4 ± 2.0
2	1.2 ± 0.04	0.46 ± 0.04	0.29 ± 0.06	13.3 ± 3.0
10	1.4 ± 0.07	0.66 ± 0.06	0.16 ± 0.03	10.6 ± 2.2
20	1.9 ± 0.1	1.1 ± 0.08	0.081 ± 0.009	8.9 ± 1.2
100	4.8 ± 0.8	25.6 ± 2.3	0.0028 ± 0.0003	7.2 ± 1.0

Table 3.1.: Coefficients of variation and parameters of Fano factor fits averaged over 200 runs. The coefficient of variation is measured in the domain of the first 6000 enzymes. The data is fitted to the Fano factor Eq. (3.4) via parameters diffusion coefficient D and correlation time τ_{corr}.

peaks and reaches values up to 10 indicating strong bursting. The peaks may hint at important points of evolutionary explosion.

This analysis is further confirmed by the comparison of C_vs determined with different sliding window sizes (compare Figs. 3.2C and A.4). The comparison clearly demonstrates that the peaks of C_v are not an effect of the limited window size since even for larger window sizes the C_v reaches comparable values Fig. A.4 and exhibits peaks. In Table 3.1 the systematic increase of the asymptotic C_v with increasing γ is given for all IEIs up to an Enzyme Time of 6000. This demonstrates the different characteristic time scales of the evolutionary process.

To substantiate this analysis we also calculated the autocorrelation function $C_{xx}(\tau)$ of the normalized IEIs for each γ as shown in Fig. 3.2D. For $\gamma = 100$ we observe strong correlations for small time lags τ indicating bursting. For unbiased evolution ($\gamma = 0$), no significant correlation on any time scale τ is observed which is in accordance with our assumed reason for bursting, the sequence information. In the inset of Fig. 3.2D, $C_{xx}(\tau)$ is plotted on a log-log scale. In this representation, the autocorrelation decreases linearly at the beginning as it is observed in other models of self-organized criticality [196]. But due to the limited enzyme pool size there is a strong cross over to the pure random behavior for large τ. Thus, our enzyme based model can quantitatively support the bursting dynamics of punctuated equilibrium.

While the Coefficient of variation allows for the analysis of dynamical variations on the scale of the average IEI μ, the Fano factor [52] characterizes variability in IEI on all accessible time scales T [123]. Therefore, the normalized time is divided in M non-overlapping windows and in each window the number of invention events N is determined. The Fano factor is defined as

$$F(T) = \frac{\langle N^2 \rangle - \langle N \rangle^2}{\langle N \rangle} \ , \tag{3.3}$$

where the time scale $T = T_{\mathrm{tot}}/M$ is given by the ratio between total time T_{tot} and the number of windows M.

For $\lim M \to \infty$, i.e. $T \to 0$, F equals 1. The dependence of $F(T)$ is shown in Fig. 3.2E and exhibits an increasing and saturating behavior. The increase is an indicator of long-range correlations. Because an increase is observed for all values of γ, these correlations are most likely a result of biochemical constraints given by the underlying metabolic network structure.

Since the analysis of the C_v has already shown the stochastic character of the expansion process, we hypothesize that the evolutionary process basically represents a diffusion process on the network. The evidence for long-range correlation suggests an Ornstein-Uhlenbeck process [198, 62] as approximative dynamics. For such a process the Fano factor can be expressed as [123]

$$F(T) = D \cdot \tau_{\mathrm{corr}} \left(1 - \frac{\tau_{\mathrm{corr}}}{T} \left[1 - \exp\left(-\frac{T}{\tau_{\mathrm{corr}}} \right) \right] \right) \ , \tag{3.4}$$

where D denotes a scaled diffusion coefficient and τ_{corr} is the correlation time. In order to characterize the dynamics on the network, we fit Eq. (3.4) to the Fano factor determined by Eq. (3.3) from simulations.

We find a very good agreement for all γ values as shown in Fig. 3.2E. From the fitting procedure, we can estimate the diffusion coefficients D and correlation times τ_{corr} for each γ. The acceleration due to the sequence information leads to an increase of D accompanied by larger C_vs. The correlation time decreases in the units of relative time. This is caused by the faster expansion for larger γ. In this case, the invention of a novel sequence, representing a new class of enzymes, triggers the discovery of related sequences in short evolutionary time and thus the correlation time is shorter. For smaller γ, new classes are invented before all enzymes with a similar sequence structure are included and thus correlation ranges over enzyme classes leading to larger τ_{corr}. From Eq. (3.4) we expect that the product $D \cdot \tau_{\mathrm{corr}}$ should stay rather constant what is verified in Table 3.1.

3.6.3. Appearance: Order of Enzymes, Compounds, and Organisms

From the observation of enzyme bursts and of the correlation time for large γ, one may expect that similar organisms appear at similar times. This would provide further understanding of punctuated equilibrium in organismic evolution. Following our previous results indicating bursting evolutionary behavior in the time series of new enzymes, we focus now on the possible biological and biochemical consequences of the appearance and order of metabolic compounds, enzymes, and even entire organisms.

Not all evolutionary paths are possible. Rather, the order of enzyme appearance is constrained by two factors. First, the selection criteria that only useful reactions are positively selected implies a chemical constraint. Some enzymes require other enzymes to be present since otherwise their required substrates could not be provided. The second constraint results from sequence similarity. It is conceivable that the sequence organization favors a certain order of enzyme evolution limiting large jumps in sequence space.

Our model allows us to distinguish between the biochemical and evolutionary constraints which have shaped the metabolic map. To achieve this, we determine for $\gamma = 10$ and $\gamma = 0$ all pairs of enzymes which appear in the same temporal order in all 200 runs, excluding the seed enzymes which appear by definition before all others. Ordered pairs found for $\gamma = 0$ can only result from biochemical constraints since in this case sequence information is ignored. To identify those ordered pairs which result as a consequence of sequence similarities, we remove the pairs found for $\gamma = 0$ from the pairs determined for $\gamma = 10$. The remaining ordered pairs define a tree with 7117 nodes and 1348709 edges. Since visualization of such a large tree is impractical, we concentrate on all paths of length three or higher from root to leaf node (see Figs. 3.3A and A.5 for a larger fraction of the tree).

Most of the enzymes on the first hierarchy level belong to essential pathways of central carbon metabolism. Enzymes in lower levels tend to belong to biosynthesis pathways of more specialized compounds. The tree gives insight into an enzyme's role in an evolutionary context. For example, enzymes 2.1.1.128 (a methyltransferase) or 1.2.1.38 (an oxidoreductase) appear only after a considerable number of precursors (5 and 32 respectively) have evolved. Apparently their sequences could have only evolved after many sequences for enzymes of central carbon metabolism had arisen. Interestingly, the opposite observation can be made for another methyltransferase, enzyme 2.1.1.116. The discovery of five enzymes directly dependent on the evolution of this particular sequence makes it plausible that this enzyme has presented an evolutionary bottleneck.

Highly important for the origin of life is the synthesis of amino acids as building blocks of proteins, and nucleotides for DNA and RNA. The amino-acids appear in good correlation (rank correlation 0.7) with previous results investigating the robustness of *E. coli's* network against reaction removal [30] (see Fig. 3.3B). This is not surprising and can be explained by stoichiometric effects. If more metabolic paths allow for the synthesis of a particular amino acid, it is likely to be discovered earlier. At the same time, one would expect that its production will be more robust against removal of reactions. The order of appearance of amino acids also reflects the commonly known biochemical synthesis pathways. Glutamate as a precursor of proline and arginine is synthesized first. In bacteria aspartate is the common precursor for lysine, threonine, and methionine. For all used γ values except $\gamma = 100$ this order is reproduced in the evolutionary scenarios. However, for $\gamma = 100$ threonine appears slightly before aspartate.

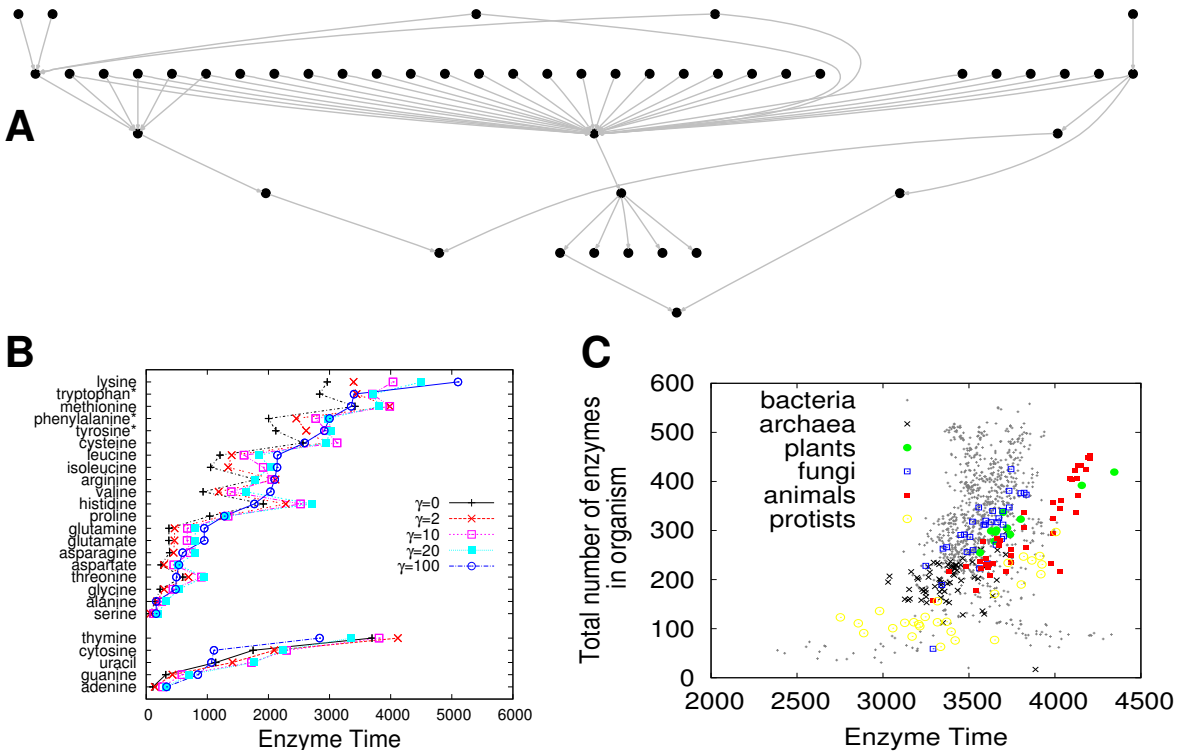


Figure 3.3: Time order of appearance of enzymes, amino acids and nucleotides, and entire organisms. **A**: Time-ordered ranking of enzyme appearance for $\gamma = 10$. From the graph of all time-ordered pairs of enzymes with $\gamma = 10$ pairs also appearing in the $\gamma = 0$-case are removed and only the paths of length 3 of higher are shown (order-precision 100%). Time runs from top to bottom; the seed enzymes as root nodes are omitted for simplicity. On the version on the enclosed CD every enzyme is linked to its entry in KEGG. **B**: Appearance of amino acids (top part) and nucleotides (bottom part) sorted by the $\gamma = 100$ appearance and averaged over 200 runs. The order is very similar (rank correlation 0.7) to the order of robustness observed in the *E. coli* network [30]. Further, aromatic amino acids (labeled by *) are synthesized late. The γ-curves look similar indicating that the order strongly originates from stoichiometry rather than from sequence relations. **C**: Every enzyme defined by its EC number is mapped to its genes and thus to the corresponding organisms. An organism is assumed to have evolved if 80% of its annotated enzymes are discovered. The x-axis depicts the mean enzyme time of birth of a new organism while the y-axis shows the size of the organisms given by the enzyme repertoire. For higher organisms, the appearance time correlates well with the size of the organisms but this is not the case for bacteria and archaea. See Table A.2 for a list of all organisms and the appearance time.

The pyruvate family of leucine, isoleucine, and valine is detected in close proximity. Furthermore, the aromatic amino acids, phenylalanine, tryptophan, and tyrosine, labeled by asterisks appear rather late (position 16, 17, and 19 for $\gamma = 100$) as a result of their more complex chemical structure.

Additionally, we investigate the relationship between the simplicity of synthesis and the actual usage of amino acids. For this, we compare the time of appearance to the frequency of the amino acids in the enzyme sequences of our consensus set and find a significant correlation for $\gamma = 100$ (Spearman 0.51, p-value 0.02, see Table A.1). The fact that metabolites detected earlier in evolution are cheaper to synthesize, supports the hypothesis that cost minimization is an important factor for amino acid usage in protein synthesis.

Different organisms have different specialized metabolic networks, which depend on their resources and living environment. Studying when the metabolic networks of various species have evolved could help refine and understand the tree of life. Clearly, the discovery of a complete set of metabolic reactions for a given organism in our evolutionary simulation does not necessarily reflect the organism's appearance during evolution. However, it presents a prerequisite for the emergence of the corresponding metabolism. Fig. 3.3C presents our simulated discovery of the metabolic enzymes of 1097 organism-specific networks retrieved from the KEGG database. The size of the networks is plotted versus the average enzyme time at which 80% of an organism's enzymes were found ($\gamma = 10$). For higher organisms the enzyme time of appearance correlates well with the network size (Pearson correlation: animals=0.88, plants=0.95, fungi=0.75, protists=0.77, archaea=0.055, bacteria=0.25). Also, similar organisms tend to appear closely together, see Table A.2 and following. For example, eight species of *Drosophila* occur from enzyme time 3571 to 3639, six *Plasmodium* species from enzyme time 3179 to 3317, or seven *Mycoplasma* from enzyme time 2956 to 3171.

3.7. Conclusion

We developed a model of metabolic evolution based on a Systems Biology approach that combines experimental data, bioinformatic tools, modeling techniques, and time series analysis. Starting from an initial seed of prebiotic metabolites and from a set of simple enzyme sequences exhibiting a large amount of conserved proteome fragments, we simulated the expansion of the metabolic network by iterative invention of novel enzymes and addition of allowed metabolites. We focused on the role of sequence information as a source of evolutionary memory. The assumption that new enzymes with a sequence similar to already explored enzymes have a higher probability to appear was implemented by the use of the inverse BLAST-based enzyme distance, Eq. (3.1), determining the corresponding propensities of the Gillespie algorithm. For a quantitative analysis, the propensities were scaled by the power of the weighting factor γ, where $\gamma = 0$ corresponds to a pure random invention process and large γ to a highly sequence similarity dependent process, respectively. Previous models have explicitly investigated the effect of gene duplications and identified these events as important components of evolutionary innovation [143, 73]. Including gene duplications in our model is difficult because we mimic the evolution of enzymes at the ecosystem-level, rather than at a single species level. We assume that the probability to evolve one gene from another is determined by sequence similarity. This dependence holds regardless of whether one assumes gene duplications as an underlying mechanism or not. In this respect, different assumptions about the frequency of gene duplication events could be reflected in our model by applying different functional dependencies of the propensities on the sequence similarities.

The model generates temporal network dynamics, where the sequence-similarity driven expansion process leads to an acceleration of evolution. The implemented process of mutation and selection may be seen as a concrete realization of the network-based reconciliation [203]. In this framework, neutral evolution with mutations which do not lead to new phenotypes are combined with positive selection. From this conceptual model, we expect that evolutionary changes often occur in cycles of neutral diversity expansion and selective diversity contraction leading to a boom and bust behavior which was shown in simulations of RNA evolution [57] and experimentally by the analysis of the evolution of the human influenza virus antigen *hemagglutinin* [104, 173]. A phylogenetic analysis of *hemagglutinin* has

revealed multiple short evolutionary branches corresponding to accumulation of neutral diversity.

This kind of behavior can be observed in our model for metabolic evolution as well. Here, neutral mutations occur whenever a new sequence is added that codes for an enzyme which is already present in the network. Sequence information leads to a bursting like behavior, where enzymes of one class are invented within short intervals whereas discovery of a new enzyme class needs more and larger mutations and thus occurs only rarely, confirming a boom and bust behavior. Therefore our model gives a first molecular description of punctuated equilibrium in metabolic evolution. This is quantified by the coefficient of variation C_v which increases with increasing sequence information dependency of the expansion process. Using a sliding window for the calculation of C_v indicates events of evolutionary explosion. A high autocorrelation function of the IEIs for small time lags in the case of large γ provides further evidence for the bursting behavior. Moreover, the good agreement of the Fano factor with the analytical result of biased Brownian motion points to the diffusive character of network evolution and allows for an estimation of typical correlation times of the evolutionary process. High sequence dependence leads to shorter correlation times, since once a functional sequence is found, all directly related enzymes are invented as well.

From our simulations, we could extract typical time orders of enzyme appearances which start with carbon metabolism that is needed for all subsequent processes. Although the model is rather elementary and neglects putative transient enzyme sequences or metabolites, the obtained order of amino acid appearance fits nicely with their robustness. This illustrates that many evolutionary paths lead to the development of simple but essential building blocks, whereas complex structures which depend on the previous discovery of simpler ones occur later. Interestingly, mapping enzymes to organisms by the EC number leads to results that match the intuition, despite the simplifying assumptions made. Bacteria appear as first species and plants and animals rather late in the artificial evolution. Moreover, occurrence time and complexity are strongly correlated for animals and plants.

Since the aim of our model is not to explain the early origins of metabolism, we assume that catalysts have already evolved and neglect mechanisms for the production of the enzymes themselves. As a consequence, we do not expect the model to produce realistic evolutionary paths for these early stages. Only after a core metabolism was assembled and the protein synthesis machinery has evolved, a closely intertwined coevolution of metabolites and enzymes can be seen as plausible. Moreover, the agreement of our mathematical model with phenomenological observations supports the relevance of our framework, which enables a quantitative description of metabolic evolution. Our approach might be used in future work for gaining deeper insights into enzymatic functions and their role in interactions between different species.

3.8. Acknowledgments

We acknowledge financial support, in particular for a five-months research visit at Boston University, from the International Research Training Group *Genomics and Systems Biology of Molecular Networks* IRTG 1360 (MS) We further acknowledge the German Federal Ministry of Education and Research, Systems Biology Research Initiative *Go*FORSYS (AS), the Scottish Universities Life Science Alliance SULSA (OE), the NASA Astrobiology Institute (NNA08CN84A), and the US Department of Energy (DS). We would like to thank Nils Christian, Zoran Nikoloski, and Thomas Handorf for useful discussions, and Anu Jayaraman for initial explorations of the enzyme-metabolite co-evolution approach.

4. Assembly of an Interactive Correlation Network for the Arabidopsis Genome Using a Novel Heuristic Clustering Algorithm[†]

4.1. Abstract

A vital quest in biology is comprehensible visualization and interpretation of correlation relationships on a genome scale. Such relationships may be represented in the form of networks, which usually require disassembly into smaller manageable units, or clusters, to facilitate interpretation. Several graph-clustering algorithms that may be used to visualize biological networks are available. However, only some of these support weighted edges, and none provides good control of cluster sizes, which is crucial for comprehensible visualization of large networks. We constructed an interactive coexpression network for the Arabidopsis (*Arabidopsis thaliana*) genome using a novel Heuristic Cluster Chiseling Algorithm (HCCA) that supports weighted edges and that may control average cluster sizes. Comparative clustering analyses demonstrated that the HCCA performed as well as, or better than, the commonly used Markov, MCODE, and k-means clustering algorithms. We mapped MapMan ontology terms onto coexpressed node vicinities of the network, which revealed transcriptional organization of previously unrelated cellular processes. We further explored the predictive power of this network through mutant analyses and identified six new genes that are essential to plant growth. We show that the HCCA-partitioned network constitutes an ideal "cartographic" platform for visualization of correlation networks. This approach rapidly provides network partitions with relative uniform cluster sizes on a genome-scale level and may thus be used for correlation network layouts also for other species.

4.2. Introduction

The complete, or partial, genome sequences from a vast number of organisms have increased our understanding of the design principles for biological systems [101]. The sequence availability has also provided platforms for various omics technologies, including transcriptomics, interactomics and proteomics [158, 111, 9]. Such techniques have generated an immense amount of data that for the most part are publicly available. One of the central ideas behind the concept of systems biology is to utilize these types of data sets to reveal functional relationships between genes, proteins, and other molecules [101]. Transcriptional coordination, or coexpression, of genes may uncover groups of functionally related genes [39, 81, 24, 142, 210, 195]. Such relationships were initially utilized to reveal functional gene modules in yeast and mammals [81] and to explore orthologous gene functions between different species and kingdoms [182, 18]. Comparable studies have also been undertaken in plants [24, 142, 74]. In addition, several Web-based tools for plants offer various forms of coexpression analyses. These include CressExpress [179], ATTED-II [139], Arabidopsis Coexpression Data Mining Tools [116], Genevestigator [220], GeneCAT [132], CSB.DB [180], CoreCarb [133] and Expression Angler of the Bio-Array Resource [193]. These tools can provide coexpressed gene lists for user specified query genes and thus represent user-friendly web resources for biologists.

While it appears useful for scientists to examine these types of coexpression lists, more information

[†]Published as:
M. Mutwil, B. Usadel, M. Schütte, A. Loraine, O. Ebenhöh, S. Persson, *Assembly of an interactive correlation network for the Arabidopsis genome using a novel heuristic clustering algorithm*, Plant Physiology 152 (1): 29–43 (2010)

is generally acquired by visualizing the relationships in the form of networks [87]. Several studies have explored the properties of such network assemblies [81, 16, 122, 115]. The distribution of connections in the networks may generally be described by power-law-related relationships (i.e. a small number of nodes appear to have a large number of connections while most nodes have very few connections [2]). Another apparent feature is that essentiality correlates with high connectivity in both coexpression and protein-protein interaction networks in several species [18, 83, 27], although this relationship is less clear in mammalian protein-protein interaction networks [61, 221].

Although features of coexpression and protein-protein interaction networks have been investigated, the output is generally not very useful for visual inspection and interpretation. One major task, therefore, is to make the networks more accessible to biologists (i.e. to produce visualizations of networks that may easily be interpreted [5]). For genome-scale networks, this requires dividing the network into smaller manageable units, or clusters. Such clustering, however, may artificially assign genes to certain clusters and therefore skew the output of the biologically "correct" network. It is important, therefore, to maintain as many relevant biological relationships as possible despite division. The ideal number, or size, of clusters to maintain these relationships is very rarely known and is generally very difficult to predict for biological networks. On the other hand, biological networks may also be viewed as clusters within clusters (i.e. as a hierarchical structure that can be viewed on different levels). For example, genes associated with photosynthesis may be viewed as a cluster that belongs to a supercluster of genes associated with functions in the chloroplast. Thus, the ideal clustering algorithm, and subsequent visualization scheme, should generate partitions of manageable sizes that can be readily reconnected into a whole network to be used for manual inspection.

Several graph-clustering algorithms are available, for example Markov Clustering (MCL) [197], Restricted Neighborhood Search Clustering [100], MCODE [8], and others, such as the recently published CAST algorithm [80, 200], but none of these can efficiently control cluster sizes. While these partitioning methods provide useful layouts for global biological and clustering interpretations, they are not particularly useful for visual inspection. To overcome this problem we developed a novel Heuristic Cluster Chiselling Algorithm (HCCA) and employed it to construct an interactive correlation network for the Arabidopsis (*Arabidopsis thaliana*) genome (Arabidopsis Gene Network [AraGenNet]; http://aranet.mpimp-golm.mpg.de/aranet). We show that the HCCA-generated cluster solutions were as good, or better than, the commonly used partition algorithms Markov, MCODE, and k-means using real world data. We also show that this type of visualization may reveal biological relationships that are not apparent from single gene coexpression approaches. Finally, we explored the network surroundings to identify essential Arabidopsis genes and present six new genes that are essential for plant growth through mutant analyses.

4.3. Results and Discussion

4.3.1. Calculation of Pearson-Based Correlation Networks

To generate a starting network for the HCCA, we calculated the degree of transcriptional coordination between all the genes present on the Arabidopsis ATH1 array (22,810 probe sets) using 351 Robust Multi-array Average (RMA)-normalized microarray data sets from The Arabidopsis Information Resource (TAIR). Prior to choosing these data sets, we removed data sets that displayed poor replication between arrays [132]. Since it is rather difficult to assess whether lowly expressed genes represent noise or real data, we chose to include all probe sets in the analysis. We then calculated an all-versus-all coexpression network matrix using a Pearson correlation coefficient cutoff of 0.8. In contrast to Spearman correlation, Pearson correlations only capture linear relationships between any two given components. However, it is anticipated that most linked expression profiles will adhere to a linear relationship [35].

To assess whether the topology of the obtained Pearson correlation network for Arabidopsis also followed such a relationship, we calculated the node degree distribution of all individual nodes in the net-

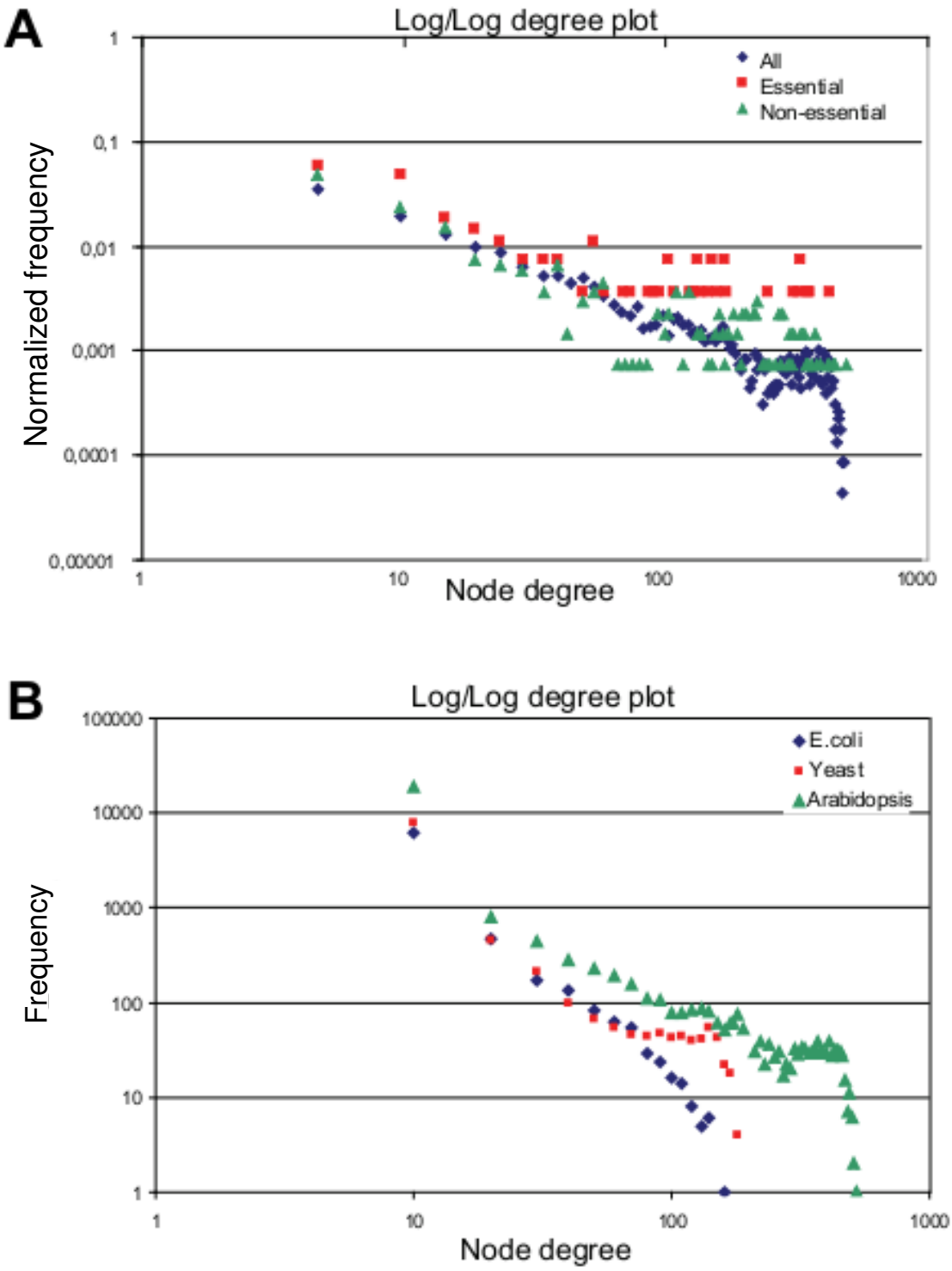


Figure 4.1: Network characteristics and mutant analyses. **A**. Log-log plot of node degree distribution for 261 essential (red points), 1224 non-essential (green points), and all genes (22810 blue points) in the Pearson correlation network ($r \geq 0.8$) for Arabidopsis. **B**. Log-log plot of node degree distribution for Pearson correlation networks ($r \geq 0.8$) from *E. coli* (blue), yeast (red), and Arabidopsis (green). The x axis represents the node degree (i.e. the number of connections a node holds), and the y axis displays the frequency (i.e. the number of genes [**B**]) or the normalized frequency (i.e. the normalized number of genes [**A**]) showing this degree.

work. Figure 4.1A shows that the node degree distribution is best described by a truncated power-law behavior. We also observed similar deviations from classical power-law behavior in Pearson correlation networks generated for yeast (*Saccharomyces cerevisiae*) and to a lesser degree for *Escherichia coli* (Fig. 4.11B), in agreement with recent reports [199].

4.3.2. Centrality vs. Essentiality

Another apparent feature in biological networks is that essentiality typically correlates positively with high node degree (i.e. mutations in highly connected nodes tend to result in more severe phenotypes compared to less well connected nodes [2, 83, 27, 221]). To assess if this type of relationship also is evident in our Pearson correlation network, we analyzed gene connectivity versus embryo lethality. We did this by linking phenotypic data from TAIR (www.arabidopsis.org) to the genes in our Pearson-based network ($r = 0.8$). Figure 4.1A shows the node degree distribution of embryo-lethal genes, genes associated with any type of phenotype, and all genes included on the ATH1 microarray. Whereas the node degree distribution for genes associated with nonlethal phenotypes did not deviate significantly compared with all genes present on the ATH1 gene chips (Fig. 4.1A), genes corresponding to embryo lethality were significantly more connected compared with nonessential genes (Fig. 4.1A; Fig. B.4B; $P < 0.05$). Similar observations have also been reported for coexpression and protein-protein interaction networks in yeast [2, 27].

4.3.3. Construction of a Highest Reciprocal Rank-Based Correlation Network in Arabidopsis

Several studies have used r value cutoffs ranging between 0.6 and 0.8 to depict coexpression correlations (for example [199]). However, different genes have different distributions of r values (i.e. at a given cutoff, some genes may correlate significantly with hundreds of genes while other genes may not correlate with any). Despite this, it is still possible that the latter may hold biologically relevant relationships. For example, the two transcription factors MYB33 (At5g06100) and MYB65 (At3g11440) regulate pollen and anther development, are expressed similarly, and are functionally redundant [124]. However, an r value cutoff of 0.8 did not associate these genes transcriptionally ($r = 0.7$; data not shown; [132]). To minimize this problem we chose to normalize the r value distributions in the calculated Pearson correlation networks using highest reciprocal rank (HRR) as they define the mutual coexpression relationship between two genes of interest. Using this approach, the MYB33 and MYB65 were readily transcriptionally linked (mutual average rank=2 using GeneCAT; [132]). With this approach, we were also able to define a connection cutoff, or maximum number of connections, for a given gene. The importance of defining such cutoff is apparent when looking at the distribution of r values among the data. For example, approximately 1500 genes are only expressed in pollen (estimated from GeneCAT; [132]). All of these genes are correlated with each other with an r value of 0.8 and should therefore be connected to each other in a Pearson-based correlation network [122]. However, it is virtually impossible to retain any information from such network structure through manual inspection. Instead, we argue that displaying these genes in close network vicinities, which is achieved by the HRR-based network, is more useful. In addition, recent results indicate that correlation rank based networks produce sounder results than networks based on correlation coefficients [139].

We set the HRR limit to 30, thus capping the maximum number of edges per node to 30. The resulting HRR network seemed a reasonable compromise between readability and richness of information. In addition, we defined three degrees of coexpression weights using highest reciprocal ranks of 10, 20, and 30 [132]. Similar approaches have also been used by several coexpression Web tools, such as GeneCAT and ATTED-II [132, 139]. The resulting weighted HRR network contained 103,587 edges between 20,785 nodes, and was used as the starting network for the HCCA. As anticipated, not all the probe sets shared strong correlation with other probe sets, resulting in 2,025 nodes that were not included in the network (data not shown). The HRR based network shared 29,956 edges and 6942 nodes with the Pearson-based coexpression network using $r \geq 0.8$ cutoff (231882 edges, 7178 nodes).

