Two Challenges in Genomics That Can Benefit from Petascale Platforms

Catherine Putonti[1,2], Meizhuo Zhang[1], Lennart Johnsson[1], and Yuriy Fofanov[1,2]

[1] University of Houston, Department of Computer Science, 218 Philip G. Hoffman Hall, Houston, Texas 77204-3058 USA
[2] University of Houston, Department of Biology and Biochemistry, Houston, Texas 77204-5001 USA
putonti@bioinfo.uh.edu, mzhang@bioinfo.uh.edu, johnsson@cs.uh.edu, yfofanov@uh.edu

Abstract. Supercomputing and new sequencing techniques have dramatically increased the number of genomic sequences now publicly available. The rate in which new data is becoming available, however, far exceeds the rate in which one can perform analysis. Examining the wealth of information contained within genomic sequences presents numerous additional computational challenges necessitating high-performance machines. While there are many challenges in genomics that can greatly benefit from the development of more expedient machines, herein we will focus on just two projects which have direct clinical applications.

1 Introduction

Recent advances in sequencing techniques have lead to an explosion in the amount of biological data available. The number of sequences made publicly available is increasing exponentially. The whole genome sequencing strategy fragments the genomic sequence into many overlapping sequences, sequences these smaller segments, and then assembles these shotguns into contiguous sequences. Assembling the shotgun sequences produced necessitates high-performance machines. It was estimated that the assembly of the human genome took approximately 20,000 hours of CPU time [1]. In addition to the human genome, one must mention the ongoing sequencing efforts of several other organisms such as chimpanzee, chicken, rat, mouse, cow, and dog, just to name a few. With the advent of the recently developed 454 sequencing technique by 454 Life Sciences (Branford, CT), one may expect the availability of sequence data to proceed even more rapidly. Sequencing of the corn genome, considered the most complex sequencing project attempted to date, is possible thanks to IBMs Blue Gene/L supercomputer, capable of a peak performance of 5.7 teraflops (http://www.iastate.edu/%7enscentral/news/2006/jan/supercomputer.shtml).

In addition to sequencing projects, identification of single nucleotide polymorphisms (SNPs), which appear throughout the genomic sequence, is still underway. SNPs are responsible for the variations among individuals and have

W. Lehner et al. (Eds.): Euro-Par 2006 Workshops, LNCS 4375, pp. 313–322, 2007.

been shown to be directly correlated with an individuals susceptibility to disease, response to vaccines and medications, as well as the success of blood, tissue, and organ transplantations. To date 11,961,761 SNPs, 5,646,244 of which have been validated, have been identified in the human genome and are publicly available from the National Center for Biotechnology Information (NCBI). The International HapMap Consortium was formed specifically to catalog genetic these variations in order to determine the similarities and differences in human beings [2]. All of the data from the HapMap project is publicly available from http://www.hapmap.org and includes the genotypes available from the Affymetrix (Santa Clara, CA) GeneChip® Human Mapping 500K Array Set which is comprised of two arrays, each of which is capable of genotyping 250,000 SNPs.

These genomic sequences and identified SNPs contain a wealth of information including indicators for an individuals immunity and susceptibility to disease. Mining such data is not a trivial task. Translating the information found to real applications in the medical and clinical arena is of great importance. In recent years, various diagnostic assays have been developed using nucleic acid-based technologies including the polymerase chain reaction (PCR), microarrays of cDNA and oligonucleotides, and nucleic acid sequence-based amplification (NASBA) assays, amongst others. Nucleic acid-based methods are founded on the principles of hybridization and include primer and/or probe sequences which are complementary to a region of the genomic material of the target region, e.g. a particular gene, mRNA, etc., such that in its presence, the primer/probe will hybridize to the targets DNA/RNA. The results of such assays also contribute to the overwhelming amount of data presently available.

The high-throughput nucleic acid-based microarray format was originally utilized exclusively for the monitoring of gene expression. As is evidenced by the Affymetrix GeneChip®, this is no longer the case. Microarrays, commercially produced as well as those produced in-house, are now being employed for gene expression profiling, sequencing and resequencing efforts, genotyping, DNA-protein interactions, pathogen–host interactions, diagnostics, etc. Countless publications have been dedicated to discussing the applications of microarrays, e.g. [3,4,5,6]. Commercially available arrays can contain thousands to hundreds of thousands of probe sequences, each of which contains information specific to the experimental design. Through gene expression experiments, one can gather information about an organisms cellular functions, regulatory mechanisms as well as biochemical pathways. Microarrays used in such experiments, however, produce only an image representing the hybridization of the microarray probe sequences and the target mRNA. The task of translating this hybridization image into the actual gene regulatory and proteomic networks is far from simple and computationally intensive. Numerous approaches and applications have been developed for inferring these networks. The power of the computational resources available is of primary concern for all such techniques. As the cost of microarrays continues to decrease and the number of probes which can be accommodated increases, the

number of assays produced and thus information generated will most certainly continue to increase.

