GCB 2012 Highlight Abstracts

Table of Contents

Dynamic modeling resolves complex hormonal crosstalk in infected plants

Muhammad Naseem, Dominik Schaack and Thomas Dandekar
Department of Bioinformatics, Biocenter, Am Hubland, D-97074 Wuerzburg, Germany

dandekar@biozentrum.uni-wuerzburg.de

Abstract

Owing to their multi faceted interactions analysis of the combined output of plant hormones is always a challenge. Hormonal crosstalk plays a pivotal role in successful system protection or plant vulnerability. We developed a dynamical model and analysed the impact of individual plant hormones in an interactive way. We first established a network combining available interaction data and then used this molecular interaction network as a substrate for dynamic simulations on hormonal aspects of plant immunity. Our analysis revealed that plant hormones such as SA, GA and CK promote immunity against the infection of *Pst* DC3000 in *Arabidopsis*. On the other hand JA, Auxin and ABA promote vulnerability of *Arabidopsis*. These findings are in line with current literature, old and new experiments. Dynamic modelling can be applied to investigate antagonism and synergism between hormonal pathways in plants. It allows to study infections and host-pathogen interactions and in general the molecular events during organismic interactions.

1 Introduction

Plant hormones are shared weaponry in pathogen infections. Pathogenic attack causes hormonal imbalances in plants. Depending upon the trophic nature of pathogen, either SA (Salicylic Acid) or JA/ET (Jasmonic acid / Ethylene) mediated defense pathways are operative in plants [Gra9]. Antagonism between JA and SA and synergism between ET and JA has long been elucidated (reviewed by: [Rob11]. Furthermore, growth regulatory hormones such as auxin promote JA responses and suppress the SA pathway of resistance [Wan7]. Similarly, ABA (Abscisic Acid) antagonizes SA mediated defense signalling while SA abolishes ABA responses [Rob11]. On the contrary, GA (Gibberellic Acid) reinforces SA accumulation [Gra9]. It is worth mentioning that *Pst* DC3000 also promotes in-planta levels of auxin and ABA. Furthermore, it is injecting a JA mimicry to suppress SA mediated defense. Taking these pathways carefully into account, we established a plant immune defence network [Nas12] and performed dynamic simulations on various aspects of plant immunity. Here we highlight key connections in this network (Figure 1) and discuss the implications of various hormones in plant immune defence.

2 Results and Discussion

We analyzed the impact of phytohormones on immune defense using a Boolean model. Boolean network models have an advantage over ODE-based kinetic models in complex networks including immune and pathogen responses. In contrast to ODE models Boolean network models can also work when kinetic information is scarce and many nodes are involved [Sch11]. SQUAD (Standardized Qualitative Dynamical Systems; [Phi9]) is a powerful modelling package which combines Boolean and ODE models. It creates a system of exponential functions that allows interpolation between the step function of Boolean models according to the sum of activating and inhibitory input [Phi9]. To generate a network we integrated hormonal nodes with protein regulatory molecules in the cell according to the Boolean logic of their interactions (Figure 1; see [Nas12] for detailed Network topology). To model the impact of plant hormones on plant immunity, we performed SQUAD simulations by taking individual hormones as input activating nodes. The activation of PR-1 over time was used as an index of plant immunity (for detailed methodology see [Nas12]). We found in the simulations that the three hormones ET, SA and GA activate PR-1 (Figure 2 D, E, H). These hormones further enhance the signal of the activity of PR-1 in the presence of *Pst*, while JA, auxin and ABA diminish even the residual activity of PR-1 manifested by *Pst* alone (Figure 2 A). Contrary to the immunity promoting effect of SA and GA, we saw that auxin, JA and ABA mediate susceptibility of *Arabidopsis* against infection by *Pst* DC3000. Moreover, our modelling suggests a promoting role of cytokinin against infection by *Pst* DC3000 in *Arabidopsis* (Figure 2 G). These simulation results are in line with literature and own new experimental data and thus qualify analysis of Boolean models [Nas12] as a step forward to investigate plant regulatory and signalling networks. Boolean models specifically resolve here hormonal crosstalk during complex host-pathogen interactions in plants.

3 Conclusion

During host pathogen interactions in plants, *Pst* DC3000 uses effectors to modulate endogenous levels of phytohormones to mediate susceptibility of the *Arabidopsis* plant. Our simulations successfully predicted responses when hormone levels were changed and compared the uninfected plant to the plant infected with *Pst* DC3000. Modelling suggests that plant hormones such as SA and GA promote resistance against infection by *Pst* DC3000, while JA, auxin and ABA enhance susceptibility for infection of the plant. These predicted hormonal effects are similar to own and previous experimental results. Furthermore, our simulations suggested promoting effects of cytokinin on plant immunity against the infection by *Pst* DC3000 which again could be verified by experiment. Crosstalk of cytokinins as well as other hormones under different conditions in plants and other organisms can advantageously be explored with the methods of dynamic modelling outlined here.

References

[Gra9] Grant MR, and Jones JDG. Hormone (Dis) harmony mould plant health and disease. Science 324: 750-752, 2009.

[Rob11] Robert-Seilaniantz A, Grant M, and Jones JDG. Hormone crosstalk in plant disease and defense: More than just salisylate-jasmonate antagonism. Annu. Rev. Phytopathol. 49: 317-43, 2011.

[Wan7] Wang D, Mukhtar KP, Culler AH, and Dong X. Salicylic Acid Inhibits Pathogen Growth in Plants through Repression of the Auxin Signalling Pathway. Current Biol. 17: 1784-1790, 2007.

[Nas12] Naseem M, Philippi N, Hussain A, Wangorsch G, Ahmed N, and Dandekar T. Integrated systems view on networking by hormones in *Arabidopsis* immunity reveals multiple crosstalk for cytokinins. Plant Cell, 5: 1793-814, 2012.

[Sch11] Schlatter R, Philippi N, Wangorsch G, Pick R, Sawodny O, Borner C, Timmer J, Ederer M, and Dandekar T. Integration of Boolean models exemplified on hepatocyte signal transduction. Brief. Bioinform. 13: 365-76, 2011.

[Phi9] Philippi N, Walter D, Schlatter R, Ferreira K, Ederer M, Sawodny O, Timmer J, Borner C, and Dandekar T. Modelling system states in liver cells: Survival, apoptosis and their modifications in response to viral infection. BMC Syst. Biology 3: 97, 2009.

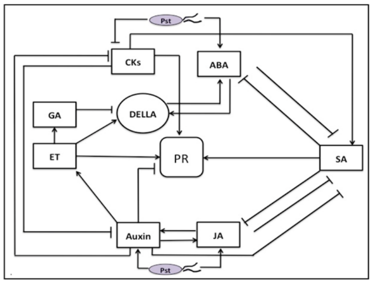


Figure 1: Logical connections in the plant pathogen hormone interaction network tested.
Infection with *Pst* DC3000 (shown with flagellae) in *Arabidopsis*. Connectivity among nodes is based either on activation (→) or inhibition (⊣).We give here only the most central backbone of interactions. For the detailed network topology modeled please see [Nas12]. PR-1 is a well-known key marker node for immunity against the infection of *Pst* in *Arabidopsis*.

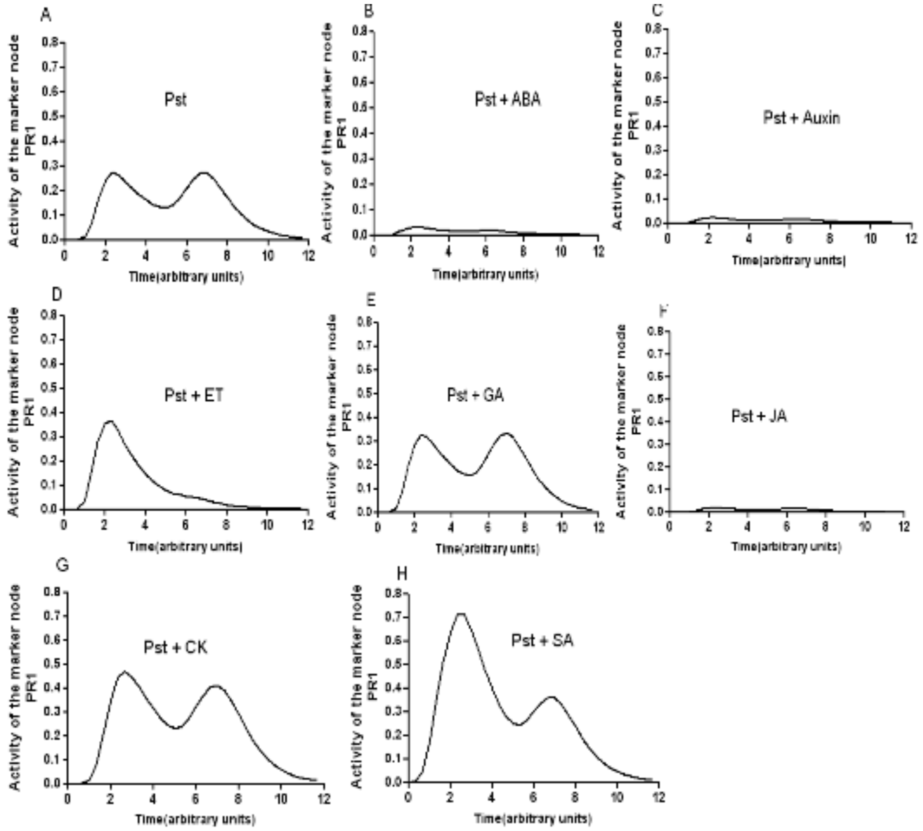


Figure 2: Modelling the impact of plant hormones on immune defense in *Arabidopsis* against by *Pst* DC3000.

Activity of PR-1 over arbitrary units of time (y-axis) is shown as immune output of the plant over time (x-axis). Activated nodes of *Pst* and plant hormones as activating input signal change the state of immunity in the host. Modelling of hormonal response for: A) Virulent *Pst* DC3000 infection and plant immune response B) *Pst* DC3000 infection after the application of plant hormone ABA C) auxin D) ethylene E) Gibberellic Acid F) Jasmonic Acid G) Cytokinin and H) Salicylic Acid. Experimental results qualitatively verified trajectories and shapes.

Two new perspectives on NAD metabolism

Toni I. Gossmann[1,2], Mathias Ziegler[1], Pål Puntervoll[3], Luis F. de Figueiredo[4,5], Stefan Schuster[4] and Ines Heiland[4]
[1]*Department of Molecular Biology, University of Bergen, Norway*
[2]*University of Hohenheim, Institute of Plant Breeding, Stuttgart* [3] *CBU, Bergen, Norway,* [4] *Department of Bioinformatics, Friedrich-Schiller-University Jena* [5] *EBI, Hinxton, Cambridge, UK*

toni.gossmann@googlemail.com

Abstract: NAD$^+$ has gained increased attention during recent years due to its involvement in cellular signalling and regulation and changes in NAD metabolism are associated with ageing, diabetes and neurodegenerative diseases. Here we review the key findings of our two recent studies on the theoretical investigation of NAD metabolism. In the first [dFGZS11], we used elementary flux mode analysis and revealed unexpected fluxes including futile cycles and NAD$^+$ signalling without net consumption of NAD$^+$. We furthermore identified essential enzymes such as NAD$^+$-kinase (NADK), converting NAD$^+$ into NADP$^+$, and the mononucleotide adenylyl transferase (NMNAT). The second study [GZP$^+$12], investigated the phylogeny of this pathway and revealed that the two NAD salvage pathways exist simultaneously in some species and that the first enzyme of this pathway in higher eukaryotes (Nam-phosphoribosyltransferase (NamPT)) seems to have been lost several times during evolution. Both analyses successfully combine bioinformatic approaches with biochemical expertise.