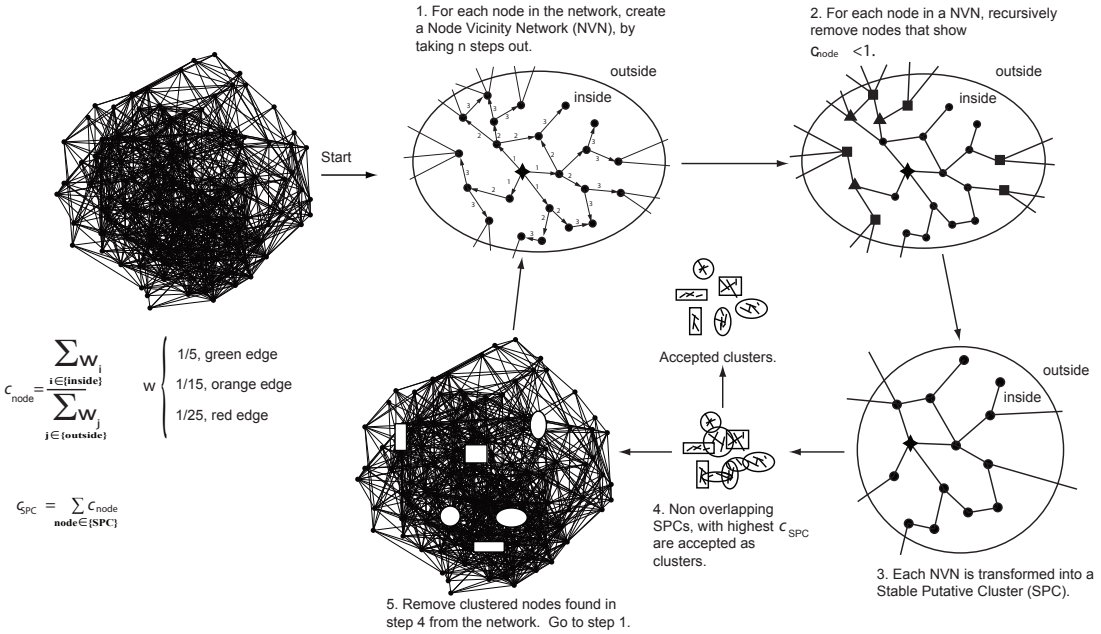


Figure 4.2: Schematic work flow of the HCCA (n = 3). The HCCA accepts a network as input. Step 1, Each of the nodes in the network is used to generate NVNs by taking n steps away from a seed node (indicated as a star). Step 2, Each NVN is then ”chiseled” by recursively removing nodes that have higher connectivity to nodes outside of an NVN than to nodes inside the NVN. In this example, squared and triangular nodes are removed in the first and second rounds of chiseling, respectively. Step 3, The chiseling either completely depletes a NVN of nodes or produces a stable putative cluster (SPC). Step 4, Nonoverlapping SPCs with highest c_{SPC} values are extracted and accepted as clusters. Step 5, Nodes that were accepted as clusters in step 4 are removed from the network. The remaining network is then transferred to step 1 and rechiseled (steps 2–5).

4.3.4. Designing the HCCA

Genome-scale coexpression networks, like other networks, consist of nodes and edges that may form a continuous structure or separate islands of clusters, depending on what cutoff one uses. While the smaller structures in such networks may be suitable for visual inspection, other regions may not be due to the number of nodes and edges in these regions. To make such regions more accessible, it is necessary to partition the network into smaller units, or clusters. Obviously, such partitioning will lead to a division of network structures that may, or may not, reflect the “real” network properties. Most biological networks do not contain sufficient data to assess whether the divisions are justifiable or not. However, the flaws in network divisions may be overcome if the different partitions can be reassembled into the structures they were initiated from. We argue that if we can visualize individual network partitions, or clusters, and put these into context with other clusters then the connectivity between the individual clusters may reflect the larger structures that were partitioned.

Many graph-clustering algorithms do not support weighted edges and do not yield cluster sizes that readily allow visual interpretations. In addition, many graph-clustering algorithms do not allow clustering of large networks (i.e. networks consisting of several thousand nodes). Therefore, we developed a novel graph-clustering algorithm (Fig. 4.2) referred to as HCCA. The HCCA algorithm takes step size (n) and desired cluster size range as parameters. The HCCA accepts a network as starting point (Fig. 4.2). For each node in the network, the algorithm generates node vicinity networks (NVNs) by collecting all nodes within n steps away from the seed node. Nodes with higher connectivity to the outside of the NVN are

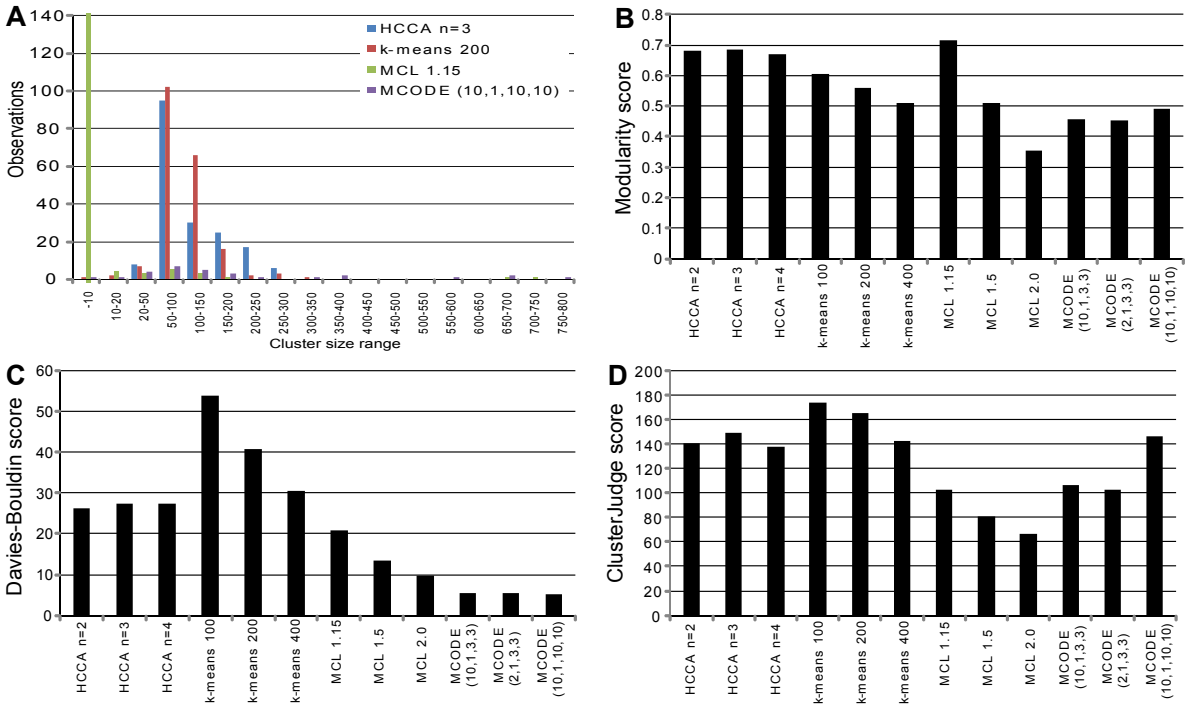


Figure 4.3: Cluster comparison of HCCA, MCL, k-means, and MCODE. **A**, Graph displaying the cluster size range (x axis) versus number of clusters (y axis; observations) for selected HCCA, MCL, k-means, and MCODE partitions of the HRR network (HRR cutoff = 30). **B**, Modularity scores for different settings for the HCCA, MCL, k-means, and MCODE algorithms. k-means 100, 200, and 400 represent desired cluster number parameters for k-means; MCL 1.15, 1.5, and 2.0 represent different inflation degrees for the MCL; HCCA n = 2, 3, and 4 represent different step size (n) as described in Figure 4.2; MCODE (A, B, C, and D) represent degree cutoff, node score cutoff, k-core, and maximum depth, respectively. High modularity values represent better clustering. **C**, Davies-Bouldin score, or index, for different settings for the HCCA, MCL, k-means, and MCODE. The settings are in accordance with **B**. Low Davies-Bouldin score represents better clustering. **D**, ClusterJudge scores of the clustering generated by HCCA, MCL, k-means, and MCODE, respectively. The settings are in accordance with **B**. High ClusterJudge score represents better clustering.

iteratively removed. The resulting clusters are then ranked by outside-to-inside connectivity ratio and filtered according to desired cluster size range. Nonoverlapping clusters are retained by the algorithm, and nodes in these clusters are removed from the network. Nodes associated with rejected clusters are returned to the network and reevaluated. The HCCA recursively creates nonoverlapping clusters until no nodes are left in the network or no more stable clusters can be obtained (Fig. 4.2). In the latter case, remaining nodes are associated with clusters to which they display the highest connectivity.

4.3.5. Visual Inspection of the Network Solutions

To partition the network, we used the HCCA with different steps (n) away from the seed node (Fig. 4.2) and desired cluster sizes ranging from 40 to 400. For example, for n = 3, the HCCA generated 181 clusters that contained approximately 40 to 300 genes per cluster (Fig. 4.3A). To assess the biological relevance of the partitioned network, we initially compared obtained connections with known biological data through visual inspection. For example, the secondary cell wall cellulose synthase genes CESA4, CESA7, and CESA8 have been used extensively for coexpression analyses [24, 142, 115].

In agreement with these analyses, we obtained genes associated with secondary cell wall synthesis, including IRX6, IRX8, IRX9, IRX12, and several transcription factors that recently have been implicated in secondary cell wall regulation [218], in the network vicinity of the three CESA genes (Fig. B.1).

4.3.6. Estimates of Clustering Solutions

A few other graph-clustering algorithms also support weighted edge graphs, such as the commonly used MCL [122, 197, 51]. To estimate the quality of the clustering solution obtained by HCCA, we clustered the HRR network using the MCL algorithm with a range of different inflation values (Table B.1). In addition, we included clustering solutions for MCODE [8, 146], performed clustering using k-means with different settings [71], and then compared the results obtained from the HCCA with the different clustering solutions for the other algorithms (Figs. 4.3A to D, Table B.1). We used two different metrics to evaluate the clustering efficiency; the commonly used quantity modularity [135], which judges partitions by comparing inside-to-outside connectivity ratios, and the Davies-Bouldin index, which measures the compactness and separation of the obtained clusters [37]. Our HCCA approach yielded better cluster partitioning compared with the MCL, k-means, and MCODE in terms of modularity (Fig. 4.3B, Table B.1). In addition, the HCCA solutions were clearly better than all the k-means partitions in terms of the Davies-Bouldin index (Fig. 4.3C, Table B.1). However, the MCL and MCODE partitions rendered better Davies-Bouldin scores compared with the HCCA (Fig. 4.3C, Table B.1). While the best overall MCL solution was the MCL 1.15, it is important to point out that this partition contains cluster sizes in the range of two to 2,500 genes per cluster (Fig. 4.3A, Table B.2), and therefore is not useful for our purposes. These results show that the HCCA performed better than k-means in terms of modularity and Davies-Bouldin index, and scored comparable index numbers as MCL and MCODE in terms of modularity.

When considering modular networks, it is generally expected that neighboring nodes fulfill related functions, which also has been recognized in social networks [207]. Hence, ideally, one coexpressed gene cluster should contain genes associated with similar biological functions. Therefore, we also tested the overlap of MapMan ontology classes with the clusters generated by the HCCA, MCL, MCODE and k-means. We used an approach similar to ClusterJudge [63], which uses mutual information between clusters and MapMan ontology terms to score clustering quality [181]. In brief, this approach scores the overlap between the ontological terms and the clusters, then subtracts the mean score obtained for randomly assigned clusters, and divides this by the standard deviation (SD) of the random clustering solutions. Therefore, a score of 0 (or even negative scores) would indicate random biological categories and clusters, whereas higher scores (which have no upper bound) indicate better concordance between biological categories and clusters. Using this assessment the HCCA-partitioned networks scored better than all of the MCL and MCODE partitions and scored nearly as well as the solutions generated by k-means (Fig. 4.3D, Table B.1). It is important to note that the latter commonly used algorithm cannot generate clusters based on graphs but must use the original expression data, which has an inherent advantage compared with HCCA, MCODE, and MCL.

We have also investigated how HCCA performs on unweighted HHR network. The HCCA-generated partitions performed slightly better in terms of modularity and ClusterJudge score and much better in terms of Davies-Bouldin score compared with the other clustering algorithms (Table B.1). However, it is important to note that the HCCA partitions of unweighted networks produced several clusters exceeding the desired maximum cluster size of 400 (Table B.2). This is most likely due to the more detailed information retained in the weighted network. It should be noted that by lowering the c_{SPC} cutoff value (see Fig. 4.2 legend), it should still be possible to generate clusters within the desired cluster range using HCCA. Also, the number of clusters obtained from the unweighted network was smaller than for the weighted network (Supplemental Table B.2).

Taken together, these tests show that the HCCA partitions scored better than k-means, MCL, and MCODE in terms of modularity and Davies-Bouldin index and outperformed the MCL and MCODE solutions in terms of biologically relevant associations.

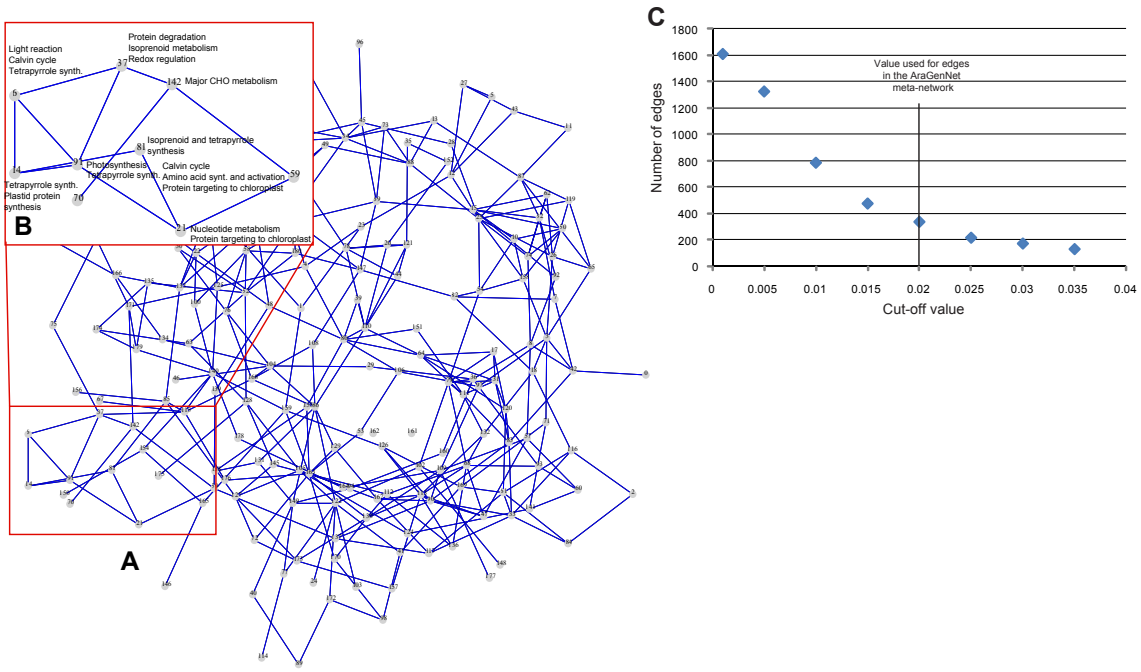


Figure 4.4: Meta-network of coexpressed gene clusters generated by HCCA (n = 3). **A**, Nodes in the meta-network, or assembled cluster-level network, represent clusters generated by HCCA. Edges between any two nodes represent interconnectivity between the nodes above threshold 0.02 (according to **C**). **B**, Enlarged region depicts part of the meta-network presumably associated with photosynthesis. Cluster annotations were inferred by MapMan terms, phenotypic, and expression data (http://aranet.mpimp-golm.mpg.de/aranet). **C**, Connectivity cutoff values $[c(A, B)]$ for edges in the meta-network. We used a cutoff of 0.02 for visualization purposes.

4.3.7. Comparisons of Partition Similarities

While the above results show that HCCA generated cluster solutions that are as good, or better than, MCL, MCODE, and k-means, the HCCA also produced clusters with relative uniform size (Fig. 4.3A, Table B.2) and therefore is well suited for cluster visualization for manual inspection. In contrast, the best performing MCL partitions resulted in cluster sizes between two and 2,500 genes (Fig. 4.3A, Table B.2), which is in good agreement with what has recently been reported [122]. Although the cluster size distribution between the different algorithms varied, we anticipated a relatively high overlap in cluster content between the different solutions. Therefore, we compared the overlap of genes associated with certain clusters for the HCCA, MCL, MCODE, and k-means solutions by adjusted Rand indices, which measure similarities between two clustering solutions (Figure B.5; [78]). Interestingly, each of the algorithms appeared to have generated clusters with different contents. For example, comparison of the HCCA (n=3) and MCL1.2 (inflation value = 1.2) solutions resulted in an average rand index of 0.2495 (identical partitions result in an index of 1; Figure B.5). However, these solutions contain different cluster sizes, which influence the outcome of the average Rand index. Comparing 1,000 k-means partitioned networks, each featuring 100 cluster centers, with a reference k-means network resulted in an average adjusted Rand index of 0.4, which is considerably lower than the index of 1 for identical partitions. Therefore, it appears that the seemingly low average adjusted Rand indices for the different solutions may in fact signify rather good agreement in cluster contents. The rather low values may be explained by unequal cluster size distributions, and by uncertain cluster partitioning for some of the genes.

4.3.8. Robustness of Clustering toward Node Removal and Different HRR Cutoffs

The ATH1 microarray chip contains 22,810 probe sets covering roughly 80% of the genes in the Arabidopsis genome. This means that approximately 5000 genes are omitted from the chip and, therefore, from our analysis. To assess whether omission of such a number of genes may significantly skew the connections in the HRR network, we randomly removed approximately 20% of the genes from our data sets and reclustered the network using HCCA. We repeated this 20 times and then assessed whether the clusters were significantly different by estimating the average adjusted Rand index. Figure B.5 shows that the average score for HCCA (n=3) was 0.3818, with only 4% SD. This value is similar to the value obtained for the comparison of 1,000 k-means clustering solution with 100 cluster centers. These data show that the network outline and HCCA clustering are robust against removal of a significant portion of randomly selected genes and therefore also should display biologically meaningful correlations despite the absence of some genes on the ATH1 chip.

To test how different HRR cutoffs influence the clustering by HCCA, we calculated adjusted Rand indices between networks generated using HRR of 10, 20, 30, 40, and 50. Table B.3 shows that the adjusted Rand index is relatively high (>0.4) for networks generated by similar HRR cutoffs (HRR20 versus HRR30, HRR30 versus HRR40, and HRR40 versus HRR50), despite the fact that the networks differ dramatically in the number of edges (Table B.3). Taken together, these results indicate that clusters obtained by HCCA are robust against the parameters used to generate the coexpression networks.

4.3.9. Construction of an Interactive Correlation Network for the Arabidopsis Genome

To illustrate the usefulness of the network partition obtained from the HCCA, we implemented an interactive coexpression network browser, which we named the AraGenNet (http://aranet.mpimp-golm.mpg.de/aranet). Since the aim of the visualization scheme was to reassemble the partitioned HRR network for manual inspection, the network works on two levels: on assembled cluster level (meta-network), and on the gene level (Figs. 4.4 and 4.5). The cluster-level network (Fig. 4.4) represents an overview of the interactions between different partitions, or clusters, and therefore depicts the coexpressed context for individual clusters. Therefore, we refer to this network as a meta-network. Any two clusters in the meta-network are connected if the combined weight of edges between them is larger than a certain threshold. We set this linkage threshold, or connectivity score, to 0.02, as this value produced a connection-rich

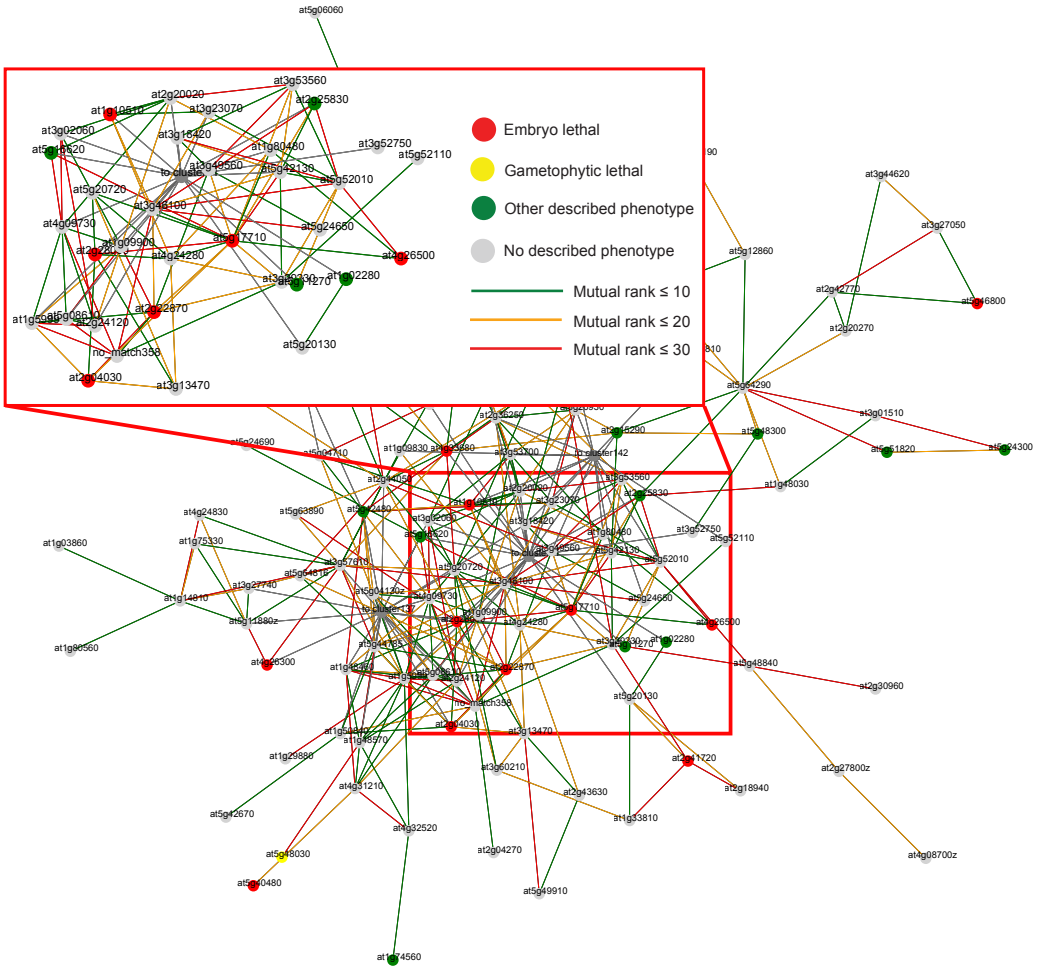


Figure 4.5: Features of HCCA (n = 3) gene cluster 59. Nodes in this cluster, or gene-level network, represent genes, while edges and edge coloration depict the HRR values between any two nodes. Red, yellow, and green node colors depict gene mutants displaying embryo-lethal, gametophyte-lethal, and other described phenotypes, respectively. Gray nodes represent genes with no described phenotype.

but readable meta-network (Figs. 4.4A and B). A node in the meta-network consists of a cluster of co-expressed genes generated from the HCCA (n=3; Fig. 4.5). This gene-level network becomes visible by clicking on a cluster node in the meta-network. All connections in the gene-level network are based on HRR, and are weighted accordingly (i.e. HRR below 10, 20, and 30 are color coded green, orange, and red, respectively (Fig. 4.5)). These visualization schemes prove the capability and functionality of the HCCA clustering approach.

4.3.10. Phenotype and Ontology Mapping onto Network

Since coexpressed genes often tend to be functionally related [39, 81, 24, 142, 210], we anticipated that connected clusters in the meta-network would share a certain degree of functional commonalities [58]. To assess this, we analyzed the genes in each cluster for MapMan ontology term enrichments. We also mapped phenotypic data (http://www.arabidopsis.org/) and tissue-dependent expression profiling for the individual genes. By combining these analyses, we then attempted to describe what biological functions are associated with the individual clusters. For example, mutations in genes associated with cluster 59 (Fig. 4.5) often result in embryo lethality or pale green plants. The dominant expression profile of genes in this cluster shows high expression in aerial tissues and low expression in roots, pollen, and seeds. MapMan ontology analysis revealed that the most significantly enriched term is amino acid metabolism ($P \leq 10^{-9}$). Taken together, these data suggest that cluster 59 is overrepresented for genes involved in amino acid metabolism in the chloroplast and that this function is important for chloroplast development, photosynthesis, and embryo development. This conclusion is supported by the fact that cluster 59 was highly enriched for genes with plastidic localization ($P < 0.001$; data not shown).

4.3.11. Prediction and Verification of Essential Genes in the Network

To expand the visual features of the network, we color-coded the severity of the phenotypic traits using red (embryo lethality), yellow (gametophytic lethality), and green (other phenotypes) nodes in the network (Fig. 4.5). Interestingly, we observed an uneven distribution of embryo-lethal genes per cluster compared with genes associated with nonlethal phenotypes (Fig. 4.6A). For example, the chloroplast-associated clusters 21, 59, and 137 showed strong enrichment for essential genes ($P < 10^{-5}$; Table B.4). This suggests that nodes in clusters associated with certain biological processes are more essential. For example, of the 111 genes associated with cluster 59, 12 are known to be essential for embryo development (Fig. 4.6A, Table B.4). As described above, this cluster may be associated with amino acid activation in the chloroplast.

We also investigated how the essentiality of a gene is determined by the number and the distances of its homologs in the network. Figure 4.6B shows that embryo-lethal genes are clearly overrepresented by single-copy genes ($P < 0.001$, Fig. B.2A). Furthermore, essential genes tend to be underrepresented for genes with family members in the network vicinity (i.e. in the node vicinity network ($P < 0.05$; Fig. B.2B-C)). Conversely, nonessential genes tend to be neighbors to their family members ($P < 0.05$; Fig. B.2E-F). Taken together, the probability of essentiality for a given gene appears to depend not only on the connectivity of the gene (Fig. 4.1A) but also on its functional uniqueness in the network vicinity and on its biological role. Interestingly, similar results have recently also been observed in protein-protein interaction studies in yeast (Zotenko et al., 2008). This study convincingly showed that essentiality corresponded to gene products that are well connected and that are associated with certain biological functions.

To explore the prediction of essentiality, we chose 20 genes associated with clusters that harbor numerous essential genes (i.e. the connected clusters 21, 59, and 137 (Figs. 4.6A and B.3)) and that are well connected to other essential genes in the network. We ordered T-DNA mutant lines corresponding to these genes and analyzed them for mutant phenotypes (Table 4.1). Out of the 20 mutant lines, two resulted in embryo lethality, one in seedling lethality, two in male gametophyte lethality, and one in dwarfed pale green plants (Figs. 4.6C to E; Table 4.1). Chlorotic cotyledon phenotypes are typically associated with chloroplastic functions (for example [56]), supporting our prediction that genes belonging

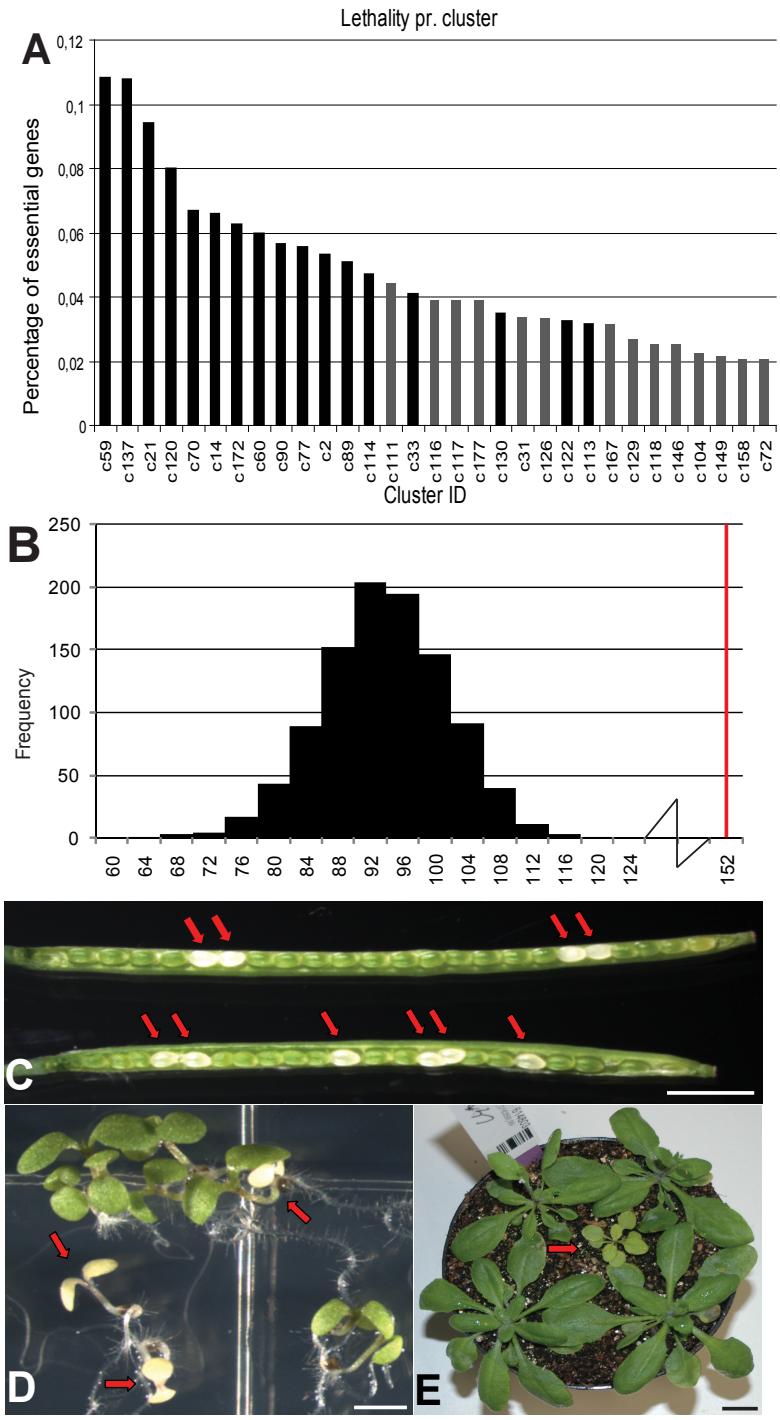


Figure 4.6: Essentiality distribution and mutant phenotypes in the HCCA (n = 3) partitioned network. **A**, The graph displays the relative distribution of essential genes per any given cluster in the network (HRR cutoff = 30). Black bars depict clusters significantly enriched ($P \leq 0.05$) for essential genes. **B**, Distribution of singlecopy genes from 1,000 samplings of 152 random nodes from the HRR network (black bars). Any given gene was referred to as being single copy if no close homolog was detected (score coverage threshold of 30 and length coverage of the protein of 70%). The observed 152 essential, single-copy genes are denoted by the red line. **C**, Siliques from a plant heterozygous for mutation in At3g14900 (cluster 137). Red arrows indicate chlorotic embryos. Bar=3mm. **D**, Mutant seedlings (At1g15510) from cluster 137 exhibiting pale cotyledons (indicated by arrows). Bar=3mm. **E**, Chlorotic dwarfed mutant (At3g57180; indicated by the arrow) from cluster 21. Bar=1cm.

Table 4.1.: *Characteristics of mutants*
Family size and family members in vicinity indicate the size of a gene family as defined by Clusters of Orthologous Groups of proteins and the number of family members in the gene network vicinity ($n = 2$), respectively.

Gene	T-DNA line	Phenotype	Family size	Family members in vicinity
At3g23940	SALK_069706	Gametophytic lethal	0	0
At1g74260	SALK_050980	Gametophytic lethal	0	0
At5g64580	SAIL_74_G12	Embryo lethal	0	0
At3g14900	SALK_123989	Embryo lethal	0	0
At1g15510	SALK_112251	Seedling lethal	182	38
At3g57180	SALK_068713	Pale green, dwarf	0	0

to these clusters (i.e. 21, 59 and 137) are functionally associated with the chloroplast. These results illustrate how a coherent and easy-to-navigate data visualization scheme, such as the AraGenNet, can predict biologically meaningful relationships. Recently, the pollen deficient mutant corresponding to the gene At1g74260 was confirmed by another study [19].

4.3.12. Associations of Functional Annotations Using MapMan Ontology

Although the visualization of coexpressed genes may give insight into functional gene patterns and arrangements, an equally relevant quest is to understand how these patterns and arrangements are organized to fulfill cellular functions. To investigate this, we explored the notion that coexpressed genes, and therefore network vicinities, often are functionally related [24, 142, 210, 81]. To assess how different ontological terms are transcriptionally connected, we used the nonclustered HRR network (HRR cutoff = 30) and calculated whether certain MapMan ontology terms were overrepresented in nonoverlapping node vicinities (NVNs in Fig. 4.2). We then identified terms that co-occurred more often than expected by chance ($P \leq 0.05$). These significantly associated terms were connected, and the resulting ontological network was visualized as an interactive network browser (Fig. 4.7; http://aranet.mpimp-golm.mpg.de/aranet/Mapman_network). To get a more complete network, we also retained connections representing parent-child relationships, which are trivial due to their mutual overlap. From this visualization, it became evident that terms that represent related processes tend to be connected; for example, photosynthesis-related processes (dark green) were connected to plastidial protein synthesis (light blue) and to "protein assembly and co-factor ligation", which comprises many proteins involved in the assembly of the plastidial apparatus (light blue). Furthermore, the chloroplast cluster (dark green) is closely associated with genes related to tetrapyrrole biosynthesis (light green; Fig. 4.7). These processes most likely reflect parts of the basal plastidial photosynthetic activity program. Other examples were mitochondrial processes linked to the tricarboxylic acid cycle cycle as well as polyamine synthesis being coupled to Arg degradation more than would be expected by the trivial link of Arg decarboxylase, which is present in both processes. Also arabinogalactan proteins were linked to abiotic stress, which is in line with their up-regulation upon salt stress [106].

