An Approach to the Determination of Optimized Oligonucleotide Sets for DNA Chips

Uta Bohnebeck Manfred Nölte Thomas Schäfer Mārīte Šīrava
Thomas Waschulzik Gerald Volkmann

University of Bremen
Center for Computing Technologies (TZI), P. O. Box 330 440, D-28834 Bremen,
Tel: ++49-421-218-7091/9384, gen_inf@informatik.uni-bremen.de

DNA chip-based hybridization analysis is a promising new technology which potentially allows rapid and cost-effective screens for all possible mutations and sequence variations in DNA as well as gene expression analysis used for diverse applications. The aim is that DNA probes (e.g. oligonucleotides) should combine with complementary sequences of single strands of DNA from a sample injected onto the chip. With respect to a specific diagnostic task, the critical point is the design of an appropriate set of oligonucleotides taking into consideration their sensitivity, specificity, and hybridization properties. This poster presents a model which was developed to select oligos to be suitable for the identification of HCV [1], it can be, however, generalized for other analogous applications.

The main idea is to provide a software system that assists and documents the process of selection and configuration of optimized oligonucleotide sets by an appropriate combination of new and standard bioinformatic methods. We split the process in two major steps. The first step begins with the retrieval of all relevant sequences from the international nucleic acid sequence databases like the EMBL data library or Genbank using for instance the taxonomy browser of NCBI complemented by new knowledge-based components in order to select the database entries regarding additional biological features. Then a data preprocessing is needed because the lenghts and the relative gene positions of the base sequences contained in the database entries can differ very strongly. The next problem is the determination of highly sensitive motivs that can be reached by multiple alignment techniques or by the application of generalized suffix trees [4]. In order to guarantee a high specifity of these motivs a control recherche is carried out by a homology search through the sequence databases in combination with an evaluation component.

The second major step consists of determining oligonucleotides based on the motivs created in step one, scoring each oligonucleotide and compiling a cummulative set of optimized oligonucleotides for hybridization. The score considers sensitivity and specifity as well as melting temperature T_M [3] and secondary structure of the oligonucleotides themselves and the target DNA. A second score evaluating the oligonucleotide sets is maximized by a combination of a gradient descent and a genetic algorithm [2]. After utilization of the DNA chip the analysis of hybridization signals of the DNA chip has to be carried out. Image processing is used to get numeric values for each hybridization signal. A designated set of controlling oligonucleotides verifies a proper hybridization experiment. Interpreting the hybridization signals the initial question (e.g. hepatitis C infection) can be answered.

References

[1] H. Gersdorf, L. Boldt, C. Niemeyer, R. Bischoff, U. Bohnebeck, W. Sälter, M. Wolf, and D. Blohm. Hepatitis-C-Nachweis mit Hilfe von DNA-Chips. In *Proceedings des Dechema-Statusseminars Chiptechnologie für DNA-Diagnostik und Sequenzanalyse in Deutschland, 25.-26.01.1999*, Frankfurt, 1999.

[2] A. Kel, A. Ptitsyn, V. Babenko, S. Meier-Ewert, and H. Lehrach. A genetic algorithm for designing gene family-specific oligonucleotide sets used for hybridization: the G protein-coupled receptor protein superfamily. *Bioinformatics*, 14(3):259–270, 1998.

[3] M. Nölte, T. Waschulzik, M. Bethke, J. Hoheisel, and D. Blohm. Bestimmung von Hybridisierungseigenschaften von Oligonukleotiden mit Hilfe künstlicher Neuronaler Netzwerke. In *Proceedings des Dechema-Statusseminars Chiptechnologie für DNA-Diagnostik und Sequenzanalyse in Deutschland, 25.-26.01.1999*, Frankfurt, 1999.

[4] J. T. L. Wang, T. G. Marr, D. Shasha, B. Shapiro, and G.-W. Chirn. Discovering Active Motifs in Sets of Related Protein Sequences and Using Them for Classification. *Nucleic Acid Research*, 22(14):2769–2775, 1996.