

Research projects for Data Engineering derived from open questions in the Life Sciences

Construction of multi-level phylogenies to interpret microbial communities

Phylogenetic trees (capturing the evolutionary relationships among species, the 'tree of life') are a highly relevant structure for the interpretation of biological data. For example, microbiome data (e.g., the composition of microbial communities in the human gut) can be interpreted by evaluating the distribution of the data on a phylogenetic tree. One technical challenge is that these trees exist on different levels of detail (called 'taxonomic ranks'). This project is about constructing a mapping between taxonomic ranks from existing databases and, based on this mapping, analyze the data distribution on multiple levels simultaneously. The potential outcome is to detect, systematics of these data on multiple levels. Developing the computational tools for such a multi-level phylogenetic analysis will help us assess the function of a microbial community and thus will allow us to understand the role of the microbiome in human health and disease.

Further reading:

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2), 217-223.

Claussen, J. C., Skieceviciene, J., Wang, J., Rausch, P., Karlsen, T. H., Lieb, W., Baines, J. F., Franke, A., and Hütt, M.-T. (2017). Boolean analysis reveals systematic interactions among low-abundance species in the human gut microbiome. *PLoS Computational Biology*, 13(6):e1005361.

Constructing data analysis pipelines for GenomeSpace

The enormous diversity of data in biology and medicine, as well as their fragmentation into a huge number of databases and formats currently prevents us from establishing a holistic view on cellular behavior, on the response of organisms to environmental influences and on human diseases. Substantial progress towards these goals can be expected from combining multiple analysis methods to a whole data analysis pipeline, which homogenizes data formats and merges different databases. A promising framework for such projects is the recently published GenomeSpace platform. In this research project, we will focus on gene expression data (the simultaneous measurement of activities for all or many genes in a biological cell). The task is to combine two major interpretation strategies of gene expression data: (1) gene set enrichment via the Gene Ontology classification system and (2) network coherences (i.e. the clustering of gene expression data in a given biological network). These analysis methods need to be combined and integrated as a full data analysis pipeline in the GenomeSpace platform.

Further reading:

Qu, K. et al. (2016). Integrative genomic analysis by interoperation of bioinformatics tools in GenomeSpace. *Nature Methods*, 13(3), 245.

Gene Ontology Consortium. (2014). Gene ontology consortium: going forward. *Nucleic Acids Research*, 43(D1), D1049-D1056.

Knecht, C., Fretter, C., Rosenstiel, P., Krawczak, M., and Hütt, M. T. (2016). Distinct metabolic network states manifest in the gene expression profiles of pediatric inflammatory bowel disease patients and controls. *Scientific Reports*, 6, 32584.

A tool for merging biological interaction databases

Representing large amounts of biological information as networks has become one of the main interpretation strategies of Systems Biology. Exhaustive lists of experimentally determined interactions among cellular components (e.g., genes or proteins) are the prominent data resource behind these networks. However, such interaction information is strongly diversified across a multitude of databases and according to the specific biological type of interaction under consideration. This project sets out to merge existing interaction databases to create a unified, holistic tool, where the user can select, which databases and which types of biological interaction should be included in the network. Examples of databases to be included are: STRING, BioGrid, IntAct. This tool can then be used to analyze the topology of the full cellular-molecular interaction network, as well as for the interpretation of 'omics' data (like gene expression profiles or metabolite profiles) using this network.

Further reading:

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., ... and Jensen, L. J. (2016). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic acids research*, gkw937.

Chatr-Aryamontri, A., Breitkreutz, B. J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., ... and Reguly, T. (2014). The BioGRID interaction database: 2015 update. *Nucleic acids research*, 43(D1), D470-D478.

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., ... and Jandrasits, C. (2011). The IntAct molecular interaction database in 2012. *Nucleic acids research*, 40(D1), D841-D846.

Analysis of biological interaction data using the framework of multilayer networks

Information on interactions among genes and proteins is scattered across a multitude of databases and, furthermore, organized according to different types of interactions (see previous project description). In this project, the data from several such databases, covering all types of interaction, will be accumulated and the resulting interaction network will be statistically analyzed using the recently established perspective of multilayer networks. The key questions are: Have different types of interaction 'co-evolved' to compensate each other or to enhance each other? Do different types of interaction operate on different scales (local vs. global)?

Further reading:

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., ... and Jensen, L. J. (2016). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic acids research*, gkw937.

Chatr-Aryamontri, A., Breitkreutz, B. J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., ... and Reguly, T. (2014). The BioGRID interaction database: 2015 update. *Nucleic acids research*, 43(D1), D470-D478.

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., ... and Jandrasits, C. (2011). The IntAct molecular interaction database in 2012. *Nucleic acids research*, 40(D1), D841-D846.

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., and Porter, M. A. (2014). Multilayer networks. *Journal of complex networks*, 2(3), 203-271.

Wiring economy of gene regulatory networks

Many biological networks are embedded in space. For the network of interacting neurons, the brain, this is an obvious and relevant feature for the functional interpretation of data. In the case of the network of interacting genes, the transcriptional regulatory network, this is less obvious, but of similar functional relevance. The nodes in this network, the genes, are embedded in the genome, an object with a highly involved spatial organization. In all cases one can ask, whether the network is optimally embedded in space (for example balancing wiring length minimization and high processing efficiency). The purpose of this project is to employ a methodology developed in Computational Neuroscience (Chen et al. 2013) and apply it to the genome embedding of the transcriptional regulatory network of the bacterial model organism *Escherichia coli*.