1 Introduction

NAD$^+$ is a key metabolite as it participates in a vast number of redox reactions. Over the past decades it gained additional attention as it is involved in many signalling reactions that play a role in cell regulation. Changes in NAD$^+$ metabolism have been found during aging and in metabolic disease such as diabetes. NAD$^+$ depending signalling reactions include the consumption of NAD$^+$ and require a constant replenishment of cellular NAD$^+$ pools. This can be done either by salvaging the product of these reaction, nicotinamide (Nam), or by synthesising NAD$^+$ *de novo* from the amino acids aspartate or tryptophan. As a systematic analysis of NAD metabolism has not been performed before, we first analysed the underlying metabolic network using elementary flux mode (EFM) analysis. As

NAD$^+$ can be synthesised and degraded by multiple routes, EFMs help to decompose the network and identify possible metabolic routes and analyse the effect of enzyme deletions. As human and yeast are by far best investigated organisms regarding NAD metabolism, but show remarkable differences in their NAD related enzyme composition, we initially reconstructed the NAD metabolism of these two species as a basis for our EFM-analysis [dFGZS11].

The results from this initial study showed that the vast majority of NAD$^+$ biosynthesising enzymes are present in yeast and human. However, the number of NAD$^+$-consuming enzymes differs substantially between the two species and there are some routes that are present in either human or yeast. This differences in the NAD metabolism called for the reconstruction of the evolutionary history of NAD metabolism. Previous studies for prokaryotes have shown that neither *de-novo* synthesis nor salvage of NAD$^+$ are universal and occur via modules of different genes [GSC$^+$09]. However little is known on how the complexity of NAD metabolism evolved over time in higher organisms. Therefore, we have analysed the phylogenetic distribution of NAD metabolism related enzymes in 45, mainly higher eukaryotic species [GZP$^+$12].

2 Major findings

We first reconstructed a metabolic network of NAD$^+$ biosynthesis in human and yeast. These were combinde to create a generalised model that comprises 113 EFMs. These are metabolic routes that are stoichiometrically and thermodynamically balanced and consists of a minimal set of enzymes that can operate at steady state. 50 EFMs can be found in human and 100 in yeast but only 40 are shared. Several of these EFMs constitute futile cycles, which are routes in the metabolic networks with no net transformation except hydrolysis of ATP. Whether these futile cycles are of physiological relevance depends on corresponding kinetics and regulation mechanisms. The much larger number of possible routes found in yeast is rather surprising. Another tendencies is that within the human network amidated forms are preferred, while yeast preferentially uses deamidated forms. Moreover, in both species elementary modes were identified that allow NAD$^+$ dependent signalling without net consumption of NAD$^+$. This was not known so far. Furthermore, Nam-mononucleotide adenelyl transferase (NMNAT) was identified to be essential for NAD$^+$ biosynthesis. This is consistent with experimental findings.

The combined human-yeast model was used as a basis for the phylogenetic analysis of NAD^+ metabolism in eukaryotes. Looking across 45 mainly eukaryotic species it becomes apparent that all investigated species are able to synthesise NAD^+ from at least one precursor and most species have more than one NAD^+ biosynthetic pathway. Some species lack the possibility to synthesise NAD^+ *de-novo* from aspartate or tryptophan and must therefore live under conditions which provide a sufficient amount of the NAD precursors nicotinic acid and Nam, commonly known as the vitamin niacin. The enzymes NADK and NMNAT are found in all species and can be considered to be essential for NAD metabolism supporting our results from the EFM analysis. Furthermore, the Preiss-Handler pathway is the most predominant NAD biosynthetic route among organisms suggesting a universal role for the generation of NAD^+.

The comparison between yeast and human showed that NamPT is an enzyme which can be found in humans but not in yeast, while the enzyme Nam-deaminase (NADA) is present in yeast but not in human. Both enzyme use Nam and provide the first step for Nam recycling to NAD^+ in the respective species. However, NamPT clearly provides a more efficient and economic route to NAD^+. It had been speculated that those two enzymes are mutually exclusive [RAGL03]. Surprisingly, the multispecies comparison reveals a scattered distribution of both enzymes across the animal kingdom. It rather suggests that NamPT enzymatic function got lost several times during evolution while the loss of NADA happened once and is common to all vertebrates. Interestingly, all species identified so far that have both NADA and NamPT have aquatic habitats. Whether, this has any physiological implications we do not know and we also do not know whether the enzymes are indeed expressed simultaneously.

Looking at the relation between NAD^+-consuming and Nam-recycling enzymes we initially assumed that increase of NAD^+-consuming enzymes should be reflected in an increase in biosynthetic routes. This is surprisingly not the case. In contrast we found a parallel phylogenetic appearance of the enzyme Nam-N-methyltransferase which is marking Nam, the product of NAD^+-consuming reactions, for further degradation and thus removing Nam from recycling to NAD^+.

3 Conclusions and future perspectives

The decomposition of the NAD metabolic network has revealed unexpected fluxes including futile cycles and NAD^+ signalling without net consumption of NAD$^+$. The physiological relevance has to be shown experimentally. Furthermore, some reactions might not occur in the cell, as our models currently neglect compartmentalisation. The investigation and integration of the subcellular localisation and the identification of metabolite transporters is therefore crucial to understand NAD metabolism. As furthermore not all routes identified will be feasible under physiological conditions we are currently building a kinetic model to better understand the kinetic constraints that limit NAD$^+$-biosynthesis and -consumption.

The results from our phylogenetic analysis have revealed several interesting aspects that raise important issues. For example, why do some organisms encode both NADA and NamPT while many others do not? It has been suggested that the lack of NADA in vertebrates is compensated by gut microbiotic flora [GSC$^+$09]. Therefore such an interplay between organisms could serve as a pool for the entry metabolite of the Preiss Handler pathway. It could also provide a possible explanation for the concurrent existence of NADA and NamPT in some species.

Another question arising from our analysis is, why higher eukaryotes require an enzyme for Nam-degradation whereas species with a low NAD$^+$-consumption do not? Again the answer might be provided by kinetic modelling as Nam is known to be a potent inhibitor of some NAD$^+$-consuming enzymes and might therefore interfere with NAD$^+$-dependent signalling.

References

[dFGZS11] Luis F de Figueiredo, Toni I Gossmann, Mathias Ziegler, and Stefan Schuster. Pathway analysis of NAD+ metabolism. *Biochem J*, 439(2):341–348, Oct 2011.

[GSC$^+$09] Francesca Gazzaniga, Rebecca Stebbins, Sheila Z Chang, Mark A McPeek, and Charles Brenner. Microbial NAD metabolism: lessons from comparative genomics. *Microbiol Mol Biol Rev*, 73(3):529–41, Table of Contents, Sep 2009.

[GZP$^+$12] Toni I Gossmann, Mathias Ziegler, Pål Puntervoll, Luis F de Figueiredo, Stefan Schuster, and Ines Heiland. NAD(+) biosynthesis and salvage - a phylogenetic perspective. *FEBS J*, Mar 2012.

[RAGL03] Anthony Rongvaux, Fabienne Andris, Frédéric Van Gool, and Oberdan Leo. Reconstructing eukaryotic NAD metabolism. *Bioessays*, 25(7):683–690, Jul 2003.

Image-based systems biology: A quantitative approach to elucidate the kinetics of fungal morphologies and virulence

Franziska Mech and Thilo Figge
Research Group Applied Systems Biology, Hans-Knöll-Institute Jena

franziska.mech@hki-jena.de

Abstract: *Aspergillus fumigatus* and *Candida albicans* are the major human-pathogenic fungi. There is a variety of experimental set-ups available to investigate the virulence and morphologies of both fungi. Imaging of these experiments using fluorescence microscopy yields vast amounts of image data which could not be analysed manually. Therefore, we applied the approach of 'image-based systems biology'. It comprises the automated image analysis with subsequent statistical feature analysis, followed by mathematical modelling. Application of 'image-based systems biology' to *A. fumigatus* phagocytosis assays and *C. albicans* epithelial invasion assays reveals important factors of the virulence of wild-type *A. fumigatus* and enables the quantitative description of the morphological transition of *C. albicans*, during invasion of the epithelium.

1 Introduction

Elucidation of communication and interplay between the fungal pathogens and the human host is of great interest. The spatio-temporal resolution of host-pathogen interactions provides vast amounts of biological data [Rit10]. For that, imaging technologies, such as fluorescence microscopy, are often utilised in microbiology [RC11, AKM$^+$11]. Thus, observation of large amounts of cell assays is possible leading to additional insights into biological processes which are, so far, not achievable with 'omics' data analysis alone. However, the very time-consuming and highly error-prone manual analysis of the large amounts of data represents the bottleneck of the analysis [JMK$^+$07, NHL$^+$06]. Therefore, an automated approach is at need since systematic studies of comprehensive mutant screenings cannot be performed otherwise. Automatic analysis of spatio-temporal data sets provides morphological features, as well as spatial and temporal dynamics of observed systems [RGH$^+$10]. These observations represent a useful source for verifying or driving new hypotheses and, thus, can be incorporated into system models [CGT+08]. Integration of image analysis is the key of the 'image-based systems biology' approach. The gained quantitative and morphological features of the system under consideration, as well as interactions in the communication between host and pathogen during fungal infections are further statistically analysed to determine important characteristics of the system. Subsequently, the quantitative features are used to build and test mathematical models of certain processes. In this paper the application of the quantitative approach 'image-based systems biology' is highlighted on two different fungal experiments: (i) *Aspergillus fumigatus* phagocytosis assays [MTG+11] and (ii) *Candida albicans* epithelial invasion assays [MWL+].