While there are many challenges in genomics that can greatly benefit from the development of more expedient machines, herein we will focus on just two projects having direct clinical applications. Both are based on the analysis of genomic sequences for the design of probe sequences for the identification of single somatic base mutations and SNPs.

2 Genomic Signatures for Monitoring the Rate of Accumulation of Somatic Mutations

In the living cell, DNA undergoes frequent chemical changes, most of which are quickly repaired. Those that are not result in a mutation. Evolution absolutely depends on mutations because this is the only way that new alleles can be created. At the same time, however, most of the mutations observed are harmful or, at best, neutral. Mutations are relatively rare events. Humans inherit about 3×10^9 base pairs of DNA from each parent. Just considering single-base substitutions, this means that each cell has approximately 6×10^9 different base pairs that can be the target of a substitution. Single-base substitutions are most likely to occur when DNA is being copied. It has been estimated that in humans and other mammals, uncorrected errors occur at the rate of 1 in every 5×10^7 nucleotides [7].

The ability to evaluate the rate at which these mutations are accumulating is advantageous for many facets of research. Firstly, it would be possible to estimate a tissues specific biological age as well as predict the risk of somatic mutation-related disorders such as those that cause certain types of cancer. For instance, retinoblastoma, cancer of the retina, typically affects children. The development of a tumor occurs as a result of somatic mutations in both copies of the RB1 gene or through the inheritance of a mutation in a single copy of the RB1 gene and a somatic mutation in the other. Thus, by monitoring the mutation rate within the RB locus of children having a hereditary predisposition could provide a means of early detection. Another important application for the estimation of the mutation accumulation rate is for monitoring the deviation of cell lines from their ancestors. Because at the present time cell lines are less expensive than laboratory animals, human (and other organisms, e.g., mouse, rat, etc.) cell lines are commonly used in cancer research, drug design, and drug screening. Because mammalian cells usually cannot perform more than 50–60 divisions, many cell lines are artificially immortalized. Technically, human cells growing in a culture are different from normal human cells in their natural environment, and this difference becomes more and more significant in time because of the mutation accumulation process. Therefore, research as well as drug development and testing would greatly benefit from early detection of the genetic drift of cell lines. The current method to estimate the mutation rate is based on analysis of a particular part, usually within the coding region, of a genome. Such an approach requires PCR amplification of the genomic region of interest followed by sequencing, which can be both time consuming and expensive.

Recently, utilizing our computational abilities we have performed analysis of subsequences of length n (n-mers) located in the area of "one mismatch distance for several microbial genomes. The one mismatch distance corresponds to sequences that are not present in the genome but given any one base change or mismatch are present in the genomic sequence. Such sequences have a higher probability of appearing as a result of a single point mutation than, for example, sequences 2, 3, etc. mismatches away. For each n-mer located in the one mismatch distance area, we were able to compute the exact number ancestor sequences in the genome which can "mutate to this n-mer. As a result of these calculations, we observed a large variation in the number of ancestors for different sequences: from 1 or 2 (expected) to thousands for microbes and presumably dozens of thousands for humans. This new observation leads us to the idea of using a set of sequences having a high number of ancestors as an indicator of the accumulations of mutations. This approach is based on the assumption that sequences with a higher number of ancestors have a much higher probability to appear as the result of random mutations.