Since biologically relevant associations were confirmed in the MapMan ontology network, we also investigated associations between other biological processes, which were previously unrelated MapMan terms and which might help to generate new functional insights. Interestingly, plant defensins were connected to sphingolipid biosynthesis in planta. As often the mode of action of plant defensins seems to be mediated by sphingolipids of the attacking pathogen [190, 191, 148], it could be speculated that plant sphingolipids might play a role in this mechanism as well. Furthermore, it might be interesting to investigate what caused the link introduced between aromatic amino acid degradation and starch breakdown (Fig. 4.7, lower left corner). Thus, the combination of coexpressed gene vicinities and ontology terms may similarly reveal new associations between different processes in the cell.

4.4. Conclusions

We have constructed an interactive correlation network for Arabidopsis using a novel HCCA. The cluster solutions obtained from this clustering algorithm performed as well as, or better than, the commonly used clustering algorithms MCL, MCODE, and k-means. More importantly, by visualizing the portioned clusters, we could reassemble the network; therefore, we were able to place the obtained partitions into larger biological contexts. We predicted that unique, well-connected genes with certain biological functions tend to be more essential than other genes and confirmed this by mutant analyses. The presented data, therefore, show that comprehensible visualization of genome-scale correlation networks may render new insights into the wiring of biological systems. We propose that this type of network visualization constitutes an easy-to-navigate framework for biologists to prioritize genes for functional analyses.

4.5. Acknowledgments

We would like to thank Ms. Christy Hipsley, and Drs. Chris Somerville, Alisdair Fernie, and Lothar Willmitzer for useful comments on the manuscript. We would also like to thank Mrs' Anja Fröhlich and Anett Döring for technical assistance.

4.6. Materials and Methods

4.6.1. Microarray Data

All calculations for this work were done using python and java scripts. Databases for Arabidopsis (*Arabidopsis thaliana*), yeast and *Escherichia coli* use Affymetrix ATH1 (22 810 probe sets), Affymetrix Yeast Genome S98 (9 335 probe sets), and Affymetrix Ecoli_ASv2 (7312 probesets) GeneChips, respectively. Arabidopsis microarray datasets consisting of 1428 ATH1 microarrays were obtained from TAIR (http://www.arabidopsis.org/). Separate Arabidopsis tissue atlas data sets containing 121 microarrays, which were used for plotting gene expression across Arabidopsis tissues, were generated by the AtGenExpress project [163] and were obtained from TAIR. The data was quality controlled by visual inspection of boxplots of raw positive match data and RMA residuals of RMA-normalized data using the RMA express program. Cel files showing artifacts on RMA residual plots or visibly deviating from the majority on the positive match box plots were removed from further analysis. In addition, we removed experiments representing very similar transcriptomic snapshots by iteratively discarding microarrays that displayed Pearson correlation [$r(A, B) \geq 0.95$] to more than three other microarrays. From these analyses, we retained 351 microarrays, which subsequently were normalized using R package simpleAffy. The 244 *E.coli* and 789 yeast microarray datasets used to generate Figure 4.1 were downloaded from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/), RMA normalized, and quality controlled as for the arrays for Arabidopsis. Names of the cel files used to construct Arabidopsis HRR networks are downloadable from the AraGenNet home page.

4.6.2. Phenotypic Data for Arabidopsis

Phenotypic data for Arabidopsis was requested and obtained from TAIR curators and were divided into essential, gametophyte lethal and nonlethal sets. All the expression data, coexpression network and phenotypic data presented in this work are downloadable from AraGenNet home page (http://aranet.mpimp-golm.mpg.de/aranet).

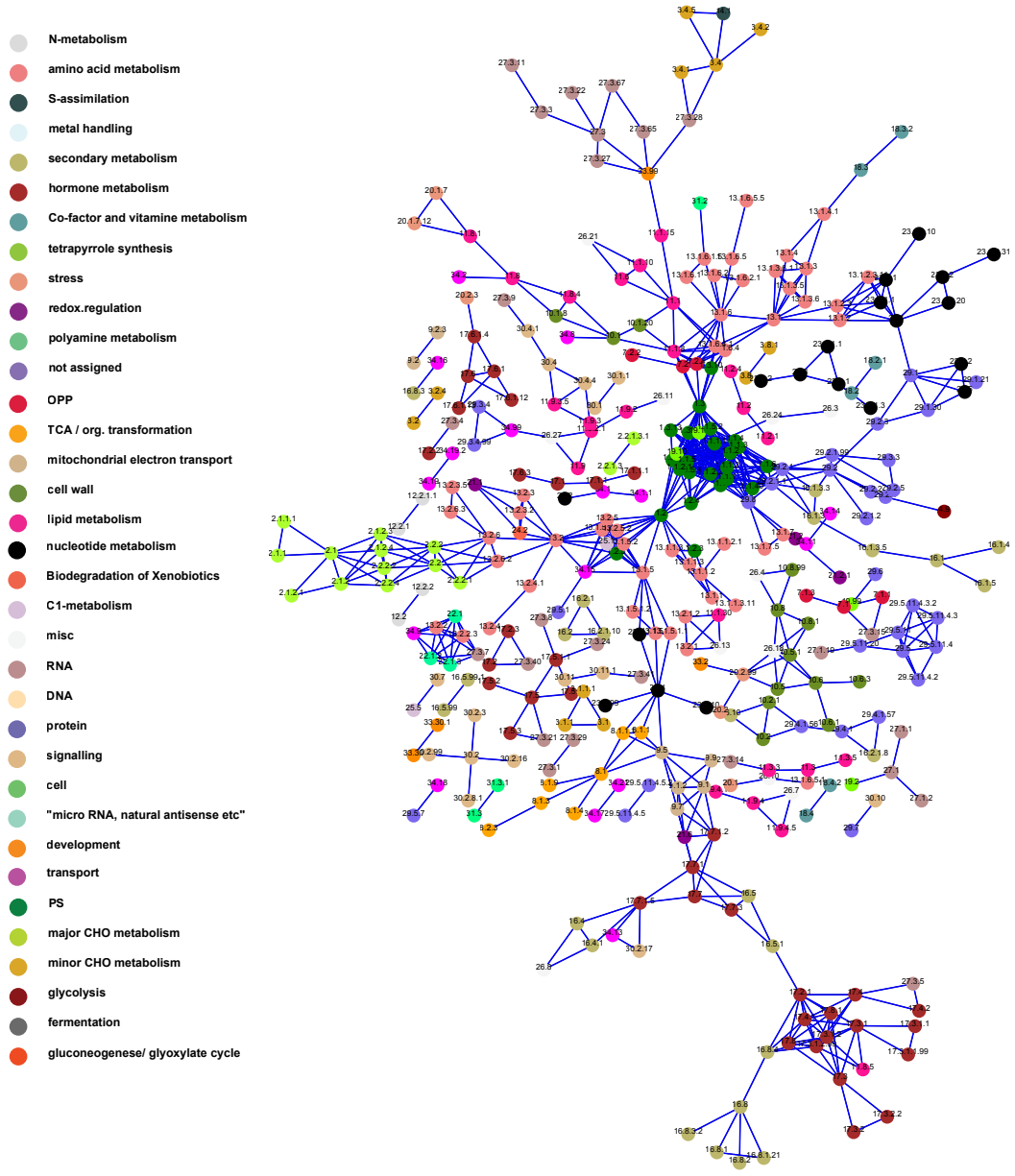


Figure 4.7: Network of coexpressed MapMan ontology terms. Nodes in this network represent biological processes as defined by MapMan ontology terms. Node colors and numbers depict the different MapMan terms (legend at left), while edges represent significant ($P \leq 0.001$) associations between the terms based on coexpression. OPP, Oxidative pentose pathway; PS, photosynthesis; CHO, carbohydrate.

4.6.3. Construction of Coexpression Networks

Pearson-based coexpression networks were used for the centrality-versus-essentiality study and for generating log-log plots. These networks were created using the 351 ATH1 microarrays described above. An edge in the network represents two genes with Pearson correlation [$r(A, B) \geq 0.8$]. All subsequent analyses were done on HRR-based networks, including the visualized interactive coexpression network used on the AraGenNet home page. HRR score between genes A and B is calculated according to:

$$HRR(A, B) = \max(r(A, B), r(B, A)) \tag{4.1}$$

where $r(A, B)$ is correlation rank of gene B in gene A's coexpression list. Any two genes that were present in each other's top 10, 20, or 30 correlation lists were connected by green, orange, or red connections, respectively. Edges representing HRR values 10, 20, and 30 were assigned weights 1/5, 1/15, and 1/25, respectively. Any two clusters in the meta-network were connected if the connectivity score exceeded 0.02 according to:

$$c(A, B) = \frac{1}{2}\left(\frac{\sum_{i \in A \to B} w_i}{\sum_{j \in Aout} w_j} + \frac{\sum_{k \in B \to A} w_k}{\sum_{l \in Bout} w_l}\right) \tag{4.2}$$

where $A \to B$ are connections from cluster A to cluster B, Aout and Bout are edges going out of cluster A,B and where

$$w = \begin{cases} 1/5 & \text{, green edge} \\ 1/15 & \text{, orange edge} \\ 1/25 & \text{, red edge} \end{cases} . \tag{4.3}$$

We used $c(A, B) \geq 0.02$, which connects clusters A and B, if the average mutual weights of edges between the two clusters exceed 0.02. The connectivity score can range from 0 (no edges between the clusters) to 1 (all outgoing connections from cluster A are connected to cluster B and vice versa).

4.6.4. Comparison of a Pearson Correlation Network and a Graphical Gaussian Network

Our Pearson correlation network ($r = 0.8$) was compared with data sets from a recently published Graphical Gaussian (GGM) network [115], and common edges were identified by set comparisons (Fig. B.4A). Approximately one-third of the edges in the GGM network were also present in our network, consistent with a previous comparison made between the GGM and a Pearson correlation network [115].

To assess the association of node degree (number of nodes a node is connected to) with phenotype characteristics (essential or non-essential), a node degree of genes showing a phenotype versus those not showing any phenotype was compared. This was done across 20 coexpression networks generated by using Pearson r values ranging from 0.9 to -0.9 (steps of 0.1). The median node degree of genes showing a phenotype was compared with the median node degree of genes not showing any phenotype at a given r value cutoff. Significant differences (Wilcoxon test; $P < 0.05$) in the median node degree between these two classes were used to indicate significant differences between the two classes.

4.6.5. HCCA clustering algorithm

The HCCA can be implemented by a pseudocode available from the AraGenNet home page, and the full source code is available upon request from the authors. A simplified description of the algorithm is depicted in Figure 4.2 and in the "Results and Discussion" section. Python implementation of HCCA, together with sample networks, is available from AraGenNet home page.

4.6.6. MCL

We used the available C code (http://micans.org/mcl/; [197]) for MCL calculations. The method simulates random walks on the graph, with the walking probability respecting the weight (i.e. HRR values) of the edges (HRR value of 10 received weight 1/5, 20 received 1/15, and 30 received 1/25). We used different inflation values, which are the Hadamard power of a stochastic matrix that gives the probabilities for the random walk. Low inflations result in slower random walks, and vise versa. The inflation parameter may range from ≥ 1 to 5, where small values generate fewer but larger clusters.

4.6.7. k-means Clustering

To partition probe sets based on the original data, the expression values for each probe set were centered, scaled, and then subjected to the k-means clustering procedure provided by R using the default algorithm [71].

4.6.8. MCODE Clustering

The MCODE plugin for Cytoscape (http://baderlab.org/Software/ MCODE; [8]) calculates the local density of nodes in a network. Based on this score, a seed node is chosen as a starting point to collect nodes as long as their scores deviate from the seed node within a certain range. After clustering, it allows postprocessing single clusters without changing the rest of the network. Since MCODE has the option to vary six or seven parameters, we attempted to make the output comparable to the HCCA, MCL, and k-means cluster solutions; therefore, we emphasized the solutions that cluster a large portion of nodes [8].

4.6.9. Comparison of Clustering Solutions

The clustering solutions were judged by modularity [135] which evaluates the graph partitioning by comparing the sum of edge weights within clusters with edge weights linking different clusters. This value is subsequently subtracted by the value that one expects for random partitions. The obtained modularity score ranges between -1 and 1, where 1 represents perfect modularity, 0 represent value expected by chance, and -1 represents value worse than expected by chance.

The partitions were also evaluated by the Davies-Bouldin (DB) index [37] using the clusterSim R-package. It is defined as:

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{i\neq j}\left\{\frac{S_n(Q_i)+S_n(Q_j)}{S(Q_i,Q_j)}\right\} \tag{4.4}$$

with n number of clusters, S_n average distance of all objects from the cluster to their cluster center and $S(Q_iQ_j)$ distance between two cluster centers. Davies-Bouldin score can range from 0 to infinity. Values close to 0 are achieved by good (distant) clustering. However, the value of zero is gained by just one big cluster.

We used adjusted Rand indices to compare two clustering solutions by pairwise affiliation of nodes [78].

The scores for biological significance of clusters were calculated using the approximate mutual information between the clustering and MapMan categories [194] having at least 10 members. In the case where the clustering solution did not assign all genes to clusters, only those that could be assigned where considered. To make the HCCA clustering comparable to k-means, genes not assigned to any cluster by HCCA were not subjected to k-means, as these genes are most likely difficult to cluster. From this mutual information value, the mean mutual information from 1,000 random assignments (denoted by $\overline{MI}$) with preserved cluster sizes was subtracted, and the result was divided by the SD (denoted by σ) of these random mutual information values according to:

$$S = \frac{MI_{\text{cluster}} - \overline{MI}_{\text{random}}}{\sigma_{\text{random}}} \tag{4.5}$$

4.6.10. Overrepresentation Analysis

In order to identify terms which might be associated, we randomly sampled approximately 700 nonoverlapping NVNs from the whole network and tested for a significant overrepresentation of MapMan terms within these clusters using a Fisher exact test ($P < 0.05$ after Benjamini-Hochberg correction). This was repeated several times to exclude random effects. Subsequently, we tested for significant cooccurrence of overrepresented terms using again a Fisher exact test.

4.6.11. Uniqueness vs. Essentiality Estimates

To group Arabidopsis genes into gene families, a BLASTCLUST analysis on Arabidopsis protein sequences obtained from TAIR was performed. Length coverage threshold of 70% and score coverage threshold were used as parameters.

We used random sampling to investigate whether there is correspondence between a gene having essential or non essential characteristics and its uniqueness in the genome or node vicinity network. So far, 261 genes are characterized as being essential (phenotypic data from TAIR), and 152 of these are single-copy genes based on the settings above. To investigate whether essential genes tend to be single copy, we sampled 261 random nodes 1,000 times and counted the number of single-copy genes acquired in each sampling. To investigate whether essential genes that do belong to gene family tend to be unique in the network vicinity, we sampled 109 (261 total – 152 single copy) random nodes 1,000 times. The number of genes unique or nonunique in the network vicinity was then counted, and represented as a histogram. The same was done for nonessential genes with characterized nonlethal phenotype (1,224 total, 422 single copy).

4.6.12. Plant Cultivation and Mutant Analysis

T-DNA knockout lines (Table B.5) were obtained from the Nottingham Arabidopsis Stock Centre [4]. The seeds were surface sterilized, sown on plates containing Murashige and Skoog medium (1x Murashige and Skoog salts, $8gL^{-1}$ Agar, $1\times$ B5 vitamins, $10.8gL^{-1}$ Sucrose) and incubated for 48 h at $4°$C in the dark. The plates were then incubated for 7 days at $21°$C with 16-h photoperiod. T-DNA insertions were confirmed using PCR (Table B.5). Images of seedlings and siliques were done using Leica MZ 16 FA stereomicroscope.

5. Analyzing Gene Coexpression Data by an Evolutionary Model[†]

5.1. Abstract

Coexpressed genes are tentatively translated into proteins that are involved in similar biological functions. Here, we constructed gene coexpression networks from collected microarray data of the organisms *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, and *Escherichia coli*. Their degree distributions show the common property of an overrepresentation of highly connected nodes followed by a sudden truncation. In order to analyze this behavior, we present an evolutionary model simulating the genetic evolution. This model assumes that new genes emerge by duplication from a small initial set of primordial genes. Our model does not include the removal of unused genes but selective pressure is indirectly taken into account by preferentially duplicating the old genes. Thus, gene duplication represents the emergence of a new gene and its successful establishment. After a duplication event, all genes are slightly but iteratively mutated, thus altering their expression patterns. Our model is capable of reproducing global properties of the investigated coexpression networks. We show that our model reflects the mean inter-node distances and especially the characteristic humps in the degree distribution that, in the biological examples, result from functionally related genes.

5.2. Introduction

The increasing amount and easy accessibility of microarray gene expression data [17, 108] allows for systematic comparison of gene expression patterns on an organism scale under a wide variety of conditions [215]. Based on this data, coexpression networks are often constructed with the goal to identify clusters of similarly expressed genes which, it is assumed, are functionally related [187]. In the construction of the coexpression networks, genes are usually represented as nodes of a graph which are connected by edges if the similarity of the respective expression patterns lies above a predefined threshold. This approach is useful to support hypotheses about the functions of unknown genes. For example, Mutwil et al. [134] reported that this approach was successfully applied and several genes were correctly predicted to result in an embryo-lethal phenotype.

The study of graph properties of biological networks has a long tradition and encompasses diverse networks such as protein-protein interaction networks, transcriptional regulatory networks, signaling networks, and metabolic networks [41, 128, 94, 83]. Interestingly, many of the studied examples exhibit characteristic features that distinguish them from random network structures. The most widely discussed properties are the scale-freeness [14, 113], which means that the degree distributions follow a power law of the form $d(n) \propto n^{-\gamma}$, and the observation that they are organized in a small-world structure, which means that the average minimal path length between nodes is shorter than expected for random networks [205, 6].

Here, we observe that the coexpression networks of *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, and *Escherichia coli* (hereafter called *A. thaliana*, *S. cerevisiae*, and *E. coli*) roughly follow a power-law degree distribution. However, for high degrees, there is a characteristic overrepresentation of highly coexpressed genes followed by a sharp truncation. To investigate these properties we intro-

[†]Published as:
M. Schütte, M. Mutwil, S. Persson, O. Ebenhöh, *Analyzing Gene Coexpression Data by an Evolutionary Model*, Genome Informatics 24, 154–163 (2010)

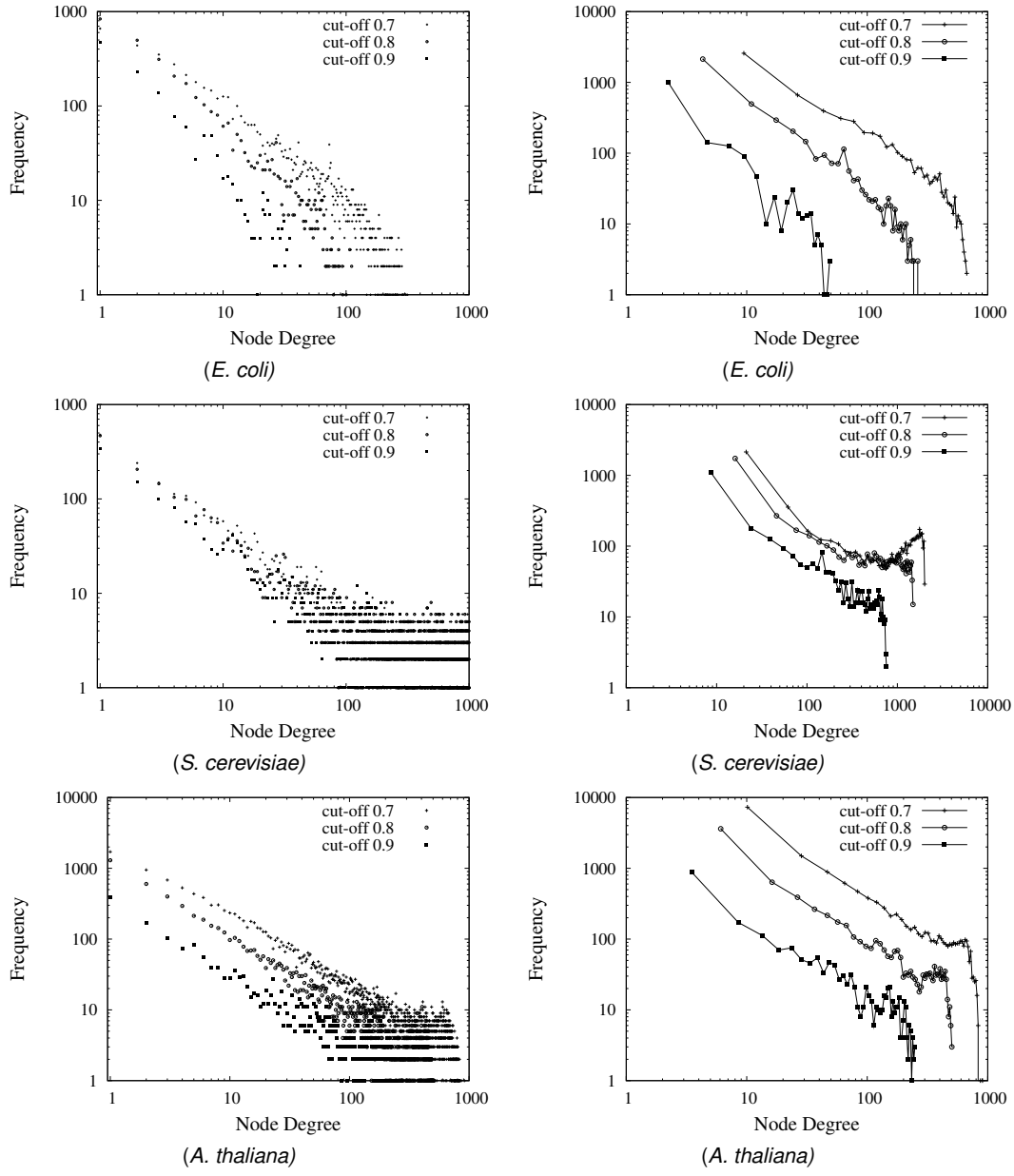


Figure 5.1: Degree distribution of data sets for different cut-offs (0.7, 0.8, 0.9). Left side: pure degree distribution, right: in a binned version. The binned distributions of *A. thaliana* and *S. cerevisiae* clearly show the same property of a concentration of highly connected nodes followed by a sharp edge.

Table 5.1.: Network characteristics of the collected coexpression networks.

Pearson cut-off	total number of genes	experiments	coexpressed pairs (edges)	connected (isolated) nodes	average path length
			A. thaliana		
0.7	22810	351	882836	15318 (7492)	5.5 ± 1.7
0.8	22810	351	231881	7178 (15632)	9.7 ± 4.4
0.9	22810	351	39119	2045 (20765)	2.9 ± 1.6
			S. cerevisiae		
0.7	9335	789	2607992	7052 (2283)	4.2 ± 2.2
0.8	9335	789	1075760	5145 (4190)	5.9 ± 3.7
0.9	9335	789	177226	2610 (6725)	2.7 ± 1.3
			E. coli		
0.7	7311	244	99188	5606 (1705)	5.6 ± 2.0
0.8	7311	244	18983	3425 (3886)	10.3 ± 5.2
0.9	7311	244	3299	1312 (5999)	2.8 ± 1.3

duce a model to mimic the growth of the coexpression network. For biological networks growth models are of special interest as they potentially allow to draw conclusions about the evolutionary pressures that have shaped these networks and thus to obtain hints about their design principles [204]. The genome's growth is mainly determined by gene duplications which account for about ninety percent of all eukaryotic genes [140, 65, 189]. Whereas existing growth models rather utilize graph theoretical methods [20, 156, 189, 199, 217] which basically follow preferential attachment [14] for the selection of duplicated nodes, we introduce a model that is based on numerical vectors as nodes in the network and connect them if their correlation values lie above a threshold. In this way, we simulate the emergence of coexpression patterns based on gene duplications and mutations, starting from a small number of initial, primordial, genes. Our simulation results support the view that the characteristic degree distribution of coexpression networks largely results from the functional and homological relatedness of the highly connected genes.

5.3. Data Sets

We downloaded microarray data for three different organisms, *A. thaliana*, *S. cerevisiae*, and *E. coli*. For *A. thaliana* we used the Affymetrix ATH1 array. Then, 1,428 ATH1 microarray data sets were collected from TAIR (http://www.arabidopsis.org/). For *S. cerevisiae* and *E. coli*, we took the Affymetrix Yeast Genome S98 and Affymetrix Ecoli_ASv2 GeneChips. The microarry data was downloaded from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/). Data of all organisms were RMA normalized and quality controlled by several steps. First, for Cel files that have shown either artifacts on RMA residual plots or deviated from the majority of the box plots of positive matches by visual inspection, data is

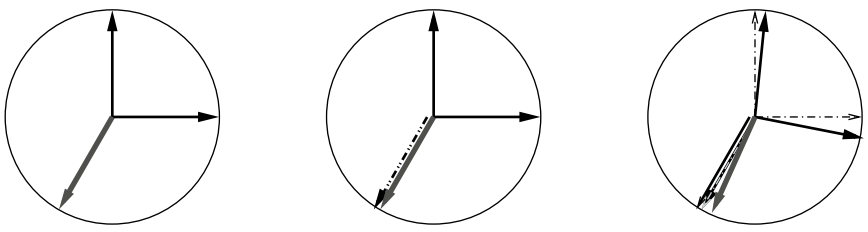


Figure 5.2: 2D scheme of the attachment procedure. From left to right: Three current genes, the thick gray one is chosen to be duplicated, then all are mutated along the circle. Right image: Final situation with now four genes (dashed ones represent previous direction).

removed. Then, those experiments that are mutually very similar were removed by dropping microarrays that have Pearson correlation coefficient higher than 0.95 with more than three other microarrays to reduce a bias in the experimental conditions [134]. From the data we constructed coexpression networks for different Pearson correlation coefficient cut-offs (called Pearson cut-off in Tab. 5.1), see Tab. 5.1 and Fig. 5.1.

Analyzing the data, the degree distributions roughly follow a power-law behavior but we observe an overrepresentation of highly connected nodes followed by a sharp edge, see Fig. 5.1. To get a biological interpretation, we exemplarily investigated genes of these nodes. In the case of *S. cerevisiae* with cut-off 0.9, we find significantly many genes that translate into proteins of the large and small ribosomal subunits. Of the total number of 2610 genes 221 are associated with ribosomes but 11 of the 23 most connected ($8\% \rightarrow 48\%$, p-value $p < 10^{-6}$). For *E. coli* 19 of the 22 most connected genes code for flagella proteins of the bacterium's chemotaxis movement, compared to 38 flagella of 1312 total genes ($3\% \rightarrow 86\%$, $p < 10^{-15}$). In *A. thaliana*, several genes of the protein family PF00069 (http://pfam.sanger.ac.uk/), protein kinases, are overrepresented, 43 of 2045 to 10 of 51 ($2\% \rightarrow 20\%$, $p < 10^{-6}$). These results indicate that highly connected genes, hubs, are also mutually highly connected [221, 83, 109] which to some extent is contrary to the idea of distinct functional modules [149]. To support this hypothesis, we calculate the clustering coefficient [208]. For a node i it is defined as $c_i = 2n/k_i \cdot (k_i - 1)$ where n is the actual number of edges between the neighbors of i and $k_i \cdot (k_i - 1)/2$ is the maximum possible number of edges between these. Complete connection leads to $c_i = 1$ and no clustering to $c_i = 0$. We obtain for the sets of high degree nodes listed above mean clustering coefficients $\bar{c}_{Yeast} = 0.49$, $\bar{c}_{E.coli} = 0.80$, and $\bar{c}_{Ara.} = 0.58$, and mean clustering coefficients only within the subnetwork of these highly connected nodes $\bar{c}^{(high)}_{Yeast} = 1$, $\bar{c}^{(high)}_{E.coli} = 0.93$, and $\bar{c}^{(high)}_{Ara.} = 0.96$.

5.4. Model

The aim of our model is to simulate the evolution of the organism's gene expression profile. We start from a small number of given initial genes. Then at every iteration step, we duplicate with high probability one of these initial genes. This is motivated from the known results of preferential attachment models [14], in which earlier nodes more likely become hubs. The chosen selectivity criterion mimics that established genes are robust in their expression patterns while their duplicates are redundant shortly after duplication but gain a new function by divergence. Due to the strong selectivity towards duplicating established genes, we can exclude gene loss as an explicit process [114, 219]. After duplication, all current genes undergo a slight random mutation. Such mutations might change the coexpression pattern of a gene. Iteratively, gene duplication and subsequent mutations are repeated, until the genome reaches a predefined size.

Technically, our model is designed along the proximity to experimental data, see Fig. 5.2 as a 2D illustration. Experiments are usually run under a variety of different conditions like stress in temperature or nutrient supply. Hence, the data consists of vectors where every entry belongs to a certain experimental condition. Therefore, we also represent genes as D-dimensional unit vectors. Randomly, we produce a set of I initial vectors. Then at every step in the process we duplicate one of the vectors with strong selectivity towards established genes. This is implemented by randomly choosing a candidate gene g for duplication according to a Fermi-Dirac distribution

$$P(g, A) = \left(1 + \exp\left\{\frac{g - I}{f(A)}\right\}\right)^{-1}, \tag{5.1}$$

where genes are indexed by their order of appearance ($g \leq I$ standing for the initial genes), A is the current genome size, and $f(A)$ a function of A, here chosen as $f(A) = c \cdot A$, where c is an adjustable parameter. In the limit $f(A) \rightarrow 0$, Eq. (5.1) is equal to a Heaviside step function with the step at I, for higher values the edge softens and approaches a Boltzmann distribution. With this particular choice of distribution, every gene in the genome that has itself emerged by a duplication event is selected

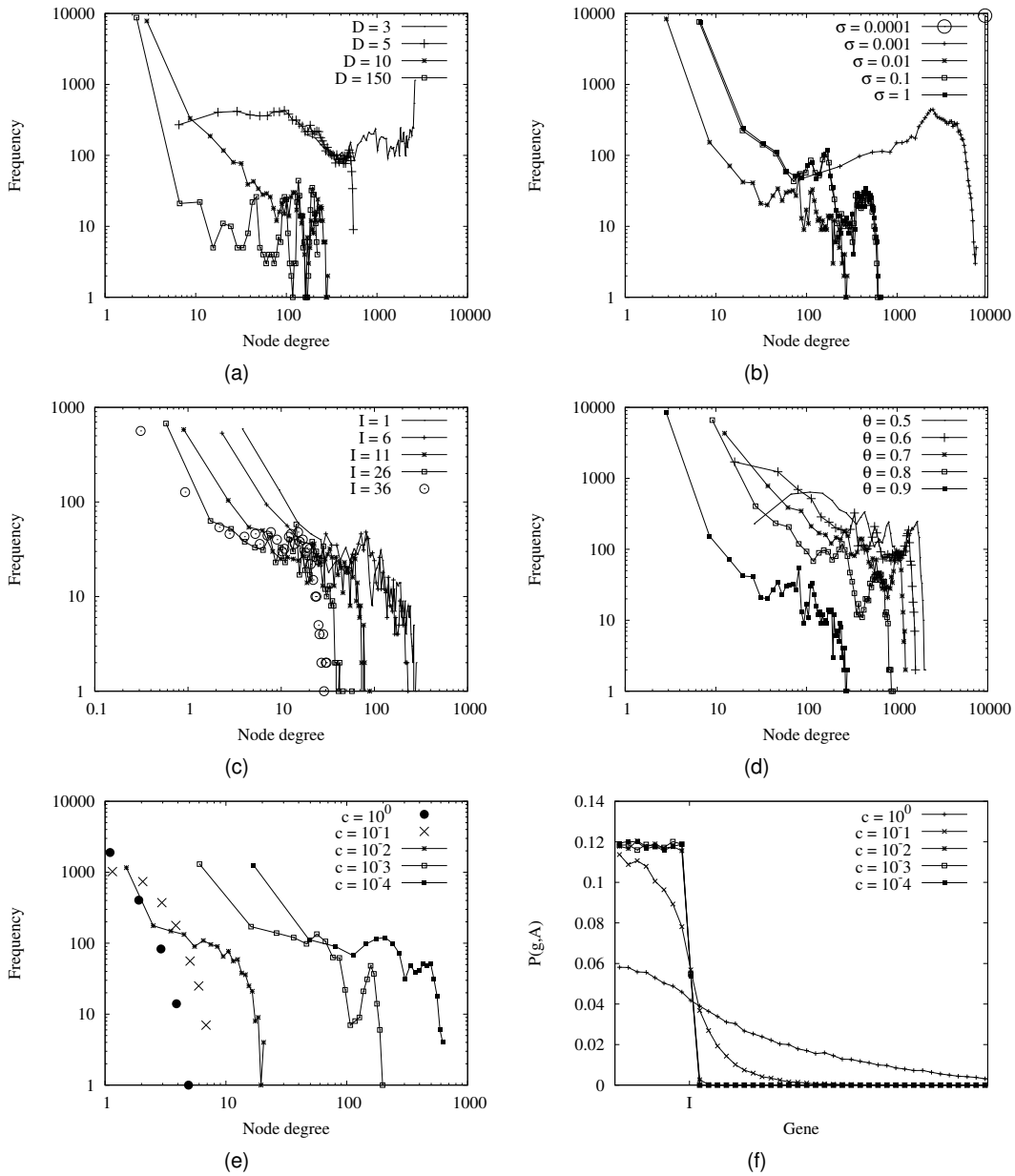


Figure 5.3: Degree distribution of simulated data depending on the different parameters. The reference set of fixed parameters is: $I = 1$, $\theta = 0.9$, $D = 16$, $\sigma_{mut} = 0.01$, $c = 10^{-4}$. (a)–(e) variation of Dimension, mutation rate σ_{mut}, number of initial nodes, Pearson correlation cut-off to create the network, parameter c of Fermi distribution. (f) Sketch of the Fermi distribution for different c.

