The Bioinformatics Network Luxembourg

Abstract Book

Methods and Applications in Bioinformatics:
From ideas to results

3rd Annual Symposium
5th November 2010

Centre Hospitalier de Luxembourg
(Auditorium)

Schedule

Timing	Name	Title of Talk	Page
9.00-9.15	Patrick Pirotte	Opening, Introduction	
	Chair: Mario Albrecht, MPI Saarbrücken		
9.15-10.00	Andreas Keller (FEBIT GmbH)	Keynote: Bioinformatics for challenging high-throughput genomics and transcriptomics: the “Iceman’s” last secrets	1
10.00-10.25	Petr V. Nazarov (CRP Santé)	CoExpress: a tool for co-regulation analysis of mRNA and miRNA	2
10.25-10.50	Olivier Keunen (CRP Santé)	Multi-modal imaging and pharmacokinetic modeling shed light on mechanism of action of anti-angiogenic therapy in malignant gliomas	3
10.50-11.15	coffee break		
	Chair: Marie-Dominique Devignes, LORIA		
11.15-12.00	Holger Stenzhorn (Uniklinikum Saarland)	Keynote: Towards personalized medicine – From clinical trial data sharing and integration to targeted treatments	5
12.00-12.25	Fidel Ramírez (MPI Saarbrücken)	BioMyn: A Web Site for Mining Human Gene and Protein Annotations	6
12.25-12.50	Adrien Coulet (LORIA, Nancy)	A computational method to extract pharmacogenomic relationships from text	7
till 14.00	Lunch break		
	Chair: Holger Stenzhorn, Uniklinikum Saarland		
14.00-14.45	Mario Albrecht (MPI Saarbrücken)	Keynote: Mining and Prioritizing Molecular Data and Processes for Integrative Systems Biomedicine	8
14.45-15.10	Francisco Azuaje (CRP Santé)	Linking molecular networks to clinical decision-making support	9
15.10-15.35	Antony Le Béchec (University of Luxembourg)	MIR@NT@N: a framework integrating transcription factors, microRNAs and their targets to identify sub-network motifs in meta-regulation network models	10
15.35-16.00	coffee break		
	Chair: Andreas Keller, FEBIT AG		
16.00-16.40	Marie-Dominique Devignes (LORIA, Nancy)	Keynote: Research in bioinformatics at the LORIA (Nancy) : high-performance algorithms and knowledge-based approaches for structural systems biology	11
16.40-17.05	Marc Lensink (IRI - CNRS USR3078,France)	Predicting Specificity in Protein-Protein Interactions	12
17.05-17.30	Paul Wilmes (CRP Gabriel Lippmann)	From molecules to ecosystems and back : bioinformatic approaches for Eco-Systems Biology	13
17.30- open	Drinks		

KEYNOTE 1

Bioinformatics for challenging high-throughput genomics and transcriptomics: the "Iceman's" last secrets

Andreas Keller

febit GmbH, Heidelberg, Germany

Novel high-throughput techniques are applied to produce huge amounts of data. Examples include the well-established microarray based gene- or microRNA expression profiling and the more recently developed second generation sequencing approaches. Especially in the light of personalized medicine, these experimental techniques bear the potential to detect diseases in time, to improve the understanding of molecular causes of the respective diseases and to enhance patients' therapies. To this end, however, sophisticated computer aided analysis is a necessary prerequisite.

One example of a high-throughput transcriptomics project is the „Human Diseases miRNome". Here, blood samples of about 1.000 individuals with a wide variety of diseases, including patients with different cancer types, autoimmune diseases, cardiovascular, and chronic inflammatory diseases, as well as healthy control individuals, were screened for all small non-coding miRNAs and the respective peripheral profiles were investigated and compared using biostatistics, systems biology and machine learning tools.