For microbial genomes, identifying the n-mers within the areas of 1, 2, and 3+ mismatch(s) is computationally feasible due to the fact that these genomic

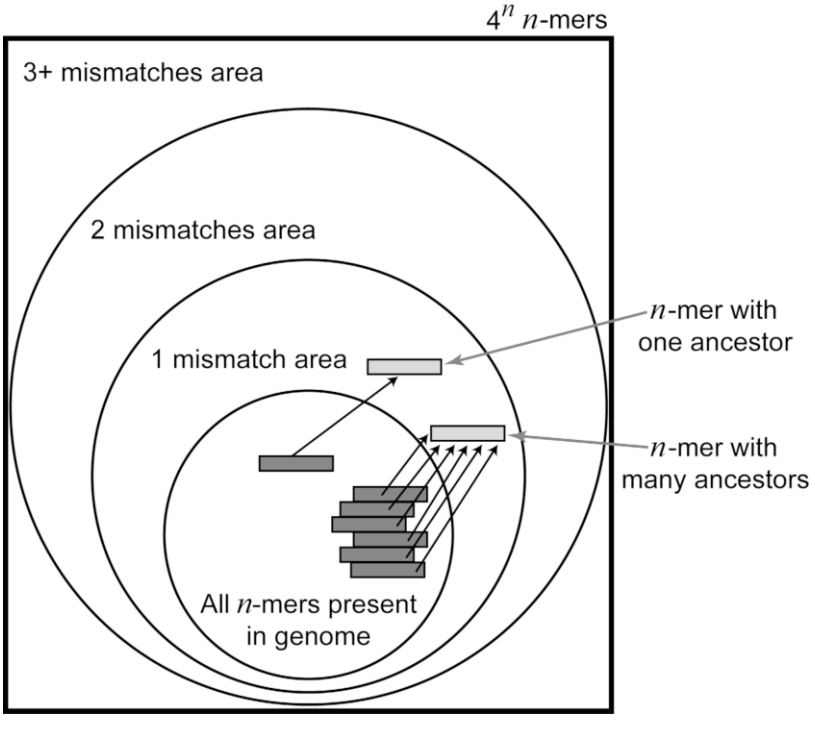

Fig. 1. Those n-mers that are located in the "one mismatch distance area have different probabilities of appearing as result of random mutation

sequences are relatively short. For instance, the *Escherichia coli* K12 genome (NCBI: NC_000913) is 4,639,675 bp. Considering both strands of this genome, all n-mers are expected to appear only once at most for $n \leq 12$. Thus, for 12-mers one may expect that some sequences are absent from the genomic sequence which may have ancestors present. This prediction is made under the assumption that the appearance of n-mers is independent which is known not to be completely true; it does, however, provide a reasonable estimation. There are in fact several 10- and 11-mers which are absent from the *E. coli* K12 genome. For a single n-mer S of length n, there are $3n$ possible n-mers which can be created by changing only one base, $3n * 3(n-1)$ possible n-mers by changing any two bases, etc. For each of the n-mers absent from the genome, there are

$$\prod_{0}^{c} 3(n-c) \tag{1}$$

possible ancestors which can be present given c base changes. For the human genome ($\approx$ 3 Gbp), much longer n-mers must be examined. As a result, many more possible combinations of mutations must be considered. It is important to mention that rather than looking exclusively at the human genome sequence, SNP information must also be included as the appearance of such a variation would be the result of the individuals particular genotype rather than the occurrence of a mutation.

Using such sequences in an assay, e.g. as microarray probes, would provide a means to rapidly and conclusively examine the rate of the accumulation of mutations in microbial and eukaryotic organisms, including microbial cultures, strains of model organisms, and human cell lines. We expect that the same approach will also be useful in distinguishing individual humans, as well as other eukaryotes (including organisms of economic importance, e.g., pigs, cows, chickens, etc.) based on small DNA samples. Furthermore, we anticipate that the sensitivity of this technique will be sufficient to monitor the rate of somatic mutations accumulating in different tissues during the lifetime of an organism. Such an opportunity is likely to be of great benefit to cancer related research. In contrast to existing methods of mutation rate monitoring, our approach takes into account entire genomes, including noncoding sequences which in the case of human cover 97% of genome. A calculation of such sequences and their ancestors becomes exponentially more difficult as the n-mer size being considered increases. Due to the large number of calculations that are required, petascale computing provides a solution in which such computations can be conducted in an acceptable amount of time.