Table 5.2.: Effect of different parameter values on simulated coexpression networks with 9335 nodes as in *S. cerevisiae*. The reference set is the same as in Fig. 5.3 and its resulting values are listed under dimension $D = 16$.

varied parameter	coexpressed pairs (edges)	connected (isolated) nodes	average path length
Dimension D			
3	7549584	9335 (0)	1.7 ± 0.5
5	859392	9335 (0)	1.9 ± 0.2
10	53982	4217 (5118)	3.1 ± 0.5
16	39697	1531 (7804)	6.3 ± 2.0
40	42869	927 (8408)	4.7 ± 1.8
150	35413	638 (8697)	4.6 ± 1.7
Mutation rate σ_{mut}			
0.0001	43566445	9335 (0)	1 ± 0
0.001	14724849	9335 (0)	1.6 ± 0.5
0.1	163742	3340 (5995)	6.9 ± 2.9
1	168799	3378 (5957)	6.8 ± 2.9
Initial nodes I			
6	42355	1907 (7428)	5.4 ± 1.6
11	15367	1534 (7801)	6.6 ± 2.1
26	7426	1265 (8070)	7.4 ± 2.4
36	7137	1293 (8042)	7.2 ± 2.1
Pearson threshold θ			
0.5	3360623	9335 (0)	1.9 ± 0.2
0.6	1908730	9333 (2)	2.0 ± 0.2
0.7	944298	8989 (346)	2.4 ± 0.5
0.8	337840	5728 (3607)	3.3 ± 0.7
Fermi parameter c			
10^{-3}	57146	5354 (3981)	6.9 ± 2.4
10^{-2}	8489	4452 (4883)	13.2 ± 4.3
10^{-1}	3317	3682 (5653)	29.2 ± 10.4
10^{0}	1508	2400 (6935)	10.2 ± 7.6

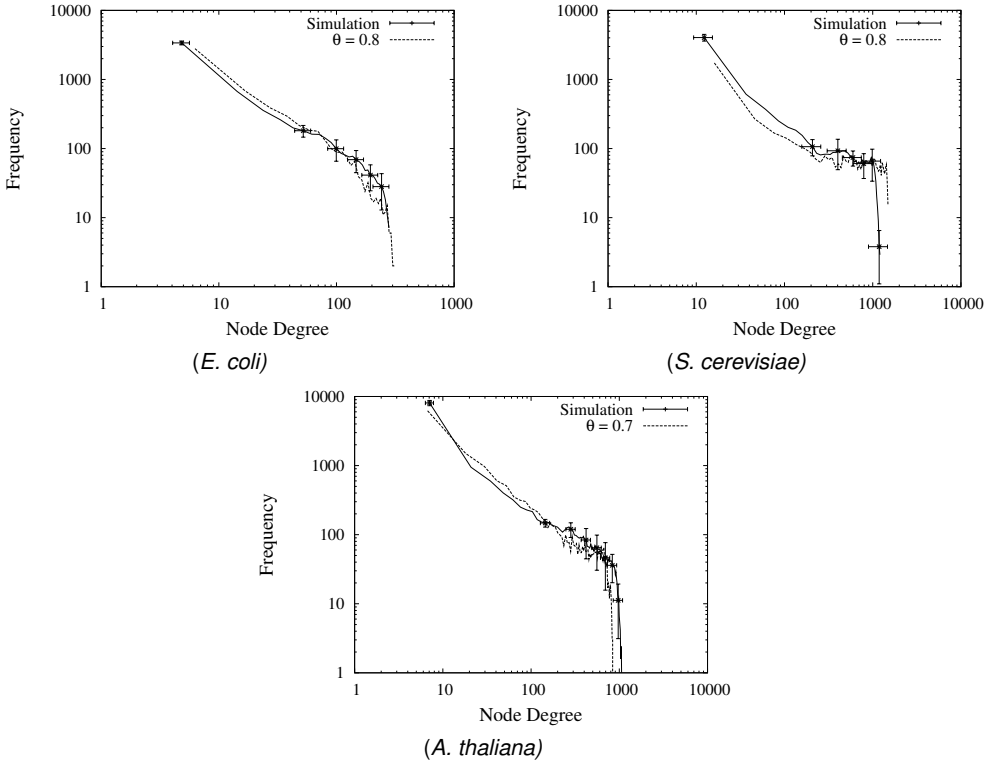


Figure 5.4: Degree distribution of simulations manually fitted to real data. Parameters chosen: *E. coli* 0.8 cut-off: $c = 10^{-3}$, $D = 14$, $I = 2$, $\sigma_{mut} = 0.05$, $\theta = 0.8$; *S. cerevisiae* 0.8: $c = 10^{-4}$, $D = 14$, $I = 1$, $\sigma_{mut} = 0.05$, $\theta = 0.8$; *A. thaliana* 0.7: $D = 14$, $I = 4$, $\sigma = 0.01$, $\theta = 0.88$, $c = 10^{-4}$. With error bars resulting from several simulations with different random initialization.

for duplication with only a small probability, whereas in most of the cases one of the initial genes is selected.

After every duplication, the second step is the mutation of all current genes. This is done by the addition of a normally distributed random number with zero mean and a small variance σ_{mut} to every entry of a vector. The size of the variance σ_{mut} is chosen quite small but still by chance and the total number of mutations it is possible to obtain a drastic change in a vector. After each mutation step all vectors are normalized to unity again. Completing the last duplication step, we construct a network form the vectors. For every vector pair, we calculate the corresponding entry in the correlation matrix using the standard definition of the Pearson correlation coefficient and connect them if they exceed a given threshold. In summary, five parameters control the model behavior: dimension D, mutation variance σ_{mut}, correlation threshold θ, the number of initial vectors I, and the constant c.

5.5. Results

The model behavior for different parameter combinations is summarized in Fig. 5.3 and Tab. 5.2. For small dimensions the behavior is drastically different from the experimental data (see Fig. 5.3a). For $D = 3$ and $D = 5$ the distributions miss the typical slope for small degrees but rather appear highly connected over a broad range. This becomes apparent in the average path length (Tab. 5.2), which is very short for low dimensions. This indicates that a critical number of experimental repetitions under different conditions is necessary to obtain reliable results. A very low mutation rate leaves all vectors almost unchanged and leads to one large spike (see Fig. 5.3b for $\sigma_{mut} = 0.0001$). Increasing the parameter this spike melts apart but still a broader maximum of highly connected nodes is observable. The

number of initial nodes causes only a shift in the curves leaving the actual shape unchanged (Fig. 5.3c). This behavior can be understood as we mainly see a superposition of similar but smaller networks. A similar observation as for the dimensions can be made for the cut-off of the Pearson correlation coefficient, Fig. 5.3d. For $\theta >= 0.7$ the distributions are similar except for the scale. However, lower cut-offs do change the shape qualitatively as they seem to allow influence of random relations between nodes. By this analysis, we conclude that the different parameters cannot be tuned independently. Dimension and cut-off have a similar impact on how strictly nodes are connected. The dimension also determines the overall effect of mutations: For a certain mutation strength, an original and a mutated duplicated gene are more likely to be connected for a smaller dimensionality of the vector.

Figures 5.3e and f depict the influence of the selective pressure. The distributions for high values of c ($c = 0.1$ and $c = 1$), which correspond to a uniform probability of picking a gene, lead to a degree distribution in which highly connected nodes are not overrepresented. The lower values ($c = 10^{-3}$ and $c = 10^{-4}$) produce qualitatively similar curves, which are shifted to higher degrees for lower values of c, while in both cases high degree nodes are overrepresented. Fig. 5.3f illustrates the softening of the Fermi edge depending on the parameter c. There is a rather strong change from $c = 10^{-1}$ to $c = 10^{-2}$ which leads to an enormous increase in probability for values higher than the edge position.

In order to demonstrate that the model is capable of reproducing important features of the correlation networks determined from experimental microarray data, we manually fit the model parameters to the data (see Fig. 5.4). The mutual dependence of parameters allows to reduce the dimensionality in comparison with the data sets by adjusting mutation strength or threshold.

5.6. Conclusion

We presented an evolutionary model based on numerical representations of gene coexpression data that can explain some observed properties of the coexpression networks of *A. thaliana*, *S. cerevisiae*, and *E. coli*. These networks contain a group of highly coexpressed genes that by data analysis code for proteins with very similar function, flagella in E. coli and ribosome in *S. cerevisiae*, or of the same protein family, PF00069 protein kinases in *A. thaliana*. Due to the capability of the model to reflect characteristic features of the experimental data, in particular the overrepresentation of highly coexpressed genes, it allows to assess which evolutionary parameters are critical for these features to emerge. A robust observation from our modelling results is that the characteristic overrepresentation of highly connected genes can only be reproduced under high selective pressures towards duplicating established genes. This finding is consistent with the notion of preferential attachment which assumes that those genes that are already highly coexpressed with other genes have a higher chance to establish a new gene by duplication. Further model results demonstrate the necessity to take great care when interpreting coexpression network properties. The importance of the dimensionality of the expression vectors, representing the number of different experimental conditions, shows that the possibility cannot be ruled out that the inclusion of even more data will lead to completely new network properties. Similarly, the choice of the threshold value used to construct a coexpression network must be critically assessed as a too low or a too high value may result in dramatically different network characteristics.

5.7. Acknowledgments

This work was supported by the International Research Training Group *Genomics and Systems Biology of Molecular Networks* IRTG 1360 (MS), and the Scottish Universities Life Science Alliance SULSA (OE).

6. Summary, Conclusion, and Future Perspectives

With improved experimental techniques, huge data sets of biological and biochemical data became accessible. These opened the door for large-scale mathematical modeling of biological phenomena. In this thesis we presented four published papers utilizing genetic sequence and expression data as well as biochemical data of the metabolic network: the first two chapters deal with a model of metabolic evolution simulating the enzyme-pathway coevolution. We investigated the generated time-courses of the evolutionary process by a time-series analysis and conclude that new enzymes appear in bursts: a situation which reminds of punctuated equilibrium and helps for a molecular understanding of this concept. The last two papers analyze gene-coexpression data from two perspectives. With a novel clustering approach we were able to create an interactive gene coexpression network of *Arabidopsis thaliana* and to assign phenotypes to six previously not annotated genes. Secondly, with an evolutionary model, we could explain characteristics in the degree distribution of the gene coexpression networks of *A. thaliana*, *E. coli*, and *S. cerevisiae*.

Metabolic Evolution

After the first establishment of a simple metabolism, the further evolution of metabolism has most likely been occurred in parallel with the further development of sequences of the catalyzing enzyme, i.e. their protein sequences. The first two chapters are founded on this hypothesis.

In Chapter 2, we checked whether there exists a relation between enzyme sequences that lie close to each other on the chemical reaction network, in our case given by the KEGG database. Enzymes are defined to be neighbors if they share a metabolite in any catalyzed reaction. Since KEGG provides numerous sequences from various organisms a complete analysis is not feasible. In order to reduce the amount of sequences, we first developed a method to derive a consensus set of sequences which conserves the structure of the original data as present in KEGG. We defined sequence similarity cut-offs by a benchmark against the COG database. The same method is also used for the analysis of Chapter 3. Using the consensus set and thus conserving the complete variety of sequence space, we find some evidence for a significant correlation mainly on nearest neighbors which tend to have a sequence similarity. Long-range correlations between enzyme sequence distances and the enzyme distance on the metabolic network can not be detected. This can be explained by the high promiscuity of enzymes [98]. A required enzyme in a new pathway needs not necessarily be newly invented but recruitment of either single enzymes at different positions in the network or of entire modules occurs alike. Further, measures of sequence similarity do not show transitivity which might be a further property hiding a higher correlation. As a small illustration, the two strings ”cumulative” and ”dissertation” do not show any significant sequence alignment, however the string ”cumulativedissertation” does with both.

We also excluded ubiquitous metabolites from the analysis that appear in very many reactions like cofactors or water. An improvement of the results of the sequence-network-distance correlation might potentially be achieved by a further removal of specific molecules from every reaction and following only main metabolic fluxes or take into account just major metabolites of every reaction, for instance given by the ”rpair” data in KEGG.

Although there is only slight correlation detectable, we follow up on these results with a larger model on chemical evolution. We assume that a new enzyme arises by mutation not necessarily by an enzyme within the close vicinity on the metabolic network but we follow the weaker assumption that a new enzyme can emerge by mutation from any member of the currently existing enzyme pool.

The model of chemical evolution, developed in Chapter 3, substantially improves the method of

network expansion by the additional usage of enzyme sequence data. Thus, the model approaches a more realistic description of evolution and further allows for the implementation of a time coordinate. The choice of the next enzyme as well as its time point was implemented by the Gillespie algorithm adapted for evolutionary processes. Further, we introduced a parameter γ which controls the influence of sequence similarity on the choice of the next enzyme. We assume that there is a higher probability for newly appearing enzymes which originate by a smaller number of mutations from existing ones. From a set of seed metabolites, the model generates possible evolutionary walks on the metabolic graph. In order to capture the variation in topology of chemical reaction as well as sequence space, we performed 200 simulations for every value of γ. The generated temporal network dynamics shows that the sequence-similarity driven expansion with high γ explores the network quicker than a random procedure ($\gamma = 0$). This resembles the biological assumption that a mutant of an already established sequence rather adds an instantly useful new enzymatic tool than it would be the case by adding any random new sequence. Our process underlies one restriction since we already searched in the set of actually realized sequences whereas nature putatively could mutate to any amino acid combination whether they can properly be folded or not [155]. While this in general reduces the time intervals especially for random new sequences, it also reminds of the Levinthal's paradox dealing with time scales in protein folding [222]. Instead of realizations of sequences the question is how linear chains of amino acids can find their native 3D fold structure within fraction of seconds, while a theoretical estimation would lead to characteristic times of 10^{24} years. Speculations assume that the reason for this speed up is evolutionary memory in terms of helper proteins as chaperones and other optimized processes like folding a protein already during its synthesis.

From the generated simulations, we obtained time series of enzyme appearance. We measured the time intervals between any two new enzymes, the interspike intervals. In order to interpret this point process, we calculated the Coefficient of variation (C_v), the autocorrelation function, and the Fano factor. The C_v and autocorrelation clearly show that new enzymes preferentially appear in bursts of close temporal vicinity while two such bursts are separated by longer time intervals. This trend increases with larger stress on the sequence similarity as a criterion to pick the next enzyme. From this, we concluded that, assuming that small mutational changes drive evolution, the evolution of metabolism follows a bursting-like behavior. The calculated Fano factor could be fitted to an analytical result of the Fano factor of Brownian motion, and allows for the interpretation that our process explores the metabolic network similar to diffusion. Since punctuated equilibrium has so far been a concept on the macro-level but resembles the same behavior, we derived the first molecular description on punctuated equilibrium behavior which could potentially help to understand the phenomenon at a species level. From the modeling approach we are not able to definitely conclude whether we observe the same dynamical behavior as in the macroscopic phenomenon or if this is really the underlying reason for punctuated equilibrium.

In order to justify our model as biologically relevant, we analyzed the temporal order in which enzymes, metabolites, and organisms occur and yielded results in accordance with previous findings or current biological knowledge. From this perspective we assume that the model provides biologically relevant metabolic evolution scenarios. To fully justify our model a comparison with enzymatic age data would be useful. Since we constructed the consensus set and thus rather arbitrarily extracted sequences, a very careful comparison with age data is necessary. Further, every evolutionary walk on the network follows a slightly different route. The MANET database provides some estimates on the age of metabolic proteins calculated through the protein folds [99]. Recently, a promising method that maps the genetic history onto a geological time scale was published [36]. These data might also be used to either find the most likely evolutionary route or optimize the seed metabolites and enzymes for a best fit and thus conclude on a prebiotic atmosphere.

An experimental setup matching the simulations could possibly be very similar to the famous Miller-Urey experiment [126]. Additionally, one could use a source of radioactive radiation to enhance mutational changes in the enzymes. The detection of single molecules is a critical issue but could be achieved by Mass spectrometry and gas chromatography. In order to obtain a time resolution, neces-

sarily one either needs to extract matter from the experiment without perturbing the setup or carry out many identical experiments in parallel and evaluate them after different times.

Our findings also depend on the quality of data in the KEGG database. In order to curate the data we removed erroneous reactions that do not obey mass balance. Also, we manually added irreversibility information. One disadvantage was observed namely that not for all EC numbers there exists a sequence in KEGG. A reason for this is that enzymes have systematically been analyzed long before sequence techniques were developed and so a function but no sequence for these enzymes is known. In order to cope with this problem, we decided to randomly add sequence distances to those enzymes. We picked a value from the distribution of all known sequence distances for every pairwise combination and repeated simulations for the procedure multiple times to reduce artifacts. This assumption seems most promising since neglect of these reactions would strongly influence the metabolic network where some pathways could not be expanded. An alternative approach that assigns the largest distance of 1 to every other enzyme lead to artifacts because all those enzymes were tentatively found at the end of the expansion process.

A further critical issue concerns the organisms in KEGG. The database mainly contains data of bacteria simply because those are sequenced. Higher organisms and especially plants are underrepresented, see Fig. 3.3. A similar argument can be made for the annotated genes and thus for the present enzymes. Tentatively, well known proteins are annotated in most organisms while some poorly understood are not (see Fig. 2.3). Thus our approach of the consensus set only captures the sequence variability given in the database. The organism-specific networks we constructed for Fig. 3.3 were generated by a simple mapping every EC number to its genes in different organisms. The generated networks are smaller than detailed reconstructed networks [48, 42] but because this should influence all networks equally we do not expect it to have a qualitative impact on the results.

For the analysis of the bursting behavior similar to punctuated equilibrium, we did not focus on a single organism but rather on a pseudo-organism comprised of all reactions in the KEGG database. This can be understood as the metabolism of the entire biosphere. The same holds true for the enzyme sequences where we generated a consensus set containing sequences from various organisms. This allows for an estimation of general evolutionary mechanisms but does not shed light on an organism's evolution. An investigation of a single organism would be an interesting special case. This could putatively allow for a characterization of specific burst loci in the genome. One particular question might be whether the bursts can be biologically interpreted. One suggestion for an investigation could be the following as to interpret bursts as reminiscent of genome duplications for which *A. thaliana* with two or three genome duplications might be a good candidate. Since the network expansion is a stochastic process with varying results one must think about a precise measure of significance whether the enzymes within or close to a burst indeed belong to the same gene fragment of a genome duplication and also about a precise measure for a burst.

For preliminary simulations we chose the organisms *Saccharomyces cerevisiae*, *Zea mays*, *Oryza sativa japonica*, *Homo sapiens*, *Populus trichocarpa*, and *Arabidopsis thaliana* and performed the same evolutionary expansion as in the model of Chapter 3 but with an organism-specific set of protein sequences and we repeated this for various seed combinations. These simulations rather show a correlation between the number of sequences and the number of bursts. Here, we defined a burst in analogy to Fig. 3.2; if the C_v for a certain window exceeds the current mean C_v of the previous windows by a factor $3/2$, we call this a burst. In order to avoid counting the same burst multiple times, we require that two bursts are separated by at least one window size.

The results can be compared to recent knowledge on genome duplications, [183, 184, 97, 67]. Since the expansion process does not explore the entire network depending on the seed and the lacking data on reactions and enzyme sequences, we also calculated an extrapolated number of bursts N_{Bur}^{tot} which should estimate the number of bursts for the actual network size of the organism. For this, we multiply the computed number N_{Bur}^{comp} following $N_{Bur}^{tot} = N_{Bur}^{comp} \cdot N_{Enz}^{(tot)} / (2 \cdot N_{Enz}^{(comp)})$, where $N_{Enz}^{(comp)}$ is the number of enzymes explored in the expansion and $N_{Enz}^{(tot)}$ is the number of enzymes in the entire organism specific network. A factor $1/2$ is added to avoid a double counting of a burst. The results look

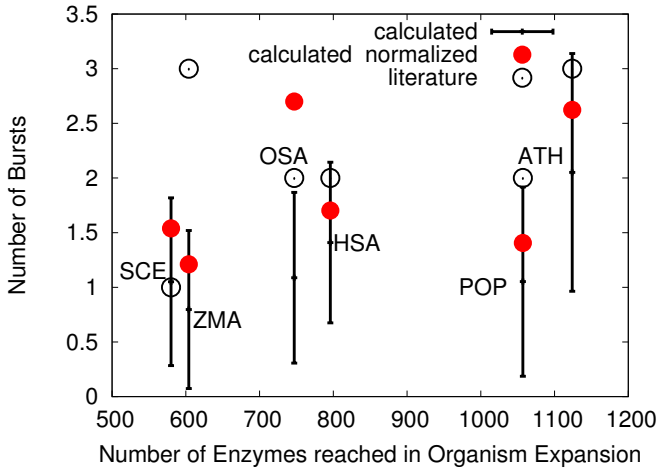


Figure 6.1: The mean number of bursts found during the expansion. We define a burst using a sliding window of 50 enzymes. If the coefficient of variation, C_v, in a window exceeds the 1.5-fold of the mean C_v of all previous windows, we call this a burst. Two bursts need to be separated by 50 enzymes. KEGG-IDs: SCE–*Saccharomyces cerevisiae*, ZMA–*Zea mays*, OSA–*Oryza sativa japonica*, HSA–*Homo sapiens*, POP–*Populus trichocarpa*, ATH–*Arabidopsis thaliana*.

promising but surely need refinement and statistical and biological validation, see Fig. 6.1.

From a theoretical point of view, an interesting question that we brought up already in the manuscript concerns the relationship between our model and the concept of self-organized criticality [11, 12] which has also been shown to produce punctuated equilibrium behavior [10]. A main finding of the concept is an explanation of power-law fluctuations, especially for $1/f$ Flicker noise. The prominent example is the sand pile which produces avalanches of various sizes depending on the slope of the pile. Similarly, we detect avalanches of new enzymes in dependence on the current status of the metabolic network. Thus, a detailed spectral analysis of fluctuations, potentially finding $1/f$ distributions, and the creation of a mathematical foundation could be a big asset yielded in future investigations.

Analysis of gene-coexpression networks

Chapter 4 and 5 introduce two techniques to analyze gene coexpression. In Chapter 4 we present a novel clustering technique called Heuristic Cluster Chiseling Algorithm (HCCA). This cluster algorithm has the feature that it roughly conserves the cluster size and is thus well applicable for the break down of large data sets for the purpose of visualization. This is in contrast to existing clustering methods, which either conserve the number of clusters, such as k-means, or generate clusters of all sizes following a certain criterion as local density, such as MCODE, or a flow on the graph, such as MCL. We could show that HCCA outperforms the existing algorithms by a variety of cluster measures. The modularity score, clusterJudge, and Davies-Bouldin address a score to every single cluster, while the adjusted Rand index compares two different cluster solutions of the same data set. Since these measures do not use any biological knowledge, the method calls for application outside biology.

Clustering is an often used approach to find interrelations in data. A characterization of nodes in the coexpression network with biological functions lead to the conclusion that essential genes (e.g. essential for plant growth) tentatively are detached from other genes of the same function within the network vicinity. We could use this information to identify twenty candidate genes for which no phenotype is known so far but which presumably, following from our analysis, might be essential for the plant. Indeed, six of these mutants harm the plant. Two result in an embryo-lethal phenotype, two male gametophytes, one seedling lethal, and one pale green dwarf which could be shown by a T-DNA mutant analysis. This

result shows the predictive power of a bioinformatic approach as clustering if it is combined with current biological knowledge. The introduced clustering method has meanwhile successfully been applied to different plant coexpression networks, see aranet.mpimp-golm.mpg.de [131]. We expect that it will help to uncover further gene properties.

For further analysis of the coexpression network we developed a model mimicking the evolution of the gene expression profile. By the very simple procedure of duplication and selection followed by mutation we could reproduce characteristics of the coexpression networks from measured data. We could confirm an over-representation of highly connected nodes which was observed in *A. thaliana*, *E. coli*, and *S. cerevisiae*. These nodes tend to be relatives in function, zinc fingers, the flagella, or the ribosome, respectively.

Although our model for the evolution of a general expression profile is rather simple, it in fact reproduces the observed properties of the degree distribution, namely the sharp truncation for high degrees and a hump right before the high degrees. An improved version could include gene deletions. These are only intrinsically incorporated because we add some random Gaussian noise to the gene vectors at every iteration step and this noise can by chance be very high and lead to a significant change of the vector. This would represent a deletion plus invention of a new random vector. We also used only single gene duplications. Realistically, duplications of either larger genome fragments, meaning multiple vectors in our model, or even the entire genome could be incorporated. However, since already the simple version leads to sufficient results, we chose not to elaborate the model because incorporation of further mechanisms would go hand in hand with more parameters and rising complexity.

The two projects are connected as both utilize data on genes: gene sequences and expression profiles. A connection of the two worlds, the metabolic network and enzyme sequence data and the expression profile of the particular protein, either regulator or catalyst, would lead to a holistic understanding of the cellular mechanisms. It has been argued that stochastic noise in gene expression has an impact on the pathway regulation [13]. Furthermore, it was shown that a connection of expression data with metabolic modeling can be used to localize the tissue-specific metabolism in flux-balance models and especially predict post-transcriptionally regulated fluxes [172]. Besides, more complex models incorporating both types of data and thus connect gene expression and metabolism have not been developed yet but could potentially uncover links between the genotype and the phenotype.

With the results of these four papers, we could enhance current knowledge, biologically as well as methodically. A potential molecular description underlying punctuated equilibrium could be identified by time-series analysis of putative evolutionary walks on the metabolic network which were obtained from an enzyme-pathway coevolution model. We could furthermore explain potential characteristics in the gene-coexpression profile as residues of the evolutionary process, and lastly identify the phenotypes of six genes which are essential for plant growth by a novel clustering technique.

Bibliography

[1] ADAMI, C. Self-organized criticality in living systems. *Phys Let A 203* (1995), 29–32.

[2] ALBERT, R. Scale-free networks in cell biology. *J Cell Sci 118*, Pt 21 (Nov 2005), 4947–4957.

[3] ALBERT, R., AND BARABÁSI, A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys. 74*, 1 (Jan 2002), 47–97.

[4] ALONSO, J. M., STEPANOVA, A. N., LEISSE, T. J., KIM, C. J., CHEN, H., SHINN, P., STEVENSON, D. K., ZIMMERMAN, J., BARAJAS, P., CHEUK, R., GADRINAB, C., HELLER, C., JESKE, A., KOESEMA, E., MEYERS, C. C., PARKER, H., PREDNIS, L., ANSARI, Y., CHOY, N., DEEN, H., GERALT, M., HAZARI, N., HOM, E., KARNES, M., MULHOLLAND, C., NDUBAKU, R., SCHMIDT, I., GUZMAN, P., AGUILAR-HENONIN, L., SCHMID, M., WEIGEL, D., CARTER, D. E., MARCHAND, T., RISSEEUW, E., BROGDEN, D., ZEKO, A., CROSBY, W. L., BERRY, C. C., AND ECKER, J. R. Genome-wide insertional mutagenesis of arabidopsis thaliana. *Science 301*, 5633 (Aug 2003), 653–657.

[5] AOKI, K., OGATA, Y., AND SHIBATA, D. Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant Cell Physiol 48*, 3 (Mar 2007), 381–390.

[6] ARITA, M. The metabolic world of escherichia coli is not small. *Proc Natl Acad Sci U S A 101*, 6 (Feb 2004), 1543–1547.

[7] BACHMANN, P. A., LUISI, P. L., AND LANG, J. Autocatalytic self-replicating micelles as models for prebiotic structures. *Nature 357* (1992), 57–59.

[8] BADER, G. D., AND HOGUE, C. W. V. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics 4* (Jan 2003), 2.

[9] BAERENFALLER, K., GROSSMANN, J., GROBEI, M. A., HULL, R., HIRSCH-HOFFMANN, M., YALOVSKY, S., ZIMMERMANN, P., GROSSNIKLAUS, U., GRUISSEM, W., AND BAGINSKY, S. Genome-scale proteomics reveals arabidopsis thaliana gene models and proteome dynamics. *Science 320*, 5878 (May 2008), 938–941.

[10] BAK, P., AND SNEPPEN, K. Punctuated equilibrium and criticality in a simple model of evolution. *Phys Rev Lett 71*, 24 (Dec 1993), 4083–4086.

[11] BAK, P., TANG, C., AND WIESENFELD, K. Self-organized criticality: An explanation of the 1/f noise. *Phys Rev Lett 59*, 4 (Jul 1987), 381–384.

[12] BAK, P., TANG, C., AND WIESENFELD, K. Self-organized criticality. *Phys Rev A 38*, 1 (Jul 1988), 364–374.

[13] BAR-EVEN, A., PAULSSON, J., MAHESHRI, N., CARMI, M., O'SHEA, E., PILPEL, Y., AND BARKAI, N. Noise in protein expression scales with natural protein abundance. *Nat Genet 38*, 6 (Jun 2006), 636–643.

[14] BARABASI, A. L., AND ALBERT, R. Emergence of scaling in random networks. *Science 286*, 5439 (Oct 1999), 509–512.

[15] BARABÁSI, A.-L., AND BONABEAU, E. Scale-free networks. *Sci Am 288*, 5 (May 2003), 60–69.

[16] BARABÁSI, A.-L., AND OLTVAI, Z. N. Network biology: understanding the cell's functional organization. *Nat Rev Genet 5*, 2 (Feb 2004), 101–113.

[17] BARRETT, T., TROUP, D. B., WILHITE, S. E., LEDOUX, P., RUDNEV, D., EVANGELISTA, C., KIM, I. F., SOBOLEVA, A., TOMASHEVSKY, M., MARSHALL, K. A., PHILLIPPY, K. H., SHERMAN, P. M., MUERTTER, R. N., AND EDGAR, R. Ncbi geo: archive for high-throughput functional genomic data. *Nucleic Acids Res 37*, Database issue (Jan 2009), D885–D890.

[18] BERGMANN, S., IHMELS, J., AND BARKAI, N. Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol 2*, 1 (Jan 2004), E9.

[19] BERTHOMÉ, R., THOMASSET, M., MAENE, M., BOURGEOIS, N., FROGER, N., AND BUDAR, F.

pur4 mutations are lethal to the male, but not the female, gametophyte and affect sporophyte development in arabidopsis. *Plant Physiol 147*, 2 (Jun 2008), 650–660.

[20] BHAN, A., GALAS, D. J., AND DEWEY, T. G. A duplication growth model of gene expression networks. *Bioinformatics 18*, 11 (Nov 2002), 1486–1493.

[21] BOLLOBÁS, B. *Modern Graph Theory*. Springer, 2002.

[22] BORENSTEIN, E., KUPIEC, M., FELDMAN, M. W., AND RUPPIN, E. Large-scale reconstruction and phylogenetic analysis of metabolic environments. *Proc Natl Acad Sci U S A 105*, 38 (Sep 2008), 14482–14487.

[23] BRENNER, S., JOHNSON, M., BRIDGHAM, J., GOLDA, G., LLOYD, D. H., JOHNSON, D., LUO, S., MCCURDY, S., FOY, M., EWAN, M., ROTH, R., GEORGE, D., ELETR, S., ALBRECHT, G., VERMAAS, E., WILLIAMS, S. R., MOON, K., BURCHAM, T., PALLAS, M., DUBRIDGE, R. B., KIRCHNER, J., FEARON, K., MAO, J., AND CORCORAN, K. Gene expression analysis by massively parallel signature sequencing (mpss) on microbead arrays. *Nat Biotechnol 18*, 6 (Jun 2000), 630–634.

[24] BROWN, D. M., ZEEF, L. A. H., ELLIS, J., GOODACRE, R., AND TURNER, S. R. Identification of novel genes in arabidopsis involved in secondary cell wall formation using expression profiling and reverse genetics. *Plant Cell 17*, 8 (Aug 2005), 2281–2295.