2 Fungal virulence and morphology

First, the host-pathogen interactions between *A. fumigatus* and macrophages shortly after infection were investigated using phagocytosis assays of different *A. fumigatus* strains. In the early stages of infection *A. fumigatus* resides in its conidial form and is phagocytosed by macrophages. An automated image analysis algorithm was developed to successfully recognise conidia and macrophages [MTG+11]. The subsequent feature analysis revealed a decreased adhesion ratio for the pksP mutant compared to the wild type. Furthermore, the phagocytosis ratio increased as well as the formation of conidial clusters. We assume that due to the lack of the outer cell wall layer (rodlet/ melanin layer) α and β1-3glucans are exposed which enhance the recognition by macrophages and the increase in the aggregation behaviour. Finally, we rigorously validated the segmentation and classification algorithm, involving a quantitative comparison with a manual analysis by experts. This showed high precision and sensitivity scores and facilitated the adaptation to further experiments. Next, we extended and applied the algorithm to epithelial invasion assays of *C. albicans* [MWL+]. These assays were carried out hourly for the first six hours of infection. During that time *C. albicans* is initially in its yeast form and starts adhering to the epithelial surface. This is followed by hyphae formation which triggers a tighter adhesion, thus facilitating active penetration of the host surface or inducing endocytosis by the epithelium. To account for the different cell morphologies of *C. albicans* the automated image analysis algorithm was extended to recognise spherical and cylindrical cells. Following the 'image-based systems biology' approach we statistically analysed the acquired features. The interpretation of these data was supported by two mathematical models, the kinetic growth model and the kinetic transition model, that were developed in terms of systems of ordinary differential equations. The kinetic growth model describes the increase in hyphal length and revealed that hyphae undergo mass invasion of epithelial cells immediately following primary hypha formation. Based on the kinetic transition model, the route of invasion was quantified in the state space of non-invasive and invasive fungal cells depending on their number of hyphae. This analysis revealed that the fungal decision to form primary hypha represents an ultimate commitment to invasive growth and suggests that *in vivo* the yeast to hypha transition must be under extremely tight negative regulation by yet unknown mechanisms that avoid the transition from commensal to invasive/pathogenic growth.

3 Conclusion

In this review we highlighted the findings of two investigations performed recently on infection processes of human-pathogenic fungi [MTG+11, MWL+]. In this context, we applied the ‘image-based systems biology’ approach for the first time. It comprises (i) analysis of large sets of microscopy image data in an automated fashion, (ii) statistical quantification of characteristic features on the basis of the high-throughput and high-content screening of image data, and (iii) integration of acquired spatio-temporal information into mathematical models. Our results are promising with regard to complementing traditional systems biology approaches based on gene-expression data and pave the way for new insights into fungal infection processes.

References

[AKM+11] Daniela Albrecht, Olaf Kniemeyer, Franziska Mech, Matthias Gunzer, Axel A Brakhage, and Reinhard Guthke. On the way toward systems biology of Aspergillus fumigatus infection. *International Journal of Medical Microbiology : IJMM*, 301(5):453–9, June 2011.

[CGT+08] Arvind K Chavali, Erwin P Gianchandani, Kenneth S Tung, Michael B Lawrence, Shayn M Peirce, and Jason a Papin. Characterizing emergent properties of immunological systems with multi-cellular rule-based computational modeling. *Trends in immunology*, 29(12):589–99, December 2008.

[JMK+07] Anne Järve, Julius Müller, Il-Han Kim, Karl Rohr, Caroline MacLean, Gert Fricker, Ulrich Massing, Florian Eberle, Alexander Dalpke, Roger Fischer, Michael Trendelenburg, and Mark Helm. Surveillance of siRNA integrity by FRET imaging. *Nucleic acids research*, 35(18), 2007.

[MTG+11] Franziska Mech, Andreas Thywissen, Reinhard Guthke, Axel A Brakhage, and Marc Thilo Figge. Automated image analysis of the host-pathogen interaction between phagocytes and Aspergillus fumigatus. *PloS One*, 6(5):e19591, April 2011.

[MWL+] Franziska Mech, Duncan Wilson, Teresa Lehnert, Bernhard Hube, and Marc Thilo Figge. Epithelial invasion outcompetes hyphal development during *Candida albicans* Infection. submitted (May 2012).

[NHL+06] Beate Neumann, Michael Held, Urban Liebel, Holger Erfle, Phill Rogers, Rainer Pepperkok, and Jan Ellenberg. High-throughput RNAi screening by time-lapse imaging of live human cells. *Nature methods*, 3(5):385–390, 2006.

[RC11] Lisa Rizzetto and Duccio Cavalieri. Friend or foe: using systems biology to elucidate interactions between fungi and their hosts. *Trends in microbiology*, (Figure 1):1–7, August 2011.

[RGH+10] Karl Rohr, WJ Godinez, Nathalie Harder, Stefan Wörz, and J. Tracking and quantitative analysis of dynamic movements of cells and particles. *Cold Spring Harbor*, 2010(6):pdb.top80, June 2010.

[Rit10] Jens Rittscher. Characterization of Biological Processes through Automated Image Analysis. *Annual review of biomedical engineering*, 12(April):315–44, August 2010.

Identifying the unknowns by aligning fragmentation trees

Florian Rasche[1], Kerstin Scheubert[1], Franziska Hufsky[1,2], Thomas Zichner[3], Marco Kai[4], Aleš Svatoš[4] and Sebastian Böcker[1]

[1] *Chair for Bioinformatics, Friedrich Schiller University, Jena, Germany*
[2] *Max Planck Institute for Chemical Ecology, Jena, Germany*
[3] *Genome Biology Research Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany*
[4] *Research Group Mass Spectrometry and Proteomics, Max Planck Institute for Chemical Ecology, Jena, Germany*

sebastian.boecker@uni-jena.de

Abstract: Mass spectrometry allows sensitive, automated and high-throughput analysis of small molecules. In principle, tandem mass spectrometry allows us to identify “unknown” small molecules not in any database, but the automated interpretation of such data is in its infancy. Some years ago, fragmentation trees have been introduced for the automated analysis of the fragmentation patterns of small molecules. We have recently presented a method for the automated comparison of such fragmentation patterns, based on aligning the compounds’ fragmentation trees. This method enables us to cluster compounds based solely on their fragmentation patterns, and resulting clusterings show a good agreement with known compound classes. We also show that fragmentation pattern similarities are strongly correlated with the chemical similarity of molecules. Finally, we presented a tool for searching a database for compounds with fragmentation pattern similar to an unknown sample compound. Our method allows fully automated computational identification of small molecules that cannot be found in any database.

Extended abstract of Rasche *et al.*, *Anal. Chem.*, 84(7):3417–3426, 2012 [RSH$^+$12].

1 Introduction

Mass spectrometry (MS) is a key analytical technology for detecting and identifying small molecules such as metabolites [CLH$^+$08]. It is orders of magnitude more sensitive than nuclear magnetic resonance (NMR). Several analytical techniques have been developed, most notably gas chromatography MS (GC-MS) and liquid chromatography MS (LC-MS). LC-MS is usually combined with a gentle ionization, that results in minimal fragmentation of the adduct ions formed. Molecules can be further analyzed using tandem MS: Molecules are mass-selected, fragmented, and the mass-to-charge ratios (m/z) of the resulting fragments recorded.

Fragmentation in LC-MS experiments (usually collision-induced dissociation (CID)) is less reproducible than fragmentation by electron ionization for GC-MS. Even the time-consuming manual analysis of such data, as well as searching in spectral libraries, are major problems. Apart from a few pioneering studies, there are few computational methods for the automated analysis of tandem MS data from small molecules.

For decades, MS experts have manually determined fragmentation pathways to explain tandem MS data and determine the molecular structure. In 2008, Böcker and Rasche [BR08] presented an automated and swift method for annotating tandem MS data using a hypothetical *fragmentation tree* (FT). Tree nodes are annotated with the molecular formulas of the fragments and the edges represent (neutral or radical) *losses*. Computing FTs does not require databases of compound structures or of mass spectra. Neither does it require, apart from lists of common and implausible losses, expert knowledge of fragmentation. Expert evaluation suggests that the FTs are of very good quality [RSM+11]. Similar FTs can be identified using visual comparison, which indicates some similarity in the structure of the underlying compounds. Unfortunately, "manual comparison of FTs is also laborious and time-consuming" [RSM+11].

In [RSH+12], we presented an automated method for comparing the FTs of two compounds. This allows us to use FTs in applications such as database searching, where we replace the direct comparison of mass spectra by the comparison of the (annotated and more informative) FTs. Our method is based on local tree alignments, generalizing local sequence alignments. We assume that structural similarity is inherently coded in the CID spectra fragments. FT similarity is defined by its edges, which represent losses and nodes, representing fragments. The local tree alignment contains those parts of the two trees where similar fragmentation cascades occurred.

Aligning FTs when the molecular structure of one compound is known can help elucidate the structure of the unknown compound. In [RSH+12], we presented three workflows based on similarity scores. First, we compute pairwise tree alignments for all compounds and so generate a pairwise similarity matrix. We then cluster the compounds based solely on this similarity measure. We find that the resulting clusters agree well with the structural properties of the compounds. Second, we showed that FT similarities and structural similarities (Tanimoto scores) are strongly correlated. Third, we determine the similarities of a fragmentation tree from an unknown compound with all trees in a database, to search for related compounds. To filter out spurious hits, we presented a statistical evaluation based on decoy database searching. We named this approach *fragmentation tree basic local alignment search tool* or FT-BLAST for short. Finally, as a proof of principle we showed how biological samples from Icelandic poppy (*P. nudicaule*) can be analyzed in this framework.

2 Methods

We shortly recall the most important principle of our FT alignment method introduced in [RSH$^+$12], see there for all details. For the automated comparison of FTs we followed the paradigm of pairwise *local alignments*. We defined a simple similarity measure on the edges (losses) and nodes (fragments) of the two FTs. We generalized this similarity measure to trees of identical topology and summed the similarity of tree edges. We also allowed for the insertion and deletion of edges. We searched for *subtrees* in the two FTs that maximized our similarity measure.

Similarity of subtrees was defined as the sum of similarities of edges which, in turn, was chosen to reward identical losses and penalize distinct losses and insertions or deletions. Edge similarities were modified based on the number of non-hydrogen atoms contained. Similarity between fragments (nodes) was also rewarded or penalized. We modified a known recurrence for the problem in three ways. First, we also considered edge similarities. Second, we computed local alignments for maximum subtree similarity by adding a “zero-case” to the recurrence, corresponding to the leaves of the subtree. Third, we scored *join nodes* where two losses were combined into one, corresponding to the non-appearance of intermediate fragmentation steps. Alignment scores will clearly be large for large trees and small for small trees, so we normalized similarities by perfect match scores. To do this we computed for each FT the alignment score against itself, then used the minimum of the two scores, taken to the power of 0.5. We refrained from using the similarity matrix directly. Instead, for each compound we viewed its similarity matrix column as a fingerprint (or feature vector), as is done with gene expression data. See Fig. 1 for an example.

