Efficient statistical computations on genome-scale data using reduced representations

Md Pavel Mahmud[1] and Alexander Schliep[1,2]

[1]Department of Computer Science, Rutgers University
[2]BioMaPS Institute for Quantitative Biology, Rutgers University
{pavelm,wiedenhoeft,schliep}@cs.rutgers.edu

1 Introduction

The amount of data generated by high-throughput technologies requires improvements to algorithms and statistics to cope with biases and noise and assess *statistical* significance of findings in reasonable running times. A recent insight was that many computations can be greatly accelerated by using a suitable reduced representation of the input. This strategy accelerates full Bayesian HMM to the point of competitiveness with Maximum-Likelihood HMM and also leads to large improvements in quality and speed of the analysis of high-throughput data(HTS), a technology to unravel gene sequences on a large scale, in a compressive genomics approach.

2 Full Bayesian Hidden Markov Models

Hidden Markov Models (HMM) are widely used for analyzing Comparative Genomic Hybridization (CGH) data to identify chromosomal aberrations or copy number variations by segmenting observation sequences. For efficiency reasons often parameters of an HMM are estimated with maximum likelihood and a segmentation is obtained with the Viterbi algorithm. This introduces considerable uncertainty in the segmentation, which can be avoided with Bayesian approaches using Markov Chain Monte Carlo (MCMC) sampling. While their advantages have been clearly demonstrated, the likelihood based approaches are preferred in practice for their lower running times; datasets coming from high-density arrays and high-throughput sequencing amplify these problems.

These computational disadvantages can be alleviated by pre-processing the input to arrive at a suitable reduced representation. Leveraging spatial relations between data points in typical data sets leads to a (partial) clustering approach, inspired by ideas from using sequence compression for HMMs with discrete observations and spatial data structures such as *kd*. Based on this reduced representation, we propose an approximative MCMC sampler. We test our approximate sampling method on simulated and biological ArrayCGH datasets and demonstrate a speed-up of 10 to 70 while achieving competitive results with the state-of-the art Bayesian approaches.

3 Compressive Genomics

The default modus operandi for analyzing high-throughput sequencing (HTS) data—which is pervasive in clinical and biological applications such as cancer research, and which is expected to gain enormous momentum in future personalized medicine applications—is individual analysis on the level of reads and on the level of genomes. This leads to non-trivial computational demands, as typical data sets consist of up to 2 billions of sequencing reads, and large studies might provided hundreds of such data sets. Core steps of HTS analysis are read error correction, mapping to the reference genome and identifying genetic variations

We propose to compute reduced representations of HTS data for a single or multiple sequencing experiments in a way that existing and future methods can directly operate on these reduced representations of the data and enable the use of advanced statistics even on very large data sets. We arrive at a reduced representation of HTS data sets by a novel highly-efficient clustering method able to cope with many billions reads from a single or multiple sample genomes, even in repetitive genomes. The reduced representation simplifies sharing and storing of data. Additionally, existing downstream analysis methods can be used at great improvements to running times; adapted or novel downstream algorithms can operate directly on the clustered representations at even larger improvements. The most important consequence of the much increased computational efficiency is not the savings in CPU cycles, but rather the ability to use much more sensitive and accurate tools. Even for read mapping, our approach makes the routine use of Stampy feasible, negating the needs for additional advances downstream tools for calling small structural variants not identified by the speedier, but tools such as Bowtie which are less accurate in the presence of indels. Our approach constitutes a novel foundation for efficient statistical tools for analyzing HTS data.