

Course Title (Svenska Benämning)

Systems Biology and the Omics-Cascade

Course Number (Kursnummer)

2143

Credit Points (Antal högskolepoäng)

1.5

Level (Nivå)

Research Level (Forskarnivå)

Course Aims

The aim of the course is to provide students with a broad spectrum overview of the field of systems biology and present the different technologies and methods employed in this type of research approach. Specific aims include the following:

- Convey to students the concept and rationale behind systems biology and omics-related research approaches and experimental design.
- Provide an overview of available web-based databases, including KEGG, Reactome database, GenBank, PDB, UniProt, EBI, MIPS, and GO.
- Explain the strengths and weaknesses behind different computer programming languages and the appropriate applications, including Perl, R, Bioruby, etc.
- Educate students on the current state of systems biology and potential future research directions to assist students in applying these methods to their research.

Learning Outcome (Kursens lärandemål)

Following successful completion of the course, the student will acquire a basic level of knowledge regarding systems biology and different omics-related technologies used in the acquisition of large-scale datasets. In addition, students will be updated on recent findings within these fields. Specifically, students will be able to:

- Reflect on the utility of omics-related research in terms of hypothesis generation and experimental design.
- Relate to principle the appropriate applications of different computer programming languages and algorithms.
- Theorize on the future development of omics-related research approaches and systems biology, as well as motivate the answer.

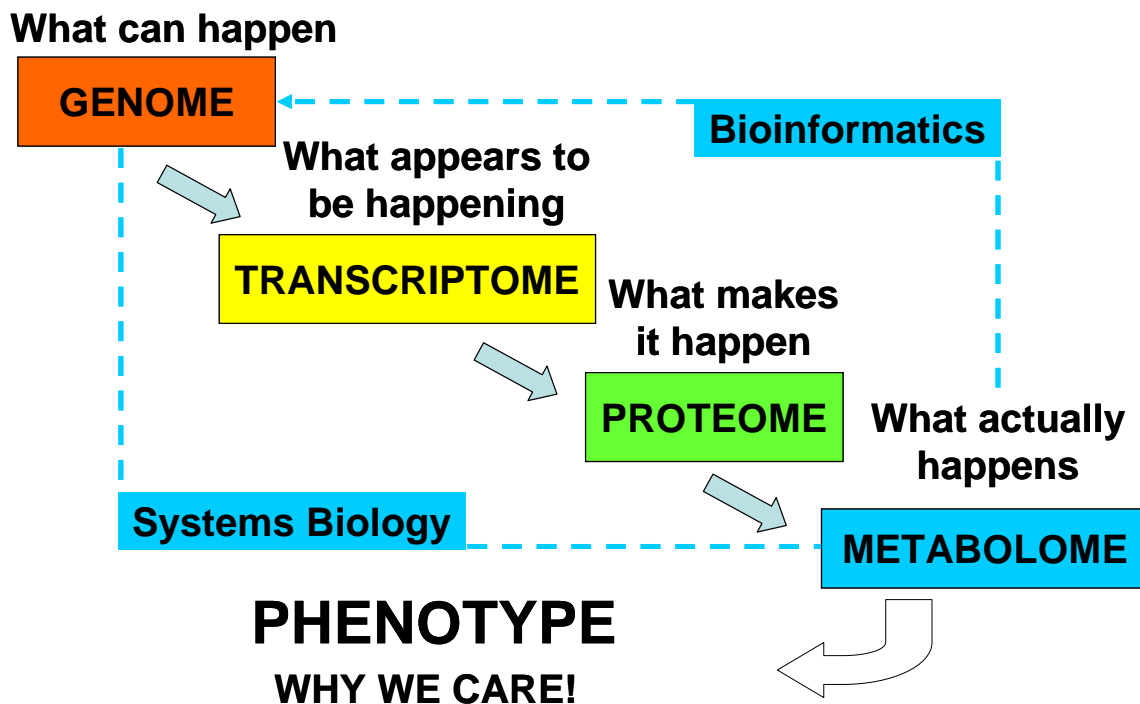
Contents of the Course (Kursens innehåll)