3 Discussion

To achieve the full potential of small molecule MS analysis and to overcome limitations of spectral libraries, we need methods for the computational analysis of fragmentation spectra from unknown compounds. Rule-based approaches for analyzing compound fragmentation spectra may suffer from the tremendous number of rules, both known and unknown. In addition, completely unknown compounds may not necessarily follow the

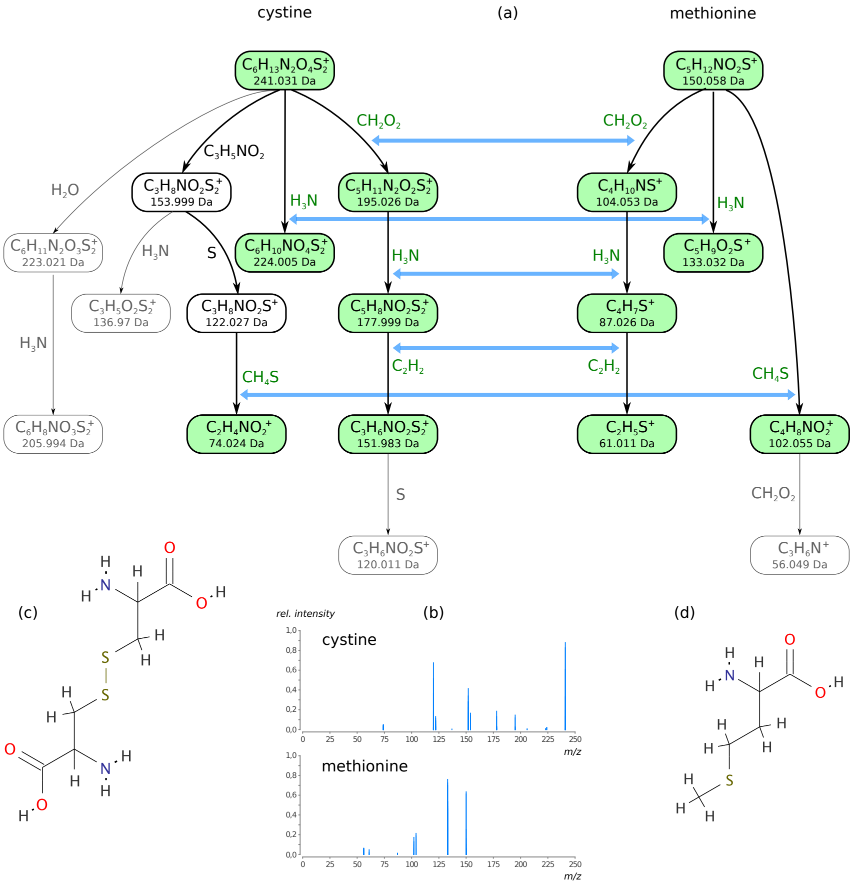

cystine
(a)
methionine
$C_6H_{13}N_2O_4S_2^+$ 241.031 Da
$C_5H_{12}NO_2S^+$ 150.058 Da
CH_2O_2
CH_2O_2
$C_3H_5NO_2$
H_2O
$C_3H_8NO_2S_2^+$ 153.999 Da
$C_5H_{11}N_2O_2S_2^+$ 195.026 Da
$C_4H_{10}NS^+$ 104.053 Da
H_3N
H_3N
$C_6H_{11}N_2O_3S_2^+$ 223.021 Da
H_3N
S
$C_6H_{10}NO_4S_2^+$ 224.005 Da
$C_5H_9O_2S^+$ 133.032 Da
H_3N
H_3N
$C_3H_5O_2S_2^+$ 136.97 Da
$C_3H_8NO_2S^+$ 122.027 Da
$C_5H_8NO_2S_2^+$ 177.999 Da
$C_4H_7S^+$ 87.026 Da
H_3N
C_2H_2
C_2H_2
CH_4S
CH_4S
$C_6H_8NO_3S_2^+$ 205.994 Da
$C_2H_4NO_2^+$ 74.024 Da
$C_3H_6NO_2S_2^+$ 151.983 Da
$C_2H_5S^+$ 61.011 Da
$C_4H_8NO_2^+$ 102.055 Da
S
CH_2O_2
$C_3H_6NO_2S^+$ 120.011 Da
$C_3H_6N^+$ 56.049 Da
(c)
(b)
(d)
rel. intensity
cystine
methionine
m/z
m/z

Figure 1: Optimal FT alignment for cystine (10 losses) and methionine (6 losses). (b) Fragmentation mass spectra used for computing FTs. Molecular structures of cystine (c) and methionine (d).

known rules of fragmentation. Unfortunately, real fragmentation patterns are extremely complicated, and new "rules" are constantly being introduced. This makes manual compound classification and structure elucidation cumbersome. In contrast, the approach presented here is fully automated and "rule-free", both when computing and aligning FTs. It only requires sufficiently information-rich fragmentation spectra.

Clustering results in [RSH$^+$12] show the potential of the method to differentiate compound classes. In many cases, large compound classes formed almost perfectly separated clusters; smaller compound classes were distributed among several clusters, but clusters contained few outliers. Hierarchical clustering was applied as a proof-of-concept and to demonstrate clustering results. Better results can possibly be achieved by other clustering methods and supervised Machine Learning. Nevertheless, our results indicate how to deduce the compound class of an unknown when a reasonable number of knowns are clustered simultaneously.

We found strong correlation between FT similarity and chemical similarity. FT similarity must not be understood as a *prediction* of chemical similarity in the sense of Machine Learning methods. However, FT similarity, expert knowledge, and other sources of information can be combined to permit the accurate prediction of chemical similarity.

Our method for searching spectral libraries (FT-BLAST) achieves a "larger profit" than classical spectral comparison methods, as it searches for similar, not identical, compounds. We achieved excellent search results for most compounds: Even when FT-BLAST returned only a single hit it was often meaningful. Cases where no hits or spurious hits were returned could often be attributed to small FTs, low quality measurements, or the absence of similar compounds from the database. FT-BLAST individually selects the size of the output for each query compound. For this purpose, we proposed a method for generating a decoy database of FTs that can be searched simultaneously [RSH$^+$12]. Database searching by spectral comparison has been in use for decades; but even today, no sensible methods for generating decoy databases for spectral comparisons have been developed.

By applying FT-BLAST and clustering to an unknown sample from poppy, we confirmed eight manual identifications and suggested compound classes for some other unknowns, as they were unquestionably members of a well-defined cluster. We also identified the biosynthetic precursor of several alkaloids, which come from mixed biosynthetic pathways.

FT alignments open a way to a fast classification/identification of metabolites, limiting work spent on ubiquitously occurring “uninteresting” molecules. Areas of application include natural product discovery, dereplication, or even inferring biosynthetic pathways and metabolic networks.

References

[BR08] Sebastian Böcker and Florian Rasche. Towards de novo identification of metabolites by analyzing tandem mass spectra. *Bioinformatics*, 24:I49–I55, 2008. Proc. of *European Conference on Computational Biology* (ECCB 2008).

[CLH+08] Qiu Cui, Ian A Lewis, Adrian D Hegeman, Mark E Anderson, Jing Li, Christopher F Schulte, William M Westler, Hamid R Eghbalnia, Michael R Sussman, and John L Markley. Metabolite identification via the Madison Metabolomics Consortium Database. *Nat Biotechnol*, 26(2):162–164, 2008.

[RSH+12] Florian Rasche, Kerstin Scheubert, Franziska Hufsky, Thomas Zichner, Marco Kai, Aleš Svatoš, and Sebastian Böcker. Identifying the unknowns by aligning fragmentation trees. *Anal Chem*, 84(7):3417–3426, 2012.

[RSM+11] Florian Rasche, Aleš Svatoš, Ravi Kumar Maddula, Christoph Böttcher, and Sebastian Böcker. Computing fragmentation trees from tandem mass spectrometry data. *Anal Chem*, 83:1243–1251, 2011.

KeyPathwayMiner - Combining OMICS data and biological networks

Josch Pauling, Nicolas Alcaraz, Alexander Junge, Jan Baumbach
Max Planck Institute for Informatics, Saarbrücken, Germany

jpauling@mpi-inf.mpg.de

Abstract: KeyPathwayMiner is a method for extracting and visualizing disease-specific key pathways. We identify sub-graphs, where most genes are dysregulated in a typical case-control study. Therefore, we extract all maximal connected sub-networks where all but K genes are differentially expressed/methylated/etc. in all but L cases. This model yields a very high interpretability of the results since K and L have real-world implications. We will exemplarily demonstrate KeyPathwayMiner's flexibility by analyzing promoter methylation as well as gene expression assays of complex diseases: Huntington's disease and colorectal cancer, respectively. Here, we identify biologically sound key pathways that highly overlap with known disease-related genes (literature research). Our KeyPathwayMiner implementation uses a combination of fixed-parameter, approximation and heuristic algorithms for tackling the underlying NP-hard problem. It is available as a Cytoscape plugin and has been downloaded and installed ~900 times since its first release in Oct. 2011 (~5x per day). Availability: http://keypathwayminer.mpi-inf.mpg.de

1 Introduction and Overview

While combining networks with OMICS data (known as network enrichment, for instance) is a long-standing problem in computational biology, little attention has been paid to interpretability of the results. We usually seek to identify a densely connected sub-graph in a given PPI network that is highly expressed in a given OMICS data set (typically a transcriptomics study). For complex diseases, such as cancer, gold standard data doesn't exist, i.e. known key pathways with many relevant genes, such that setting the parameters, thresholds, etc. for the underlying combined statistics is tricky and still unsolved. When computing such statistics, we need at least one such parameter that balances network density and correlation in the expression data, even when we neglect modeling the noise

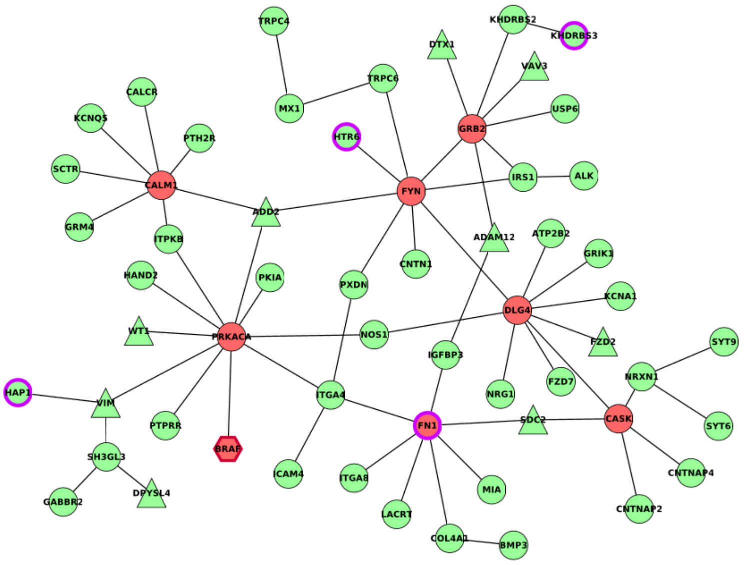


Figure 1: Largest subnetwork found containing the *BRAF* gene for K=8 and L=25. Red nodes represent exception nodes, triangle nodes are hypermethylated genes that also show significant decrease in gene expression levels, and nodes with a purple border are genes with promoters classified as CIMP.

levels in the two data types. We circumvent this problem by providing the end user with an easy-to-interpret model that asks for two parameters with a strong real-world meaning: K and L. KeyPathwayMiner computes all maximal connected sub-networks where all genes but K are expressed/differentially expressed/methylated/active/etc. in all patients but at most L. For the colorectal cancer data set, for instance, we find a 58-genes-key pathway (Figure 1) in the human interactome (approx. 10k proteins, 40k interactions) where all genes but K=8 have a hypermethylation event in the promoter in all 128 patients but at most L=25. In another example we studied Huntington's disease with gene expression data (see Figure 2 for the corresponding key pathway). We applied KeyPathwayMiner to many more data sets and compared it to similar tools obtaining equal or better results (see [AFK$^+$12, BFK$^+$12]). Since our first publication in Oct. 2011, the community downloaded and installed the Cytoscape plugin ~900x (~5x per day).

