BMC Bioinformatics

Poster presentation

Optimising oligonucleotide array design for ChIP-on-chip

Fiona Nielsen*[1], Stefan Graef[2], Xinmin Zhang[3], Stefan Kurtz[4], Sergei Denissov[1], Roland Green[3], Ewan Birney[2], Paul Flicek[2], Martijn Huynen[1] and Henk Stunnenberg[1]

Address: [1]Nijmegen Centre for Molecular Life Sciences, the Netherlands, [2]EMBL-European Bioinformatics Institute, Hinxton, Cambridge, UK, [3]NimbleGen Systems Inc., Madison, USA and [4]Center for Bioinformatics, University of Hamburg, Germany

Email: Fiona Nielsen* - fnielsen@cmbi.ru.nl

* Corresponding author

from Third International Society for Computational Biology (ISCB) Student Council Symposium at the Fifteenth Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)
Vienna, Austria. 21 July 2007

Published: 20 November 2007

BMC Bioinformatics 2007, **8**(Suppl 8):P4 doi:10.1186/1471-2105-8-S8-P4

This abstract is available from: http://www.biomedcentral.com/1471-2105/8/S8/P4

The sequencing of whole genomes has allowed for custom-made genome-wide microarray assays such as the ChIP-on-chip. With this technology we can detect e.g. transcription factor binding sites over an entire genome. In principle, an accurate detection is only limited by the resolution of the chipdesign, i.e. the tiling density of the oligonucleotides. However, the inherent noise of the DNA hybridisation severely hampers the interpretation of the results.

We mined existing ChIP-on-chip datasets to identify the main sources of noise arising from the sequence selection. We found that limiting intervals must be imposed on 1)the melting temperature, 2)the lengths of the probes, 3)palindromic sequences and 4)the sequence uniqueness relative to the rest of the genome. Based on this knowledge we developed an oligonucleotide array design algorithm [1] to generate a genome-wide array design for any given genome at a given tiling density. To obtain unique sequences we invented a novel approach for selecting the sequences. Using an augmented suffix-array implementation we score sequences by their content of sequence-unique subsequences and select preferentially the sequences with the highest content of unique subsequences.

We have tested our design algorithm using different parameter settings in a fractional factorial test setup, in effect testing eight different parameter combinations. The tests were designed for the mouse genome on the 2.18 M feature array from Nimblegen and performed under true ChIP-on-chip experimental conditions using mouse TBP ChIP samples for the hybridisations.

Test hybridisations were performed for three biological replicas, each hybridised three times, to estimate the variance across both biological and technical replicas.

From the tested designs we deduce the effect of each parameter on the resulting signal and coverage of the design. We correlate the effects and interactions of the probe properties on the probe level (signal intensities) as well as on the design level (quality measures for the whole data set). From this analysis we quantify the effect of each parameter, thus allowing us to choose the design parameter settings that optimise the signal-to-noise ratio, while maintaining a high coverage of the genome. Using our design algorithm and the optimised parameter settings we can produce a genome-wide microarray design with low noise and high coverage for any sequenced genome.

References

1. Graf Stefan, Nielsen Fiona GG, Kurtz Stefan, Huynen Martijn, Birney Ewan, Stunnenberg Henk, Flicek Paul: **Optimized design and assessment of whole genome tiling arrays.** *Bioinformatics* 2007, **23(13):**i195-i204.