Even more challenging are whole genome sequencings. Here, gigabases of genomic information can be generated within few days, detailing the complete genomic code of individuals. As one example we sequenced Ötzi, the Icemann, a mummy, which is over 5000 years old but very well conserved. By analyzing over 3 billion reads generated by SOLiD 4 paired-end sequencing we found more than 1.7 Single Nucleotide Polymorphisms that could be matched well to the phenotype and helped to understand the last secrets of the iceman.
In summary, bioinformatics approaches and tools that are tailored for high-throughput applications are essential to grasp the full potential of the promising experimental techniques and thus important to improve personalized medicine.

CoExpress: a tool for co-regulation analysis of mRNA and miRNA

Nazarov P V (1,), Khutko V, Muller A (1), Schmitz S (2), Kreis S (2), Vallar L (1)*

(1) Microarray Center, CRP-Santé, Luxembourg
(2) Life Sciences Research Unit, University of Luxembourg

Short abstract:
Here we present the user-friendly stand alone software tool CoExpress and apply it for the analysis of co-regulation between miRNA and mRNA in 14 cell lines. Co-regulation effects were validated using numerical methods, bioinformatics and qPCR experiments.

Long abstract:
Small noncoding miRNAs influence most fundamental biological processes by ultimately altering the expression levels of proteins either through degradation of mRNA or through interference with mRNA translation. miRNAs tend to have long half lives and therefore represent promising candidates to be used as disease markers and therapeutic targets. Insights into miRNA functions and miRNA target genes can be obtained from simultaneous analyses of full genome transcription profiles and miRNA levels derived from the same sample types. Therefore, there is a need for effective and user-friendly tools for fast analysis of co-regulation between miRNAs and mRNAs.

The co-regulation study was performed using the developed stand-alone software tool CoExpress, which allows for interactive analysis of mRNA and miRNA co-expression by using microarray data. The software performs microarray data pre-processing, building and visualization of co-expression between mRNA and miRNA using correlation or mutual information metrics. Taken together, the software facilitates the analysis and visualization of miRNA-mRNA, mRNA-mRNA and miRNA-miRNA co-expression events, some of which were confirmed by real-time quantitative PCR.

The proper functioning of the software was tested using public mRNA and miRNA expression data from 14 various cell lines. Data from 42 Affymetrix® HGU133plus2 arrays and 14 miRNA custom microarray experiments were downloaded from public repositories, normalized and analyzed. We have detected 7423 co-expression events between 2533 mRNAs and 199 miRNAs with r2>0.6. 22 of the most prominent mRNA-miRNA co-expression events were validated by qPCR, which showed good concordance with the results of co-expression analyses.

The computational validation of the results was performed by permutation of the samples in the mRNA data set. For the threshold of r2 = 0.6 the estimated p-value < 10–7. Then, the lists of co-expressed genes were compared to lists of potential miRNA targets, predicted by combination of 6 commonly used algorithms EIMMo, DIANA, Pictar, TargetScan, PITA, and miRanda. Despite the small concordance between predicted targets and co-expressed gene lists, we were able to find negatively regulated genes with significant p-values for most of the 199 considered miRNAs, which showed inverse correlation between the selected miRNA and its predicted target gene suggestive of a functional interaction between a given miRNA/mRNA pair.

Web site or supplementary information: http://www.bioinformatics.lu/

Multi-modal imaging and pharmacokinetic modeling shed light on mechanism of action of anti-angiogenic therapy in malignant gliomas.

*Olivier Keunen (1, *), Torfinn Taxt (2), Michal Bartos (3), Radovan Jirik (3), Frits Thorsen (2), Mikael Johansson (1,4), Anaïs Oudin (1), Morgane Sanzey (1), Siti A. Abdulrahim (1), Fred Fack (1), Rolf Bjerkvig (1,2), Simone P. Niclou (1)*

(1) NorLux Neuro Oncology, Laboratory, CRP-Santé, Luxembourg
(2) Department of Biomedicine, University of Bergen, Norway
(3) Brno University of Technology, Czech Republic
(4) Department of Radiation Sciences, Oncology, Umeä University, Sweden