3 Optimal Combinations of Genomic Signatures for Human HLA Typing

The HLA (human leukocyte antigen) system, the group of genes in the human MHC (major histocompatibility complex) located on chromosome VI, encodes

the cell-surface antigen-presenting proteins. HLA antigens are the major determinants used by the body's immune system for recognition and differentiation of self from nonself (foreign) substances. This system consists of numerous SNPs encoding 2435 known alleles according to the latest statistics available from the IMGT/HLA Database as of August 3, 2006 [8]. Since the previous release (2.13) a month earlier, 125 new alleles were added [8]. The allelic composition in the HLA loci, or the HLA type, varies significantly within the population. A direct correlation has been observed between the variation present in the HLA system and ones genetic susceptibility to diseases [9,10,11,12], response to infection [13], response to drugs and vaccines [14,15], as well as the success of blood and tissue transplants [16]. Significant research has been devoted to the development of HLA typing techniques for determining the combination of alleles responsible for particular responses. Numerous nucleic acid-based approaches have been employed for genotyping the HLA loci by designing primer/probe sequences complementary to the SNPs present in particular alleles. The design of an assay which can type all of the known alleles as well as be used for the discovery of new HLA alleles would dramatically improve current typing methods. The high-throughput microarray format is ideal, offering the convenience of miniaturization and the ability to perform thousands of hybridizations in a single experiment. Previously documented microarray-based typing methods were intended to provide a low resolution typing and are therefore able to identify only 6% to 33% of the allelic variations in the loci of interest [17,18,19,20,21]. In order to achieve higher resolution typing, more alleles must be targeted. At the same time, however, one most consider the complexity of the assay design as it may impact the expediency in which a diagnosis can be made as well as the cost and the level of expertise required for technicians to correctly interpret the results. Thus, a simplistic design is preferable. Minimizing the number of probes in the assay, while maintaining or improving the assays resolution and reliability, further complicates the task of designing typing tests thus necessitating rigorous computations.

Alleles	P_1	P_2	P_3	P_4
A_1	✓		✓	
A_2		✓		✓
A_3	✓	✓	✓	✓
A_4		✓	✓	✓
A_5	✓	✓		

Fig. 2. Determining the optimal probe set. Each allele is expected to hybridize with a subset of the probes $\{P_1, P_2, P_3, P_4\}$. Not all probes, however, may be necessary in order to maintain the same resolution. Thus, the optimal set $\{P_1, P_3, P_4\}$ provides the same resolution while reducing the complexity and cost of the assay.

Different probe sets for HLA typing can be created providing variable resolutions. Given a set of alleles, all of the sequences containing the polymorphisms can be selected as candidate probes. While all of the candidate probes can be included in the design of the assay, it is likely that some of the probes do not contribute to the informativity or increase the resolution of the typing. To illustrate such an instance, Figure 2 shows four probes (P_1, P_2, P_3, and P_4) which can be used to distinguish between five different alleles (A_1, A_2, A_3, A_4, and A_5). In order to be able to distinguish one allele from another, the set of probes which hybridizes to each allele must be unique resulting in a unique hybridization pattern on the array. The inclusion of P_2, however, is not necessary as all five alleles will still be expected to hybridize with a unique set of probes. By removing any of the other probes, it will no longer be possible to distinguish between the alleles and thus the resolution of typing is decreased. Therefore, an "optimal set containing only three probes can be used to distinguish between the alleles. Although a rather simple example, the optimal assay provides the same resolution at a reduced cost and complexity. As one can imagine, identifying the minimum number of probes necessary for distinguishing between 2000+ HLA alleles is a significantly more complex problem.

To optimize an assay, one can either search for the maximum coverage of the targets using: (1) a predetermined number of probes or (2) the minimum number of probes. If there are only a few candidate probes, it is feasible that one can iterate through each possible combination in order to identify the combination of probes with the maximum coverage. However, as the number of candidate probes increases so too does the number of possible combinations. For instance, to identify the set of probes having the maximum coverage for an assay of the predetermined size of 60 probes from 100 candidate probes, 1.37×10^{28} different combinations must be examined. Approximately 1.27×10^{30} combinations exist for the same set of 100 candidate probes for all possible probe set sizes. If analysis of each combination requires 1 millisecond, iterating through all of the combinations to find the optimal set using the minimum number of probes will take 4.02×10^{19} years!