The amount of data generated from biological experiments is steadily increasing, as is the body of information in public databases vital to the interpretation of many experiments. Sometimes commercial or open source software or web services developed for the analysis of routine tasks is sufficient. However, analysis of advanced or truly novel studies often requires innovative and unconventional solutions. Sometimes ad hoc solutions to data analysis and

management are sufficient, e.g. huge spreadsheets and carefully organized file lists. However, often it is necessary to have more advanced and rational approaches to data analysis. It is therefore beneficial for a biologist to have a basic understanding of the available bioinformatics tools. Analyses that require hours of painstaking and sometimes error prone manual repetition can be conducted quickly, accurately, and consistently with a few lines of beginner-friendly computer code in a nearly natural language syntax.

Traditional approaches to biological research have focused on identifying specific genes, proteins and metabolites, and then attempting to determine their endogenous function. However, that type of approach is limited in that it examines each variable in isolation, thereby potentially missing broader effects on the systems level. This course will provide an overview of systems biology and omics-related research approaches. The course will be divided into three main sections: genome informatics, on-line databases with focus on KEGG and applications of different programming languages. The intergration of the so-called “omics-cascade” will be discussed within the context of systems biology and a number of resources will be covered.

The Omics-Cascade



The term “Systems Biology” is being used with increasing incidence in the scientific literature. However, a review of the papers published with the keyword “systems biology” quickly demonstrates that this discipline is extremely broad and that in fact many people are unclear as to what exactly a “systems biology approach” entails. Systems biology can be defined as “the study of an organism, viewed as an *integrated* and *interacting network* of genes, proteins and biochemical reactions which give rise to life”. The key distinction in a systems biology approach to research is that instead of analyzing individual components or aspects of an organism, such as lipids or a specific gene/protein, the focus is placed on examining how all of the components interact as a single system. It is thought that by focusing on the entire organism as a single unit that biological mechanisms can be elucidated.

For example, the disease atherosclerosis exhibits a complex development that involves multiple components of the inflammatory cascade as well as the immune system. Accordingly, it is difficult to attribute any one specific marker (i.e., metabolite, gene or protein) with risk for disease. Instead, by examining the system as a whole and exploring multiple pathways simultaneously, a greater indication of organism state and current phenotype can be acquired.

Systems biology emerged out of the Human Genome Project, which provided the initial impetus to think of science in terms of large-scale systemic datasets. The science of systems biology has been greatly facilitated and convergently evolved with increased access to the internet, on-line databases, data storage facilities and computing power. Of particular interest is the development that systems biology is truly interdisciplinary in that no single discipline can lay claim to being “systems biology”. Instead, it is a conglomeration of biologists, statisticians, computer scientists, mathematicians, physicists, and engineers, among others. Systems biology studies are providing the research community with a growing understanding of how transcribed genes and their resulting translated proteins give rise to a suite of metabolites that determine biological form and function. In a sense, systems biology can be considered to be an investigation into the factors that determine the phenotype of the organism, which is what most studies are interested in determining.

In this course, specific emphasis will be placed on use of the KEGG database (Kyoto Encyclopedia of Genes and Genomes). KEGG is a database of biological systems, consisting of genetic building blocks of genes and proteins (KEGG GENES), chemical building blocks of both endogenous and exogenous substances (KEGG LIGAND), molecular wiring diagrams of interaction and reaction networks (KEGG PATHWAY), and hierarchies and relationships of various biological objects (KEGG BRITE). KEGG provides a reference knowledge base for linking genomes to biological systems and also to environments by the processes of PATHWAY mapping and BRITE mapping. The increasing amount of genomic information is the basis for understanding principles of how higher-order biological systems, such as the cell, the organism, and the biosphere, are formed, as well as for medical, industrial, and other practical applications. However, current informatics technologies cannot readily uncover higher-level complexity of such biological systems, although they are effective to find and characterize building blocks of genes, proteins, and chemical substances. Since 1995 KEGG has been developing knowledge-based methods for uncovering higher-order systemic behaviors of the cell and the organism from genomic information. The reference knowledge is stored in KEGG and associated bioinformatics technologies are developed both for basic research and practical applications.