Introduction

Glioblastoma multiforme (GBM) is the most common primary brain tumour in adults. Despite important progress in the molecular characterization of GBMs in recent years, treatment efficacy has only improved marginally and median survival after diagnosis is still limited to about 14 months. GBMs are highly vascularised tumors and endothelial cell proliferation is one of the neuropathological hallmarks of the disease. Therefore the concept of interfering with blood vessel formation has long been proposed as a promising strategy against GBM. Recent clinical trials have shown good radiological response rates with bevacizumab, an antibody against vascular endothelial growth factor (VEGF) [1]. Yet, it is currently not known through which biological mechanisms bevacizumab exerts its effect, and whether it has a true anti-tumor effect [2].

Methods

In this study, we assessed the response to bevacizumab of highly angiogenic GBM xenografts, obtained by passaging human GBM biopsies in nude rats. Multi-modal magnetic resonance imaging (MRI) was used, followed by genomic, proteomic and metabolomic studies as well as immunohistochemical analysis. Animals were scanned on a 7T Pharmascan Bruker MRI scanner, using T1 and T2-weighted sequences to assess tumour morphology, diffusion weighted imaging (DWI) to assess cellularity, and dynamic contrast enhancement MRI (DCE-MRI) to assess tumour physiology. Advanced pharmacokinetic modeling [3] provided access to perfusion and permeability parameters important to understand the mechanism of action of the therapy. Magnetic resonance spectroscopy (MRS) was used to investigate tumor metabolism in vivo.

Results

Bevacizumab reduces contrast enhancement in GBM xenografts and only causes a minor reduction in tumor progression. More interestingly, DCE-MRI and pharmacokinetic modeling revealed a significant reduction of the vascular supply, evidenced by a decrease in intratumoral blood flow and blood volume. Accordingly, immunohistochemistry and electron microscopy revealed a strong reduction of large and medium sized blood vessels, and a reduced number of mitochondria in the tumor cells. Importantly, this was accompanied by a dramatic increase in infiltrating tumor cells. At the molecular level, MRS revealed an increase in lactate and alanine metabolites, while molecular analyses further evidenced an induction of HIF1α protein, a marker of hypoxia and an activation of the PI3K pathway.

Discussion

Using multi-modal imaging and advanced pharmacokinetic modeling, we were able to show that vascular remodeling induced by anti-VEGF treatment leads to reduced intratumoral blood supply and a more hypoxic tumor microenvironment. This may favor a glycolytic metabolism, leading to enhanced tumor cell invasion into the normal brain, a phenotype that has been reported in the clinic as well [4]. These findings are of major clinical relevance for the treatment of GBMs since tumor hypoxia is known to have a negative impact on the efficiency of radiotherapy and chemotherapy treatments. Gaining insight into the mechanisms of action of anti-VEGF therapies should also prove useful in the design of new combined therapies. Follow-up studies will further explore the mechanism of actions of these therapies, this time completing the MRI protocols with Positron Emission Tomography (PET) molecular imaging approaches, which provide markers for hypoxia, proliferation, necrosis and metabolism.

References

[1] Vredenburgh, J.J., et al., Bevacizumab plus irinotecan in recurrent glioblastoma multiforme. J Clin Oncol, 2007. 25(30): p. 4722-9.

[2] Wick, W., et al., Bevacizumab and recurrent malignant gliomas: a European perspective. J Clin Oncol, 2010. 28(12): p. e188-9; author reply e190-2.

[3] Bartos, M., et al., Perfusion analysis of DCE-MRI images using a fully continuous tissue homogeneity model with mean transit time dispersion and frequency domain estimation of the signal delay, Biosignal 2010 conference, Brno, Czech Republic.

[4] Gerstner, E.R., M.P. Frosch, and T.T. Batchelor, Diffusion magnetic resonance imaging detects pathologically confirmed, nonenhancing tumor progression in a patient with recurrent glioblastoma receiving bevacizumab. J Clin Oncol, 2010. 28(6): p. e91-3.