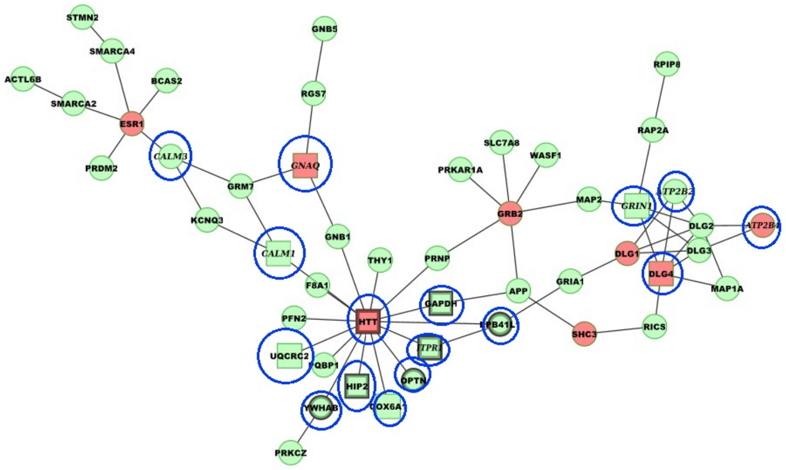


Figure 2: Huntington's disease (HD) key pathway. Here, our KeyPathwayMiner Cytoscape plug-in also used the human interactome network and genome-wide gene expression studies for 38 HD patients (and 32 healthy persons in control group) as input. The illustrated network is the maximal connected sub-network where all genes/proteins but K=8 are differentially expressed in all 38 HD patients but L=6. Red nodes represent exception genes, nodes with blue circles are genes known to be HD-related (from literature).

2 Methods and Model Summary

We provide two slightly varying models for the above introduced problem of finding key pathways:

1. INES: For all genes that have been measured in the case-control study, the profile over all cases is attached. All genes that are not dysregulated in all cases but L are considered "exception genes". We find all maximal, connected sub-networks containing at most K such "exception-genes".

2. GLONE: This is a slightly modified, alternative model. Now we identify all maximal, connected sub-networks where all but at most K nodes are expressed in all cases but in total (!) at most L, i.e. accumulated over all cases and all nodes in a solution. While INES tends to prefer solutions with many hub nodes as exception genes, GLONE circumvents this potential drawback (see [BFK$^+$12]).

Since the underlying optimization problems are computationally hard, we developed a set of three different algorithmic strategies: an exact fixed-parameter algorithm (INES only, fast for $K < 3$), a greedy approximation (INES only, fast but less accurate for higher values of K and L), as well as two Ant Colony Optimization schemes (INES and GLONE, fast and accurate for medium to high values of K, generally accurate for all tested K and L values).

3 Conclusion

Overall, KeyPathwayMiner tackles the problem of finding biomedically relevant pathways by directly combining biological networks with different types of OMICS data. In contrast to existing methods, we ensure interpretability and usability while still being robust, accurate and fast on real world application cases. For details, please refer to the three corresponding papers: [AKW+11, AFK+12, BFK+12]

References

[AFK+12] Nicolas Alcaraz, Tobias Friedrich, Timo Kötzing, Anton Krohmer, Joachim Mueller, Josch Pauling, and Jan Baumbach. Efficient key pathway mining - Combining networks and OMICS data. *Integr Biol*, 4(7):756–764, 2012.

[AKW+11] Nicolas Alcaraz, Hande Kucuk, Jochen Weile, Anil Wipat, and Jan Baumbach. KeyPathwayMiner - Detecting case-specific biological pathways using expression data. *Internet Mathematics*, 7(4):299–313, 2011.

[BFK+12] Jan Baumbach, Tobias Friedrich, Timo Kötzing, Anton Krohmer, Joachim Müller, and Josch Pauling. Efficient algorithms for extracting biological key pathways with global constraints. In *Proceedings of the fourteenth international conference on Genetic and evolutionary computation conference*, GECCO '12, pages 169–176, New York, NY, USA, 2012. ACM.

MetaboLights: Towards a new COSMOS of metabolomics data management

Kenneth Haug[1], Reza M. Salek[1,2], Pablo Conesa[1], Paula de Matos[1], Eamonn Maguire[3], Tejasvi Mahendraker[1], Philippe Rocca-Serra[3], Susanna-Assunta Sansone[3], Julian L. Griffin[2] and Christoph Steinbeck[1,*]
[1]*European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD;* [2]*Elsie Widdowson Laboratory, Fulbourn Road, Cambridge, CB1 9NL, UK, University of Cambridge, Department of Biochemistry, Cambridge CB2 1QW ;* [3]*Oxford e-Research Centre, University of Oxford, Oxford, UK.*
[*]steinbeck@ebi.ac.uk

Abstract: Exciting funding initiatives are emerging in Europe and the US for metabolomics data production, storage, dissemination and analysis. This is based on a rich ecosystem of resources around the world, which has been build during the past ten year, including but not limited to resources such as MassBank in Japan and the Human Metabolome Database (HMDB) in Canada. Now, the European Bioinformatics Institute (EBI) has launched MetaboLights a database for metabolomics experiments and the associated metadata (http://www.ebi.ac.uk/metabolights). It is the first comprehensive, cross-species, cross-platform metabolomics database maintained by one of the major open access data providers in molecular biology. In October, the European COSMOS consortium will start its work on Metabolomics data standardization, publication and dissemination workflows. The NIH in the US is establishing 6-8 metabolomics services cores as well as a national metabolomics repository. This paper reports about MetaboLights as a new resource for Metabolomics research, summarises the related developments and outlines how they may consolidate the knowledge management in this third large omics field next to proteomics and genomics.

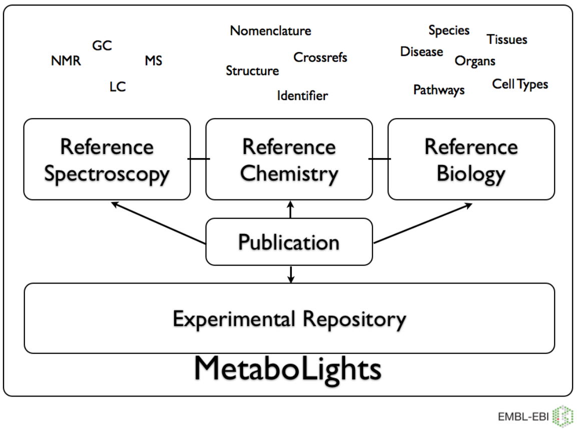
1 Introduction

Metabolomics has become an important phenotyping technique for molecular biology and medicine. It assesses the molecular state of an organism or collections of organisms through the comprehensive quantitative and qualitative analysis of all small molecules in cells, tissues, and body fluids. Metabolic processes are at the core of physiology. Consequently, metabolomics is ideally suited as a medical tool to characterize disease states in organisms, as a tool to assessment of organism for their suitability in, for example, renewable energy production or for biotechnological applications in general. In addition application of metabolomics in environmental science, toxicology, food and medical industry is well established, growing and documented. Metabolomics studies generate large amounts of analytical data (Giga- to Terabytes depending on the size of the study) and therefore impose significant challenges for biomedical and life science e-infrastructures to cope with such data volumes and ensure that the data is captured, stored and disseminated based on open and widely accepted community standards. Years after the first standardisation exercises [FRGea07, TFSea08], metabolomics is now reaching the state of a mature analytical technique as indicated by the establishment of 6-8 Regional Comprehensive Metabolomics Resource Cores (RCMRCs) by the NIH in the United States. In addition, we are now facing a rich ecosystem of specialised metabolomics databases as well as the first general metabolomics repositories and databases emerging. In Europe, the COSMOS consortium of 14 leading laboratories in metabolomics will begin its work on standards, data management and dissemination in Metabolomics. Here, we outline these developments and show how they may consolidate the knowledge management in this third large omics field next to proteomics and genomics.

2 MetaboLights – A cross-species repository for metabolomics experiments

The European Bioinformatics Institute (EBI) has recently launched MetaboLights, a database for metabolomics experiments and the associated metadata. It is the first comprehensive, cross-species, cross-platform metabolomics database maintained by one of the major open access data providers in molecular biology. MetaboLights lives at http://www.ebi.-ac.uk/metabolights. For their convenience, users can also use metabolights.org, metabolights.net and metabolights.eu. The EBI ensures long-term stability and maintenance of the resource. Like all other EBI resources, the MetaboLights database is completely open to the public, including open access to the data. Data are made available in publicly accepted open standards compliance with MIBBI (The Minimum Information for Biological and Biomedical Investigations) [TFSea08]. The software is open source and adheres to the promotion of open source file formats, such as mzML and nmrML. One of the main submission channels for MetaboLights' use is the ISA Tools Suite [SRSFea12]. MetaboLights is not intended to replace specialist resources for Metabolomics. Rather, it will build on prior art and collaborate. We are dedicated to close collaboration with all major parties involved in the creation of this prior art, such as the Metabolomics Society, Metabomeeting and the Metabolomics Standards Initiative (MSI). MetaboLights aims to agree on formal data sharing agreements with major resources such as the Human Metabolome Database, the Golm Metabolome Database and the Rikken Metabolomics Platform. Currently we house selection of experimental raw data and their associated metadata for different platforms such as NMR, GC-MS and LC-MS.

Figure 1: MetaboLights general outline



3 Outlook

In October, the European COSMOS (COordination of Standards in MetabOlomicS) consortium will start its work on Metabolomics data standardization, publication and dissemination workflows. It is the aim of COSMOS to develop efficient policies to ensure that Metabolomics data is

1. Encoded in open standards to allow barrier-free and widespread analysis.

2. Tagged with a community-agreed, complete set of metadata (mini-

mum information standard).

3. Supported by a communally developed set of open source data management and capturing tools.
4. Disseminated in open-access databases adhering to the above standards.
5. Supported by vendors and publishers, who require deposition upon publication
6. Properly interfaced with data in other biomedical and life science e-infrastructures, such as
    - ELIXIR (http://www.elixir-europe.org/),
    - BioMedBridges (http://www.biomedbridges.eu/),
    - EU-OPENSCREEN (http://www.eu-openscreen.de/) and
    - BBMRI (http://www.bbmri.eu/).

During 2012, MetaboLights' repository layer will be expanded by a reference layer with chemical, spectroscopic and biological reference information about individual metabolites (Figure 1).

The NIH in the US is establishing 6-8 metabolomics services cores as well as a national metabolomics repository. Together with similar initiatives in Australia, Japan and hopefully more emerging over time, this opens the door for a global network of metabolomics data collection, exchange and dissemination.

References

[FRGea07] Oliver Fiehn, Don Robertson, Jules Griffin, and et al. The metabolomics standards initiative (MSI). *Metabolomics*, 3(3):175–178, 2007.

[SRSFea12] Susanna-Assunta Sansone, Philippe Rocca-Serra, Dawn Field, and et al. Toward interoperable bioscience data. *Nat Genet*, 44(2):121–126, January 2012.

[TFSea08] Chris F Taylor, Dawn Field, Susanna-Assunta Sansone, and et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature Biotechnology*, 26(8):889–896, 2008.