[25] BURGARD, A. P., AND MARANAS, C. D. Optimization-based framework for inferring and testing hypothesized metabolic objective functions. *Biotechnol Bioeng 82*, 6 (Jun 2003), 670–677.

[26] BURGARD, A. P., PHARKYA, P., AND MARANAS, C. D. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol Bioeng 84*, 6 (Dec 2003), 647–657.

[27] CARLSON, M. R. J., ZHANG, B., FANG, Z., MISCHEL, P. S., HORVATH, S., AND NELSON, S. F. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics 7* (2006), 40.

[28] CHAU, H. F. Scaling behavior of the punctuated-equilibrium model of evolution. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 49*, 5 (May 1994), 4691–4692.

[29] CHEN, Q. W., AND CHEN, C. L. The role of inorganic compounds in the prebiotic synthesis of organic molecules. *Current Organic Chemistry 9*, 10 (2005), 989–998.

[30] CHRISTIAN, N., MAY, P., KEMPA, S., HANDORF, T., AND EBENHÖH, O. An integrative approach towards completing genome-scale metabolic networks. *Mol Biosyst 5*, 12 (Dec 2009), 1889–1903.

[31] CHURCH, G. M. Genomes for all. *Sci Am 294*, 1 (Jan 2006), 46–54.

[32] COPLEY, S. D., SMITH, E., AND MOROWITZ, H. J. The origin of the rna world: co-evolution of genes and metabolism. *Bioorg Chem 35*, 6 (Dec 2007), 430–443.

[33] COX, D., AND ISHAM, V. *Point Processes*. Chapman and Hall, London, 1980.

[34] COX, D., AND LEWIS, L. *The Statistical Analysis of Series and Events*. Wiley, New York, 1966.

[35] DAUB, C. O., STEUER, R., SELBIG, J., AND KLOSKA, S. Estimating mutual information using b-spline functions–an improved similarity measure for analysing gene expression data. *BMC Bioinformatics 5* (Aug 2004), 118.

[36] DAVID, L. A., AND ALM, E. J. Rapid evolutionary innovation during an archaean genetic expansion. *Nature 469*, 7328 (Jan 2011), 93–96.

[37] DAVIES DL, B. D. A cluster separation measure. *IEEE Trans. Pattern Anal. Machine Intell. 1* (1979), 224–227.

[38] DAWKINS, R. *The Blind Watchmaker*. Penguin, London, 2000.

[39] DERISI, J. L., IYER, V. R., AND BROWN, P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science 278*, 5338 (Oct 1997), 680–686.

[40] DIJKSTRA, E. A note on two problems in connexion with graphs. *Numerische Mathematik 1* (1959), 269–271.

[41] DOKHOLYAN, N. V., SHAKHNOVICH, B., AND SHAKHNOVICH, E. I. Expanding protein universe and its origin from the biological big bang. *Proc Natl Acad Sci U S A 99*, 22 (Oct 2002), 14132–

14136.

[42] DUARTE, N. C., BECKER, S. A., JAMSHIDI, N., THIELE, I., MO, M. L., VO, T. D., SRIVAS, R., AND PALSSON, B. O. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci U S A 104*, 6 (Feb 2007), 1777–1782.

[43] DYSON, F. *Origins of Life*. Cambridge University Press, 1999.

[44] EBENHÖH, O., HANDORF, T., AND HEINRICH, R. Structural analysis of expanding metabolic networks. *Genome Inform 15*, 1 (2004), 35–45.

[45] EBENHÖH, O., HANDORF, T., AND HEINRICH, R. A cross species comparison of metabolic network functions. *Genome Inform 16*, 1 (2005), 203–213.

[46] EBENHÖH, O., HANDORF, T., AND KAHN, D. Evolutionary changes of metabolic networks and their biosynthetic capacities. *Syst Biol (Stevenage) 153*, 5 (Sep 2006), 354–358.

[47] EDWARDS, J. S., IBARRA, R. U., AND PALSSON, B. O. In silico predictions of escherichia coli metabolic capabilities are consistent with experimental data. *Nat Biotechnol 19*, 2 (Feb 2001), 125–130.

[48] EDWARDS, J. S., AND PALSSON, B. O. The escherichia coli mg1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc Natl Acad Sci U S A 97*, 10 (May 2000), 5528–5533.

[49] ELDREDGE, N., AND GOULD, J. G. *Models in Paleobiology*. T.Schopf, 1972, ch. Punctuated equilibria: an alternative to phyletic gradualism, pp. 82–115.

[50] ELENA, S. F., COOPER, V. S., AND LENSKI, R. E. Punctuated evolution caused by selection of rare beneficial mutations. *Science 272*, 5269 (Jun 1996), 1802–1804.

[51] ENRIGHT, A. J., DONGEN, S. V., AND OUZOUNIS, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res 30*, 7 (Apr 2002), 1575–1584.

[52] FANO, U. Ionization yield to radiation. ii. the fluctuation of the random number. *Phys. Rev. 72*, 1 (1947), 26–29.

[53] FIERS, W., CONTRERAS, R., DUERINCK, F., HAEGEMAN, G., ISERENTANT, D., MERREGAERT, J., JOU, W. M., MOLEMANS, F., RAEYMAEKERS, A., DEN BERGHE, A. V., VOLCKAERT, G., AND YSEBAERT, M. Complete nucleotide sequence of bacteriophage ms2 rna: primary and secondary structure of the replicase gene. *Nature 260*, 5551 (Apr 1976), 500–507.

[54] FINN, R. D., MISTRY, J., TATE, J., COGGILL, P., HEGER, A., POLLINGTON, J. E., GAVIN, O. L., GUNASEKARAN, P., CERIC, G., FORSLUND, K., HOLM, L., SONNHAMMER, E. L. L., EDDY, S. R., AND BATEMAN, A. The pfam protein families database. *Nucleic Acids Res 38*, Database issue (Jan 2010), D211–D222.

[55] FLEISCHMANN, R. D., ADAMS, M. D., WHITE, O., CLAYTON, R. A., KIRKNESS, E. F., KERLAVAGE, A. R., BULT, C. J., TOMB, J. F., DOUGHERTY, B. A., AND MERRICK, J. M. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science 269*, 5223 (Jul 1995), 496–512.

[56] FLORES-PÉREZ, U., SAURET-GÜETO, S., GAS, E., JARVIS, P., AND RODRÍGUEZ-CONCEPCIÓN, M. A mutant impaired in the production of plastome-encoded proteins uncovers a mechanism for the homeostasis of isoprenoid biosynthetic enzymes in arabidopsis plastids. *Plant Cell 20*, 5 (May 2008), 1303–1315.

[57] FONTANA, W., AND SCHUSTER, P. Continuity in evolution: on the nature of transitions. *Science 280*, 5368 (May 1998), 1451–1455.

[58] FREEMAN, T. C., GOLDOVSKY, L., BROSCH, M., VAN DONGEN, S., MAZIÈRE, P., GROCOCK, R. J., FREILICH, S., THORNTON, J., AND ENRIGHT, A. J. Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS Comput Biol 3*, 10 (Oct 2007), 2032–2042.

[59] FÖRSTER, J., FAMILI, I., FU, P., PALSSON, B. O., AND NIELSEN, J. Genome-scale reconstruction of the saccharomyces cerevisiae metabolic network. *Genome Res 13*, 2 (Feb 2003), 244–253.

[60] GALPERIN, M. Y., WALKER, D. R., AND KOONIN, E. V. Analogous enzymes: independent inventions in enzyme evolution. *Genome Res 8*, 8 (Aug 1998), 779–790.

[61] GANDHI, T. K. B., ZHONG, J., MATHIVANAN, S., KARTHICK, L., CHANDRIKA, K. N., MOHAN, S. S., SHARMA, S., PINKERT, S., NAGARAJU, S., PERIASWAMY, B., MISHRA, G., NANDAKUMAR, K., SHEN, B., DESHPANDE, N., NAYAK, R., SARKER, M., BOEKE, J. D., PARMIGIANI, G., SCHULTZ, J., BADER, J. S., AND PANDEY, A. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet 38*, 3 (Mar 2006), 285–293.

[62] GARDINER, C. *Handbook of stochastic methods*. Springer, Berlin, 1985.

[63] GIBBONS, F. D., AND ROTH, F. P. Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res 12*, 10 (Oct 2002), 1574–1581.

[64] GILLEPSIE, D. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem. 8* (1977), 2340–2354.

[65] GOUGH, J., KARPLUS, K., HUGHEY, R., AND CHOTHIA, C. Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. *J Mol Biol 313*, 4 (Nov 2001), 903–919.

[66] GRANICK, S. Speculations on the origins and evolution of photosynthesis. *Ann N Y Acad Sci 69*, 2 (Aug 1957), 292–308.

[67] GU, X., WANG, Y., AND GU, J. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nat Genet 31*, 2 (Jun 2002), 205–209.

[68] HANDORF, T., CHRISTIAN, N., EBENHÖH, O., AND KAHN, D. An environmental perspective on metabolism. *J Theor Biol 252*, 3 (Jun 2008), 530–537.

[69] HANDORF, T., AND EBENHÖH, O. Metapath online: a web server implementation of the network expansion algorithm. *Nucleic Acids Res 35*, Web Server issue (Jul 2007), W613–W618.

[70] HANDORF, T., EBENHÖH, O., AND HEINRICH, R. Expanding metabolic networks: scopes of compounds, robustness, and evolution. *J Mol Evol 61*, 4 (Oct 2005), 498–512.

[71] HARTIGAN, J., AND WONG, M. A k-means clustering algorithm. *Applied Statistics 28* (1979), 100–108.

[72] HEINRICH, R., AND SCHUSTER, S. *The Regulation of Cellular Systems*. Chapman & Hall, New York, 1996.

[73] HINTZE, A., AND ADAMI, C. Evolution of complex modular biological networks. *PLoS Comput Biol 4*, 2 (Feb 2008), e23.

[74] HIRAI, M. Y., SUGIYAMA, K., SAWADA, Y., TOHGE, T., OBAYASHI, T., SUZUKI, A., ARAKI, R., SAKURAI, N., SUZUKI, H., AOKI, K., GODA, H., NISHIZAWA, O. I., SHIBATA, D., AND SAITO, K. Omics-based identification of arabidopsis myb transcription factors regulating aliphatic glucosinolate biosynthesis. *Proc Natl Acad Sci U S A 104*, 15 (Apr 2007), 6478–6483.

[75] HOROWITZ, N. H. On the evolution of biochemical syntheses. *Proc Natl Acad Sci U S A 31*, 6 (Jun 1945), 153–157.

[76] HTTP://BLAST.NCBI.NLM.NIH.GOV/.

[77] HTTP://WWW.PERSONALGENOMES.ORG/.

[78] HUBERT, L., AND ARABIE, P. Comparing partitions. *Journal of Classification 2* (1985), 193–218.

[79] HUTHMACHER, C., GILLE, C., AND HOLZHÜTTER, H.-G. A computational analysis of protein interactions in metabolic networks reveals novel enzyme pairs potentially involved in metabolic channeling. *J Theor Biol 252*, 3 (Jun 2008), 456–464.

[80] HUTTENHOWER, C., FLAMHOLZ, A. I., LANDIS, J. N., SAHI, S., MYERS, C. L., OLSZEWSKI, K. L., HIBBS, M. A., SIEMERS, N. O., TROYANSKAYA, O. G., AND COLLER, H. A. Nearest neighbor networks: clustering expression data based on gene neighborhoods. *BMC Bioinformatics 8* (2007), 250.

[81] IHMELS, J., LEVY, R., AND BARKAI, N. Principles of transcriptional control in the metabolic network of saccharomyces cerevisiae. *Nat Biotechnol 22*, 1 (Jan 2004), 86–92.

[82] JENSEN, R. A. Enzyme recruitment in evolution of new function. *Annu Rev Microbiol 30* (1976), 409–425.

[83] JEONG, H., MASON, S. P., BARABÁSI, A. L., AND OLTVAI, Z. N. Lethality and centrality in protein

networks. *Nature 411*, 6833 (May 2001), 41–42.

[84] JOHNSTON, W. K., UNRAU, P. J., LAWRENCE, M. S., GLASNER, M. E., AND BARTEL, D. P. Rna-catalyzed rna polymerization: accurate and general rna-templated primer extension. *Science 292*, 5520 (May 2001), 1319–1325.

[85] JOSHI, T., AND XU, D. Quantitative assessment of relationship between sequence similarity and function similarity. *BMC Genomics 8* (2007), 222.

[86] JOU, W. M., HAEGEMAN, G., YSEBAERT, M., AND FIERS, W. Nucleotide sequence of the gene coding for the bacteriophage ms2 coat protein. *Nature 237*, 5350 (May 1972), 82–88.

[87] JUPITER, D. C., AND VANBUREN, V. A visual data mining tool that facilitates reconstruction of transcription regulatory networks. *PLoS One 3*, 3 (2008), e1717.

[88] KACSER, H., AND BEEBY, R. Evolution of catalytic proteins or on the origin of enzyme species by means of natural selection. *J Mol Evol 20*, 1 (1984), 38–51.

[89] KANEHISA, M., ARAKI, M., GOTO, S., HATTORI, M., HIRAKAWA, M., ITOH, M., KATAYAMA, T., KAWASHIMA, S., OKUDA, S., TOKIMATSU, T., AND YAMANISHI, Y. Kegg for linking genomes to life and the environment. *Nucleic Acids Res 36*, Database issue (Jan 2008), D480–D484.

[90] KANEHISA, M., AND GOTO, S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res 28*, 1 (Jan 2000), 27–30.

[91] KANEHISA, M., GOTO, S., FURUMICHI, M., TANABE, M., AND HIRAKAWA, M. Kegg for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res 38*, Database issue (Jan 2010), D355–D360.

[92] KANEHISA, M., GOTO, S., HATTORI, M., AOKI-KINOSHITA, K. F., ITOH, M., KAWASHIMA, S., KATAYAMA, T., ARAKI, M., AND HIRAKAWA, M. From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res 34*, Database issue (Jan 2006), D354–D357.

[93] KARP, P. D., OUZOUNIS, C. A., MOORE-KOCHLACS, C., GOLDOVSKY, L., KAIPA, P., AHRÉN, D., TSOKA, S., DARZENTAS, N., KUNIN, V., AND LÓPEZ-BIGAS, N. Expansion of the biocyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res 33*, 19 (2005), 6083–6089.

[94] KARTAL, O., AND EBENHÖH, O. Ground state robustness as an evolutionary design principle in signaling networks. *PLoS One 4*, 12 (2009), e8001.

[95] KAUFFMAN, S. Question 1: origin of life and the living state. *Orig Life Evol Biosph 37*, 4-5 (Oct 2007), 315–322.

[96] KAUFFMAN, S. A. *The origins of order: self-organization and selection in evolution*. Oxford University Press, 1993.

[97] KELLIS, M., BIRREN, B. W., AND LANDER, E. S. Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature 428*, 6983 (Apr 2004), 617–624.

[98] KHERSONSKY, O., ROODVELDT, C., AND TAWFIK, D. S. Enzyme promiscuity: evolutionary and mechanistic aspects. *Curr Opin Chem Biol 10*, 5 (Oct 2006), 498–508.

[99] KIM, H. S., MITTENTHAL, J. E., AND CAETANO-ANOLLÉS, G. Manet: tracing evolution of protein architecture in metabolic networks. *BMC Bioinformatics 7* (2006), 351.

[100] KING, A. D., PRZULJ, N., AND JURISICA, I. Protein complex prediction via cost-based clustering. *Bioinformatics 20*, 17 (Nov 2004), 3013–3020.

[101] KITANO, H. Systems biology: a brief overview. *Science 295*, 5560 (Mar 2002), 1662–1664.

[102] KLAMT, S., AND GILLES, E. D. Minimal cut sets in biochemical reaction networks. *Bioinformatics 20*, 2 (Jan 2004), 226–234.

[103] KLAMT, S., AND STELLING, J. Combinatorial complexity of pathway analysis in metabolic networks. *Mol Biol Rep 29*, 1-2 (2002), 233–236.

[104] KOELLE, K., COBEY, S., GRENFELL, B., AND PASCUAL, M. Epochal evolution shapes the phylodynamics of interpandemic influenza a (h3n2) in humans. *Science 314*, 5807 (Dec 2006), 1898–1903.

[105] KRAUSE, A., STOYE, J., AND VINGRON, M. Large scale hierarchical clustering of protein se-

quences. *BMC Bioinformatics 6* (2005), 15.

[106] LAMPORT, D. T. A., KIELISZEWSKI, M. J., AND SHOWALTER, A. M. Salt stress upregulates periplasmic arabinogalactan proteins: using salt stress to analyse agp function. *New Phytol 169*, 3 (2006), 479–492.

[107] LAZCANO, A., AND MILLER, S. L. On the origin of metabolic pathways. *J Mol Evol 49*, 4 (Oct 1999), 424–431.

[108] LEE, H. K., HSU, A. K., SAJDAK, J., QIN, J., AND PAVLIDIS, P. Coexpression analysis of human genes across many microarray data sets. *Genome Res 14*, 6 (Jun 2004), 1085–1094.

[109] LEE, R. E. C., AND MEGENEY, L. A. The yeast kinome displays scale free topology with functional hub clusters. *BMC Bioinformatics 6* (2005), 271.

[110] LERCHER, M. J., AND PÁL, C. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Mol Biol Evol 25*, 3 (Mar 2008), 559–567.

[111] LI, S., ARMSTRONG, C. M., BERTIN, N., GE, H., MILSTEIN, S., BOXEM, M., VIDALAIN, P.-O., HAN, J.-D. J., CHESNEAU, A., HAO, T., GOLDBERG, D. S., LI, N., MARTINEZ, M., RUAL, J.-F., LAMESCH, P., XU, L., TEWARI, M., WONG, S. L., ZHANG, L. V., BERRIZ, G. F., JACOTOT, L., VAGLIO, P., REBOUL, J., HIROZANE-KISHIKAWA, T., LI, Q., GABEL, H. W., ELEWA, A., BAUMGARTNER, B., ROSE, D. J., YU, H., BOSAK, S., SEQUERRA, R., FRASER, A., MANGO, S. E., SAXTON, W. M., STROME, S., HEUVEL, S. V. D., PIANO, F., VANDENHAUTE, J., SARDET, C., GERSTEIN, M., DOUCETTE-STAMM, L., GUNSALUS, K. C., HARPER, J. W., CUSICK, M. E., ROTH, F. P., HILL, D. E., AND VIDAL, M. A map of the interactome network of the metazoan c. elegans. *Science 303*, 5657 (Jan 2004), 540–543.

[112] LIGHT, S., AND KRAULIS, P. Network analysis of metabolic enzyme evolution in escherichia coli. *BMC Bioinformatics 5* (Feb 2004), 15.

[113] LIMA-MENDEZ, G., AND VAN HELDEN, J. The powerful law of the power law and other myths in network biology. *Mol Biosyst 5*, 12 (Dec 2009), 1482–1493.

[114] LYNCH, M., AND CONERY, J. S. The evolutionary fate and consequences of duplicate genes. *Science 290*, 5494 (Nov 2000), 1151–1155.

[115] MA, S., GONG, Q., AND BOHNERT, H. J. An arabidopsis gene network based on the graphical gaussian model. *Genome Res 17*, 11 (Nov 2007), 1614–1625.

[116] MANFIELD, I. W., JEN, C.-H., PINNEY, J. W., MICHALOPOULOS, I., BRADFORD, J. R., GILMARTIN, P. M., AND WESTHEAD, D. R. Arabidopsis co-expression tool (act): web server tools for microarray-based gene expression analysis. *Nucleic Acids Res 34*, Web Server issue (Jul 2006), W504–W509.

[117] MARTIN, W., AND RUSSELL, M. J. On the origin of biochemistry at an alkaline hydrothermal vent. *Philos Trans R Soc Lond B Biol Sci 362*, 1486 (Oct 2007), 1887–1925.

[118] MASLOV, S., KRISHNA, S., PANG, T. Y., AND SNEPPEN, K. Toolbox model of evolution of prokaryotic metabolic networks and their regulation. *Proc Natl Acad Sci U S A 106*, 24 (Jun 2009), 9743–9748.

[119] MATTHÄUS, F., SALAZAR, C., AND EBENHÖH, O. Biosynthetic potentials of metabolites and their hierarchical organization. *PLoS Comput Biol 4*, 4 (Apr 2008), e1000049.

[120] MAXAM, A. M., AND GILBERT, W. A new method for sequencing dna. *Proc Natl Acad Sci U S A 74*, 2 (Feb 1977), 560–564.

[121] MELÉNDEZ-HEVIA, E., AND ISIDORO, A. The game of the pentose phosphate cycle. *J Theor Biol 117*, 2 (Nov 1985), 251–263.

[122] MENTZEN, W. I., AND WURTELE, E. S. Regulon organization of arabidopsis. *BMC Plant Biol 8* (2008), 99.

[123] MIDDLETON, J. W., CHACRON, M. J., LINDNER, B., AND LONGTIN, A. Firing statistics of a neuron model driven by long-range correlated noise. *Phys Rev E Stat Nonlin Soft Matter Phys 68*, 2 Pt 1 (Aug 2003), 021920.

[124] MILLAR, A. A., AND GUBLER, F. The arabidopsis gamyb-like genes, myb33 and myb65, are microrna-regulated genes that redundantly facilitate anther development. *Plant Cell 17*, 3 (Mar

2005), 705–721.

[125] MILLER, S. L. A production of amino acids under possible primitive earth conditions. *Science 117*, 3046 (May 1953), 528–529.

[126] MILLER, S. L., AND UREY, H. C. Origin of life. *Science 130*, 3389 (Dec 1959), 1622–1624.

[127] MILO, R., ITZKOVITZ, S., KASHTAN, N., LEVITT, R., SHEN-ORR, S., AYZENSHTAT, I., SHEFFER, M., AND ALON, U. Superfamilies of evolved and designed networks. *Science 303*, 5663 (Mar 2004), 1538–1542.

[128] MILO, R., SHEN-ORR, S., ITZKOVITZ, S., KASHTAN, N., CHKLOVSKII, D., AND ALON, U. Network motifs: simple building blocks of complex networks. *Science 298*, 5594 (Oct 2002), 824–827.

[129] MITHANI, A., PRESTON, G. M., AND HEIN, J. A stochastic model for the evolution of metabolic networks with neighbor dependence. *Bioinformatics 25*, 12 (Jun 2009), 1528–1535.

[130] MOROWITZ, H. J. *Beginnings of Cellular Life*. Yale University Press, 2004.

[131] MUTWIL, M., KLIE, S., TOHGE, T., GIORGI, F. M., WILKINS, O., CAMPBELL, M. M., FERNIE, A. R., USADEL, B., NIKOLOSKI, Z., AND PERSSON, S. Planet: Combined sequence and expression comparisons across plant networks derived from seven species. *Plant Cell (under revision)* (2011).

[132] MUTWIL, M., OBRO, J., WILLATS, W. G. T., AND PERSSON, S. Genecat–novel webtools that combine blast and co-expression analyses. *Nucleic Acids Res 36*, Web Server issue (Jul 2008), W320–W326.

[133] MUTWIL, M., RUPRECHT, C., GIORGI, F. M., BRINGMANN, M., USADEL, B., AND PERSSON, S. Transcriptional wiring of cell wall-related genes in arabidopsis. *Mol Plant 2*, 5 (Sep 2009), 1015–1024.

[134] MUTWIL, M., USADEL, B., SCHÜTTE, M., LORAINE, A., EBENHÖH, O., AND PERSSON, S. Assembly of an interactive correlation network for the arabidopsis genome using a novel heuristic clustering algorithm. *Plant Physiol 152*, 1 (Jan 2010), 29–43.

[135] NEWMAN, M. E. J., AND GIRVAN, M. Finding and evaluating community structure in networks. *Phys. Rev. E 69* (2004), 026113.

[136] NOOR, E., EDEN, E., MILO, R., AND ALON, U. Central carbon metabolism as a minimal biochemical walk between precursors for biomass and energy. *Mol Cell 39*, 5 (Sep 2010), 809–820.

[137] NOTEBAART, R. A., KENSCHE, P. R., HUYNEN, M. A., AND DUTILH, B. E. Asymmetric relationships between proteins shape genome evolution. *Genome Biol 10*, 2 (2009), R19.

[138] NOWAK, M. A. Five rules for the evolution of cooperation. *Science 314*, 5805 (Dec 2006), 1560–1563.

[139] OBAYASHI, T., HAYASHI, S., SAEKI, M., OHTA, H., AND KINOSHITA, K. Atted-ii provides coexpressed gene networks for arabidopsis. *Nucleic Acids Res 37*, Database issue (Jan 2009), D987–D991.

[140] OHNO, S. *Evolution by gene duplication*. Springer-Verlag, 1970.

[141] PACZUSKI, M.AND S. MASLOV, S., AND BAK, P. Avanlance dynamics in evolution, growth, and depinning models. *Phys Rev E 53* (1995), 414–443.

[142] PERSSON, S., WEI, H., MILNE, J., PAGE, G. P., AND SOMERVILLE, C. R. Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proc Natl Acad Sci U S A 102*, 24 (Jun 2005), 8633–8638.

[143] PFEIFFER, T., SOYER, O. S., AND BONHOEFFER, S. The evolution of connectivity in metabolic networks. *PLoS Biol 3*, 7 (Jul 2005), e228.

[144] PHALAKORNKULE, C., LEE, S., ZHU, T., KOEPSEL, R., ATAAI, M. M., GROSSMANN, I. E., AND DOMACH, M. M. A milp-based flux alternative generation and nmr experimental design strategy for metabolic engineering. *Metab Eng 3*, 2 (Apr 2001), 124–137.

[145] PRICE, N. D., REED, J. L., AND PALSSON, B. O. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol 2*, 11 (Nov 2004), 886–897.

[146] PRIETO, C., RISUEÑO, A., FONTANILLO, C., AND LAS RIVAS, J. D. Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles. *PLoS One 3*, 12 (2008),

e3911.

[147] PÁL, C., PAPP, B., AND LERCHER, M. J. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet 37*, 12 (Dec 2005), 1372–1375.

[148] RAMAMOORTHY, V., CAHOON, E. B., THOKALA, M., KAUR, J., LI, J., AND SHAH, D. M. Sphingolipid c-9 methyltransferases are important for growth and virulence but not for sensitivity to antifungal plant defensins in fusarium graminearum. *Eukaryot Cell 8*, 2 (Feb 2009), 217–229.

[149] RAVASZ, E. Detecting hierarchical modularity in biological networks. *Methods Mol Biol 541* (2009), 145–160.

[150] RAYMOND, J., AND SEGRÈ, D. The effect of oxygen on biochemical networks and the evolution of complex life. *Science 311*, 5768 (Mar 2006), 1764–1767.

[151] RIEHL, W. J., KRAPIVSKY, P. L., REDNER, S., AND SEGRÈ, D. Signatures of arithmetic simplicity in metabolic network architecture. *PLoS Comput Biol 6*, 4 (Apr 2010), e1000725.

[152] RISON, S. C. G., TEICHMANN, S. A., AND THORNTON, J. M. Homology, pathway distance and chromosomal localization of the small molecule metabolism enzymes in escherichia coli. *J Mol Biol 318*, 3 (May 2002), 911–932.

[153] ROSEN, B. P., AJEES, A. A., AND MCDERMOTT, T. R. Life and death with arsenic: Arsenic life: An analysis of the recent report "a bacterium that can grow by using arsenic instead of phosphorus". *Bioessays* (Mar 2011).

[154] ROST, B. Enzyme function less conserved than anticipated. *J Mol Biol 318*, 2 (Apr 2002), 595–608.

[155] RUSS, W. P., LOWERY, D. M., MISHRA, P., YAFFE, M. B., AND RANGANATHAN, R. Natural-like function in artificial ww domains. *Nature 437*, 7058 (Sep 2005), 579–583.

[156] RZHETSKY, A., AND GOMEZ, S. M. Birth of scale-free molecular networks and the number of distinct dna and protein domains per genome. *Bioinformatics 17*, 10 (Oct 2001), 988–996.

[157] SANGER, F., NICKLEN, S., AND COULSON, A. R. Dna sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A 74*, 12 (Dec 1977), 5463–5467.

[158] SCHENA, M., SHALON, D., DAVIS, R. W., AND BROWN, P. O. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science 270*, 5235 (Oct 1995), 467–470.

[159] SCHILLING, C. H., EDWARDS, J. S., AND PALSSON, B. O. Toward metabolic phenomics: analysis of genomic data using flux balances. *Biotechnol Prog 15*, 3 (1999), 288–295.

[160] SCHILLING, C. H., LETSCHER, D., AND PALSSON, B. O. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J Theor Biol 203*, 3 (Apr 2000), 229–248.

[161] SCHILLING, C. H., AND PALSSON, B. O. The underlying pathway structure of biochemical reaction networks. *Proc Natl Acad Sci U S A 95*, 8 (Apr 1998), 4193–4198.

[162] SCHILLING, C. H., SCHUSTER, S., PALSSON, B. O., AND HEINRICH, R. Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnol Prog 15*, 3 (1999), 296–303.

[163] SCHMID, M., DAVISON, T. S., HENZ, S. R., PAPE, U. J., DEMAR, M., VINGRON, M., SCHÖLKOPF, B., WEIGEL, D., AND LOHMANN, J. U. A gene expression map of arabidopsis thaliana development. *Nat Genet 37*, 5 (May 2005), 501–506.

[164] SCHMIDT, S., SUNYAEV, S., BORK, P., AND DANDEKAR, T. Metabolites: a helping hand for pathway evolution? *Trends Biochem Sci 28*, 6 (Jun 2003), 336–341.

[165] SCHUSTER, S., FELL, D. A., AND DANDEKAR, T. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat Biotechnol 18*, 3 (Mar 2000), 326–332.

[166] SCHUSTER, S. C. Next-generation sequencing transforms today's biology. *Nat Methods 5*, 1 (Jan 2008), 16–18.

[167] SCHÜTTE, M., KLITGORD, N., SEGRÈ, D., AND EBENHÖH, O. Co-evolution of metabolism and protein sequences. *Genome Inform 22* (Jan 2010), 156–166.

[168] SEGRÈ, D., BEN-ELI, D., DEAMER, D. W., AND LANCET, D. The lipid world. *Orig Life Evol Biosph 31*, 1-2 (2001), 119–145.

[169] SEGRÈ, D., DELUNA, A., CHURCH, G. M., AND KISHONY, R. Modular epistasis in yeast metabolism. *Nat Genet 37*, 1 (Jan 2005), 77–83.

[170] SEGRÈ, D., VITKUP, D., AND CHURCH, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A 99*, 23 (Nov 2002), 15112–15117.

[171] SHLOMI, T., BERKMAN, O., AND RUPPIN, E. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci U S A 102*, 21 (May 2005), 7695–7700.

[172] SHLOMI, T., CABILI, M. N., HERRGARD, M. J., PALSSON, B. O., AND RUPPIN, E. Network-based prediction of human tissue-specific metabolism. *Nat Biotechnol 26*, 9 (Sep 2008), 1003–1010.

[173] SMITH, D. J., LAPEDES, A. S., DE JONG, J. C., BESTEBROER, T. M., RIMMELZWAAN, G. F., OSTERHAUS, A. D. M. E., AND FOUCHIER, R. A. M. Mapping the antigenic and genetic evolution of influenza virus. *Science 305*, 5682 (Jul 2004), 371–376.

[174] SMITH, E., AND MOROWITZ, H. J. Universality in intermediary metabolism. *Proc Natl Acad Sci U S A 101*, 36 (Sep 2004), 13168–13173.

[175] SOBOLEVSKY, Y., FRENKEL, Z. M., AND TRIFONOV, E. N. Combinations of ancestral modules in proteins. *J Mol Evol 65*, 6 (Dec 2007), 640–650.

[176] SOBOLEVSKY, Y., AND TRIFONOV, E. N. Conserved sequences of prokaryotic proteomes and their compositional age. *J Mol Evol 61*, 5 (Nov 2005), 591–596.

[177] SOBOLEVSKY, Y., AND TRIFONOV, E. N. Protein modules conserved since luca. *J Mol Evol 63*, 5 (Nov 2006), 622–634.

[178] SPIRIN, V., GELFAND, M. S., MIRONOV, A. A., AND MIRNY, L. A. A metabolic network in the evolutionary context: multiscale structure and modularity. *Proc Natl Acad Sci U S A 103*, 23 (Jun 2006), 8774–8779.

[179] SRINIVASASAINAGENDRA, V., PAGE, G. P., MEHTA, T., COULIBALY, I., AND LORAINE, A. E. Cressexpress: a tool for large-scale mining of expression data from arabidopsis. *Plant Physiol 147*, 3 (Jul 2008), 1004–1016.

[180] STEINHAUSER, D., USADEL, B., LUEDEMANN, A., THIMM, O., AND KOPKA, J. Csb.db: a comprehensive systems-biology database. *Bioinformatics 20*, 18 (Dec 2004), 3647–3651.

[181] STEUER, R., HUMBURG, P., AND SELBIG, J. Validation and functional annotation of expression-based clusters based on gene ontology. *BMC Bioinformatics 7* (2006), 380.