KEYNOTE 2

Towards personalized medicine
From clinical trial data sharing and integration to targeted treatments

Holger Stenzhorn

University Hospital Saarland, Germany

Medicine is currently seeing a revolutionary transformation of the very nature of healthcare from reactive to preventive: The changes – catalyzed by a novel systems approach to disease focusing on integrated diagnosis, treatment and prevention in individuals - lead existing approaches to medicine to personalized predictive treatments over the coming years.

To face this challenge, we propose the creation of an open, standards-compliant and modular framework of tools and services that includes efficient secure sharing and handling of large personalized data sets enabling demanding multiscale in-silico simulations. Important aspects at this are privacy, non-discrimination, and access policies to maximize protection of and benefit for patients.

The developed tools and technologies will be validated within concrete, advanced clinical research settings: Pilot cancer trials are selected based on clear research objectives, emphasizing the need for multilevel dataset integration. To sustain a self-supporting infrastructure real-world use case will be set-up to highlight tangible results for clinicians.

BioMyn: A Web Site for Mining Human Gene and Protein Annotations

Fidel Ramírez

Department of Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics, Saarbrücken, Germany.
E-mail: fidel.ramirez@mpi-inf.mpg.de

The BioMyn web site aims at facilitating complex bioinformatics analyses by offering centralized access to human gene and protein annotations. BioMyn is subdivided in two interrelated areas: one containing tools for studying sets of genes and proteins and the other consisting of a search engine for identifying genes and proteins matching the annotations of a biological query. BioMyn integrates many well-known, publicly available, data sources for human genes and proteins; these include molecular functions, disease and drug associations, sequence family classifications, protein domain architectures, metabolic and signaling pathways, ortholog information, protein interactions and protein complexes, and tissue expression. Our data warehouse already contains over 3 million annotations and continues to grow due to monthly data updates. By providing access to important data sources, BioMyn allows researchers to mine and discover new biological knowledge as well as to test existing hypotheses and formulate new ones - while saving time with, otherwise, necessary visits to many external web sites.

We also developed a search engine capable of quickly identifying similarly annotated genes or proteins. To quantitatively estimate and rank the functional similarity between genes or proteins, we created a new functional similarity measure called BioSim. In contrast to most existing measures based solely on the Gene Ontology (GO), this new measure uses all annotations present in BioMyn. Our evaluations demonstrate that using multiple annotation sources greatly improves the precision of the search engine results, compared to using GO annotations only. The use of our search engine can lead to the discovery of novel gene and protein functions as well as new disease associations. Our functional similarity measures can also be used for the quality assessment of existing protein-protein interactions. The current version of BioMyn can be found at http://biomyn.de, a newer version will become available soon.

A computational method to extract pharmacogenomic relationships from text

Adrien Coulet (*), Nigam Shah, Yael Garten, Mark Musen and Russ Altman

Laboratoire Lorrain de Recherche en Informatique et ses Applications (Nancy), France
E-mail: Adrien.Coulet@loria.fr

Most pharmacogenomics knowledge is contained in the text of published studies, and is thus not available for automated computation. Natural Language Processing (NLP) techniques for extracting relationships in specific domains often rely on hand-built rules and domain-specific ontologies to achieve good performance. In a new and evolving field such as pharmacogenomics (PGx), rules and ontologies may not be available. Recent progress in syntactic NLP parsing in the context of a large corpus of pharmacogenomics text provides new opportunities for automated relationship extraction.

We describe an ontology of PGx relationships built starting from a lexicon of key pharmacogenomic entities and a syntactic parse of more than 87 million sentences from 17 million MEDLINE abstracts. We used the syntactic structure of PGx statements to systematically extract commonly occurring relationships and to map them to a common schema. Our extracted relationships have a 70-87.7% precision and involve not only key PGx entities such as genes, drugs, and phenotypes (e.g., VKORC1, warfarin, clotting disorder), but also critical entities that are frequently modified by these key entities (e.g., VKORC1 polymorphism, warfarin response, clotting disorder treatment).