How Little Do We Actually Know? – On the Size of Gene Regulatory Networks.

Richard Röttger, Jan Baumbach
Max Planck Institute for Informatics, Saarbrücken, Germany

roettger@mpi-inf.mpg.de

Abstract: Nowadays, we have whole-genome sequences for more than more than two thousand species available for download from the NCBI databases. Ongoing improvement of DNA sequencing technology will further feed this trend. However, the availability of sequence information is only the first step in understanding how cells survive, reproduce and adjust their behavior. The molecular biological mechanisms, which control organized development and adaptation of complex organisms still remain widely undetermined. Transcriptional gene regulation is one of the key players here. The direct juxtaposition of the total number of sequenced species to the handful of model organisms with known regulations is astonishing. Recently, we investigated how little we even know about these model organisms. Our aim was to predict the sizes of the whole-organism regulatory networks of seven species. In particular, we provided a statistical lower bound for the expected number of regulations. For *E. coli* we estimate at most 37% of the expected gene regulatory interactions to be already discovered, 24% for *B. subtilis*, and < 3% for human respectively. We conclude that even for our best-researched model organisms we still lack substantial understanding of fundamental molecular control mechanisms, at least on a large scale.

Pubmed: http://www.ncbi.nlm.nih.gov/pubmed/22585140
Web: http://dx.doi.org/10.1109/TCBB.2012.71

1 Introduction

In 2010, whole-genome sequences of more than 1200 microbes, 3600 viruses and 39 eukaryotic species, as well as over 2400 sequences for eukaryotic organelles were available for download at the web site of the National Center for Biotechnology Information, NCBI [ea11]. The continuously improving next-generation sequencing techniques will further increase the number of sequenced species dramatically. The heavily reduced cost for DNA sequencing by over two orders of magnitude allows us to utilize

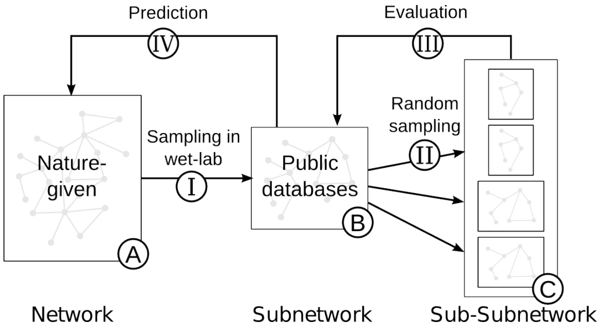


Figure 1: Overview: We aim to infer the size of the nature-given transcriptional gene regulatory network (A) of a specific organism. Given is a known subnetwork (B), which was sampled by biologists in laboratories and stored in public databases (I). In order to study the reliability of our prediction, we randomly sample (II) even smaller networks, subsubnetworks (C) and subsequently test if we can predict the known network size and evaluate the robustness of the approach (III). After verifying the estimator to be bias-free and robust, we are able to predict (IV) the size of the nature-given network (A).

this technology as routine method for unraveling the genetic repertoire of numerous species of varying complexity and ecological, economic and medical significance [Met10].

However, the availability of genome sequences, even complete ones, is only the first step towards a comprehensive understanding of how cells survive, reproduce and adapt to changing environmental conditions. One major molecular control mechanism of cells is transcriptional gene regulation. Key players are the so-called transcription factors (TFs), proteins that possess DNA-docking domains to bind certain regions within the DNA sequence. Thereby they influence the expression of numerous target genes (TGs) and control genetic programs like growth, survival, reproduction, digestion, immune responses, etc. Transcriptional gene regulatory networks (GRN) emerge, with nodes corresponding to genes and directed edges that represent regulatory interactions [PS92].

Understanding this fundamental molecular control mechanism on a large scale is one of the most important goals in systems biology and a prerequisite for the subsequent modeling and analysis of cell response and behavior [BWKT09]. However, our current knowledge is limited and the

Organism	$\mathbf{E^S}$	$\widehat{\mathbf{E}}^{\mathbf{N}}$	**95% Lo.**	**Ratio**	**Ratio CI**
H. sapiens	3,902	165,807	130,326	2.35%	2.99%
M. musculus	1,730	190,359	157,251	0.91%	1.08%
R. norvegicus	804	421,819	313,890	0.19%	0.26%
*A. thaliana**	1,852	569,013	111,026	0.31%	1.67%
E. coli	3,946	15,399	10,562	25.63%	37.36%
B. subtilis	1,391	9,716	5,780	14.31%	24.07%
C. glutamicum	806	9,114	5,696	8.84%	14.15%

*this is a slightly modified dataset. Please refer to the main paper for an exhaustive discussion.

Table 1: The table shows a summary of our results for four species. Most important are the last columns Ratio and Ratio CI, which give the fraction of known regulatory interactions in relation to the predicted nature-given network size $\widehat{E}^N$, i.e. an estimation of how much we have discovered yet. In the column Ratio, we use the estimation $\widehat{E}^N$ for the comparison, whereas in Column Ratio CI the lower bound of the 95% confidence interval was used (95% Lo.). The column E^S gives the number of known regulations for each species.

reconstruction of the gene regulatory networks is far from being complete. Even for *E. coli*, the model organism with the largest currently available experimentally validated data for any free-living organism, we have information about the transcriptional regulation of only around one third of the genes [S. 08].

With this short highlight paper, we briefly describe (1) a robust model to estimate the expected size of transcriptional gene regulatory networks, i.e. the number of edges, and (2) conclude that we actually know very little about one of the most important mechanisms that controls genetic activity.

2 Model and Results

Our estimator model essentially works on graph invariants. GRNs are directed graphs with two types of nodes (genes): the transcription factors (TFs) and the target genes (TGs). With our network size prediction model we account for the two types of nodes but also for three types of regulatory interactions (edge types): (TF→TF) regulations between two transcription factors, (TF→TG) regulations between a transcription factor and target gene and (TF-self) self regulations of transcription factors.

Subsequently, for each species with a partially known GRN, we calculated the three edge probabilities. As these probabilities form graph invariants, we can use them to estimate the total size of the GRN. Two challenges arise: (1) We lack a gold standard, i.e. a fully known GRN for at least one organism, that we could use for validating our model. (2) We usually have only one reference database per species that describes the known GRN of this organism. Thus assessing the variability of our method is not straight forward. Figure 1 gives an overview about the utilized methodology. For robustness assessment we use bootstrapping methods to receive confidence intervals for our estimations. Finally, we were interested in the lower bounds, i.e. how much of an organism's GRN can we expect to know in the very best case. We applied our model to seven species, among them human, mouse, *E. coli* and *B. subtilis*. See Table 1 for our major results for these four species. We found it quite astonishing that even for *E. coli* only 37% of the expected GRN is known, in the best case. For mammals our current knowledge is too limited to even give trustworthy estimations.

For formal definitions and exhaustive description of the mentioned methods, please refer to the main paper's method section [RRTB12].

References

[BWKT09] J. Baumbach, T. Wittkop, C.K. Kleindt, and A. Tauch. Integrated analysis and reconstruction of microbial transcriptional gene regulatory networks using CoryneRegNet. *Nature Protocols*, 4(6):992–1005, 2009.

[ea11] E. W. Sayers et al. Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, 39(suppl 1):D38–D51, 2011.

[Met10] M. L. Metzker. Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11(1):31–46, 2010.

[PS92] C.O. Pabo and R.T. Sauer. Transcription factors: structural families and principles of DNA recognition. *Annual Review of Biochemistry*, 61(1):1053–1095, 1992.

[RRTB12] R. Rottger, U. Ruckert, J. Taubert, and J. Baumbach. How Little Do We Actually Know? – On the Size of Gene Regulatory Networks. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, PP(99):1, 2012.

[S. 08] S. Gama-Castro et al. RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Research*, 36(Database issue):D120–4, 2008.

Out of (transcriptional) control? Design principles of the regulatory network controlling metabolic pathways in Escherichia coli

Frank Wessely[1,2,†], Martin Bartl[3], Reinhard Guthke[4], Pu Li[3],
and Christoph Kaleta[1,2s]

[1] *Research Group Theoretical Systems Biology, Friedrich Schiller University Jena, Jena, Germany,* [2] *Department of Bioinformatics, Friedrich Schiller University Jena, Jena, Germany,* [3] *Institute for Automation and Systems Engineering, Ilmenau University of Technology, Ilmenau, Germany,* [4] *Systems Biology/Bioinformatics Group, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Jena, Germany,* [†] *present address: School of Veterinary Medicine and Science, University of Nottingham, Sutton Bonington Campus, Loughborough, UK*

*christoph.kaleta@uni-jena.de

Abstract: While a large number of previous studies has explored the link between the structure of metabolism and its regulation, the extent to which transcriptional regulation controls metabolism has not yet been fully elucidated. We address this problem by integrating a large number of experimental data sets with a genome-scale metabolic model of *Escherichia coli* metabolism. We find that there is a strong connection between the extent of transcriptional regulation in a metabolic pathway and the protein investment into this pathway. While pathways associated to a low protein cost tend to be controlled only in key steps, pathways associated to a high protein cost are controlled by fine-tuned transcriptional regulatory programs. These different strategies for the control of metabolic pathways can be explained by a trade-off between the conflicting requirements to minimize protein investment and to maintain the ability to quickly respond to changes in environmental conditions.

1 Introduction

The increasing availability and decreasing prices of experimental techniques have led to an explosion in the number of available experimental data sets [LVW+07, BKG+09]. These data sets provide an increasingly comprehensive view on the principles that influence the evolution of the regulatory network controlling metabolism [NTSP08]. In this submission we discuss the results of a previous work [WBG$^+$11], in which we have used different types of OMICs data sets in order to identify these global principles of regulatory network evolution in the model organism *Escherichia coli*.

2 Results

In order to understand to which extent transcriptional regulation controls metabolism, we investigated the coexpression of enzymes within the pathways of all biochemically annotated subsystems of *E. coli* metabolism. This analysis was based on the concept of elementary flux patterns [KdFS09], which allowed us to identify pathways in all subsystems of metabolism. By mapping gene expression data to the corresponding pathways, we found that pathways in many subsystems of metabolism show a large degree of coexpression. However, pathways in the subsystems cofactor and prosthetic group biosynthesis, glycerophospholipid metabolism, murein recycling, nucleotide salvage pathway and pentose phosphate pathway show only weak coexpression of pathways. We call these subsystems with a low coexpression of pathways "transcriptionally sparsely regulated subsystems".

To provide an explanation for these distinct patterns of transcriptional regulation, we constructed a simplified model of a linear metabolic pathway that converts a substrate s via four intermediates into a product p. Dynamic optimization was used to identify specific regulatory programs (representing time-courses of enzyme concentrations) that allow the cell to precisely adjust the concentration of the product in a changing environment while obeying a set of physiological constraints. As objective function we used the minimization of the change of enzyme concentrations from initial concentrations and protein costs.