[182] STUART, J. M., SEGAL, E., KOLLER, D., AND KIM, S. K. A gene-coexpression network for global discovery of conserved genetic modules. *Science 302*, 5643 (Oct 2003), 249–255.

[183] TANG, H., BOWERS, J. E., WANG, X., MING, R., ALAM, M., AND PATERSON, A. H. Synteny and collinearity in plant genomes. *Science 320*, 5875 (Apr 2008), 486–488.

[184] TANG, H., WANG, X., BOWERS, J. E., MING, R., ALAM, M., AND PATERSON, A. H. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res 18*, 12 (Dec 2008), 1944–1954.

[185] TATUSOV, R. L., FEDOROVA, N. D., JACKSON, J. D., JACOBS, A. R., KIRYUTIN, B., KOONIN, E. V., KRYLOV, D. M., MAZUMDER, R., MEKHEDOV, S. L., NIKOLSKAYA, A. N., RAO, B. S., SMIRNOV, S., SVERDLOV, A. V., VASUDEVAN, S., WOLF, Y. I., YIN, J. J., AND NATALE, D. A. The cog database: an updated version includes eukaryotes. *BMC Bioinformatics 4* (Sep 2003), 41.

[186] TATUSOV, R. L., GALPERIN, M. Y., NATALE, D. A., AND KOONIN, E. V. The cog database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res 28*, 1 (Jan 2000), 33–36.

[187] TATUSOV, R. L., KOONIN, E. V., AND LIPMAN, D. J. A genomic perspective on protein families. *Science 278*, 5338 (Oct 1997), 631–637.

[188] TATUSOV, R. L., NATALE, D. A., GARKAVTSEV, I. V., TATUSOVA, T. A., SHANKAVARAM, U. T., RAO, B. S., KIRYUTIN, B., GALPERIN, M. Y., FEDOROVA, N. D., AND KOONIN, E. V. The cog database: new developments in phylogenetic classification of proteins from complete genomes.

Nucleic Acids Res 29, 1 (Jan 2001), 22–28.

[189] TEICHMANN, S. A., AND BABU, M. M. Gene regulatory network growth by duplication. *Nat Genet 36*, 5 (May 2004), 492–496.

[190] THEVISSEN, K., CAMMUE, B. P., LEMAIRE, K., WINDERICKX, J., DICKSON, R. C., LESTER, R. L., FERKET, K. K., EVEN, F. V., PARRET, A. H., AND BROEKAERT, W. F. A gene encoding a sphingolipid biosynthesis enzyme determines the sensitivity of saccharomyces cerevisiae to an antifungal plant defensin from dahlia (dahlia merckii). *Proc Natl Acad Sci U S A 97*, 17 (Aug 2000), 9531–9536.

[191] THEVISSEN, K., IDKOWIAK-BALDYS, J., IM, Y.-J., TAKEMOTO, J., FRANÇOIS, I. E. J. A., FERKET, K. K. A., AERTS, A. M., MEERT, E. M. K., WINDERICKX, J., ROOSEN, J., AND CAMMUE, B. P. A. Skn1, a novel plant defensin-sensitivity gene in saccharomyces cerevisiae, is implicated in sphingolipid biosynthesis. *FEBS Lett 579*, 9 (Mar 2005), 1973–1977.

[192] TIAN, W., AND SKOLNICK, J. How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol 333*, 4 (Oct 2003), 863–882.

[193] TOUFIGHI, K., BRADY, S. M., AUSTIN, R., LY, E., AND PROVART, N. J. The botany array resource: e-northerns, expression angling, and promoter analyses. *Plant J 43*, 1 (Jul 2005), 153–163.

[194] USADEL, B., NAGEL, A., STEINHAUSER, D., GIBON, Y., BLÄSING, O. E., REDESTIG, H., SREENIVASULU, N., KRALL, L., HANNAH, M. A., POREE, F., FERNIE, A. R., AND STITT, M. Pageman: an interactive ontology tool to generate, display, and annotate overview graphs for profiling experiments. *BMC Bioinformatics 7* (2006), 535.

[195] USADEL, B., OBAYASHI, T., MUTWIL, M., GIORGI, F. M., BASSEL, G. W., TANIMOTO, M., CHOW, A., STEINHAUSER, D., PERSSON, S., AND PROVART, N. J. Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant Cell Environ 32*, 12 (Dec 2009), 1633–1651.

[196] USHER, M., STEMMLER, M., AND OLAMI, Z. Dynamic pattern formation leads to 1/f noise in neural populations. *Phys Rev Lett 74*, 2 (Jan 1995), 326–329.

[197] VAN DONGEN, S. *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, 2000.

[198] VAN KAMPEN, N. *Stochastic processes in physics and chemistry*. North-Holland, Amsterdam, 2001.

[199] VAN NOORT, V., SNEL, B., AND HUYNEN, M. A. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep 5*, 3 (Mar 2004), 280–284.

[200] VANDEPOELE, K., QUIMBAYA, M., CASNEUF, T., VEYLDER, L. D., AND DE PEER, Y. V. Unraveling transcriptional control in arabidopsis using cis-regulatory elements and coexpression networks. *Plant Physiol 150*, 2 (Jun 2009), 535–546.

[201] VASAS, V., SZATHMÁRY, E., AND SANTOS, M. Lack of evolvability in self-sustaining autocatalytic networks: A constraint on the metabolism-first path to the origin of life. *Proc Natl Acad Sci U S A* (Jan 2010).

[202] VITKUP, D., KHARCHENKO, P., AND WAGNER, A. Influence of metabolic network structure and function on enzyme evolution. *Genome Biol 7*, 5 (2006), R39.

[203] WAGNER, A. Neutralism and selectionism: a network-based reconciliation. *Nat Rev Genet 9*, 12 (Dec 2008), 965–974.

[204] WAGNER, A. Robustness and evolvability: a paradox resolved. *Proc Biol Sci 275*, 1630 (Jan 2008), 91–100.

[205] WAGNER, A., AND FELL, D. A. The small world inside large metabolic networks. *Proc Biol Sci 268*, 1478 (Sep 2001), 1803–1810.

[206] WALDE, P., WICK, R., FRESTA, M., MANGONE, A., AND LUISI, P. L. Autopoietic self-reproduction of fatty acid vesicles. *Journal of the American Chemical Society 116*, 26 (1994), 11649–11654.

[207] WASSERMAN, S., AND FAUST, K. *Social Network Analysis*. Cambridge Univ. Press, 1994.

[208] WATTS, D. J., AND STROGATZ, S. H. Collective dynamics of 'small-world' networks. *Nature 393*, 6684 (Jun 1998), 440–442.

[209] WAY, J., AND SILVER, P. Systems engineering without an engineer: Why we need systems biology. *Complexity 13* (2007), 22–29.

[210] WEI, H., PERSSON, S., MEHTA, T., SRINIVASASAINAGENDRA, V., CHEN, L., PAGE, G. P., SOMERVILLE, C., AND LORAINE, A. Transcriptional coordination of the metabolic network in arabidopsis. *Plant Physiol 142*, 2 (Oct 2006), 762–774.

[211] WEINHOLD, N., SANDER, O., DOMINGUES, F. S., LENGAUER, T., AND SOMMER, I. Local function conservation in sequence and structure space. *PLoS Comput Biol 4*, 7 (2008), e1000105.

[212] WILKINSON, E., AND WILLEMSEN, J. Invasion perculation: a new form of perculation theory. *J Phys A 16* (1983), 3365–3376.

[213] WOLFE-SIMON, F., BLUM, J. S., KULP, T. R., GORDON, G. W., HOEFT, S. E., PETT-RIDGE, J., STOLZ, J. F., WEBB, S. M., WEBER, P. K., DAVIES, P. C. W., ANBAR, A. D., AND OREMLAND, R. S. A bacterium that can grow by using arsenic instead of phosphorus. *Science* (Dec 2010).

[214] YCAS, M. On earlier states of the biochemical system. *J Theor Biol 44*, 1 (Mar 1974), 145–160.

[215] ZAMPIERI, M., SORANZO, N., BIANCHINI, D., AND ALTAFINI, C. Origin of co-expression patterns in e. coli and s. cerevisiae emerging from reverse engineering algorithms. *PLoS One 3*, 8 (2008), e2981.

[216] ZELDOVICH, K. B., CHEN, P., SHAKHNOVICH, B. E., AND SHAKHNOVICH, E. I. A first-principles model of early evolution: emergence of gene families, species, and preferred protein folds. *PLoS Comput Biol 3*, 7 (Jul 2007), e139.

[217] ZHANG, Z., LUO, Z. W., KISHINO, H., AND KEARSEY, M. J. Divergence pattern of duplicate genes in protein-protein interactions follows the power law. *Mol Biol Evol 22*, 3 (Mar 2005), 501–505.

[218] ZHONG, R., AND YE, Z.-H. Regulation of cell wall biosynthesis. *Curr Opin Plant Biol 10*, 6 (Dec 2007), 564–572.

[219] ZHU, J., SANBORN, J. Z., DIEKHANS, M., LOWE, C. B., PRINGLE, T. H., AND HAUSSLER, D. Comparative genomics search for losses of long-established genes on the human lineage. *PLoS Comput Biol 3*, 12 (Dec 2007), e247.

[220] ZIMMERMANN, P., HIRSCH-HOFFMANN, M., HENNIG, L., AND GRUISSEM, W. Genevestigator. arabidopsis microarray database and analysis toolbox. *Plant Physiol 136*, 1 (Sep 2004), 2621–2632.

[221] ZOTENKO, E., MESTRE, J., O'LEARY, D. P., AND PRZYTYCKA, T. M. Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput Biol 4*, 8 (2008), e1000140.

[222] ZWANZIG, R., SZABO, A., AND BAGCHI, B. Levinthal's paradox. *Proc Natl Acad Sci U S A 89*, 1 (Jan 1992), 20–22.

7. Contributions

Every chapter of this thesis has been published in a peer-reviewed journal.

I performed this work in collaboration with the research group of Daniel Segrè at Boston University, Chapters 2 and 3, and with the group of Staffan Persson at the Max Planck Institute of Molecular Plant Physiology in Potsdam, Chapters 4 and 5.

The underlying research idea of Chapters 2 and 3 was designed by Daniel Segrè and Oliver Ebenhöh.

For the results of Chapter 2 I collected the data, performed the sequence comparisons, and implemented the analysis tools. Niels Klitgord introduced me to sequencing techniques using BLAST and I subsequently performed all sequence alignments. All authors analyzed the results and wrote the article.

The simulations of Chapter 3 extend the method of network expansion [44, 45, 46, 70]. I implemented an upgraded version that also takes the sequence information of the enzymes by the Gillespie algorithm into account and I previously parsed all the data into suitable formats. Using this program I performed the simulations and further wrote analysis tools for the results in Chapter 3. Alexander Skupin supported me with his expertise in time-series analysis and gave advice for further analyses which I then performed. All authors wrote the article.

For the work in Chapter 4, Marek Mutwil created the HCCA clustering algorithm, the corresponding web tool, and performed the mutant analysis experiments. I contributed the cluster comparisons for which I used the existing methods MCL and MCODE to cluster the data. Further, I wrote analysis tools or used already available R-packages for modularity score, Davies-Bouldin score, and adjusted Rand index and analyzed the clustering solutions, Figs. 4.3 and B.5 and Tables B.1, B.2, B.3.

The research idea of Chapter 5 was designed by myself. I also implemented the model and analyzed the results together with Oliver Ebenhöh. Marek Mutwil and Staffan Persson supported me with the organism-specific data and biological expertise. With support from Oliver Ebenhöh I wrote the paper.

A. Supplementary Materials: Metabolic Evolution

Along with the manuscript we publish additional files supporting our findings.

Supplementary figure A.1 provides further results that help understanding the expansion process. A shows the number of metabolites as a function of the enzyme time. The curves are very similar to Figure 1A of the main text except for scaling. B Histograms of the final time, i.e. the absolute time when all enzymes are found during the expansion. The distributions for different values of γ look similar, however the time differs by orders of magnitude. The mean of the particular distributions was used to normalize the absolute time to normalized time for every γ. C This corresponds to Figure 1B of the main text, where here the number of possible reactions is shown. The most apparent difference to Figure 1B (counting numbers of enzymes) is the change in skewness. D Complementing Figure 1C of the main text, in which we showed how quickly border metabolites are explored during the expansion; this histogram shows how many metabolites (y axis) appear in how many reactions (x axis). Top: the subnetwork of KEGG covered through the expansion process; bottom: entire KEGG database. In the top chart one metabolite (water) takes part in 1202 reactions. Most of the metabolites occur in only a few reactions. Overall, the expanded subnetwork is better connected than KEGG (29.5% vs. 40% border metabolites).

Supplementary figure A.2 and A.3: Here, we investigated the appearance of enzyme classes. The heat maps show for one simulation (the same as in Figure 3.2A-C in the main text) in what functional relation the enzymes appear. We grouped all enzymes according to the first two digits of their EC number (EC a.b.x.x) and thus obtained 54 groups (y axis). We binned the appearing enzymes (x axis) and counted how many belong to the different EC classes, using two bin sizes (100 and 200 enzymes) and two values of γ (100 left column, 0 right column). Further, we used three measures for the enzymes in a bin and EC class: top (A1 and B1) gives the pure number belonging to a class, middle (A2 and B2) the fraction in the particular bin, and bottom (A3 and B3) the fraction of enzymes in a EC class with respect to the total number of enzymes belonging to the class (here, some classes are very small of only 1 or 2 enzymes, see numbers on the right y axis). Overall, the maps for $\gamma = 100$ show a higher clustering whereas for $\gamma = 0$ most of the map is equally colored.

Supplementary figure A.4: This figure shows the Coefficient of variation in sliding windows as in Figure 3.2C in the main text but for different window sizes. This demonstrates that the high C_v is not an effect of the limited window size.

Supplementary figure A.5: In Figure 3.3A in the main text we show a small fraction of the tree (long paths) that we obtain by a ranking of the appearance of enzymes. This figure shows a larger fraction. We still omitted terminal nodes on the first and second hierarchy level. The nodes are color-coded by the first EC digit and the enclosed CD contains a version with nodes linked to the entries in the KEGG database.

Supplementary table A.1: We counted the appearance of the 20 amino acid in the consensus set of enzyme sequences that we used for the simulations (column 2). This correlates with the appearance time of the amino acids during the expansion (see text in the manuscript and Figure 3B). The columns on the right show how many different amino acids the sequences contain. 83% contain all different amino acids and 95% require at least 19.

Supplementary table A.2 and following: Figure 3C shows the appearance of organisms. We defined the completeness of an organism if 80% of its enzymes are found during expansion. Due to the incompleteness of the database and the fact that not all of today's enzymes were a priori necessary, higher cut-offs seem over-precise. This table gives a specific list of the enzyme time and the name of an organism grouped by the kingdoms (in this order: animals, plants, fungi, protists, archaea, bacteria).

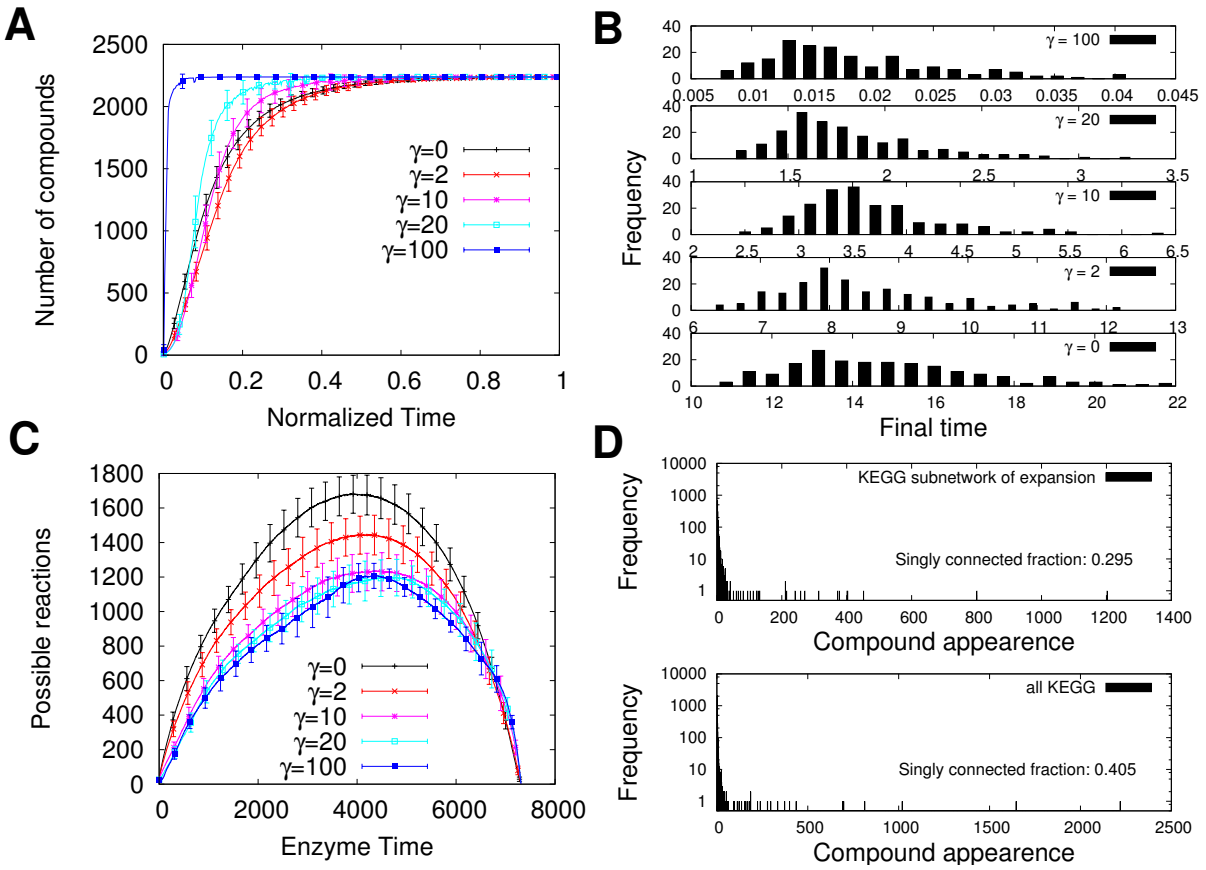


Figure A.1: Supporting figures of the expansion process. **A** Number of compounds for different γ shown over time. The curves look similar to the ones for the number of enzymes in the main paper except for scaling. **B** Distribution of the final times of 200 runs for all γ. The different time scales are clearly observable. The mean of these distributions is used to normalize the time. **C** Evolvability as a function of the number of possible reactions. **D** Distribution of the frequency of metabolites with that they appear in reactions. This is used for the border metabolite analysis.

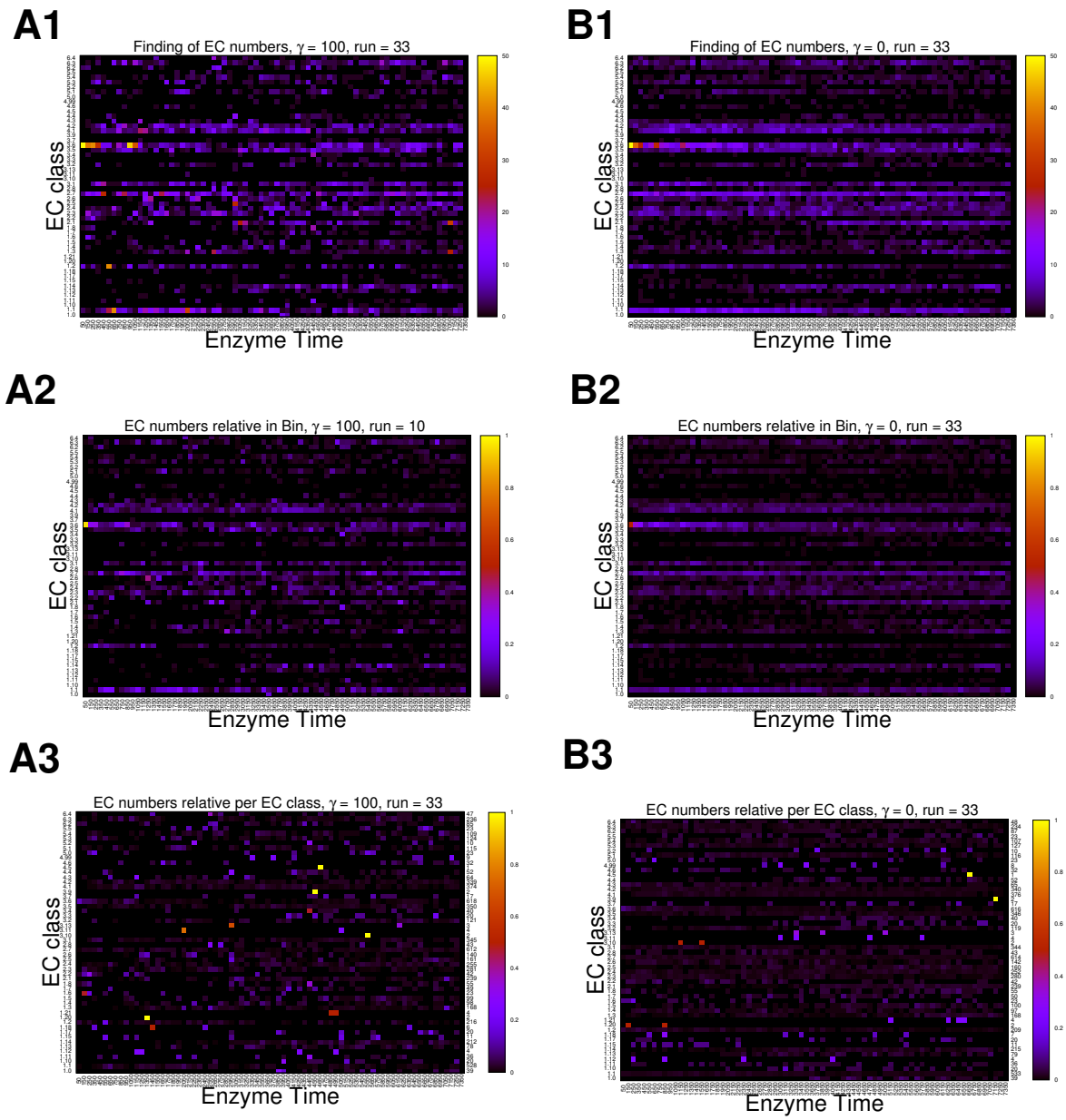


Figure A.2: Appearance of EC classes as a comparison of $\gamma = 100$ versus $\gamma = 0$ for one single run. The time scale in Enzyme Time is binned with a bin size of 100 enzymes and grouped by the first two EC digits. **A1**, **B1** Heatmap of the total number of enzymes in a certain class. **A2**, **B2** The number of enzymes per bin is normalized to the bin size. **A3**, **B3** The number of enzymes is normalized to the total number of enzymes within a EC class. The right y-tics depict the total number of enzymes in each class. Here, the clustering of enzymes of the same class clearly differs, especially for classes 1.X and 2.X, from the random case ($\gamma = 0$).

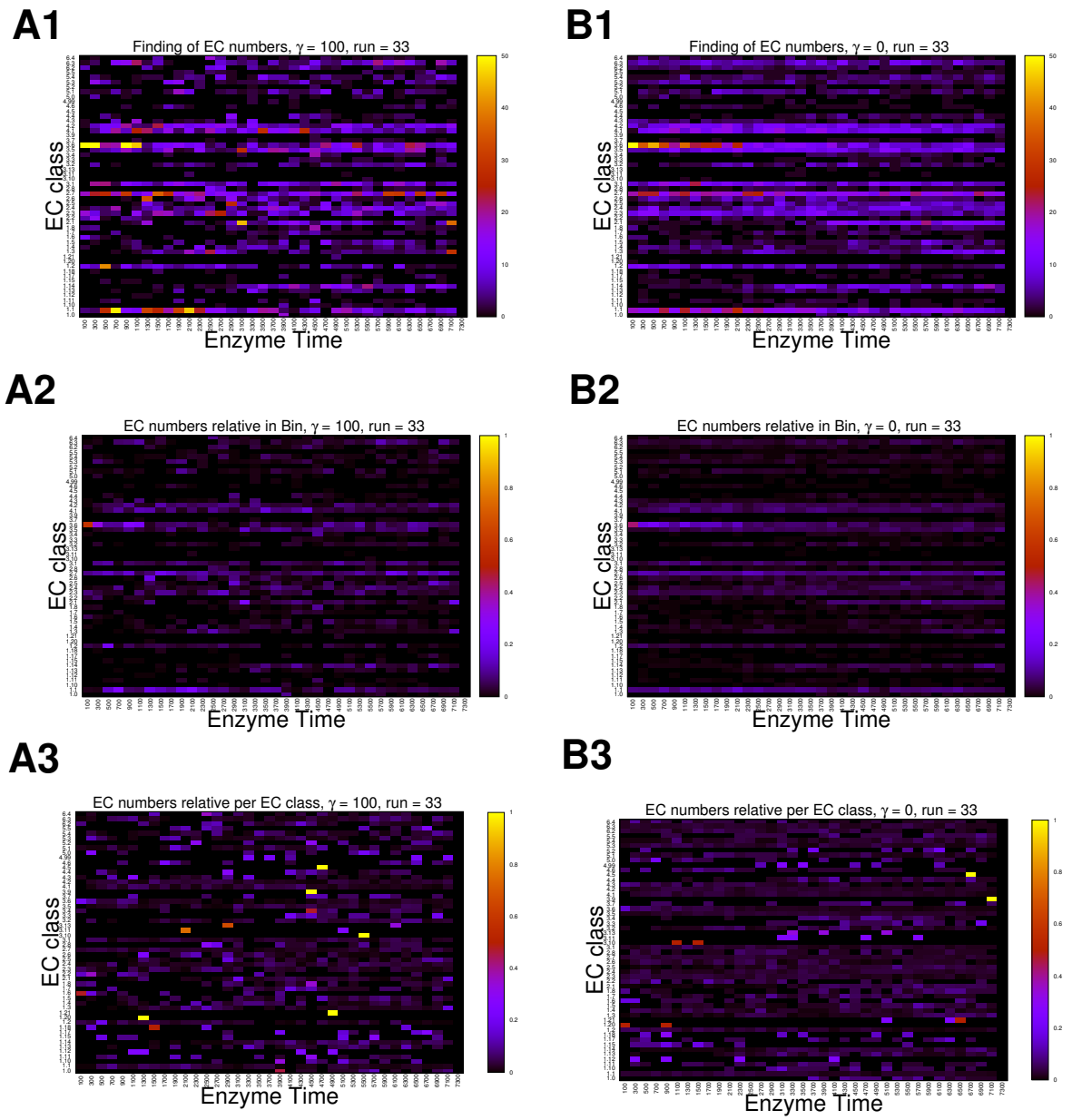


Figure A.3: The same as Fig. A.2c but with bin size 200. Appearance of EC classes as a comparison of $\gamma = 100$ versus $\gamma = 0$ for one single run. The time scale in Enzyme Time is binned with a bin size of 200 enzymes and grouped by the first two EC digits. **A1**, **B1** Heatmap of the total number of enzymes in a certain class. **A2**, **B2** The number of enzymes per bin is normalized to the bin size. **A3**, **B3** The number of enzymes is normalized to the total number of enzymes within a EC class. The right y-tics depict the total number of enzymes in each class. Here, the clustering of enzymes of the same class clearly differs, especially for classes 1.X and 2.X, from the random case ($\gamma = 0$).

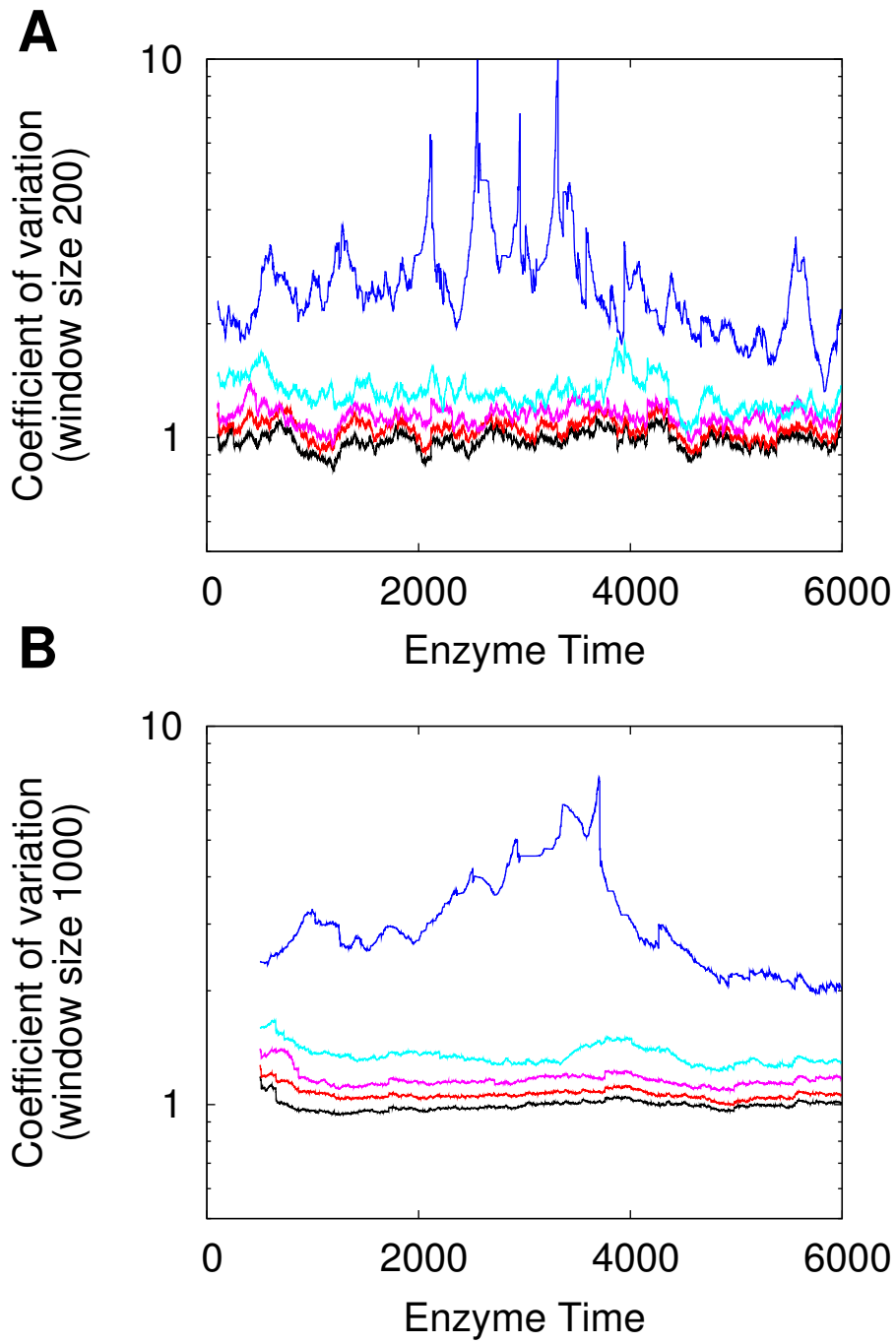


Figure A.4: Supporting figures of the coefficient of variation in sliding windows as in Fig. 3.2C of the manuscript. **A** For a window of size 200. **B** The same for a window of size 1000. The curves indicate that the bursting behavior is not a result of the limited window size since it is consistent for various window sizes.