Computing the minimum number of probes in a realistic time is nontrivial (an instance of the minimum set cover problem which is NP-complete). Optimization of the probe set design problem was first discussed in 2000 by Herwig et al. [22] in which a greedy heuristic was introduced based on clustering and entropy. Formulation of the problem was further refined by Borneman et al. [23] to the Minimum Cost Probe Set (MCPS) and Maximum Distinguishing Probe Set (MDPS); MCPS searches for the minimum number of probes necessary to distinguish all target sequences while MDPS maximizes the number of distinguished pairs of target sequences for a set of k probes. Here Borneman *et al.* [23] developed a Lagrangian relaxation algorithm to approximate the MCPS problem and a simulated annealing algorithm for the MDPS problem. While successful in designing a smaller probe set, certain sacrifices were made for efficiency by considering only one length of probes ($n = 8$) and predetermining the set size. Two approaches were also developed based on the Integer Linear Program (ILP) formulation. The method of Rash et al. [24] uses suffix trees to

solve the minimization problem. Their solution is based upon the concept of a unique barcode. This barcode is a binary vector consisting of 0s and 1s where 0 means that the probe sequence will not hybridize with the sequence of interest and 1 means that the probe sequence will hybridize with the sequence of interest. In this ILP implementation, each sequence (genome) being considered must be uniquely identified by at least one probe under the assumption that only one target sequence is present in the sample [24]. The second approach of Klau et al. [25] consists of three steps: (1) computing the target–probe incidence matrix, (2) computing a design matrix, and (3) decoding the result for identification of the sequence(s) present in the sample [26]. The design is computed using a branch-and-cut algorithm (http://www.inf.fu-berlin.de/inst/ag-bio). This algorithm proves more robust than that proposed by Rash et al. [24] by taking into consideration during design the set size, the probability of hybridization errors, and the case in which multiple targets are simultaneously present in the sample (*d-separability*) [25]. All of these approaches [22,23,24,25,26], can only approximate the best solution within the space and time allotted to the probe set design process. Therefore, the optimal set identified by any such approach may not in fact be the true optimal set having the minimum number of probes for the maximum coverage. With respect to designing the optimal probe set for high resolution HLA typing, certain alleles may occur with very low probability. If it is possible to identify all alleles except this rare allele with a higher level of resolution and a smaller probe set, such a solution may be preferred, thus adding another dimension to the optimization problem.

4 Conclusions

The rate in which new data is becoming available far exceeds the rate in which one can perform analysis. Sequence data as well as the results of microarray experiments of gene expression profiling, genotyping, and diagnostics further contribute to amount of data to be examined. Analysis of large, complex genomic sequences such as the human genome necessitates high-performance computing resources. The projects discussed here are just two of many that are currently underway in research laboratories throughout the world. Due to the limitations of current systems, it has only been possible to analyze a fraction of the vast amount of biological data currently available. The development of cutting edge computational resources, both in terms of the memory available and the precision and speed in which calculations can be performed, is likely to dramatically impact biotechnology, human health as well as our general understanding of mechanisms of disease development, vaccine development, aging, and evolution.

References

1. Walgate, R.: Weapons lab to develop Celeras new supercomputer. Genome Biol. (2001)
2. The International HapMap Consortium: A haplotype map of the human genome. Nature 437 (2005) 1299–1320