The results of this optimization procedure showed that for a full control of flux through a pathway, transcriptional regulation of initial and terminal positions of a pathway is sufficient (sparse transcriptional regulation). The role of the control of the first enzyme of a pathway is to regulate the flux into the pathway and avoid the accumulation of intermediates. In contrast, the control of the terminal reaction of a pathway allows the cell to precisely adjust the rate of synthesis of the product. Performing the same optimization for a large number of pathways with randomized kinetic parameters, we found that these principles hold true regardless of kinetic parameters. Moreover, we found that with increasing cost of enzymes of a pathway (i.e. increasing enzyme concentrations) there is a shift from the sparse transcriptional regulation of a metabolic pathway to the coordinated transcriptional control of all enzymes in a pathway (pervasive transcriptional regulation).

We validated these predictions by an analysis of the position-specific frequency of regulatory events in the pathways of transcriptionally sparsely regulated subsystems. We confirmed that there is a significant increase in the frequency of transcriptional regulation at the beginning and end of pathways. Moreover, we found a significant increase of the frequency of post-translational regulation at the beginning of pathways. Thus, the control at the initial positions of pathways is achieved through a combination of transcriptional as well as post-translational regulation, while control at the end of pathways is achieved through transcriptional regulation. In other subsystems that were not identified as being transcriptionally sparsely regulated by the expression analysis, we did not find this pattern of transcriptional regulation, while the pattern of post-translational regulation prevailed. Investigating data of protein costs (defined as the total mass of a particular protein in the cell) for different subsystems we found that in particular subsystems with a small cost of proteins show a pattern of transcriptional sparse regulation.

3 Discussion

Since we were able to confirm the predictions of the optimization, there appears to be an evolutionary mechanism favoring sparse transcriptional regulation in pathways with low-cost enzymes. We propose an evolutionary trade-off between the two conflicting objectives of the minimization of protein investment and the minimization of response time. The optimal strategy to reduce protein investment is to transcriptionally control proteins and express them only if they are needed. However, response times on a transcriptional level are usually very slow. Optimal response times can be achieved through a constitutive expression of most enzymes in a pathway and a transcriptional control of key steps. The interplay between both objectives results in a pervasive transcriptional control of all enzymes within a pathway if they are associated to a high cost. In pathways with low-cost enzymes, transcriptionally sparse regulation prevails. In support of these results, we found that even costly pathways such as the pentose phosphate pathway, for which rapid response times are required, are sparsely regulated due to a strong advantage of a rapid response time. Finally, if there is only a small fitness advantage of both cellular objectives, sparse transcriptional regulation is a minimum requirement to precisely control the flux through a pathway.

These results demonstrate that, in contrast to the classical picture of regulation, the control of key positions of metabolic pathways is sufficient to achieve a full control over the flux through a pathway. Such a pattern of sparse transcriptional regulation is useful if a higher fitness advantage can be achieved through rapid response times in comparison to the fitness advantage of a reduced protein cost.

References

[BKG$^+$09] Bryson D Bennett, Elizabeth H Kimball, Melissa Gao, Robin Osterhout, Stephen J Van Dien, and Joshua D Rabinowitz. Absolute metabolite concentrations and implied enzyme active site occupancy in *Escherichia coli*. *Nat Chem Biol*, 5(8):593–599, Aug 2009.

[KdFS09] Christoph Kaleta, Luís Filipe de Figueiredo, and Stefan Schuster. Can the whole be less than the sum of its parts? Pathway Analysis in genome-scale metabolic networks using Elementary Flux Patterns. *Genome Res*, 19(10):1872–1883, Oct 2009.

[LVW$^+$07] Peng Lu, Christine Vogel, Rong Wang, Xin Yao, and Edward M Marcotte. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol*, 25(1):117–124, Jan 2007.

[NTSP08] Richard A Notebaart, Bas Teusink, Roland J Siezen, and Balázs Papp. Co-regulation of metabolic genes is better explained by flux coupling than by network distance. *PLoS Comput Biol*, 4(1):e26, Jan 2008.

[WBG$^+$11] Frank Wessely, Martin Bartl, Reinhard Guthke, Pu Li, Stefan Schuster, and Christoph Kaleta. Optimal regulatory strategies for metabolic pathways in *Escherichia coli* depending on protein costs. *Mol Syst Biol*, 7:515, 2011.

Readjoiner: a fast and memory efficient string graph-based sequence assembler

Giorgio Gonnella and Stefan Kurtz
Center for Bioinformatics (ZBH), University of Hamburg

gonnella@zbh.uni-hamburg.de

In a recently published paper [GK12] we describe a new fast and memory efficient string graph-based sequence assembler: *Readjoiner*. In this extended abstract, we summarize the background, methods and results.

Background

The amount of data delivered by next-generation DNA sequencing technologies challenges the current generation of *de novo* sequence assemblers based on De Bruijn graphs.

An alternative framework of growing interest is the assembly string graph [Mye05]. As the classical overlap graph, the string graph represents sequencing reads by vertices and overlaps between reads by edges: however, in the string graph only *irreducible* suffix-prefix matches are considered.

The string graph combines the strengths of the classical overlap-layout-consensus paradigma with a compact representation suitable for the next-generation sequencing datasets. The main advantage over the De Bruijn graph is that it does not require to artificially split the reads into k-mers, thus improving the assembly of sequences containing short repeats. Furthermore, the string graph is more compact than the De Bruijn graph, thus allowing to efficiently handle larger datasets.

To construct the string graph, fast and space efficient algorithms for the computation of all suffix-prefix matches are required. Previous approaches use a suffix array (Edena, [HFF$^+$08]) or an FM-index (SGA, [SD12]) or a compact representation of the overlap graph (Leap, [DR11]).

Methods

We developed efficient methods for the construction of a string graph from a set of sequencing reads. We use suffix sorting and scanning methods to compute suffix-prefix matches; furthermore, transitive edges are recognized early in the process and excluded from the graph.

The first step of our assembly approach is to eliminate reads that are prefixes or suffixes of other reads: these are recognized by lexicographically sorting all reads and their reverse complements, using a modified radixsort for strings [KR09].

In the following step, suffix-prefix matches longer than ℓ_{min} are computed, where ℓ_{min} is an user-defined parameter. The method consists of two main algorithms. The first algorithm identifies and lexicographically sorts all *SPM-relevant suffixes*: these are suffixes of reads, sharing a prefix of length $k \leq \ell_{min}$ with some read in the readset. Here k is a parameter allowing for time / space tradeoffs in the computation. The second algorithm enumerates the suffix-prefix matches given a sorted list of SPM-relevant suffixes.

SPM-relevant suffixes are sorted using a strategy borrowed from the counting sort algorithm [CLR90]. An efficient solution is achieved by combining the use of sorted buffers for the elements to be counted/inserted, a filter based on substrings of the initial k-mers of the reads and a partitioning strategy considerably reducing the space peak of the implementation.

The suffix-prefix matches are computed using an algorithm based on a bottom-up traversal of the lcp-inverval tree. This is obtained by processing the buckets of SPM-relevant suffixes with a variant of the algorithm presented in [AKO04], additionally delivering the leaf edges of the virtual lcp-interval tree.

In order to only output irreducible suffix-prefix matches, we maintain an additional trie-data structure and exploit a novel characterization of transitive suffix-prefix matches.

The assembly string graph is constructed from the list of all irreducible suffix-prefix matches, as described in [Mye05]. Heuristically, bubbles and short dead-end paths likely arising from sequencing errors, are optionally removed from the graph. Finally, the sequence corresponding to all unbranched paths in the graph is output as a collection of contigs.

Results and Conclusion

We implemented our methods in a new open source sequence assembler, called *Readjoiner*, as part of the *GenomeTools* [GEN] genome analysis suite. Readjoiner is freely available at `http://www.zbh.uni-hamburg.de/readjoiner`.

We extensively evaluated our assembler on simulated error-free sequencing read sets based on human genomic sequences. We compared the performance of *Readjoiner* with that of the previous string graph-based tools: Edena [HFF$^+$08], SGA [SD12] and Leap [DR11]. The results were evaluated using metrics developed by the Assemblathon project [EBSJ$^+$11] and using the Plantagora assessment tool [BMRY11].

Our tests show that *Readjoiner* is faster and more space efficient than previous string graph-based tools. *Readjoiner* was $13 - 14\times$ faster than Edena, $19 - 20\times$ faster than SGA and $1.6 - 1.8\times$ faster than LEAP. Furthermore it uses about $9.1 - 9.3\times$ less memory than Edena, $1.1 - 1.2\times$ less memory than SGA and $1.6 - 3.0\times$ less memory than LEAP. Furthermore, it scales well for large datasets. For example, a 40 × coverage human genome dataset (100 nt reads for a total of 115 Gb) can be assembled on a single core in 51 hours using 52 Gb RAM.

Readjoiner is actively developed and improvements over the version described in [GK12] have been achieved. For example, suffix-prefix matches derived from independent parts of our data structures are computed in threads. We plan to integrate an error correction algorithm and incorporate mate pairs information during the assembly phase.

We would like to remark that our paper, published less than a month ago, attracts considerable interest: besides acquiring the “Highly accessed” designation by the publisher (BioMed Central), the paper was, as of May 31th, the most viewed paper for BMC Bioinformatics during May 2012 [BMC12].

References

[AKO04] M.I. Abouelhoda, S. Kurtz, and E. Ohlebusch. Replacing Suffix Trees with Enhanced Suffix Arrays. *Journal of Discrete Algorithms*, 2:53–86, 2004.

[BMC12] BMC Bioinformatics, Most Viewed, Last 30 days. http//www.biomedcentral.com/bmcbioinformatics/mostviewed, accessed on May 31st, 2012.

[BMRY11] Roger Barthelson, Adam J. McFarlin, Steven D. Rounsley, and Sarah Young. Plantagora: Modeling Whole Genome Sequencing and Assembly of Plant Genomes. *PLoS ONE*, 6(12):e28436, 12 2011.

[CLR90] T.H. Cormen, C.E. Leiserson, and R.L. Rivest. *Introduction to Algorithms*. MIT Press, Cambridge, MA, 1990.

[DR11] Hieu Dinh and Sanguthevar Rajasekaran. A memory-efficient data structure representing exact-match overlap graphs with application for next-generation DNA assembly. *Bioinformatics*, 27(14):1901–1907, Jul 2011.

[EBSJ$^+$11] Dent A. Earl, Keith Bradnam, John St. John, Aaron Darling, Dawei Lin, Joseph Faas, Hung On Ken Yu, Buffalo Vince, et al. Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Research*, 2011.

[GEN] GenomeTools - The versatile open source genome analysis software. `http://genometools.org`.

[GK12] Giorgio Gonnella and Stefan Kurtz. Readjoiner: a fast and memory efficient string graph-based sequence assembler. *BMC Bioinformatics*, 13(1):82, 2012.

[HFF$^+$08] David Hernandez, Patrice Franois, Laurent Farinelli, Magne Osters, and Jacques Schrenzel. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res*, 18(5):802–809, May 2008.

[KR09] Juha Kärkkäinen and Tommi Rantala. Engineering Radix Sort for Strings. In Amihood Amir, Andrew Turpin, and Alistair Moffat, editors, *String Processing and Information Retrieval*, volume 5280 of *Lecture Notes in Computer Science*, pages 3–14. Springer Berlin / Heidelberg, 2009.