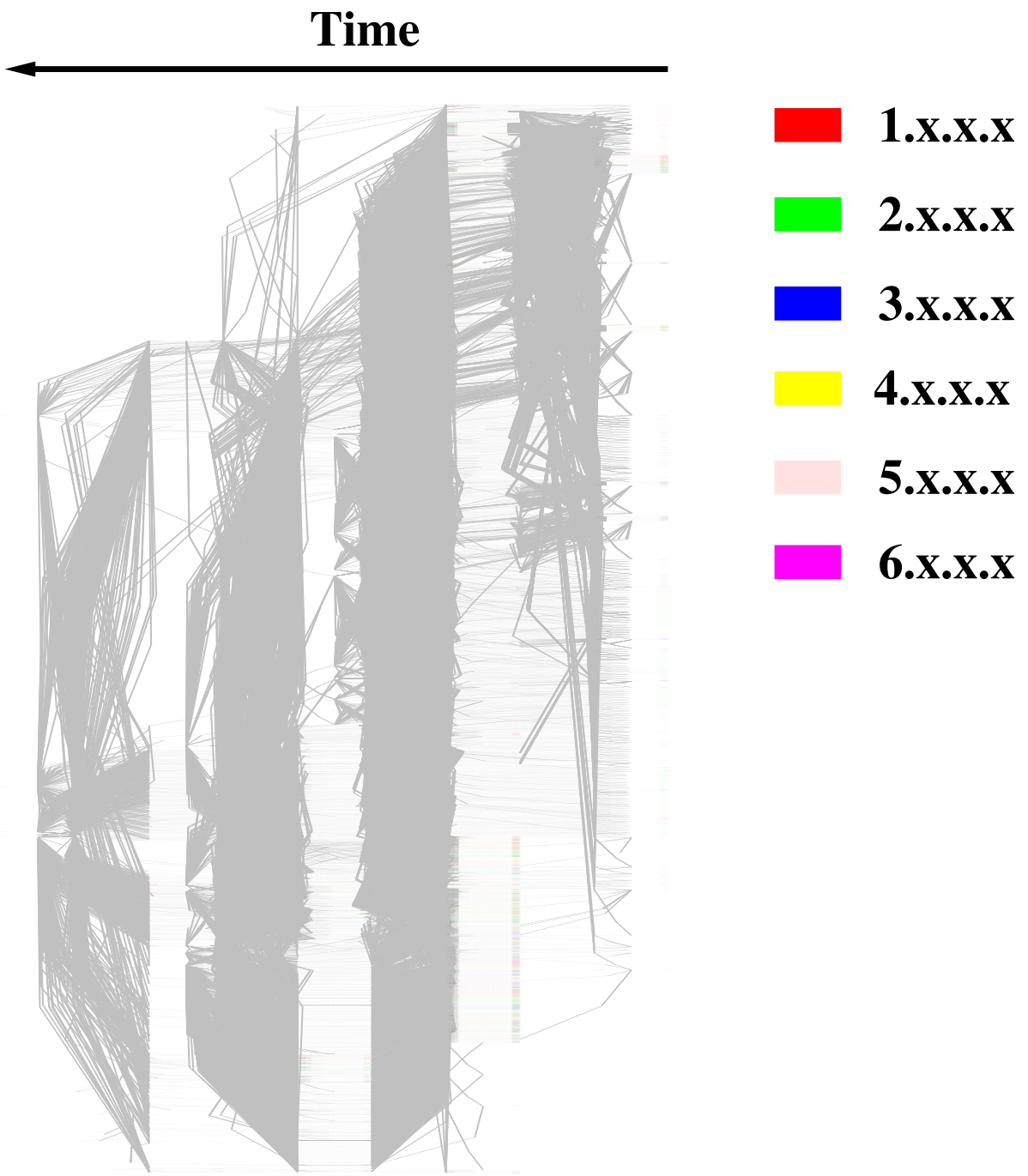


Figure A.5: Time-ordered ranking of enzyme appearance for $\gamma = 10$. For simplicity we removed some of the intermediate leafs that do not provide any further path. Nodes are color-coded by the first EC number (see the enclosed CD for a version with nodes linked to KEGG via their EC number).

amino acid	frequency in consensus set	# of different amino acids	sequences with # aa of column 3
Tryptophan	72325	1	0
Cysteine	86786	2	0
Methionine	129650	3	0
Histidine	133672	4	0
Tyrosine	175131	5	0
Glutamine	213605	6	0
Asparagine	226700	7	0
Phenylalanine	227920	8	1
Proline	279303	9	0
Threonine	296583	10	3
Lysine	303362	11	0
Aspartate	310855	12	6
Arginine	315381	13	11
Isoleucine	320837	14	18
Glutamate	371811	15	35
Serine	384861	16	55
Valine	389467	17	126
Glycine	407043	18	318
Alanine	475003	19	1433
Leucine	547509	20	9866
		21	53

Table A.1.: Frequency of amino acids in the consensus set of enzyme sequences. And the number of different amino acids in sequences and their frequency. Note: 53 sequences have 21 amino acids since an additional ”U” amino acid has been found in the sequences.

#	Enzyme	KEGG ID	Full name
1	3288	SMM	Schistosoma mansoni
2	3383	API	Acyrthosiphon pisum (pea aphid)
3	3486	NVI	Nasonia vitripennis (jewel wasp)
4	3541	BMY	Brugia malayi (filaria)
5	3562	CQU	Culex quinquefasciatus (southern house mosquito)
6	3571	DSE	Drosophila sechellia
7	3584	AAG	Aedes aegypti (yellow fever mosquito)
8	3598	DGR	Drosophila grimshawi
9	3602	DAN	Drosophila ananassae
10	3608	DPE	Drosophila persimilis
11	3616	DMO	Drosophila mojavensis
12	3618	DYA	Drosophila yakuba
13	3623	DSI	Drosophila simulans
14	3639	DER	Drosophila erecta
15	3664	CEL	Caenorhabditis elegans (nematode)
16	3677	AME	Apis mellifera (honey bee)
17	3680	AGA	Anopheles gambiae (mosquito)
18	3715	HMG	Hydra magnipapillata
19	3743	TCA	Tribolium castaneum (red flour beetle)
20	3743	TAD	Trichoplax adhaerens
21	3746	CBR	Caenorhabditis briggsae
22	3819	DME	Drosophila melanogaster (fruit fly)
23	3820	DPO	Drosophila pseudoobscura pseudoobscura
24	3834	CIN	Ciona intestinalis (sea squirt)
25	3976	SSC	Sus scrofa (pig)
26	3987	XLA	Xenopus laevis (African clawed frog)
27	3987	NVE	Nematostella vectensis (sea anemone)
28	3991	XTR	Xenopus tropicalis (western clawed frog)
29	4027	BFO	Branchiostoma floridae (Florida lancelet)
30	4031	OAA	Ornithorhynchus anatinus (platypus)
31	4034	TGU	Taeniopygia guttata (zebra finch)
32	4086	MDO	Monodelphis domestica (opossum)
33	4098	MCC	Macaca mulatta (rhesus monkey)
34	4119	DRE	Danio rerio (zebrafish)
35	4121	SPU	Strongylocentrotus purpuratus (purple sea urchin)
36	4131	PTR	Pan troglodytes (chimpanzee)
37	4134	CFA	Canis familiaris (dog)
38	4137	GGA	Gallus gallus (chicken)
39	4152	BTA	Bos taurus (cow)
40	4182	ECB	Equus caballus (horse)
41	4189	MMU	Mus musculus (mouse)
42	4202	RNO	Rattus norvegicus (rat)
43	4206	HSA	Homo sapiens (human)

Table A.2.: Appearance of animals as calculated by mapping enzymes to organisms. The first column simply accumulates the organisms, the second gives the mean appearance time of 200 runs, the third and fourth the KEGG ID and the name.

#	Enzyme	KEGG ID	Full name
1	3564	ZMA	Zea mays (maize)
2	3629	CRE	Chlamydomonas reinhardtii
3	3649	SBI	Sorghum bicolor (sorghum)
4	3655	OLU	Ostreococcus lucimarinus
5	3660	VVI	Vitis vinifera (wine grape)
6	3698	PPP	Physcomitrella patens subsp. patens
7	3722	CME	Cyanidioschyzon merolae
8	3739	RCU	Ricinus communis (castor bean)
9	3801	POP	Populus trichocarpa (black cottonwood)
10	4156	OSA	Oryza sativa japonica (Japanese rice)
11	4345	ATH	Arabidopsis thaliana (thale cress)

Table A.3.: Appearance of plants as calculated by mapping enzymes to organisms. The first column simply accumulates the organisms, the second gives the mean appearance time of 200 runs, the third and fourth the KEGG ID and the name.

#	Enzyme	KEGG ID	Full name
1	3249	MGL	Malassezia globosa
2	3291	ECU	Encephalitozoon cuniculi
3	3342	MPR	Moniliophthora perniciosa
4	3349	LBC	Laccaria bicolor
5	3370	CNB	Cryptococcus neoformans B-3501A
6	3452	BFU	Botryotinia fuckeliana
7	3474	SSL	Sclerotinia sclerotiorum
8	3489	CIM	Coccidioides immitis
9	3506	UMA	Ustilago maydis
10	3520	KLA	Kluyveromyces lactis
11	3524	VPO	Vanderwaltozyma polyspora
12	3557	ANI	Aspergillus nidulans
13	3561	CAL	Candida albicans
14	3587	SPO	Schizosaccharomyces pombe (fission yeast)
15	3595	CNE	Cryptococcus neoformans JEC21
16	3598	NFI	Neosartorya fischeri
17	3608	LTH	Lachancea thermotolerans
18	3614	PPA	Pichia pastoris
19	3628	CGR	Candida glabrata
20	3638	SCE	Saccharomyces cerevisiae (budding yeast)
21	3657	PCS	Penicillium chrysogenum
22	3667	FGR	Fusarium graminearum
23	3673	YLI	Yarrowia lipolytica
24	3690	AFV	Aspergillus flavus
25	3697	NCR	Neurospora crassa
26	3698	AGO	Ashbya gossypii (Eremothecium gossypii)
27	3701	PAN	Podospora anserina
28	3733	MGR	Magnaporthe grisea
29	3735	DHA	Debaryomyces hansenii
30	3745	AFM	Aspergillus fumigatus

31	3802	AOR	Aspergillus oryzae
32	3821	ANG	Aspergillus niger
33	3834	PIC	Pichia stipitis

Table A.4.: Appearance of fungi as calculated by mapping enzymes to organisms. The first column simply accumulates the organisms, the second gives the mean appearance time of 200 runs, the third and fourth the KEGG ID and the name.

#	Enzyme	KEGG ID	Full name
1	2752	EHI	Entamoeba histolytica
2	2856	TGO	Toxoplasma gondii
3	2888	PFD	Plasmodium falciparum Dd2
4	2980	TVA	Trichomonas vaginalis
5	3057	EDI	Entamoeba dispar
6	3127	TAN	Theileria annulata
7	3169	BBO	Babesia bovis
8	3179	PKN	Plasmodium knowlesi
9	3209	PBE	Plasmodium berghei
10	3214	PCB	Plasmodium chabaudi
11	3245	PYO	Plasmodium yoelii
12	3317	PFA	Plasmodium falciparum 3D7
13	3317	PFH	Plasmodium falciparum HB3
14	3335	GLA	Giardia lamblia
15	3370	PVX	Plasmodium vivax
16	3397	TPV	Theileria parva
17	3420	CPV	Cryptosporidium parvum
18	3649	CHO	Cryptosporidium hominis
19	3649	PTM	Paramecium tetraurelia
20	3728	MBR	Monosiga brevicollis
21	3804	TBR	Trypanosoma brucei
22	3817	PTI	Phaeodactylum tricornutum
23	3866	TPS	Thalassiosira pseudonana
24	3912	TET	Tetrahymena thermophila
25	3918	TCR	Trypanosoma cruzi
26	3925	LMA	Leishmania major
27	4006	DDI	Dictyostelium discoideum (cellular slime mold)

Table A.5.: Appearance of protists as calculated by mapping enzymes to organisms. The first column simply accumulates the organisms, the second gives the mean appearance time of 200 runs, the third and fourth the KEGG ID and the name.

#	Enzyme	KEGG ID	Full name
1	3030	TKO	Thermococcus kodakaraensis
2	3038	PFU	Pyrococcus furiosus
3	3127	TGA	Thermococcus gammatolerans
4	3147	KCR	Candidatus Korarchaeum cryptofilum
5	3152	MMZ	Methanococcus maripaludis C7

6	3156	MFE	Methanocaldococcus fervens
7	3164	RCI	Uncultured methanogenic archaeon RC-I
8	3198	PIS	Pyrobaculum islandicum
9	3203	MHU	Methanospirillum hungatei
10	3223	MVU	Methanocaldococcus vulcanius
11	3228	PAB	Pyrococcus abyssi
12	3235	MVN	Methanococcus vannielii
13	3247	MMX	Methanococcus maripaludis C6
14	3248	MLA	Methanocorpusculum labreanum
15	3251	MTP	Methanosaeta thermophila
16	3254	PHO	Pyrococcus horikoshii
17	3264	HLA	Halorubrum lacusprofundi
18	3264	MBN	Candidatus Methanoregula boonei
19	3272	HUT	Halorhabdus utahensis
20	3295	MBA	Methanosarcina barkeri
21	3302	MST	Methanosphaera stadtmanae
22	3316	MAC	Methanosarcina acetivorans
23	3327	SIS	Sulfolobus islandicus L.S.2.15
24	3327	AFU	Archaeoglobus fulgidus
25	3330	MJA	Methanococcus jannaschii
26	3333	MPL	Candidatus Methanosphaerula palustris
27	3337	MMQ	Methanococcus maripaludis C5
28	3337	MMP	Methanococcus maripaludis S2
29	3338	TON	Thermococcus onnurineus
30	3343	MAE	Methanococcus aeolicus
31	3349	DKA	Desulfurococcus kamchatkensis
32	3358	MTH	Methanobacterium thermoautotrophicum
33	3367	IHO	Ignicoccus hospitalis
34	3374	MSI	Methanobrevibacter smithii ATCC 35061
35	3378	SID	Sulfolobus islandicus M.16.4
36	3384	TSI	Thermococcus sibiricus
37	3386	PAS	Pyrobaculum arsenaticum
38	3389	APE	Aeropyrum pernix
39	3393	MMA	Methanosarcina mazei
40	3399	SIA	Sulfolobus islandicus M.14.25
41	3401	PAI	Pyrobaculum aerophilum
42	3407	MEM	Methanoculleus marisnigri
43	3412	SIY	Sulfolobus islandicus Y.G.57.14
44	3412	SIM	Sulfolobus islandicus M.16.27
45	3413	MBU	Methanococcoides burtonii
46	3416	CMA	Caldivirga maquilingensis
47	3423	HMU	Halomicrobium mukohataei
48	3424	SIN	Sulfolobus islandicus Y.N.15.51
49	3427	STO	Sulfolobus tokodaii
50	3431	PCL	Pyrobaculum calidifontis
51	3435	NMR	Nitrosopumilus maritimus
52	3452	SAI	Sulfolobus acidocaldarius
53	3454	TNE	Thermoproteus neutrophilus
54	3463	SSO	Sulfolobus solfataricus
55	3464	TPE	Thermofilum pendens
56	3520	MSE	Metallosphaera sedula

57	3533	HSL	Halobacterium salinarum R1
58	3553	TAC	Thermoplasma acidophilum
59	3570	NPH	Natronomonas pharaonis
60	3578	SMR	Staphylothermus marinus
61	3589	HBU	Hyperthermus butylicus
62	3600	HAL	Halobacterium sp. NRC-1
63	3637	MKA	Methanopyrus kandleri
64	3661	TVO	Thermoplasma volcanium
65	3666	HWA	Haloquadratum walsbyi
66	3714	PTO	Picrophilus torridus
67	3717	HMA	Haloarcula marismortui
68	3887	NEQ	Nanoarchaeum equitans

Table A.6.: Appearance of archaea as calculated by mapping enzymes to organisms. The first column simply accumulates the organisms, the second gives the mean appearance time of 200 runs, the third and fourth the KEGG ID and the name.

#	Enzyme	KEGG ID	Full name
1	2387	SMS	Candidatus Sulcia muelleri SMDSEM
2	2428	SMG	Candidatus Sulcia muelleri GWSS
3	2553	MHJ	Mycoplasma hyopneumoniae J
4	2573	BCC	Buchnera aphidicola Cc
5	2619	AYW	Phytoplasma AYWB
6	2631	MCO	Mycoplasma conjunctivae
7	2659	MHY	Mycoplasma hyopneumoniae 232
8	2666	MHP	Mycoplasma hyopneumoniae 7448
9	2731	PAL	Candidatus Phytoplasma australiense
10	2790	CRP	Candidatus Carsonella ruddii
11	2828	MAT	Mycoplasma arthritidis
12	2952	POY	Phytoplasma OY
13	2956	MPN	Mycoplasma pneumoniae
14	2987	MGA	Mycoplasma gallisepticum
15	2990	PML	Candidatus Phytoplasma mali
16	3002	MAA	Mycoplasma agalactiae
17	3053	MCP	Mycoplasma capricolum
18	3069	MHO	Mycoplasma hominis
19	3069	MFL	Mesoplasma florum
20	3120	HOR	Halothermothrix orenii
21	3137	MSY	Mycoplasma synoviae
22	3171	MGE	Mycoplasma genitalium
23	3173	TMZ	Thauera sp. MZ1T
24	3186	BMI	Brucella melitensis ATCC 23457
25	3197	DEH	Dehalococcoides sp. CBDB1
26	3197	DEB	Dehalococcoides sp. BAV1
27	3198	UUE	Ureaplasma urealyticum serovar 10 ATCC 33699
28	3201	UUR	Ureaplasma parvum serovar 3 ATCC 700970
29	3207	DET	Dehalococcoides ethenogenes
30	3212	APR	Anaerococcus prevotii
31	3212	HIP	Haemophilus influenzae PittEE

32	3217	MMY	Mycoplasma mycoides
33	3219	BLT	Bifidobacterium animalis subsp. lactis DSM 10140
34	3220	MPU	Mycoplasma pulmonis
35	3223	BLN	Bifidobacterium longum subsp. infantis ATCC 15697
⋮	⋮	⋮	⋮
136	3398	ECE	Escherichia coli O157:H7 EDL933 (EHEC)
137	3399	SAK	Streptococcus agalactiae A909 (serotype Ia)
138	3400	EOJ	Escherichia coli O26:H11 11368
139	3400	MIN	Methylacidiphilum infernorum
140	3401	TPT	Thermotoga petrophila
141	3402	JDE	Jonesia denitrificans
142	3402	PPH	Pelodictyon phaeoclathratiforme
143	3402	NMI	Neisseria meningitidis alpha14
144	3404	CMS	Clavibacter michiganensis subsp. sepedonicus
145	3406	SCA	Staphylococcus carnosus
146	3406	LPJ	Lactobacillus plantarum JDM1
147	3406	GUR	Geobacter uraniumreducens
148	3407	ECM	Escherichia coli SECEC
149	3408	ECY	Escherichia coli O152:H28 SE11 (commensal)
150	3408	LRH	Lactobacillus rhamnosus
151	3410	GBM	Geobacter bemidjiensis
152	3410	SAA	Staphylococcus aureus USA300
153	3412	SHA	Staphylococcus haemolyticus
154	3413	CJR	Campylobacter jejuni RM1221
155	3413	ECQ	Escherichia coli ED1a
156	3413	TYE	Thermodesulfovibrio yellowstonii
157	3414	SAB	Staphylococcus aureus RF122
158	3414	ECO	Escherichia coli K-12 MG1655
159	3414	ECR	Escherichia coli IAI1 (commensal)
⋮	⋮	⋮	⋮
434	3552	BCG	Bacillus cereus G9842
435	3552	TLE	Thermotoga lettingae
436	3553	TWS	Tropheryma whipplei TW08/27
437	3553	LBR	Lactobacillus brevis
438	3553	HOH	Haliangium ochraceum
439	3553	HAU	Herpetosiphon aurantiacus
440	3554	LMC	Listeria monocytogenes Clip81459
441	3554	SED	Salmonella enterica subsp. enterica serovar Dublin
442	3554	LPP	Legionella pneumophila Paris
443	3554	CBF	Clostridium botulinum F Langeland
444	3554	MAQ	Marinobacter aquaeolei
445	3554	ECA	Erwinia carotovora
446	3555	HDU	Haemophilus ducreyi
447	3555	CGL	Corynebacterium glutamicum ATCC 13032 (Kyowa Hakko)
448	3555	BCQ	Bacillus cereus Q1
449	3555	EAT	Exiguobacterium sp. AT1b
450	3555	MSL	Methylocella silvestris
451	3555	ETA	Erwinia tasmaniensis
452	3555	SEH	Salmonella enterica subsp. enterica serovar Heidelberg
453	3556	EBA	Aromatoleum aromaticum EbN1
454	3556	LBA	Leptotrichia buccalis

455	3556	GYM	Geobacillus sp. Y412MC10
456	3556	FBA	Flavobacteriaceae bacterium
⋮	⋮	⋮	⋮
787	3761	SCL	Sorangium cellulosum
788	3762	PDE	Paracoccus denitrificans
789	3763	MAB	Mycobacterium abscessus ATCC 19977
790	3763	MXA	Myxococcus xanthus
791	3764	MTC	Mycobacterium tuberculosis CDC1551
792	3765	TEL	Thermosynechococcus elongatus
793	3766	BML	Burkholderia mallei NCTC 10229
794	3766	AAV	Acidovorax avenae
795	3767	BGL	Burkholderia glumae
796	3771	LBJ	Leptospira borgpetersenii JB197
797	3771	PAG	Pseudomonas aeruginosa LESB58
798	3771	XAU	Xanthobacter autotrophicus
799	3772	PSB	Pseudomonas syringae pv. syringae B728a
800	3774	TPP	Treponema pallidum subsp. pallidum SS14
801	3775	FAL	Frankia alni
802	3775	MJL	Mycobacterium sp. JLS
803	3778	RPT	Rhodopseudomonas palustris TIE-1
804	3779	MTB	Mycobacterium tuberculosis KZN 1435
805	3781	MPA	Mycobacterium avium paratuberculosis
806	3783	PAE	Pseudomonas aeruginosa PAO1
807	3783	AOE	Alkaliphilus oremlandii
808	3784	BJA	Bradyrhizobium japonicum
809	3784	PPG	Pseudomonas putida GB-1
⋮	⋮	⋮	⋮
900	3962	PMC	Prochlorococcus marinus MIT 9515
901	3970	OTT	Orientia tsutsugamushi Ikeda
902	3974	SYX	Synechococcus sp. WH7803
903	3974	TPA	Treponema pallidum subsp. pallidum Nichols
904	3994	PMT	Prochlorococcus marinus MIT 9313
905	4012	PMF	Prochlorococcus marinus MIT 9303
906	4023	OTS	Orientia tsutsugamushi Boryong
907	4031	BGA	Borrelia garinii
908	4037	RCM	Rickettsia canadensis
909	4079	BAF	Borrelia afzelii
910	4200	BTU	Borrelia turicatae
911	4206	BHR	Borrelia hermsii
912	4211	BBZ	Borrelia burgdorferi ZS7
913	4228	BDU	Borrelia duttonii
914	4281	BBU	Borrelia burgdorferi B31
915	4328	BRE	Borrelia recurrentis

Table A.7.: Appearance of bacteria as calculated by mapping enzymes to organisms. The first column simply accumulates the organisms, the second gives the mean appearance time of 200 runs, the third and fourth the KEGG ID and the name. In contrast to the other Tables, here we only show an extract. The full list can be found on the enclosed CD.

B. Supplementary Materials: Analysis of Gene Coexpression Data

The following figures are published as supplementary material for the publication of Chapter 4.

Supplementary figure B.1 shows one single cluster, number 20, with genes preferentially involved in cell wall cellulose synthesis.

Supplementary figure B.2 provides statistical evidence for the identification of essential single-copy genes. A sampling procedure is carried out to investigate the relationship between single-copy genes, essential and non-essential, and back-ups of the same function in the network vicinity. This was used to identify candidate genes for further mutant analysis.

Supplementary figure B.3 is a close-up of the clusters in which we identified the previously unknown mutants.

Supplementary figure B.4 compares the coexpression network using Pearson correlation with a similar network based on Graphical Gaussian method. There is an overlap of about one third of the genes between the two networks.

Supplementary figure B.5 compares the cluster solutions of the same data set by the adjusted Rand index. A value of 0 means no accordance between two clusterings while a value of 1 depicts complete agreement. As a further investigation, we compared how the clustering using HCCA changes if one removes 20% of the nodes from the network. This figure is changed from the form of a table in the online supplementary into a heatmap to fit the format of the thesis.

Supplementary table B.1 lists the values of the cluster scores ClusterJudge, modularity, and Davies-Bouldin for a variety of parameter variations of the different clustering methods.

Supplementary table B.2 shows the variation in cluster size and cluster number between the methods. One observes very different distributions depending on the algorithm and parameters used. The table was slightly changed from the online supplementary.

Supplementary table B.3 gives the adjusted Rand index for HCCA when the top 10,20,30,40,50 nodes in the mutual rank network are taken into account. This value mainly changes the density of connections in the network.

Supplementary table B.4 shows statistical significance using Fisher's exact test to determine whether essential genes are enriched in the cluster. This is checked against random appearance using clusters of similar size, similar number of essential genes, and the total number of clustered genes.

Supplementary table B.5 provides detailed information on the T-DNA knock-out lines and primers used for the characterization of 20 candidate genes.

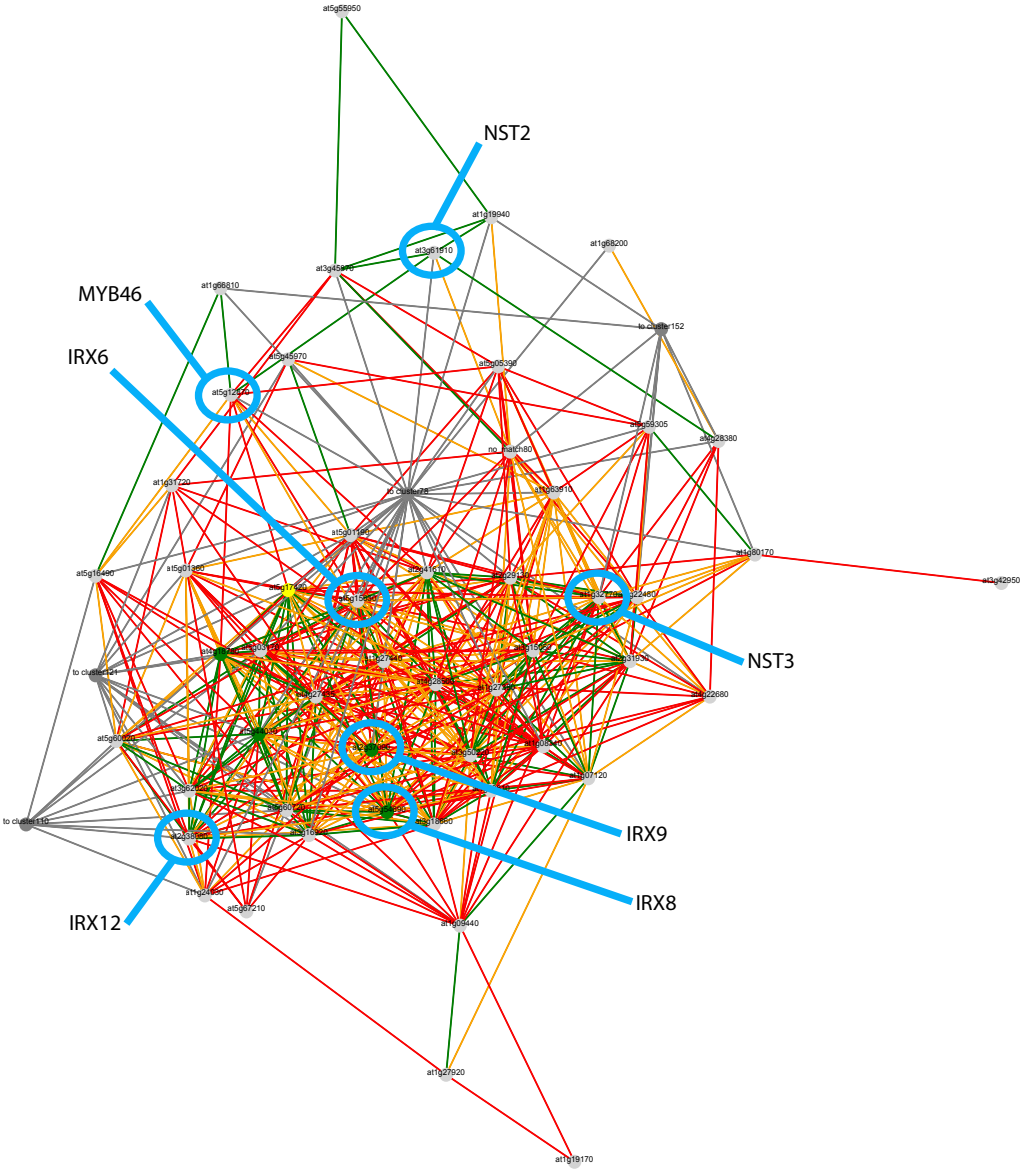


Figure S1. Cluster 20 containing genes involved in secondary cell wall cellulose synthesis. Nodes representing IRX6, IRX8, IRX9, IRX12, MYB46, NST2 and NST3 are marked by blue circles.

Figure B.1: Cluster 20 containing genes involved in secondary cell wall cellulose synthesis. Nodes representing IRX6, IRX8, IRX9, IRX12, MYB46, NST2 and NST3 are marked by blue circles.

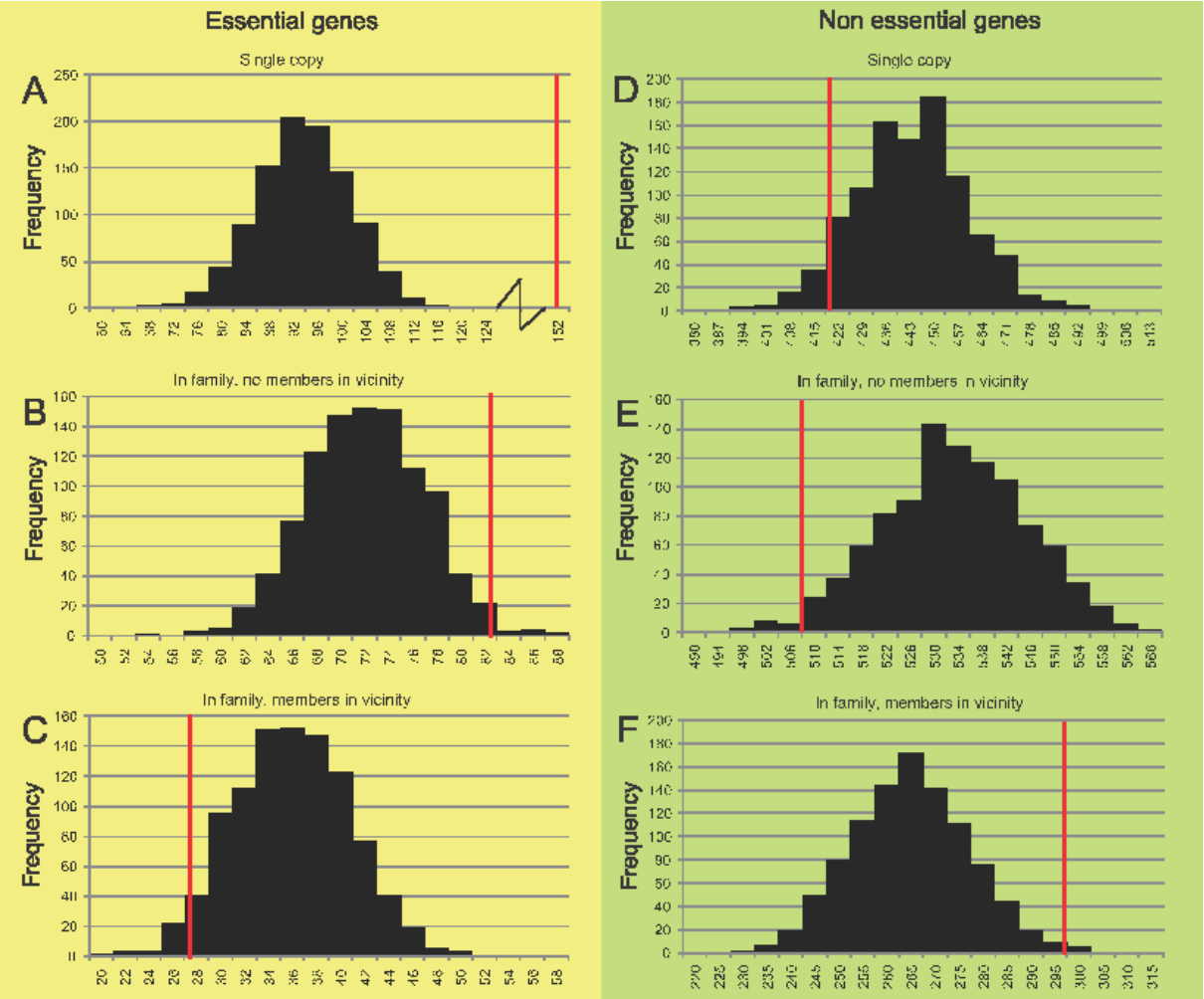


Figure B.2: Distribution of 1000 random samplings of essential and non-essential genes from the mutual rank network. **A**. Distribution of single-copy genes from sampling of 261 random genes 1,000 times. The number (152) of essential, single-copy genes observed in our network is denoted by a red bar. **B**. Distribution of genes shown to be in a family but unique in the node vicinity network (n=2) from sampling 109 random nodes 1,000 times. The observed number (82) of essential genes in family, but unique in the node vicinity network is denoted by red bar. **C**. Distribution of genes shown to be in a family with family members in node vicinity network (n=2) from sampling of 109 random nodes 1000 times. The observed number (27) of essential genes in family with family members in the node vicinity network is denoted by red bar. **D**, **E**, and **F** correspond to **A** (1,224 nodes sampled), **B** (802 nodes sampled), and **C** (802 nodes sampled), respectively, but show distribution for non-essential genes. The observed numbers of non-essential, single copy (422), non-essential, in gene family, but unique in vicinity network (507), and non-essential with family members in vicinity network (295), are denoted by red bars in the figure.

