Contribution ID: **50** Type: **Oral**

Distributed System Based Strategies for Search, Design and Evaluation of COMBO-FISH probe sets

Tuesday, 13 April 2010 14:20 (20 minutes)

A prominent means to detect genetic aberrations is the method of fluorescence in situ hybridization FISH. To avoid labelling large genomic regions by one polynucleotide like in standard FISH, for COMBO-FISH we search for a set of about 30 colocalizing short sequences with the requirement that no more than 4 of these stretches colocalize within 250 kb anywhere else in the genome. The exact search is parallelized and applied to certain subsets motivated by kinetics and stability considerations.

URL for further information

http://services.medigrid.de/5thegee

Impact

As opposed to algorithms like BLAST, we have to perform an exact search. In the pU and pY cases, the whole genome of the respective species is scanned for pU and pY sequences of minimally 15 bases by a primitive comparison algorithm, which takes one day on a desktop computer. The detected pU and complementary pY sequences are stored in an ASCII data base for each chromosome with a total of 120 Mb ASCII strings. For the design of a specifically labelling COMBO-set, the pU sequences of the gene are extracted. In a second parallelized step, the sequences are located in the whole genome, transferring the partial data bases to the distributed system. The newly developed cluster detection algorithm and deletion process regarding the side conditions runs on one processor using the collected location information to reduce the set. Improving a refined search algorithm, the time for the automatic construction of one COMBO-set was reduced to 10 minutes on a 8 processor system.

Keywords

Services@MediGRID, personalized medicine, COMBO-FISH, gene labelling, genome cluster search

Detailed analysis

The design of an oligonucleotide set colocalizing within a genomic region of typically 80 to 250 kb and respecting the requirement that no further clusters of more than 4 oligos exist calls for a parallel search on a distributed system. Taking binding and stability considerations into account, three biochemically different oligo systems can be distinguished: a) polypurines pU, b) polypyrimidines pY, c) mixed sequences. Cases a) and b) are dual in the sense that they both also allow triple helical conglomerates and the complementary sequences have to be located, too. About 2 to 4 percent of the genome (species specific) consist of pU and pY sequences of more than 15 bases, which we all collect into a special data base. The genetic region of interest is investigated for pU and pY sequences and these oligos are located within the whole genome in a parallelized distributed search. In an automatized deletion process, the set is reduced until no further clusters of more than 4 oligos exist.

Conclusions and Future Work

The design of specifically labelling COMBO-sets has been extended to 150 locations with 3 sublocations of genes at one time. The combined search is being accelerated by cross-referencing of search results, which calls for additional run time correlation algorithms. Furthermore, the search for mixed sequences must use the whole genome as data base. The work presented here and its extension can be a valuable contribution to personalized medicine using individual genetic profiles.

Primary authors: Dr SCHMITT, Eberhard (University of Heidelberg, 69120 Heidelberg, Im Neuenheimer Feld 227); Prof. HAUSMANN, Michael (University of Heidelberg, 69120 Heidelberg, Im Neuenheimer Feld 227)

Presenter: Dr SCHMITT, Eberhard (University of Heidelberg, 69120 Heidelberg, Im Neuenheimer Feld 227)

Session Classification: Bioinformatics

Track Classification: Scientific results obtained using distributed computing technologies