The result of our analysis is a network of 40,000 relationships between more than 200 entity types with clear semantics. This network is used to guide the curation of PGx knowledge and provide a computable resource for knowledge discovery.

KEYNOTE 3

Mining and Prioritizing Molecular Data and Processes for Integrative Systems Biomedicine

Mario Albrecht

Department of Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics, Saarbrücken, Germany

Sophisticated bioinformatics methods are necessary to integrate, analyze, and visualize rapidly increasing amounts of molecular data for systems biomedicine. This talk will present novel computational approaches to gaining biological insight into cellular processes by mining large-scale datasets and networks. Possible solutions to current quality and prioritization challenges of integrative data exploration will be described.

Linking molecular networks to clinical decision-making support

*Francisco Azuaje (1, *), Yvan Devaux (1), Daniel R. Wagner (1,2)*

(1) Laboratory of Cardiovascular Research, CRP-Santé, Luxembourg
(2) Division of Cardiology, Centre Hospitalier, Luxembourg
Contact: Francisco.Azuaje@crp-sante.lu

The extraction of new knowledge encoded in complex molecular networks for prognostic purposes is a crucial need in biomedical translational research. This approach has not been widely investigated in the cardiovascular research area. A typical application scenario that may be benefitted from advanced computational, systems-level strategies is the prediction of clinical outcomes after suffering a heart attack.

Here we overview recent advances in this direction, which are being developed at the Laboratory of Cardiovascular Research, CRP-Santé. Our lab have investigated different disease-driven computational approaches in areas ranging from heart failure after myocardial infarction, prediction of clinical outcome after cardiac arrest to heart repair therapeutic interventions. In this presentation we will focus on strategies that combine different types of "omics" and clinical data originating from the Luxembourg Acute Myocardial Infarction Registry (LUCKY) and several external public databases.

We have put emphasis on two inter-dependent biomarker discovery strategies. The first one uses information encoded in molecular networks to guide the search for potential biomarkers. The second emphasizes the notion of "network-based biomarkers", which are inferred from non-obvious relationships across patients and between genes in the context of molecular interaction networks.

These strategies are based on the multilayered integration of blood-derived gene expression data and different types of validated molecular interactions. We have also investigated disease-driven applications of clinically-meaningful networks extracted from different types of biological information resources: gene expression associations, protein-protein interactions and functional similarity networks computed from ontology-based relationships between genes.

Different examples will be illustrated, which demonstrate advantages regarding prognostic classification capability and the dissection of relevant mechanisms explaining clinical class-specific phenotypes. Some of these strategies have undergone independent evaluations. Our methods can be adapted to other clinical domains, including those constrained by small molecular datasets and limited domain knowledge. These strategies can pinpoint clinical-meaningful synergistic effects that cannot be identified by standard statistical and bioinformatic analyses.

MIR@NT@N: a framework integrating transcription factors, microRNAs and their targets to identify sub-network motifs in meta-regulation network models

Antony Le Béchec (1,), Elodie Portales-Casamar (2), Guillaume Vetter (1), Michèle Moes (1), Pierre-Joachim Zindy (3), Anne Saumet (4), Charles Theillet (4), Wyeth W. Wasserman (2), Charles-Henri Lecellier (5) and Evelyne Friederich (1)*

(1) Cytoskeleton and Cell Plasticity lab, Life Sciences Research Unit, University of Luxembourg, L-1511 Luxembourg
(2) Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute, University of British Columbia, 950 West 28th Avenue, Vancouver, BC V5Z 4H4, Canada.
(3) Structure and Function of the Cell Nucleus, Institute for Research in Immunology and Cancer (IRIC), Université de Montréal, Montréal (Québec), Canada
(4) Institut de Recherche en Cancérologie de Montpellier INSERM U896, Université Montpellier1, CRLC Val d'Aurelle Paul Lamarque, Montpellier, F-34298, France.
(5) Institut de Génétique Moléculaire de Montpellier UMR 5535 CNRS, 1919 route de Mende, F-34293 Montpellier cedex 5, France