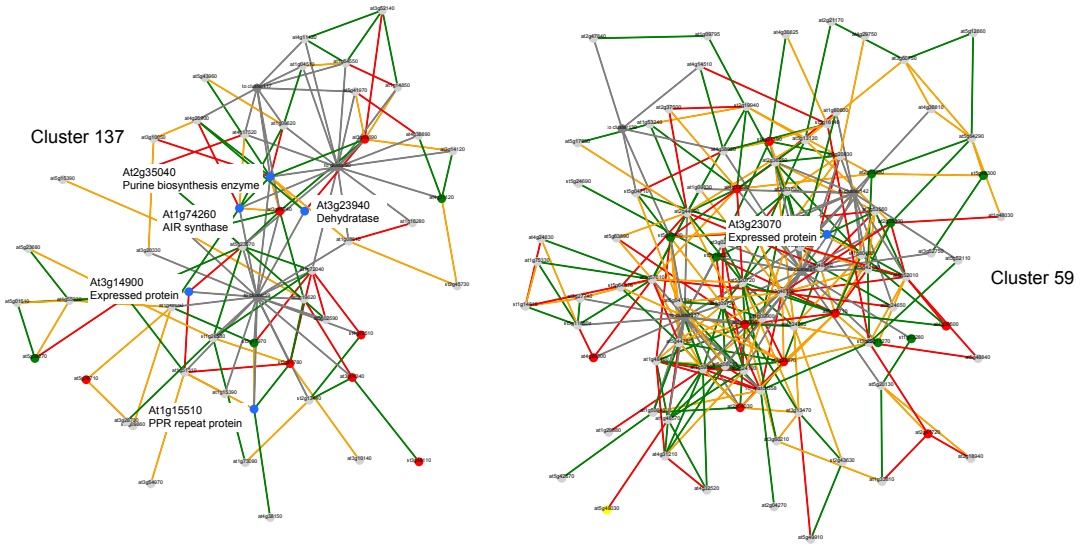


Figure S3. Cluster 21, 59 and 137. Mutants characterized in this study are marked with blue nodes.

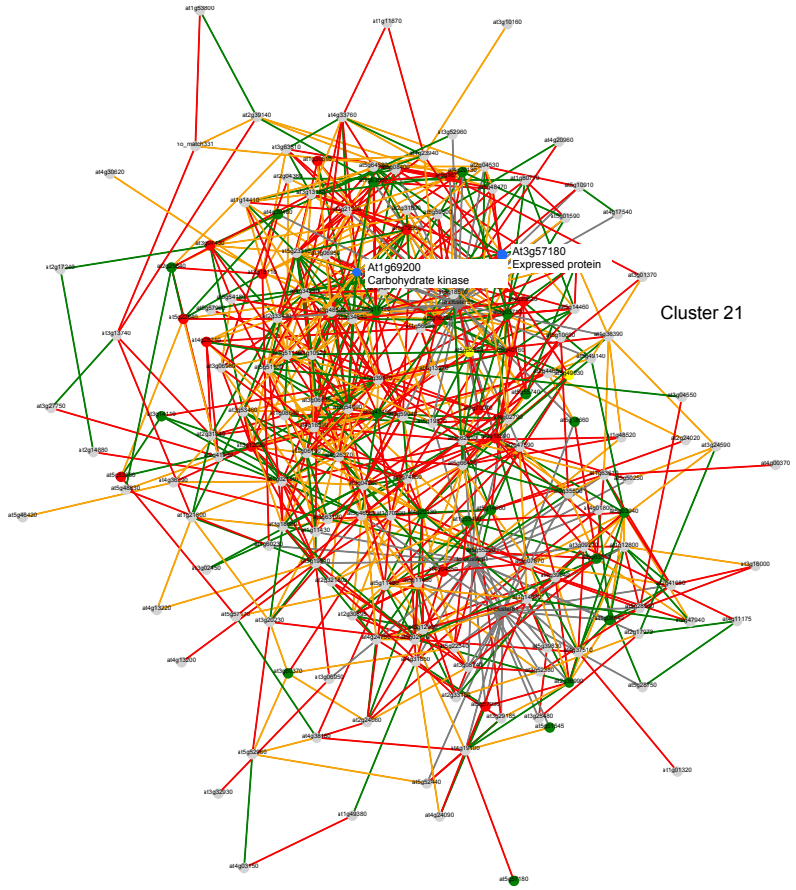


Figure B.3: Cluster 21, 59 and 137. Mutants characterized in this study are marked with blue nodes.

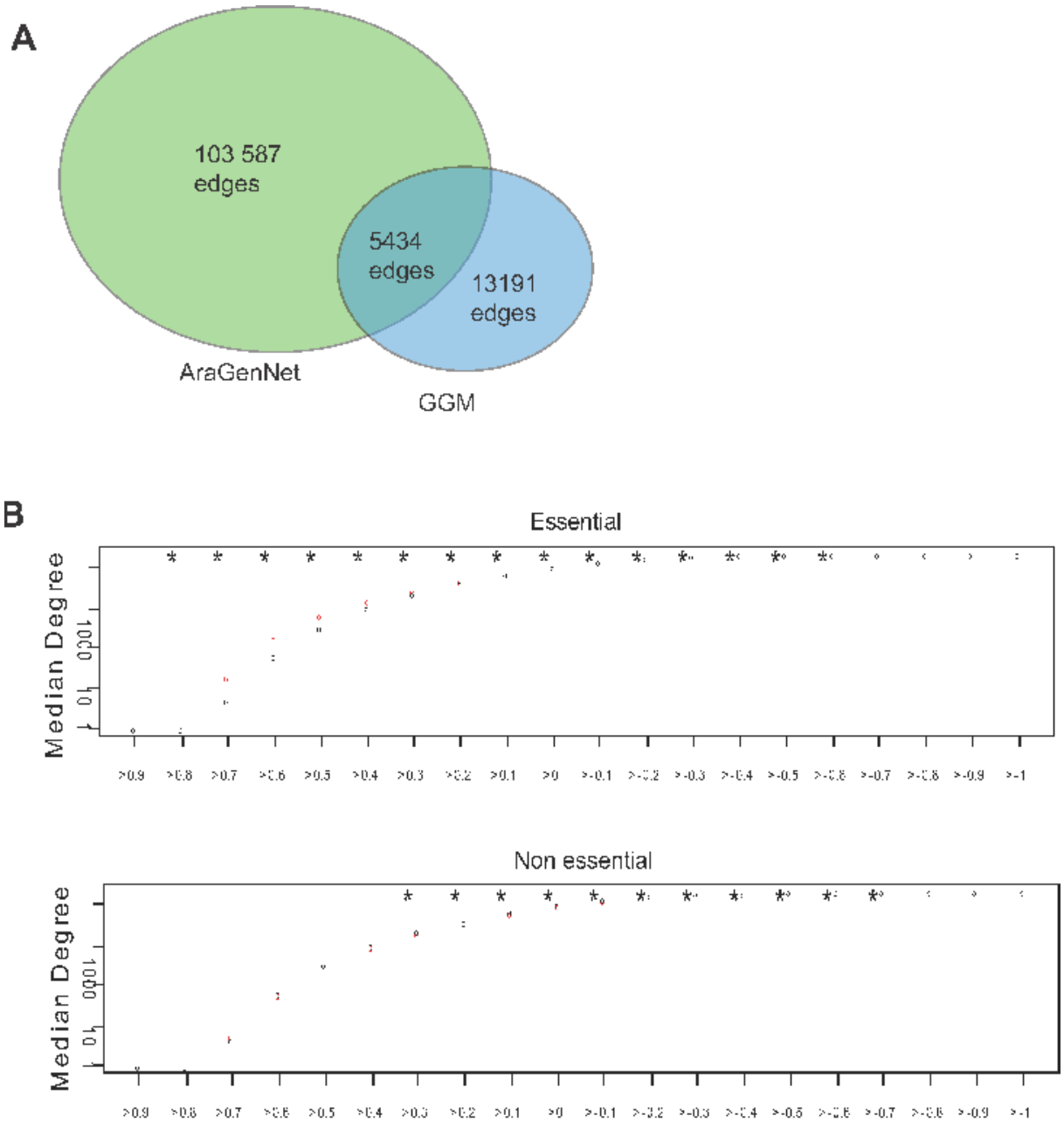


Figure B.4: Comparison of a Pearson and GGM generated network. A. Venn diagram of edges present in a Pearson ($r > 0.8$), and a GGM network [115]. B. Median Degree, or node degree, for genes using a correlation threshold as indicated on the x axis. The median degree for genes that are essential (upper panel) or non-essential (lower panel) is shown by red dots, the median degree for genes not showing this characteristic is given in black. Significant differences (Wilcoxon test $P < 0.05$) in the median degree between these two classes at a given correlation threshold are marked by an asterisk.

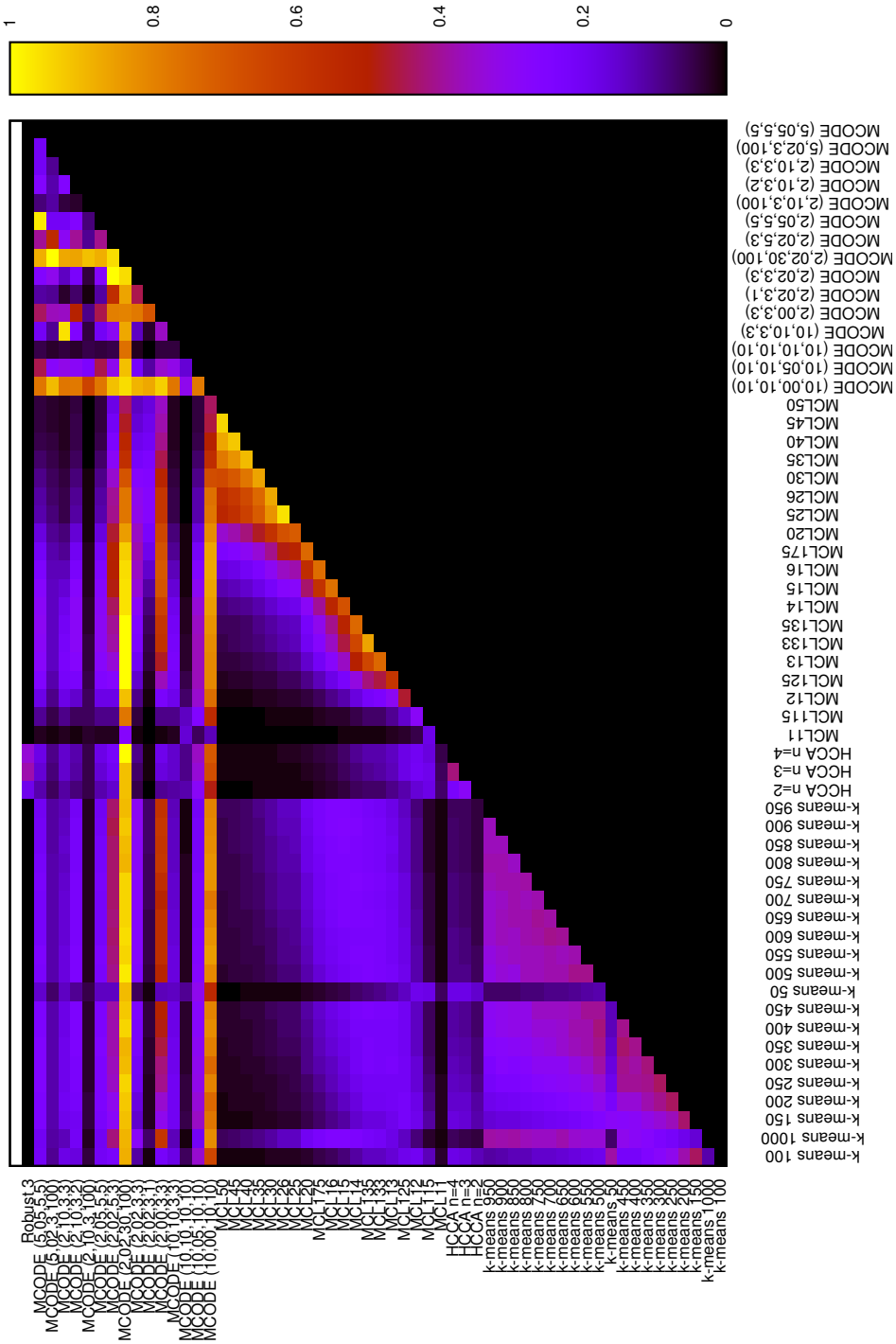


Figure B.5: Adjusted Rand index analysis of clustering solutions generated by the MCL, k-means and HCCA algorithms. To further compare the different clustering algorithms, we used the adjusted Rand index to score similarities between the clustering solutions. Robust 3 labels the comparison with a set of twenty networks of the Arabidopsis clustered with HCCA3 but with 20% of the nodes randomly deleted. It is only calculated for the comparisons with the other HCCA results and is given as the mean for these twenty networks.

Clustering algorithm	clusterJudge	Davies-Bouldin Davies-Bouldin	Modularity Modularity	Nodes clustered	Number of of clusters	Largest cluster	Smallest Cluster
HCCA n=2	141	26.36	0.68	20785	209	1214	59
HCCA n=3	149	27.53	0.69	20785	305	295	69
HCCA n=4	138	27.39	0.67	20785	317	253	48
HCCA unweighted n=2	140.7	6.29	0.71	20785	55	1925	59
HCCA unweighted n=3	141.2	6.2	0.71	20785	70	1234	69
HCCA unweighted n=4	144.8	6.22	0.71	20785	83	893	48
k-means50	186.7	71.44	0.64	20785	50	797	139
k-means100	173.9	53.86	0.6	20785	100	499	71
k-means150	166.7	46.16	0.59	20785	150	354	18
k-means200	165	40.96	0.56	20785	200	302	13
k-means250	152.9	36.92	0.55	20785	250	218	18
k-means300	155.4	34.48	0.54	20785	300	229	18
k-means350	147.1	32.16	0.52	20785	350	204	8
k-means400	142.8	30.5	0.51	20785	400	182	11
k-means450	151.2	28.74	0.51	20785	450	198	7
k-means500	135.1	27.51	0.5	20785	500	159	11
k-means550	137.4	26.24	0.49	20785	550	123	6
k-means600	136.9	25.25	0.48	20785	600	117	5
k-means650	127.6	24.6	0.47	20785	650	112	4
k-means700	129.9	24.01	0.46	20785	700	111	4
k-means750	129.2	23.07	0.47	20785	750	101	6
k-means800	132.2	22.45	0.45	20785	800	101	3
k-means850	122.9	21.81	0.45	20785	850	93	3
k-means900	119.7	21.19	0.44	20785	900	94	4
k-means950	121.3	20.8	0.43	20785	950	86	4
MCL1.1	99.8	12.01	0.62	20785	146	7676	2
MCL1.15	102.8	20.9	0.71	20785	234	2546	2
MCL1.2	109.4	21.59	0.67	20785	427	830	2
MCL1.25	100.6	19.49	0.64	20785	680	427	2
MCL1.3	96.2	17.55	0.61	20785	976	294	2
MCL1.4	84.9	14.64	0.55	20785	1630	138	2
MCL1.5	80.7	13.35	0.51	20785	2155	103	2
MCL1.75	74.1	11.44	0.42	20785	3510	49	1
MCL2.0	66.5	9.85	0.35	20785	4960	34	1
MCL2.5	54.4	memory problems	0.26	20785	7442	33	1
MCL3.0	46.7	memory problems	0.21	20785	9088	30	1
MCL3.5	42	memory problems	0.18	20785	10205	20	1
MCL4.0	40	memory problems	0.16	20785	10988	19	1
MCL4.5	38.4	memory problems	0.14	20785	11632	18	1
MCL5.0	35	memory problems	0.13	20785	12089	17	1
MCODE (2,1,3,100)	65.3	5.55	0.3	14221	272	1382	3
MCODE (2,1,3,3)	102.9	5.54	0.45	15914	450	345	3
MCODE (10,1,3,3)	106.6	5.57	0.46	15870	420	345	3
MCODE (2,00,3,3)	34.6	3.07	0.03	1795	378	24	3
MCODE (2,02,5,3)	74.5	4.03	0.12	4145	374	33	4
MCODE (2,10,3,2)	98.1	5.16	0.41	14701	704	161	3
MCODE (2,02,3,1)	21.4	memory problems	0.12	22810	18061	24	1
MCODE (5,02,3,100)	64.5	4.92	0.17	8942	612	336	3
MCODE (2,02,30,100)	29	1.13	0.02	285	12	32	18
MCODE (2,02,3,3)	70.1	4.56	0.16	7920	952	38	3
MCODE (5,05,5,5)	103.8	5.29	0.27	8601	248	155	4
MCODE (2,05,5,5)	107.7	5.37	0.27	8869	265	155	4
MCODE (10,1,10,10)	146.7	5.04	0.49	14782	33	3416	7
MCODE (10,00,10,10)	25	2.02	0.01	337	34	24	6
MCODE (10,05,10,10)	112.7	3.76	0.2	4260	77	210	7

Table B.1.: ClusterJudge, Modularity, and Davies-Bouldin scores for HCCA, k-means, MCL and MCODE clustering solutions.

Algorithm	0-10	10-20	20-50	50-100	100-150	150-200	200-250	250-300	300-350	350-400	400-450	450-500	500-550	550-600	600-650	650-700	700-750	750-800	800-850	850-900	900-950	950-1000	Total number of clusters
HCCA n=2	0	0	0	20	15	15	6	7	7	4	2	2	0	0	2	0	1	0	1	1	0	1	84
HCCA n=3	0	0	8	95	30	25	17	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	181
HCCA n=4	0	0	22	90	33	29	18	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	193
HCCA n=2 unweighted	0	0	0	7	9	9	2	4	4	0	0	3	4	1	1	1	3	3	1	1	0	1	54
HCCA n=3 unweighted	0	0	0	4	13	9	8	10	5	6	6	3	0	2	0	1	0	0	1	0	1	1	70
HCCA n=4 unweighted	0	0	1	16	14	13	6	10	3	6	4	1	4	1	1	0	2	0	0	1	0	0	83
MCL 1.1	124	4	3	5	3	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	142
MCL 1.15	141	13	27	14	10	1	3	6	4	0	0	2	2	2	1	2	0	0	0	1	0	3	232
MCL 1.2	201	50	74	42	26	11	4	3	3	4	4	1	1	1	1	0	0	0	1	0	0	0	427
MCL 1.25	307	123	144	53	25	16	6	3	2	0	1	0	0	0	0	0	0	0	0	0	0	0	680
MCL 1.3	491	202	180	64	30	5	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	976
MCL 1.33	634	244	195	74	20	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1171
MCL 1.35	763	233	217	74	13	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1302
MCL 1.4	1071	279	217	58	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1630
MCL 1.5	1552	360	227	15	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2155
MCL 1.75	3076	322	112	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3510
MCL 2	4694	211	55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4960
MCL 2.5	7315	114	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7442
MCL 3.5	10150	55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10205
MCL 4.0	10956	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10988
MCL 4.5	11608	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11632
MCL 5	12071	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12089
k-means 50	1	0	0	0	1	1	2	6	5	14	5	3	6	1	1	2	1	1	0	0	0	0	50
k-means 100	1	0	0	5	11	32	32	9	7	0	2	0	1	0	0	0	0	0	0	0	0	0	100
k-means 150	1	1	1	26	72	35	6	7	0	1	0	0	0	0	0	0	0	0	0	0	0	0	150
k-means 200	1	2	7	102	66	16	2	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	200
k-means 250	1	2	31	156	48	10	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	250
k-means 300	1	1	66	199	29	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	300
k-means 350	2	1	147	179	17	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	350
k-means 400	1	6	220	158	14	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	400
k-means 450	5	18	293	118	13	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	450
k-means 500	1	21	379	88	10	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	500
k-means 550	4	42	410	86	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	550
k-means 600	5	74	458	55	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	600
k-means 650	6	119	462	61	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	650
k-means 700	12	161	478	46	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	700
k-means 750	19	236	444	49	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	750
k-means 800	28	263	470	38	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	800
k-means 850	46	307	461	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	850
k-means 900	50	397	424	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	900
k-means 950	73	425	428	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	950
k-means 1000	102	488	386	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1000
MCODE (2,1,3,100)	146	42	38	22	5	5	2	3	0	0	1	0	1	1	1	0	1	0	0	1	0	1	270
MCODE (2,1,3,3)	138	114	101	63	19	10	2	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	450
MCODE (10,1,3,3)	125	91	106	62	21	10	2	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	420
MCODE (2,00,3,3)	371	6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	378
MCODE (2,02,5,3)	216	122	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	374
MCODE (2,10,3,2)	277	199	160	61	6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	704
MCODE (2,02,3,1)	18008	43	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18061
MCODE (5,02,3,100)	364	144	82	15	3	3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	612
MCODE (2,02,30,100)	0	2	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12
MCODE (2,02,3,3)	730	180	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	952
MCODE (5,05,5,5)	23	53	124	42	5	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	248
MCODE (2,05,5,5)	33	55	131	40	5	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	265
MCODE (10,1,10,10)	1	1	4	7	5	3	1	0	1	2	0	0	0	1	0	2	0	1	0	0	0	0	29
MCODE (10,00,10,10)	26	6	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34
MCODE (10,05,10,10)	2	10	31	23	8	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	77

Table B.2.: Cluster size distributions for for HCCA, k-means, MCL and MCODE clustering solutions. 20-400 is the desired cluster size range.

Mutual rank	HRR10	HRR20	HRR30	HRR40	HRR50
HRR10	1	0.3643	0.1595	0.0871	0.053
HRR20		1	0.4763	0.2844	0.1833
HRR30			1	0.4802	0.3257
HRR40				1	0.4407
HRR50					1

Table B.3.: Adjusted Rand index analysis of clustering solutions generated by HCCA using HRR cutoffs. Sizes of the networks compared: HRR10=26770 edges, HRR20=63491 edges, HRR30=103587 edges, HRR40=145644 edges and HRR50=189291 edges. The networks contain 22810 nodes each.

Cluster number	cluster size	non essential	essential	characterized	p-value of enrichment of essential genes in a cluster	p-value of enrichment of characterized mutants in a cluster
0	84	0	0	0	0.63	1
1	132	3	2	5	0.67	0.21
2	132	5	5	10	0.02†	0.02†
3	114	0	0	0	0.65	1
4	148	6	0	6	0.43	0.6
5	135	0	0	0	0.42	1
6	205	22	1	23	0.52	0.1
7	65	0	0	0	1	1
8	264	1	1	2	0.38	0.32
9	167	11	3	14	0.46	0.72
10	67	1	0	1	1	1
11	73	1	0	1	1	1
12	157	5	1	6	1	1
13	178	4	0	4	0.28	1
14	76	7	5	12	0†	0.04†
15	155	0	0	0	0.27	1
16	163	8	0	8	0.27	0.36
17	103	3	2	5	0.35	0.21
18	80	3	1	4	0.62	0.54
19	93	2	0	2	0.63	1
20	52	4	0	4	1	1
21	181	19	17	36	6.86E-11†	3.17E-05†
22	87	6	0	6	0.63	0.6
23	73	0	0	0	1	1
24	64	1	0	1	1	1
25	180	2	0	2	0.28	1
26	275	2	0	2	0.08	1
27	145	5	0	5	0.43	0.59
28	64	6	0	6	1	0.6
29	71	4	0	4	1	1
30	190	7	0	7	0.18	0.61
31	118	18	3	21	0.17	1
32	295	3	1	4	0.27	0.54
33	196	3	7	10	0.01†	0†
34	254	7	0	7	0.08	0.61
35	121	20	0	20	0.41	0.04‡
36	59	6	0	6	1	0.6
37	236	30	4	34	0.37	0.5
38	216	6	1	7	0.53	1
39	115	10	0	10	0.65	0.23
40	66	12	1	13	0.55	0.49
41	160	5	3	8	0.45	0.15
42	86	2	0	2	0.63	1
43	54	0	0	0	1	1
44	72	4	0	4	1	1
45	221	16	1	17	0.53	0.34
46	75	8	0	8	1	0.36
47	108	2	1	3	1	0.44

48	180	22	2	24	1	0.29
49	77	2	0	2	1	1
50	80	1	0	1	1	1
51	76	3	0	3	1	1
52	221	23	0	23	0.12	0.02‡
53	113	3	1	4	1	0.54
54	122	3	1	4	1	0.54
55	69	1	0	1	1	1
56	250	27	1	28	0.38	0.05‡
57	191	14	3	17	0.5	1
58	170	6	0	6	0.28	0.6
59	111	10	12	22	9.37E-09†	8.46E-05†
60	67	6	4	10	0.01†	0.08
61	118	3	1	4	1	0.54
62	88	0	0	0	0.63	1
63	65	2	0	2	1	1
64	209	22	3	25	0.74	0.6
65	102	1	0	1	0.64	1
66	111	7	0	7	0.65	0.61
67	114	8	0	8	0.65	0.36
68	167	7	1	8	0.73	1
69	150	11	0	11	0.43	0.23
70	120	11	8	19	0†	0.01†
71	70	3	1	4	0.57	0.54
72	97	6	1	7	1	1
73	50	2	0	2	1	1
74	65	0	0	0	1	1
75	63	4	0	4	1	1
76	63	3	0	3	1	1
77	72	7	4	11	0.01†	0.11
78	112	6	1	7	1	1
79	118	4	2	6	0.65	0.29
80	239	25	2	27	1	0.21
81	64	4	1	5	0.54	1
82	49	2	0	2	1	1
83	161	11	3	14	0.45	0.72
84	55	0	0	0	1	1
85	196	5	0	5	0.18	0.59
86	206	22	1	23	0.52	0.1
87	76	1	0	1	1	1
88	54	1	0	1	1	1
89	59	7	3	10	0.03†	0.39
90	248	10	12	22	4.83E-05†	8.46E-05†
91	76	4	1	5	0.6	1
92	79	0	0	0	1	1
93	84	4	1	5	1	1
94	62	5	0	5	1	0.59
95	195	16	1	17	0.73	0.34
96	70	0	1	1	0.57	0.18
97	71	3	1	4	0.58	0.54
98	81	3	1	4	0.62	0.54

99	64	2	0	2	1	1
100	75	6	0	6	1	0.6
101	52	2	0	2	1	1
102	59	2	0	2	1	1
103	53	5	0	5	1	0.59
104	88	7	2	9	0.28	0.66
105	222	5	0	5	0.12	0.59
106	62	5	1	6	0.53	1
107	83	10	0	10	0.63	0.23
108	68	1	0	1	1	1
109	179	13	3	16	0.48	1
110	226	23	2	25	1	0.29
111	45	1	2	3	0.1	0.08
112	126	4	2	6	0.67	0.29
113	252	18	8	26	0.01†	0.11
114	64	6	3	9	0.04†	0.2
115	233	16	0	16	0.12	0.09
116	51	1	2	3	0.12	0.08
117	51	1	2	3	0.12	0.08
118	78	1	2	3	0.24	0.08
119	57	0	0	0	1	1
120	50	1	4	5	0†	0†
121	76	4	1	5	0.6	1
122	246	11	7	18	0.03†	0.03†
123	57	0	1	1	0.5	0.18
124	140	9	2	11	0.69	1
125	47	7	0	7	1	0.61
126	60	4	2	6	0.16	0.29
127	247	17	1	18	0.38	0.23
128	151	7	0	7	0.27	0.61
129	222	25	6	31	0.05	0.81
130	173	18	6	24	0.02†	0.41
131	73	0	0	0	1	1
132	72	1	1	2	0.58	0.32
133	153	6	3	9	0.44	0.2
134	77	2	0	2	1	1
135	162	12	0	12	0.27	0.14
136	50	4	1	5	0.45	1
137	65	3	7	10	1.26E-05†	0†
138	80	12	0	12	1	0.14
139	69	4	1	5	0.57	1
140	228	17	1	18	0.53	0.23
141	86	5	1	6	1	1
142	52	9	0	9	1	0.37
143	68	4	1	5	0.56	1
144	67	7	0	7	1	0.61
145	172	9	3	12	0.46	0.45
146	79	4	2	6	0.24	0.29
147	119	4	0	4	0.41	1
148	55	3	1	4	0.49	0.54
149	139	12	3	15	0.23	0.74

150	88	6	0	6	0.63	0.6
151	70	5	0	5	1	0.59
152	90	1	1	2	1	0.32
153	87	5	0	5	0.63	0.59
154	183	12	1	13	0.73	0.49
155	62	3	0	3	1	1
156	73	6	1	7	0.59	1
157	78	2	1	3	0.61	0.44
158	48	3	1	4	0.44	0.54
159	150	5	1	6	1	1
160	81	2	1	3	0.62	0.44
161	66	4	1	5	0.55	1
162	61	4	1	5	0.52	1
163	268	14	4	18	0.57	0.54
164	149	16	3	19	0.27	1
165	60	7	1	8	0.52	1
166	89	6	0	6	0.63	0.6
167	63	7	2	9	0.17	0.66
168	70	3	0	3	1	1
169	55	2	1	3	0.49	0.44
170	69	2	1	3	0.57	0.44
171	61	2	0	2	1	1
172	48	5	3	8	0.02†	0.15
173	132	7	1	8	1	1
174	67	4	0	4	1	1
175	153	7	0	7	0.27	0.61
176	61	7	1	8	0.52	1
177	51	3	2	5	0.12	0.21
178	111	4	0	4	0.65	1
179	60	3	0	3	1	1
180	51	3	0	3	1	1

Table B.4.: Fisher's exact test for enrichment of characterized and essential genes in HCCA n=3 obtained clusters. P-value of essential genes: For each Cluster the probability was calculated that the number of essential genes could be due to chance. To this aim a Fisher's exact test was performed for each cluster taking into account the cluster size, the number of the essential genes in total and the total number of genes in clusters (20509). P-value of characterized mutants: The same procedure was repeated for the clusters only considering all mutants in the clusters. Thus the number of essential genes in a cluster was compared to all characterized genes in this cluster and tested if this was likely due to chance alone.
† more essential than expected.
‡ less essential than expected.

Gene ID	T-DNA line	Left primer	Right primer	T-DNA insert segregation (WT:Het:Homo)
At2g35040	SALK_010254.49.50.x	TTAGCCGCAAAATCACAAAAC	CGGGTACAAATTGACAACCAC	05:22:00 PM
At1g74260	SALK_050980.55.50.x	AGGAAGTAGAGGCCGTAGCTG	CTGATGAGCTACCAAGCTTGG	No transmission through pollen
At3g23940	SALK_069706.55.25.x	GTCCATATCCGAGAACACACC	GAACGTCAAGCTTATTCACGC	08:18:00 PM
At3g14900	SALK_123989.29.90.x	GCGTTTAGTGTCAAAAGGTCG	CACCCTAACTCCATCTCCTCC	27:53:00
At3g23070	SALK_125653.40.20.x	TGAAAGGCGGTTCATAATCAG	GGTGTGTGGCTAGAGAATTGG	07:21:00 PM
At1g15510	SALK_112251.52.15.x	CTCGTCTGGCTTCACAGAATC	GGAATTTTGGAATTCGGAGAC	Pale green seedlings are homozygous
At1g69200	SAIL_253_C11	TTGGAACATTGAGTTTTTGGC	TCATCGTCACTGCAGTTTCAC	Pale green seedlings are homozygous
At3g57180	SALK_068713.46.60.x	TAACCGAGCTGGTGAAATCTG	CTGTCGACGGAGTTGCTTATC	Homozygous plant is pale dwarf
At4g38150	SALK_008311.49.55.x	GGTTTCTCGGTTGGTTTCTTC	TTATCCACACGGAGGAGTGAG	Homozygous plants have wildtype phenotype
At4g24280	SALK_140810.44.20.x	CAAGCTGTTGTTAATCCCGAG	GAAGCCAGATGCTTGTGAAAG	Homozygous plants have wildtype phenotype
At1g72040	SALK_060944.41.85.x	CCGGGAGTGAGAAATCTAACC	CGTCTTCATCTTCTCCGTCAG	Homozygous plants have wildtype phenotype
At1g48570	SALK_020144.28.95.x	ATGATGACATTTCACCTTCGG	CGAGCATCAGCTTCTGTATCC	Homozygous plants have wildtype phenotype
At1g80480	SALK_144276.55.75.x	CTCGTTTGCATCTAGACGAGG	ATGTCTTCACTGCGTTCCATC	Homozygous plants have wildtype phenotype
At5g23070	SALK_074256.54.25.x	CGATTCAGAGAGCGATTCATC	ATCGCAATCATCAAATCCAAC	Homozygous plants have wildtype phenotype
At5g13120	SALK_024971.49.95.x	TCACCTACCATTTTTGCCAAC	ACAGCTAGTTGCATACCCACG	Homozygous plants have wildtype phenotype
At2g39140	SALK_144707.54.25.x	CGTCTTTTGGTGGTTAGGTTC	CCTCTGAGTGTGAAGCTGGAG	Homozygous plants have wildtype phenotype
At5g24650	SALK_136524.17.85.x	TTTGTGTCAAAAGCAATCTAACG	TCCTGTCATGGAATCAATCTTG	No insert detected
At1g28530	SALK_007063.54.20.x	ATCTGGTCGGAAATCTAAGCG	ATCTGCGCAAGATATCGCTAC	No insert detected
At1g15510	SALK_112265.54.70.x	CTCGTCTGGCTTCACAGAATC	GGAATTTTGGAATTCGGAGA	No insert detected
At1g09900	SALK_009183.21.50.x	TGATATGCAAACTCCCCAAAC	AAGTTTTTGAATGACATGCCG	No insert detected

Table B.5.: T-DNA knock-out lines and primers used.