Course Title (Kursbenämning)

Multi- and mega variate analysis in omics research

(Sv: Multi- och megavariat analys för omikforskning)

Course Number (Kursnummer)

2239

Next scheduled occation

February 9-13, 2009 (1 week full time, 8.30-5 pm daily + assigments)

Credit Points (Antal högskolepoäng)

1.5

Level (Nivå)

Graduate student level (Doktorandkurs)

Course Prerequisites (Förkunskapskrav)

Basic knowledge in univariate statistics, such as provided by e.g. graduate course 1383-85 (Basic Course in Medical Statistics) or equivalent course (i.e. understanding of basic statistical concepts such as hypothesis testing, type I and type II error, power, significance, confidence interval, regression analysis and standard deviation) is required.

Selection Criteria

Little or no prior knowledge in multivariate data analysis is expected, but familiarity of Omics research approaches is a merit. If the number of applicants exceeds the number of seats in the course, the selection of course participants will be based on the motivation that the applicant hands in at the time of application.

Course Aims

The aim of the course is to provide students with an overview of the methods available for multi- and megavariate data analysis, with specific focus on practical aspects of analysis of omics-related research using the Simca P+ software. Specific course aims include:

- Educate students on the theoretical aspects of state-of-the-art multivariate methods used in omics research, including principal component analysis (PCA) and Orthogonal Projections to Latent Structures (OPLS)
- Explain the strengths and weaknesses with multi- and mega variate analysis methods as compared to standard univariate methods

- Convey to students how PCA and OPLS can be used in practical applications to detect outliers, trends and patterns, and otherwise to interpret models to gain scientific insights
- Provide the students with practical experience in using MVA tools for putative biomarker identification, comparison of multiple treatments, and validation of models for robustness and predictability

Learning Outcome (Kursens lärandemål)

Following successful completion of the course, students will acquire a basic level of knowledge regarding multivariate analysis of large-scale datasets aquired in omics research. Students will be sufficiently proficient users to be able to apply learned methodologies in their own research projects. Specifically, students will:

- Have knowledge on how to use and interpret PCA and OPLS-DA models in Omics studies as well as for classification
- Have knowledge of how to apply multivariate tools for putative biomarker identification
- Understand how to compare multiple treatments, and understand the importance of uncorrelated information
- Be able to apply PCA to detect outliers, trends, patterns and to classify groups within complex data
- Be able to relate multiple responses to multiple inputs using PLS (Partial Least Squares) modeling
- Be able to analyze experimental data using sound statistical principles
- Have gained basic competence in how to use the SIMCA software for MVA, including reporting results in simplified graphical formats
- Be able to reflect on the utility of multivariate methods in experimental design
- Be able to theorize on the principles of multivariate data analysis, specifically in the field of systems biology, as well as motivate the answer.

Contents of the Course (Kursens innehåll)

Modern data sets contain many variables where the most interesting trends are hidden within a combination of the variables. Omics related data sets are characterized by thousands of variables (e.g. genes or proteins) being analyzed simultaneously with very few replicates. The difficulties in extracting relevant trends are thus amplified by the high number of false positives produced when using traditional univariate statistical methods for these types of data sets. As such, multivariate data analysis (MVA) is essential in the process of extracting information from the complex data sets involved in Omics studies. In this course, participants will learn how to interpret complex data sets quickly and confidently by use of the latest multivariate techniques. The students will learn how to build valid and robust predictive models based on data from transcriptomic, proteomic and metabolomic studies, and discover the secrets of overviewing data. In addition, participants will learn how to use state-of-the-art multivariate tools to extract putative biomarkers or other relevant findings in a statistically significant way.

As an additional module, the course will provide basic insight into how MVA can be used for design of experiments for optimization of the effect of experimental variables, and thereby minimize time and resources spent in the laboratory. These increased throughput and reliability of the experimental outcome will be demonstrated through hands-on analysis of real data using the software package MODDE.

The course is composed of lectures, demonstrations and computer exercises using SIMCA-P+12 software on real-life datasets, which will illustrate the principles and practice of MVA. The focus will thus be on practical applications and development of software skills, and attendance at all computer laboratory sessions is mandatory.

Each day of the course is divided into a morning and an afternoon session. Every session will start with a theoretical lecture followed by practical, real-life exercises using the SIMCA or MODDE software. During the practical sessions, the students will have the possibility to work on their own data sets under guidance of the instructors to assure that the knowledge gained during the course will come into practical use in their own research.

Examination

The course examination will consist of a written report of a MVA analysis performed on the student’s own data sets (if own data set not are available, they can be provided by the course instructor). The data set and choice of analysis method must be approved by the course organizer. The final report must be 3 A4 pages or less (12 point font, 2.5 cm margins) including references and must be turned in within 2 weeks of course completion in order to receive course credit. The report should be based on analyses performed with the SIMCA software during the course, and discuss the results in the light of the the concepts discussed in the course as well as in comparison to traditional analysis methods used in the relevant field of research. 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, as satisfactory completion of the course will also be assessed by completion and understanding of in-course computer exercises. 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)

Multi- and Megavariate Data Analysis Part I: Basic Principles and Applications, Second revised and enlarged edition. L. Eriksson, E. Johansson, N. Kettaneh-Wold, J.Trygg, C. Wikström, and S. Wold (ISBN-10: 91-973730-2-8)

Multi- and Megavariate Data Analysis Part II: Advanced Applications and Method Extensions, Second revised and enlarged edition L. Eriksson, E. Johansson, N. Kettaneh-Wold, J.Trygg, C. Wikström, and S. Wold ISBN-10: 91-973730-3-6 (ISBN-13: 978-91-973730-3-6)

Design of Experiments: Principles and Applications. L. Eriksson, E. Johansson, N. Kettaneh-Wold, C. Wikström, and S. Wold. ISBN 91-973730-4-4

Course Directors (Kursansvarig)

Åsa Wheelock
Lung Research Lab L4:01
Div. of Respiratory Medicine
Department of Medicine
Karolinska Institutet
171 76 Stockholm
Phone: 08-517 76557 or 70664
URL: www.pulmonomics.net

Craig Wheelock
Division of Physiological Chemistry II
Department of Medical Biochemistry and Biophysics
Karolinska Institutet
Scheeles väg 2
171 77 Stockholm
Phone: 08-524 87630
URL: www.metabolomics.se