Further reading:

Chen, Y., Wang, S., Hilgetag, C. C., and Zhou, C. (2013). Trade-off between Multiple Constraints Enables Simultaneous Formation of Modules and Hubs in Neural Systems. *PLoS Computational Biology*, 9(3):e1002937.

Marr, C., Geertz, M., Hütt, M.-T., and Muskhelishvili, G. (2008). Dissecting the logical types of network control in gene expression profiles. *BMC Syst Biol*, 2(1):18.

3D predictions of bacterial chromosomes and the interpretation of gene expression profiles

Understanding bacterial gene regulation is a fundamental challenge in Systems Biology. One relevant step along the way is to ask, how patterns of gene activity (gene expression profiles) come about. It has been argued that the two main contributions to gene activity are from the network of interacting genes ('digital control') and from the organization of the genome ('analog control'). Recently, new methods for predicting the 3D organization of the bacterial genome (i.e., the circular chromosome) have been published (Hacker et al. 2017). These methods can be used as a more refined model of analog control for the interpretation of gene expression profiles. This is the goal of the present project.

Further reading:

Marr, C., Geertz, M., Hütt, M.-T., and Muskhelishvili, G. (2008). Dissecting the logical types of network control in gene expression profiles. *BMC Syst Biol*, 2(1):18.

Hacker, W. C., Li, S., and Elcock, A. H. (2017). Features of genomic organization in a nucleotide-resolution molecular model of the *Escherichia coli* chromosome. *Nucleic acids research*, 45(13), 7541-7554.

Functional interpretation of disease-associated genes

Over the last few years information on disease-associated genes has been accumulated and organized in databases. The information mainly comes from large patient cohorts and the identification of small genomic variations in these cohorts compared to reference genomes from the whole population (genome-wide association studies, GWAS). A functional interpretation of these gene lists characterizing a disease has proven to be extremely challenging. Biological networks (metabolic networks, protein-protein interaction networks, signaling networks) should provide an appropriate framework for such a functional interpretation. In this project the statistical and computational tools for analyzing the distribution of disease-associated genes in biological networks will be developed and tested.

Further reading:

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., and Barabasi, A. L. (2015). Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224), 1257-601.

Hütt, M. T. (2014). Understanding genetic variation—the value of systems biology. *British journal of clinical pharmacology*, 77(4), 597-605.

Microbial interaction networks derived from the Human Microbiome Project database

The relevance of the human microbiome (i.e., the microbial communities in the human gut, lung, mouth, etc.) for health and disease is becoming more and more apparent. Recently (Claussen et al. 2017) we published a new method for estimating a microbial interaction network from microbiome data (i.e., from a large number of microbial abundance patterns). The main goal of the present project is apply this method to the emerging databases of microbiome compositions, the Human Microbiome Project (hmpdacc.org), set up a database of microbial interaction networks derived from the HMP database and study, how these networks differ under different conditions.

Further reading:

Human Microbiome Project Consortium (2012) A framework for human microbiome research. *Nature*, 486, 215-221

Claussen, J. C., Skieceviciene, J., Wang, J., Rausch, P., Karlsen, T. H., Lieb, W., Baines, J. F., Franke, A., and Hütt, M.-T. (2017). Boolean analysis reveals systematic interactions among low-abundance species in the human gut microbiome. *PLoS Computational Biology*, 13(6):e1005361.

The human microbiome as a metabolic meta-network

The recent publication of a large number of microbial metabolic models (Magnusdottir et al. 2017) allows us to compare microbial interaction networks (computed from microbiome data) with interaction networks predicted from considering the microbiome as a collection of interacting metabolisms (or a 'metabolic meta-network'). The goal of the project is to apply the interaction indices defined in Levy and Borenstein (2013) to the metabolic models published in Magnusdottir et al. (2017) and compare the resulting network prediction with the network computed from microbiome data via the method from Claussen et al. (2017). This comparison

will provide key insights in the functional organization of the human microbiome.

Further reading:

Magnusdottir, S., Heinken, A., Kutt, L., Ravcheev, D. A., Bauer, E., Noronha, A., ... and Fleming, R. M. (2017). Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nature biotechnology*, 35(1), 81.

Levy, R., and Borenstein, E. (2013). Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. *Proceedings of the National Academy of Sciences*, 110(31), 12804-12809.

Claussen, J. C., Skieceviciene, J., Wang, J., Rausch, P., Karlsen, T. H., Lieb, W., Baines, J. F., Franke, A., and Hütt, M.-T. (2017). Boolean analysis reveals systematic interactions among low-abundance species in the human gut microbiome. *PLoS Computational Biology*, 13(6):e1005361.

Eigenvector of gene regulatory networks and gene expression profiles

One of the principal steps towards an understanding of bacterial gene regulation and therefore cellular function is to quantitatively assess, how well a given gene regulatory network 'explains' a given gene expression data set (i.e., the activity profile of all/many genes). An interesting mathematical approach for addressing this question is to compare spectral properties of the gene regulatory network (in particular, the eigenvectors of the graph) with the activity patterns of genes (which are vectors, where each node of the network is characterized by a real number). This is the task of the present project. Gene expression data will be taken from the GEO database. The gene regulatory network will be downloaded from RegulonDB.

Further reading:

Clough, E., and Barrett, T. (2016). The gene expression omnibus database. *Statistical Genomics: Methods and Protocols*, 93-110.

Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muniz-Rascado, L., Garcia-Sotelo, J. S., ... and Medina-Rivera, A. (2015). RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Research*, 44(D1), D133-D143.