The computer programming portion of the course will introduce the basic concepts of a number of different languages. Emphasis will be placed on Perl and in particular the use of the BioPerl library. Perl is a powerful language, designed to allow fast prototyping of ideas, with short development time, and in particular to process and generate reports from structured file formats. The language philosophy states that "there is more than one way to do it", and that the interpreter should "do the right thing". In practice this means that there are several ways to write the same idea using a phrasing that is usually similar to “natural language”. While not often endorsed in computer science education, Perl is a pragmatic compromise between different theoretical models of software development and is a very useful language for beginners. It is the preferred language for systems administrators, web developers and bioinformaticans. The strength of a language for an application is derived from the availability of good solution libraries to common problems. In BioPerl, libraries of ready-to-

use modules, drawing on the knowledge of experts from different informatics areas, are available for tasks such as running multiple BLAST searches, interpreting BLAST-result reports, finding genes, constructing and handling multiple alignments, taxonomies, etc (<http://bioperl.org/wiki/Bptutorial.pl>). While covering many areas, it is well suited for sequence analysis and genetics.

Type of Teaching (Arbetsformer)

The course will employ a combination of teaching resources including lectures, laboratory exercises and seminars. A significant portion of the course will focus on hands-on learning using computers in the presence of instructors and IT-personel.

Examination

There will be no course examine. Instead, satisfactory completion of the course will be assessed by written examination consisting of a research report on an aspect of systems biology. This exact subject of the report may be chosen by the student, but must be approved by the course organizer. The final report must be 3 A4 pages or less (12 point font, 2.54 cm margins) including references and must be turned in within 2 weeks of course completion in order to receive course credit (there is no minimum length). The report should make use of the journal articles and concepts discussed during the course, which are expected to generate ideas for the reports. In the event of an unsatisfactory report, as determined by the course organizer, the student will be given one additional week to re-write and resubmit the report. Attendance at all laboratory sessions is required. In the case of necessary absence, the student will be required to perform the laboratory exercises outside of class before a grade is issued.

Course Reading Material (Kurslitteratur)

Given the nature of this type of course, no single textbook will be used. Instead a series of on-line resources will be employed in throughout the course. The algorithms lecture will be based on Pavel Pevzner's book, *An Introduction to Bioinformatics Algorithms* (2004), which is available in Google Books. Some material will be taken from David Mount's textbook *Bioinformatics Sequence and Genome Analysis* (<http://www.bioinformatics.org/>), which also covers the advantages and disadvantages of the respective algorithms and tools. In addition, topics such as Systems Biology Workbench (<http://sbw.kgi.edu/>), Systems Biology Markup Language (<http://sbml.org/index.psp>), and V-Cell (www.vcell.org) will be presented. Coursework will also be supplemented with appropriate journal articles at the discretion of the instructors.

Course Director (Kursansvarig)

Craig Wheelock
Division of Physiological Chemistry II
Department of Medical Biochemistry and Biophysics
Karolinska Institutet
Scheeles väg 2
SE-171 77 Stockholm, Sweden
Phone: 08 5248-7630
Fax: 08 736-0439
URL: <http://metabolomics.se/>

Instructors**Karolinska Institutet**

Craig Wheelock
Åsa Wheelock

Kyoto University

Minoru Kanehisa
Susumu Goto
Masahiro Hattori
Wataru Honda
David Leangen
Diego Diez

Soka University

Kiyoko Aoki-Kinoshita
Takako Yamaguchi
Shuichi Ikeda

Additional guest lecturers and keynote speakers will be included in the final course schedule.