COURSE OVERVIEW

GOAL OF THE COURSE

Illustrate how data-integration is crucial to systems biology where systems biology is defined as acquiring a systems view on an organism by integrating organism-specific omics data. The systems view is here represented as an interaction network.

Illustrate how an acquired systems view (network) can be used to interpret simple 'in house data' e.g. the list of expression data you generate during a master thesis.

Illustrate by one of the at this moment most relevant problems in bioinformatics (genotype phenotype mapping) the added value of network-based data-integration. Genotype-phenotype mapping results in a personalized systems view.

Note that it is impossible to give an overview of 'bioinformatics tools' and this for different reasons: tools change fastly and in general for each bioinformatics problem, several 10s of alternative tools exist. Therefore the solution of any bioinformatics problem, starts with a literature study in which one studies which protocols are available and which one would be most suitable to the problem at hand. In this course works by examples, tools/algorithms that are frequently used in different problems (also those not presented in the course) and that are relevant in the bioinformatics domain will be illustrated. The goal of the course is thus to illustrate how datamining and statistical protocols that were explained in other more fundamental courses can be applied in the domain of bioinformatics (e.g. regression based methods, itemset mining, graph based methods etc)

NETWORK BIOLOGY & NETWORK INFERENCE

This course starts with two chapters on constructing molecular interaction networks. In a first chapter we will give an overview of the molecular data that exist to measure interactions between genes at different molecular levels (protein-protein interactions, transcriptional interactions, metabolic interactions, signaling interactions). It will become clear that these molecular data are noisy and contain missing values. To infer more reliable and more complete interactions between genes it is therefore mandatory to integrate molecular data that assess a specific set of molecular interactions (for instance protein interactions) with complementary information sources (such as functional omics data, genomics data etc). This integrative network inference is the topic of the second chapter. Depending on the type of interactions one wants to infer supervised versus unsupervised methods are more appropriate. Combining the networks that are inferred at each of the different layers eventually results in an integrated network, which can be viewed as a comprehensive summary of all available molecular interaction information on an organism of interest. Exercises on this topic aim at making students aware of one of the most famous inferred functional networks STRING and will illustrate how based on expression profiling data a coexpression network can be built (coexpression networks are still very popular in applied bioinformatics studies). In a third chapter we will illustrate how the inferred networks can be used to aid with the analysis and interpretation of in house generated expression data. A high level overview of the major

classes of algorithms for network-based dataintegration are given (Significant-area-search methods, diffusion, pathfinding). As an illustrative example we will show how an example of a pathfinding tool phenetic can be used to analyze expression data. Illustrative examples of algorithms of the other classes will be described in further chapters in the context of genotype phenotype mapping and systems genetics.

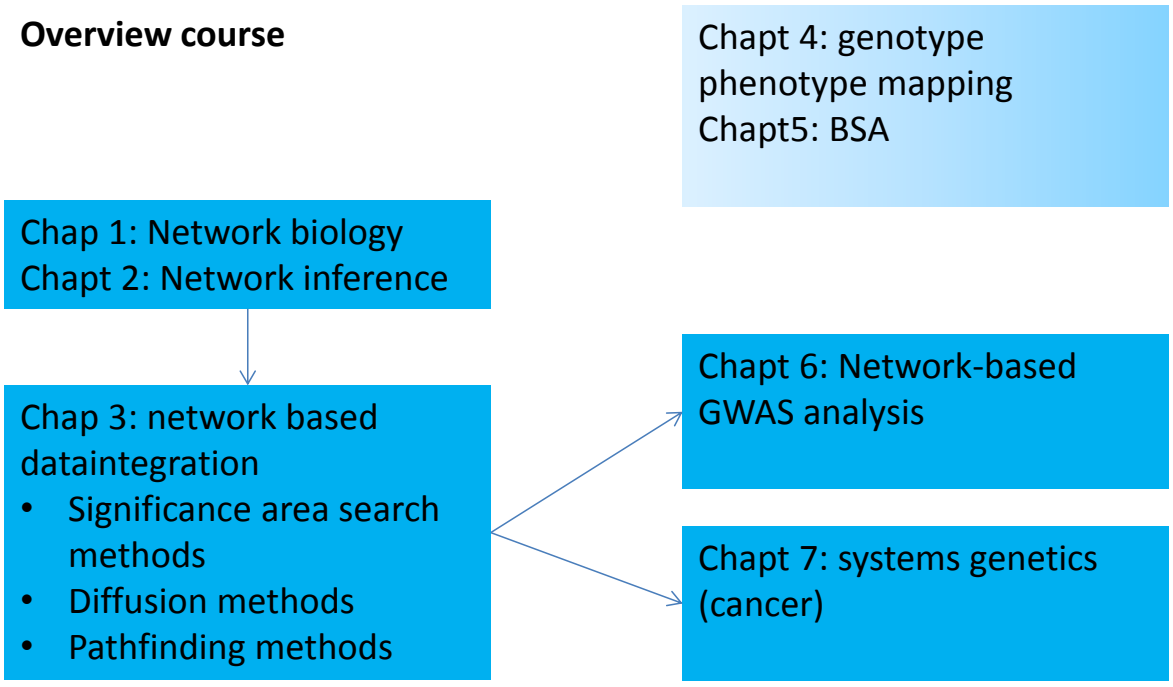
GENOTYPE PHENOTYPE MAPPING

Subsequent chapters relate to the very popular problem of genotype-phenotype mapping. With the advances in sequencing technology having access to genome data at the level of an individual becomes routine, even for very large genomes. This allows exploiting naturally occurring variation (different alleles, mutations) to better understand trait variation. Coupling genotypic to phenotypic trait variation is the area of QTL analysis and GWAS and is customarily being used in higher eukaryotes. Input data are genomic variations and trait information and the goal is to identify the genetic loci that are determining a trait value. Standard protocols for QTL and GWAS studies are mainly based on statistical procedures, which will be outlined in the first two chapters. Subsequently we will outline the bottlenecks of the current approaches (curse of dimensionality, missing heritability) and show how again network-based protocols are increasingly being used to cope with these issues, using protocols that are very similar in set up than those that were used for interpreting expression data (using as an example a Significant-area-search methods).

SYSTEMS GENETICS

Because of the increasing availability of not only individual-based genomics information but also expression information it becomes possible to consider functional data, such as expression data (e.g. eQTL analysis) as traits. Associating genomic variation to molecular traits allows to better understand the molecular mechanisms by which genomic variation results in altered phenotypes. However, because of the complexity of the problem (need to integrate multiple datasources, even larger curse of dimensionality) network-based approaches become almost mandatory. In this chapter we will illustrate how network-based approach are used to tackle the analysis of cancer data (using examples of diffusion based methods).

Overview course



LEERDOELSTELLINGEN

Show which molecular data are used to map interactions between genes

Show why data-integration is mandatory to construct relatively reliable networks

Explain why networks contain missing data

Explain why network representations are overconnected

Illustrate why network-based data integration is useful (in the context of analyzing expression data, in the context of GWAS, in the context of cancer analysis)

Illustrate why solving a bioinformatics problem requires deep biological insight (exploiting LD in BSA, understanding of clonal behavior of cancer is key to the successful analysis of cancer data)

Understand the basic principle/bottlenecks of genotype-phenotype mapping (LD, population stratification, multiple testing problem, missing inheritability)

Provide the basic methods used for genotype/phenotype mapping

Explain why integrative analysis of cancer is complex