BIOINFORMATICS – WHAT THE DATA REALLY MEAN

Munich, Spring 2005. In seminar room 5 in the Grosshadern Clinic, 26 young scientists are sitting engrossed in their work at computer terminals. In four days the participants are to learn as much as possible about the statistical analysis of DNA microarray data. On this Wednesday morning the topics are molecular diagnosis, classification by nearest shrunken centroids and support vector machines, and model assessment and selection.

Elena Reiffel, Stefanie Scheid, Florian Markowetz and Stefan Bentink

"Just a few years ago everything revolved around questions of basic research," says Dr. Rainer Spang of the Max Planck Institute for Molecular Genetics (MPIMG), in his introduction to the third course day. "Today the main concern is utilizing the molecular data for the diagnosis and therapy of patients." The data analysis often makes decisions necessary that directly affect the patient and must therefore be made very carefully. However, it is still uncommon for a medical doctor or biologist to master algorithms, plot analyses and perform hierarchical clustering. "That is why we decided to offer seminars on selected topics even back in the first NGFN funding phase," says Dr. Florian Markowetz of the MPIMG in Berlin. The courses in microarray data analysis take place four times a year, twice in Munich and twice in Heidelberg. All NGFN members are eligible to register – and if there are any openings left – scientists from other institutions can register as well. "Ideally the participants should be familiar with at least one programming language. However, it is especially important for them to refresh their knowledge of statistics prior to the course, because without statistics you cannot evaluate data," Florian Markowetz recommends. "People interested in participating can find a number of links to reference literature and software to help get acquainted with the topic at **www.compdiag.molgen.mpg.de/ngfn/**.

THREE BIG "I's" – INTEGRATE, INTERPRET AND INFORM

The courses on the analysis of microarray data are part of the module "Service, Training, Consultation and Quality Management" of the Systematic-Methodological Platform (SMP) Bioinformatics. They are only one of many activities which have been initiated to transfer know-how to the partners of the NGFN. The scientists also develop software programs, for example, and provide consulting services for intricate problems. Apart from that, they work closely with numerous projects from the Disease-oriented Genome Networks and help with statistical and mathematical problems. "In the previous NGFN funding phase we were involved in more than 30 projects from all of the clinical areas," says Professor Roland Eils of the German Cancer Research Center (DKFZ) in Heidelberg, who coordinates the SMP Bioinformatics. However, the achievements of the SMP Bioinformatics extend far beyond specific support for individual projects. Besides the module "Service, Training, Consultation and Quality Management" the bioinformaticians are concerned with data management, the standardization of data and data analysis for the entire NGFN. "We speak of the "3-I" approach: integrate, interpret and inform," Professor Eils explains.

iCHIP

"Integrate" is a keyword for data management, an area in which bioinformaticians conceive, develop and install databases and standards for the various NGFN findings. "A complex research network like the NGFN can only work efficiently when all of its members use the same vocabulary and can also access the results of the other groups," Professor Eils explains. An example is the database iCHIP. It was conceived and introduced specifically for clinical research, already back in the last NGFN phase. iCHIP helps scientists manage their findings from investigations with DNA chips, Affymetrix chips and molecular cytogenetic experiments and to link them to clinical data. iCHIP is already firmly established in several Disease-oriented Genome Networks. Next, Roland

COORDINATOR
Professor Roland Eils
German Cancer Research Center,
Heidelberg
r.eils@dkfz-heidelberg.de

Florian Markowetz and Martin Lange

Eils and his colleagues want to extend iCHIP so that findings from cellular assays can also be combined with data from RNAi experiments, proteomics research, matrix-CGH and tissue microarrays (TMA). "Clinical research is changing constantly. New methods are being established and thus new data is generated," Roland Eils explains. "That is why generally valid standards must continually be developed further." To always meet the current quality demands, the SMP Bioinformatics works closely with international committees, for instance with the European Bioinformatics Institute.

INTERPRETING DATA CORRECTLY

Dr. Jörg Rahnenführer of the Max Planck Institute for Informatics in Saarbrücken is active in the field of data analysis. Together with his colleagues, he works out the bases for the statistical analysis of microarray data. In addition, as a mathematician he is involved in clinically oriented projects. Thus, in collaboration with medical doctors from the University of Düsseldorf, the Charité in Berlin and the Saarbrücken University Clinic, Jörg Rahnenführer has developed a prognostic marker with which the survival time of tumor patients and the period until a relapse can be estimated. Such time predictions are very important to classify a tumor and to select the suitable therapy," says Dr. Bernd Wullich of the Clinic for Urology and Children's Urology in Saarbrücken, who is significantly involved in the project. Up to now, physicians have primarily relied on clinical and histological parameters in assessing a tumor. But interest in genetic markers is increasing. With their help, scientists are hoping to predict the course of cancer in a patient much more accurately. "A considerable number of research findings already exist that are based on individual genetic mutations in order to assess a tumor," Jörg Rahnenführer says. "But actually we need more complex analyses that take all genetic changes in the course of a cancer disease into consideration. Moreover, these analyses must be capable of evaluating the probabilities for possible disease courses." Together with clinicians, he developed a method to calculate this, the so-called genetic progression score (GPS). Genetic changes that have already been recorded are also included in the calculation of this score. "GPS allows us to estimate how the tumor, with the greatest probability, will develop in the patient," Bernd Wullich explains. GPS has already proved to be a suitable prognostic marker for both glioblastomas and prostate cancer to estimate the further course of the disease. Since these are two tumor types with quite different genetic backgrounds, the scientists are convinced that their approach can be applied to other kinds of cancer.

"The primary purpose of our SMP is to simplify the utilization of data generated in the NGFN initiative. We want to facilitate the transformation of data into biomedical knowledge. Our platform has the infrastructure and know-how for this 'data to knowledge' transfer."
Professor Roland Eils

Back to Munich. At 5:00 p.m., after seven hours crammed full of exercises, lectures and discussions, the scientists turn off their computers. Many a head is spinning, but most of the participants seem satisfied as they leave the Grosshadern Clinic. Anja Weigmann from Hannover is also happy that she took the time for this course: "For me, today was very rewarding. This afternoon we evaluated and discussed my data, which I brought with me from Hannover because we simply had not got any further. Now, at last, I have reliable results."

References

Ralser M et al. (2005); An integrative approach to gain insights into the cellular function of human ataxin-2. J Mol Biol 346 (1): 203-214

Brors B (2005); Microarray annotation and biological information on function. Methods Inf Med 44 (3): 468-472

Lottaz C, Spang R (2005); Molecular decomposition of complex clinical phenotypes using biologically structured analysis of microarray data. Bioinformatics 21 (9): 1971-1978