3. Jares, P.: DNA microarray applications in functional genomics. Ultrastruct. Pathol. 30 (2006) 209–219
4. Lockhart, D.J., Winzeler, E.A.: Genomics, gene expression and DNA arrays. Nature 405 (2000) 827–836
5. Peeters, J.K., Van der Spek, P.J.: Growing applications and advancements in microarray technology and analysis tools. Cell Biochem. Biophys. 43 (2005) 149–166
6. Geschwind, D.H.: DNA microarrays: translation of the genome from laboratory to clinic. Lancet Neurol. 2 (2003) 275–282
7. Kimball, J.W.: Biology. 6th edn. Wm. C. Brown, Iowa (1994)
8. Robinson, J., Waller, M.J., Parham, P., de Groot, N., Bontrop, R., Kennedy, L.J., Stoehr, P., Marsh, S.G.E.: IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. Nucleic Acids Res. 31 (2003) 311–314
9. Diepstra, A., Niens, M., Te Meerman, G.J., Poppema, S., van den Berg, A.: Genetic susceptibility to Hodgkins lymphoma associated with the human leukocyte antigen region. Eur. J. Haematol. Suppl. 75 (2005) 34–41
10. Saftlas, A.F., Beydoun, H., Triche, E.: Immunogenetic determinants of preeclampsia and related pregnancy disorders: A systematic review. Obstet. Gynecol. 106 (2005) 162–167
11. Ahmedov, G., Ahmedova, L., Sedlakova, P., Cinek, O.: Genetic association of type 1 diabetes in an Azerbaijanian population: the HLA-DQ, -DRB1*04, the insulin gene, and CTLA4. Pediatr. Diabetes 7 (2006) 88–93
12. Listi, F., Candore, G., Balistreri, C.R., Grimaldi, M.P., Orlando, V., Vasto, S., Colonna-Romano, G., Lio, D., Licastro, F., Franceschi, C., Caruso, C.: Association between the HLA-A2 allele and Alzheimer disease. Rejuvenation Res. 9 (2006) 99–101
13. Keet, I.P., Tang, J., Klein, M.R., LeBlanc, S., Enger, C., Rivers, C., Apple, R.J., Mann, D., Goedert, J.J., Miedema, F., Kaslow, R.A.: Consistent associations of HLA class I and II and transporter gene products with progression of human immunodeficiency virus type 1 infection in homosexual men. J. Infect. Dis. 180 (1999) 299–309
14. Ovsyannikova, I.G., Vierkant, R.A., Poland, G.A.: Importance of HLA-DQ and HLA-DP polymorphisms in cytokine responses to naturally processed HLA-DR-derived measles virus peptides. Vaccine 24 (2006) 5381–5389
15. Ovsyannikova, I.G., Pankratz, V.S., Vierkant, R.A., Jacobson, R.M., Poland, G.A.: Human leukocyte antigen haplotypes in the genetic control of immune response to measles–mumps–rubella vaccine. J. Infect. Dis. 193 (2006) 655–663
16. Morishima, Y., Sasazuki, T., Inoko, H., Juji, T., Akaza, T., Yamanoto, K., Ishikawa, Y., Kato, S., Sao, H., Sakamaki, H., Kawa, K., Hamajima, N., Asano, S., Kodera, Y.: The clinical significance of human leukocyte antigen allele compatibility in patients receiving a marrow transplant from serologically HLA-A, HLA-B, and HLA-DR matched unrelated donors. Blood 99 (2002) 4200–4206
17. Haddock, S.H., Quartararo, C., Cooley, P., Dao, D.D.: Low-resolution typing of HLA-DQA1 using DNA microarray. Methods Mol. Biol. 170 (2001) 201–210
18. Consolandi, C., Frosini, A., Pera, C., Ferrara, G.B., Bordoni, R., Castiglioni, B., Rizzi, E., Mezzelani, A., Bernardi, L.R., De Bellis, G., Battaglia, C.: Polymorphism analysis within the HLA-A locus by universal oligonucleotide array. Hum. Mutat. 24 (2004) 428–434
19. Palmisano, G.L., Delfino, L., Fiore, M., Longo, A., Ferrara, G.B.: Single nucleotide polymorphisms detection based on DNA microarray technology: HLA as a model. Autoimmun. Rev. 4 (2005) 510–514

20. Bang-Ce, Y., Xiaohe, C., Ye, F., Songyang, L., Bincheng, Y., Peng, Z.: Simultaneous genotyping of DRB1/3/4/5 loci by oligonucleotide microarray. J. Mol. Diagn. 7 (2005) 592–599
21. Wells, D.: Advances in preimplantation genetic diagnosis. Eur. J. Obstet. Gynecol. Reprod. Biol. 115 (2004) S97-S101
22. Herwig, R., Schmidt, A., Steinfath, M., OBrian, J., Seidel, H., Meier-Ewert, S., Lehrach, H., Radelof, U.: Information theoretical probe selection for hybridization experiments. Bioinformatics 16 (2000) 890–898
23. Borneman, J., Chrobak, M., Vedova, C.D., Figueroa, A., Jiang, T.: Probe selection algorithms with applications in the analysis of microbial communities. Bioinformatics 17 (2001) S39–S48
24. Rash, S., Gusfield, D.: String barcoding: uncovering optimal virus signatures. In: Myers, G., Hannenballi, S., Istrail, S., Perzner, P., Waterman, M. (eds): RECOMB '02: Proceedings of the Sixth Annual International Conference on Computational Biology. ACM Press, New York (2002) 254–261
25. Klau, G.W., Rahmann, S., Schliep, A., Vingron, M., Reinert, K.: Optimal robust nonunique probe selection using integer linear programming. Bioinformatics 20 (2004) I186–I193
26. Schliep, A., Torney, D.C., Rahmann, S.: Group testing with DNA chips: generating designs and decoding experiments. Proc. IEEE Comput. Soc. Bioinform. Conf. 2 (2003) 84–91