Lecture

L5.1

High throughput sequencing technologies — an algorithm for genome assembly

Aleksandra Swiercz1,2, Jacek Blazewicz1,2, Marek Figlerowicz2, Piotr Gawron1, Marta Kasprzak1,2, Edward Kirton1, Darren Platt3, Lukasz Szajkowski3

1Institute of Computing Science, Poznan University of Technology, Poznan, Poland; 2Institute of Bioorganic Chemistry, Academy of Sciences, Poznan, Poland; 3Lawrence Livermore National Laboratory, Joint Genome Institute, USA

Recently, new approaches of massively parallel sequencing have been developed. They obtain large amount of data in a short time. The set of short DNA fragments, being the output of the sequencer, later on has to be merged together to get the original sequence. The problem of combining DNA fragments and searching for the original sequence is called DNA assembly. It is well known from its high complexity both on the biological and computational levels. The new approaches to sequencing result in shorter sequences, but with relatively good qualities and in a shorter time. Thus, reading a DNA sequence became cheaper and more accessible.

A new SR-ASM algorithm has been proposed for the short fragments assembly. The method takes into account the specificity of the data coming out from the sequencer. DNA fragments may come from different strands of DNA helix, thus apart from the input fragments the algorithm also has to consider the reverse complementary fragments to the input ones. In the DNA fragments errors can appear. The fragments contain insertions, deletions and substitutions of the nucleotides. Therefore, analyzing fragments overlapping, one has to allow possible errors in the alignment of two fragments.

SR-ASM algorithm is based on the modified concept of DNA graph constructed in the following way. Each vertex corresponds to an input DNA fragment in the graph. The number of the input fragments is increased at the beginning by adding the reverse and complementary fragments. Two vertices are connected if two respective fragments overlap. Due to errors in DNA sequences coming from the sequencing step, few errors are admitted in the alignment of the sequences. The algorithm reduces the time of computation, by reducing the number of computed alignments. In the next step, the algorithm searches for the longest path in the graph, which passes only by one fragment from the pair: either by an input fragment or by its reverse and complementary counterpart. If one fragment is chosen, than the other one from the pair is also excluded from the set of available and unused fragments, and is not considered any more in the creation of the path. Usually, it is not possible to find one path which passes by all the fragments due to errors in the input set and repetitions of subsequences in the original sequence. Therefore, several paths are returned.

In the last part of the algorithm, the consensus sequence is determined on the basis of the alignment. The usefulness of the algorithm was proven while performing computational tests on the real data coming from the sequencing experiment of the genome of Prochlorococcus marinus bacteria of length 2 Mbp. The sequencing experiment was performed at the Joint Genome Institute (JGI) in the Lawrence Livermore National Laboratory. The SR-ASM algorithm was compared with other assemblers freely available, like PHRAP, CAP3 and TIGR and with the results obtained by methods at the JGI together with expert’s finishing. Our algorithm resulted in the small number of long contigs, and all of the contigs were of high quality. The outcomes were comparable to the outcomes obtained at the JGI, and were much better than for the other assembly methods.
L5.2

Proteomics as a useful method for biomarker discovery

Magdalena Łuczak

Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland
e-mail: Magda Łuczak <magdalu@ibch.poznan.pl>

Biomarkers are widely used as a valuable source of information in all types of clinical studies (diagnostics, disease assessment and monitoring, prediction of therapeutic outcome). Proteomics provides a unique chance for the searching of new biomarkers. In the recent ten years, great improvement has been made in proteomic technology such as the development of quantitative proteomic methods, high-throughput mass spectrometry, protein-chips, and advanced bioinformatics programs for data analysis. However, the separation of proteins by 2D-electrophoresis is still the most widely used proteomic method. The comparative 2D gel approach which means the comparison between the two conditions can be used for the search of new biomarkers in different pathological conditions. Improvement has also been made in the field of gel-free proteomics. Proteome screening has been significantly improved by the development of highly reproducible capillary chromatography and quantitative analysis by stable isobaric labeling. Moreover, mass spectrometers have been enriched with new types of analyzers and ion sources and it further increased the resolution of proteome analysis. In general, the lecture will present different proteomic techniques and methods used for the identification of protein biomarkers and the applications of the latter in a clinical practice.

L5.3

DNA microarrays as tools for screening genome and transcriptome diversity

Luiza Handschuh

Department of Hematology, Poznan University of Medical Sciences, Poznan, Poland
Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland
e-mail: Luiza Handschuh <luizahan@ibch.poznan.pl>

Microarrays are one of the basic tools used in genomics. Since early 1990s the tremendous progress in microarray technology has been observed. However, among DNA, protein, cell and tissue arrays, DNA microarrays are still the most popular. The small slides or chips covered with tightly packed DNA molecules, complementary to specific target sequences, are currently applied in many research areas. In the high-throughput sequencing era, DNA microarrays serve as fast screening tools for assessment of genome and transcriptome diversity. Comparative genomic hybridization (CGH) arrays, consisted of BACs, cDNAs or long oligonucleotides, are mainly used for identification of large deletions or duplications. Fragmentized genome of interest and the reference genome, fluorescently labeled with different dyes are co-hybridized with a CGH array. The strength of signals obtained for both samples is compared using specialized software to detect copy number alterations. SNP (single nucleotide polymorphism) is analyzed using short oligonucleotide microarrays, which ensure much higher resolution than CGH arrays. High-density commercial arrays are designed to detect more than a million human SNPs. Another possible although less common application of DNA oligonucleotide microarrays is genome re-sequencing via hybridization. DNA arrays composed of probes complementary to bacterial or viral sequences are used for pathogen identification. Transcriptome diversity can be studied with support of various microarray platforms, including DNA chips produced by Affymetrix or more flexible slide arrays, spotted with long oligomers or cDNAs. Contrary to gene chips, slide arrays enable direct comparison of two samples due to their co-hybridization with the same slide, as practiced in CGH. Target samples are obtained by reverse transcription of total RNA extracted from the cells/tissues of interest. Although data analysis is much more challenging than in case of CGH arrays, gene expression profiles were determined for the high number of human diseases. Despite the fact that microarrays can provide a huge amount of information of medical relevance, currently, only a small percent of results generated during genome-wide studies are introduced into the clinical practice. The reason is the lack of sufficient accuracy and reproducibility of microarray-based tests and large cohort-studies essential for their validation. Presumably, the first microarrays, which will find wider practical application, are those enabling the analysis of human genome variants. They will be used in clinics to assess personal predispositions to diseases and sensitivity to particular therapeutics.
L5.4

High throughput sequencing platforms and their application in genome research

Agnieszka Żmieńko

Institute of Bioorganic Chemistry, Polish Academy of Sciences, Center of Excellence for Nucleic Acid Technologies (CENAT), Poland

e-mail: Agnieszka.Zmienko@ibch.poznan.pl

The high demand for low-cost sequencing has driven the development of new sequencing technologies, three of which will be presented. Sequencing-by-synthesis is utilized in Illumina Genome Analyzer which uses reversible dye-terminators, and in Roche 454 instrument, which uses pyrosequencing method. The third presented platform, SOLID (Applied Biosystems) is based on sequencing-by-ligation. Each technology parallelizes the sequencing process, producing thousands or millions of sequences at once. Their current use includes genome de novo sequencing, targeted resequencing, transcriptomic and epigenetic analysis, small RNA discovery and profiling and many more. The so called next-generation technologies speeded-up genome research, already resulting in hundreds of publications in top scientific journals.