To understand biological processes and diseases, it is crucial to unravel the concerted interplay of transcription factors (TFs), microRNAs (miRNAs) and their targets within regulatory networks and fundamental sub-networks. An integrative computational resource generating a comprehensive view of these regulatory molecular interactions at a genome-wide scale would be of great interest to biologists, but is not available to date. MIR@NT@N is a novel an integrative approach based on a meta-regulation model and a large-scale database aimed at identifying and analyzing molecular interaction networks. MIR@NT@N uses a graph-based approach to predict novel molecular actors across multiple regulatory processes (i.e. TFs acting on protein-coding or miRNA genes, or miRNAs acting on messenger RNAs).

Based on these predictions, the user can generate networks and further analyze them to identify sub-networks, including motifs such as feedback and feed-forward loops (FBL and FFL). In addition, networks can be built from lists of molecular actors with an a priori role in a given biological process to predict novel and unanticipated interactions. Analyses can be contextualized and filtered by integrating additional information, such as microarray expression data, and using prediction scores on regulations. All results, including generated graphs, can be visualized, saved and exported into various formats, for further analysis. To demonstrate regulatory motif detection, we analyzed regulations from published data (Qui et al., 2010; Wang et al., 2010). For the 19 common human TFs, we predicted 43% of published interactions (TFBS score≥0.85), and generated relevant FBL and FFL networks, providing new insights through a graph based approach. MIR@NT@N has been validated using published data and was applied to the regulatory program underlying epithelium to mesenchyme transition (EMT), an evolutionary-conserved process which is implicated in embryonic development and disease.
Our aim was to construct a core sub-network and to detect regulatory motifs by including published miRNAs, TFs identified from binding site (TFBS) detection, and genes selected from microarray data. In conclusion, MIR@NT@N is an effective computational approach to identify novel molecular regulations and to predict gene regulatory networks and sub-networks including conserved motifs within a given biological context. Taking advantage of the M@IA environment, MIR@NT@N is a user-friendly web resource (http://mironton.uni.lu) which will be updated on a regular basis.

KEYNOTE 4

Research in bioinformatics at the LORIA (Nancy): high-performance algorithms and knowledge-based approaches for structural systems biology.

Marie-Dominique Devignes and Dave Ritchie

Laboratoire Lorrain de Recherche en Informatique et ses Applications (Nancy), France

This talk will present some recent work and perspectives on bioinformatics research at the LORIA. The first part of the talk concerns one of the current challenges in computational biology, which is to predict how a pair of proteins fit together, or "dock", to form a protein complex. To tackle this problem, we have developed a novel FFT-based protein docking algorithm using spherical polar Fourier correlations of protein shape and electrostatic properties. This approach has performed well in several of the “CAPRI” blind docking experiments. In addition to correlating protein shapes, our approach can also be used to encode the electrostatic potential, ionisation potential, electron affinity, and polarisability of small molecules. This provides a fast way to perform shape-based and property-based virtual screening of chemical databases. This part of the talk will give an overview of using polar Fourier correlations for docking and virtual screening, and it will describe some recent work at the LORIA on mapping these calculations to modern graphics processors.*

The second part of the talk will present our vision of knowledge-based approaches in bioinformatics. Knowledge discovery from databases (KDD) can be considered as a process involving three main steps : (i) data preparation – i.e. integrating, cleaning, selecting and formatting the data, (ii) data mining – i.e. extracting regularities, patterns, classes, or rules (also called knowledge units) from the data, and (iii) interpretation – i.e. recognizing knowledge units that are new, valid, non-trivial, and reusable. Thus, the KDD process is an experimental, iterative process which is guided by the expert with domain knowledge. Applying KDD to biological data has become necessary due to the huge amount of data produced by today's high-throughput experimental platforms. However, KDD is still hampered by various difficulties at every step of the process. Some solutions will be presented on the basis of selected research projects currently under way at the LORIA.**

In conclusion, high-performance algorithms and knowledge-based approaches bring new ways to enhance our understanding of biological systems at the structural level. This should be of benefit to many applications of systems biology in the therapeutic and biotechnological domains.