[Mye05] E. W. Myers. The fragment assembly string graph. *Bioinformatics*, 21 Suppl 2:79–85, Sep 2005.

[SD12] Jared T Simpson and Richard Durbin. Efficient de novo assembly of large genomes using compressed data structures. *Genome Research*, 22(3):549–556, 2012.

Footprints of modular evolution in a dense taxonomic clade

Andrew D. Moore & Erich Bornberg-Bauer
Evolutionary Bioinformatics Group, Institute for Evolution and Biodiversity, University of Muenster

radmoore@uni-muenster.de, ebb@uni-muenster.de

Abstract: True novelty, of any form, is rare. Most systems, including a number of biological systems, can be reduced to a set of of core units which are reused in varying contexts. These core units can be seen as modules, and their harboring system as modular. Here, we explore various aspects of modularity in protein evolution within a dense clade of 20 arthropods. By employing a simple model of protein evolution, we study how the rearrangements of domains - the modules of protein evolution, structure and function - creates novelty in few steps and at surprising speeds. We find that we can explain between 64% - 81% of all novel protein domain arrangements, and that arrangements that cannot be explained contain curious patterns of domain repeats. Furthermore, we explore the speed of module turnover - the frequency of domain gain and loss - and find that while only few new domains occur, they spread swiftly and seem associated with environmental adaptation.

1 Introduction

A primary factor in the evolution of proteins is the rearranging of protein domains, their functional, structural and evolutionary modules. Using modular rearrangement, functional diversification can occur without the formation of novel domains, simply by adding, removing or rearranging domains in proteins [MBE$^+$08]. Previous studies have illustrated that, in particular along the metazoan lineage, increased rates of domain rearrangement can be found [MBE$^+$08]. Indeed, while the number of identified domains grows very slowly, the number of combinations of these domains continues to grow with no end in sight [Lev09].

As opposed to the often slow variation at the sequence level, events such as gene fusion/fission or the shuffling of exons, which are among the genetic protagonists driving modular domain recombination [BFB10], can swiftly produce selectable phenotypes [RH12, PGWL10]. While a series

of mutations can govern selectable phenotypes, a number of mutations remain unseen to the eye of selection. In contrast, large events such as the fusion of two genes is likely to produce a phenotype, some of which may even be favored by selection [RH12]. Autonomously functioning domains used in a modular system, where functionalities can be recombined easily, provide a powerful mechanism for evolutionary innovation.
From numerous previous studies we know that the dominant mechanisms creating novel arrangements are gene duplication, fusion and terminal losses [WBBB06]; that age, function and structure of a domain do not influence their versatility [WMBB08] and that strings of domains are well suited for designing algorithms for homology search [TGW^{+}12].

While rare, evidence for novelty does exist e.g. in the large number of orphan genes, many of which are presumed to be vital for species-specific quirks [KHF^{+}09]. Beyond genes, changes in domain content between species, and species groups, can be observed [ZG11]. This indicates that novel domains do emerge - albeit at low frequency. It seems plausible that certain molecular innovation, such as required in the wake of strong environmental shifts, may be out of reach by the rearrangement of existing domains alone and may require the emergence of novel domains.

We have recently explored various aspects of modular evolution using a small, well described clade of 20 arthropods. In this data set, we have derived branch-specific rates of events in modular evolution and have assessed the evolutionary dynamics and functional impact of changes at the level of the domain repertoire. Beyond the exploration of various aspects of protein evolution, our approach illustrates the strength of domain-based analysis: the great accuracy of HMMs in identifying homologous sequences and the low rates of domain turnover helps capture functional shifts and evolutionary dynamics at a rather coarse grained level and across evolutionary long time scales of tens to hundreds of million years.

2 Results

Within the arthropods, a total of 30 domain are found to be emergent (that is, occur only within this clade) [MBB12]. By functionally annotating all proteins which harbor an emerging domain (1,291 proteins across 20 arthropods), we assessed the functional impact of novel domains. Domains that emerge within arthropods are found significantly more often in terms related to environmental adaptation (e.g. response to heat, drought, UV and other abiotic stresses), than expected by chance (see figure 1).

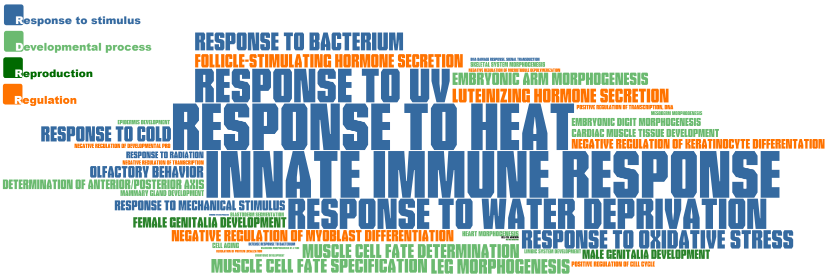


Figure 1: **TermLogo of functional groups with emerging domains.** Overrepresentation analysis of Gene Ontology terms from proteins which contain at least one emerging domain. The size of the font corresponds to the strength of obtained significance.

The majority of arrangements are unique to one, or very few species (Figure 2). The majority of arrangements are unique to one, or very few species, facilitating a roughly bimodal distribution of shared arrangements. This indicates that modular rearrangement is frequent enough to create a large diversity of arrangements, even in evolutionarily small timescales. Furthermore, while the largest proportion (∼80%) of arrangements shared by all species are single domain proteins, species-specific arrangements tend to be multi-domain indicating that older arrangements tend to be single-domain, while newly formed arrangements are more likely multi-domain.

After ancestral reconstruction of arrangement presence/absence states, we derive rates of arrangement gain for all branches. We then, for each new arrangement, investigate how new arrangements can be formed by recombining ancestral arrangements (e.g a new arrangement (A,B,C) can be formed by the fusion of the ancestral arrangements (A,B) and (C)). We consider the fusion of two arrangements, the fission of an arrangement, as well as the gain or loss of parts of arrangements. We find that we can explain up to 81% of all new arrangements by a single-step event while some new arrangements have conflicting solutions; a total of 64% of all new arrangements have only possible solution.
The evolutionary dynamics of the events are intriguing: while fusion and gain dominate early in the tree, fission and loss frequencies increase over time. A possible interpretation concerns arrangement length: recombination events that give rise to novel (viable) arrangements are likely to act between domains as to not disrupt functional domains. The smaller the number of domains that are present in an arrangement, the lower the

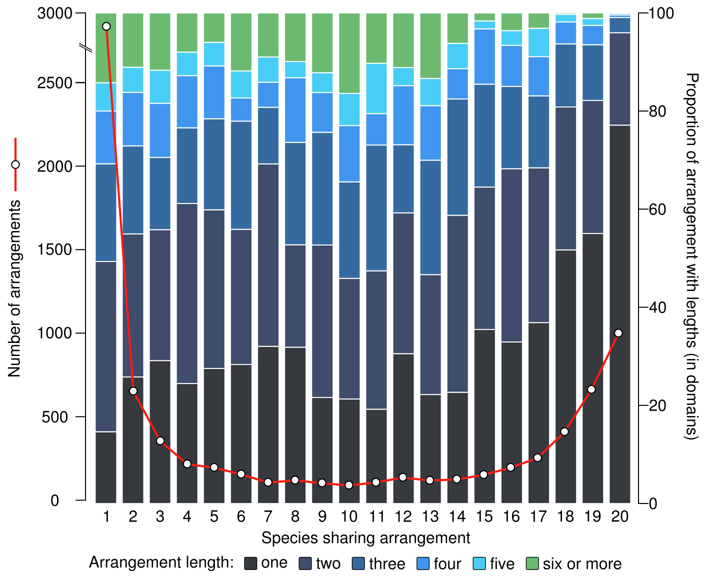


Figure 2: **Unique arrangements and arrangement length in 20 pancrustacean species.** Unique arrangements were grouped by the number of species in which they can be found. The x-axis indicates the number of species which share arrangements, the y-axis indicates the number of arrangements. For each group of shared arrangements, the arrangement length measured as the number of domains was determined and normalized to 100% (z-axis). The red line plot illustrates that the distribution of unique arrangements is roughly bimodal, with the majority of arrangements shared by either few or all species.

chance for successful fission or loss. In contrast, fusion and gain events seem more likely detrimental the longer an arrangement gets.

New arrangements that cannot be explained by one of the considered events contain complex, multi-domain repeat patterns ("supra-repeats") and are significantly enriched in domain-repeats. Such domain-repeats are essential to protein-protein interaction and DNA-binding making them key players in regulatory networks. Beyond the analysis of arthropods, we find that the overall signals in plant species are similar [KBMG12].

In summary, our results provide a detailed account of the mechanisms with which domain rearrangement events create novel proteins, and provide an excellent starting point for further analysis ranging from mathematical modeling to additional cross-species comparisons.

References

[BFB10] Marija Buljan, Adam Frankish, and Alex Bateman. Quantifying the mechanisms of domain gain in animal proteins. *Genome Biol*, 11(7):R74, 2010.

[KBMG12] Anna R Kersting, Erich Bornberg Bauer, Andrew D Moore, and Sonja Grath. Dynamics and adaptive benefits of protein domain emergence and arrangements during plant genome evolution. *Genome Biol Evol*, Feb 2012.

[KHF+09] Konstantin Khalturin, Georg Hemmrich, Sebastian Fraune, René Augustin, and Thomas C G Bosch. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet*, 25(9):404–413, Sep 2009.

[Lev09] Michael Levitt. Nature of the protein universe. *Proc Natl Acad Sci U S A*, 106(27):11079–11084, Jul 2009.

[MBB12] Andrew D Moore and Erich Bornberg-Bauer. The dynamics and evolutionary potential of domain loss and emergence. *Mol Biol Evol*, 29(2):787–796, Feb 2012.

[MBE+08] Andrew D. Moore, Åsa K. Björklund, Diana Ekman, Erich Bornberg-Bauer, and Arne Elofsson. Arrangements in the modular evolution of proteins. *Trends Biochem Sci*, 33(9):444–451, Sep 2008.

[PGWL10] Sergio G Peisajovich, Joan E Garbarino, Ping Wei, and Wendell A Lim. Rapid diversification of cell signaling phenotypes by modular domain recombination. *Science*, 328(5976):368–372, Apr 2010.

[RH12] Rebekah L Rogers and Daniel L Hartl. Chimeric Genes as a Source of Rapid Evolution in Drosophila melanogaster. *Mol Biol Evol*, 29(2):517–529, Feb 2012.

[TGW+12] Nicolas Terrapon, Sonja Grath, January Weiner, Andrew Moore, and Erich Bornberg-Bauer. Fast Homology Search Using Domain-Architecture Alignment. *JOBIM Conference proceedings*, 2012.

[WBBB06] January Weiner, Francois Beaussart, and Erich Bornberg-Bauer. Domain deletions and substitutions in the modular protein evolution. *FEBS J*, 273(9):2037–2047, May 2006.

[WMBB08] January Weiner, Andrew D Moore, and Erich Bornberg-Bauer. Just how versatile are domains? *BMC Evol Biol*, 8:285, 2008.

[ZG11] Christian M Zmasek and Adam Godzik. Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. *Genome Biol*, 12(1):R4, Jan 2011.