* HPASSB (High-Performance Algorithms for Structural Systems Biology), ANR Chaire d’Excellence, Dave Ritchie

** MBI (Modeling Biomolecules and their Interactions) project, funded by the Region Lorraine, INRIA and European FEDER, coordinated by MD Devignes. http://misn.loria.fr/spip.php?article2

Predicting Specificity in Protein-Protein Interactions

Marc Lensink

Interdisciplinary Research Institute - CNRS USR3078
Villeneuve d'Ascq, France

Biological processes in living cells are the result of the interactions between macromolecular assemblies that include various cellular components. Genomics and structural genomics projects are rapidly filling both sequence and structure space, but the interactome is still only sparsely sampled. The CAPRI protein-protein docking and scoring experiment is a community-wide effort aimed at improving the prediction of the threedimensional structure of macromolecular complexes from their individual components. Over its ten-year existence, CAPRI has catalyzed the development of docking methods. For selected systems docking has become a routine affair, even if there is no solution structure available, but there's still progress to be made.

References

*Docking and scoring protein complexes: CAPRI 2009, MF Lensink and SJ Wodak, Proteins 2010;78:3073-3084.

*Blind predictions of protein interfaces by docking calculations in CAPRI, MF Lensink and SJ Wodak, Proteins 2010; 78:3085-3095.

From molecules to ecosystems and back: bioinformatic approaches for Eco-Systems Biology

Paul Wilmes

Department of Environment and Agro-Biotechnologies, Centre de Recherche Public – Gabriel Lippmann, Belvaux, Luxembourg.

We are embedded in a microbial world. Naturally occurring microbial communities play fundamental roles in the Earth's biogeochemical cycles as well as in human health and disease, and provide essential services to mankind, e.g. preservation of food stuffs, treatment of waste or provision of raw materials for manufacturing. Microbial consortia represent highly dynamic systems that exhibit enormous complexity at all levels. With the advent of modern molecular biology methodologies, we are now able to probe deeper into the inner-workings of microbial communities and unravel some of the inherent complexity.

Excitingly, eco-systems-level molecular overviews have the potential to inform community-wide multiscale molecular models that will facilitate the predictable steering of microbial communities towards particular end points, e.g. enhanced carbon dioxide sequestration, reduction of pathogenic infections or the concomitant treatment of wastewater and production of bioenergy. However, in order to build such models, we first need robust bioinformatic approaches that allow us to extract relevant parameters from high-resolution molecular datasets. I will discuss some basic approaches using an extensive tandem high-throughput proteomic and metabolomic dataset that was generated from 14 distinct microbial biofilm samples that were collected from within a disused mine. Due to limited species richness, such biofilms represent ideal model communities for in-depth molecular characterization and placing the resulting molecular patterns into their respective ecological and evolutionary contexts.

Distinct molecular signatures were found for each analyzed sample. Simple correlation analysis between protein and metabolite abundances allowed the deconvolution of the complex molecular dataset into shared and organism-specific contingents. In particular, the observed patterns are reflective of the functional differentiation of two bacterial species that share the same genus and that co-occur in the sampled microbial communities. Our analyses indicate that both species have similar ecological niche breadths and are not in strong competition with one another. This in turn suggests that evolutionary divergence is associated with the restructuring of cellular metabolic networks which allows species to occupy distinct ecological niches. Furthermore, the apparent lack of interspecific competition may explain extensive population-level genetic heterogeneity observed extensively within different microbial communities.

The findings have broad implications for the in-depth investigation of the ecology and evolution of distinct microbial community members and for leveraging the solution of cryptic metabolic processes in the future. Furthermore, they bring us closer to devising multi-scale models of such communities for the provision of efficient microbial services in the future.