A FRAMEWORK FOR GENOMIC REARRANGEMENT ANALYSIS

Péter LACZKÓ
Advisors: Béla FEHÉR, István MIKLÓS

I. Introduction

Cancer is a class of diseases in which cells display uncontrolled growth through division beyond the normal limits. The exact molecular biologic mechanisms of such severe alterations in the regulatory processes are widespread, however, in every case they can be traced back to somatic mutations of the tumor cell genomes [1]. These mutations range from single nucleotide polymorphisms to larger scale structural variations, in extreme cases even to chromosome aneuploidity. An important subclass of somatic mutations is genomic rearrangement, which involve the relocation of various blocks of the genome, sometimes in a rather complicated way. Understanding the dynamics behind these rearrangement events allows for the assessment of their oncogenic effects and the subsequent classification of tumor cell samples based solely on their genomic landscape, leading eventually to better diagnostic and prognostic methods.

In the era in which next-generation sequencing (NGS) technologies are becoming ubiquitous and the cost of whole human genome sequencing is decreasing rapidly, there is a clear rationale for the inception of a tumor genome classification framework which would make the aforementioned diagnostics possible. Our research aims at developing such a framework, incorporating every data processing step necessary to obtain the classification of a tumor cell from its raw next-generation sequencing data.

In this paper I briefly describe the nature of genomic rearrangements that we aim to detect and discuss other methods available for similar purposes. I describe our proposed framework and the algorithm behind the components where applicable. Finally, I introduce a simulator tool that we developed and which allows for rapid evaluation of prospective rearrangement detection methods.

II. The cancer genome

DNA in healthy human cells is continuously damaged by mutagens of both internal and external origins, leading to the accumulation of somatic mutations over the lifetime of the individual. These mutations range from single nucleotide polymorphisms to large scale structural variations. While their contribution to the oncogenesis can be deduced in some cases, and perhaps even more remarkably, their signatures sometimes give away the mutagenic factors which caused them [2], making the distinction between “driver” and “passenger” mutations is by no means trivial, especially in the case of large scale structural variations which do not always have a direct genetic linkage and therefore must be analyzed on the genomic scale.

Even simple structural variations such as insertions, deletions or inversions may have radical phenotypic effects, for instance, the formation of fusion (onco-) genes or loss of function in cancer suppressor genes [3]. The rearrangement of sometimes distant parts of the genome may join the sequences of two different genes to create a fusion gene or it may position the cancer gene adjacent to regulatory elements from elsewhere in the genome, resulting in abnormal expression patterns [1]. For example, clones obtained from the MCF-7 breast cancer cell line genome were found to have a complex internal structure [4], with some genomic regions extensively scrambled around a remarkable number of breakpoints. The rearrangement process is frequently associated with gene amplification, a somatically acquired increase in copy number of restricted genomic regions also known as amplicons.

III. Rearrangement analysis

A. *End-sequence profiling*

While several techniques outside the scope of this paper exist for genomic rearrangement analysis (e.g. comparative genomic hybridization, CGH), high resolution mapping of rearrangements is made possible only by sequencing techniques. The paper [4] describes a widely used method known as end-sequence profiling (ESP). Such an analysis begins with the cloning of the genome of interest into large vectors called bacterial artificial chromosomes (BAC). Subsequently, both ends of these clones are sequenced using traditional Sanger sequencing, and the sequence obtained is mapped back to the reference genome. The size distribution of the clones is known, therefore outliers (clones whose ends map unexpectedly close or far from each other) can be identified. These clones, containing a putative genomic rearrangement, can then be sequenced in their full length and rearrangements contained therein can be analyzed.

As Sanger-sequencing a full human genome is prohibitively expensive, the ability to sequence only the clones that are likely to contain a rearrangement breakpoint is a significant advantage of the ESP technique. A further advantage is due to the relatively long read lengths inherent to Sanger sequencing: the sequence of the clones can be determined with high confidence, resulting in base-pair level identification of breakpoints. On the other hand, rearrangements significantly smaller in scale than the size of the BAC clones may remain undetected as their effect in end mapping distance remains within the distribution of normal BAC end distance. Furthermore, full sequencing of the BACs is expensive, so only genomic regions of interest are mapped to a high resolution. The mapping distances of *all* the BAC ends *together* could be used to reconstruct at least large-scale rearrangements genome-wide; however, this was found to be a challenging computational problem [5].

B. *Next-generation sequencing methods*

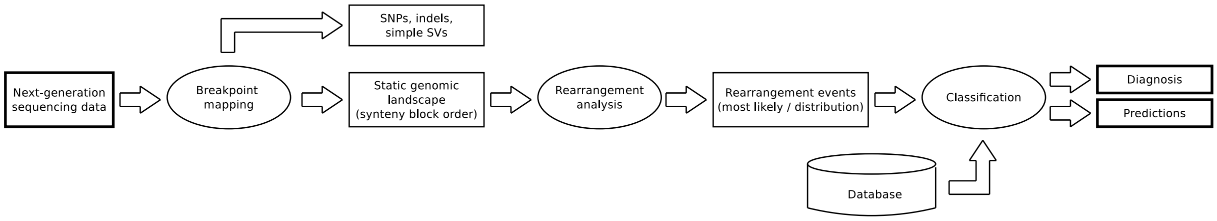
Complex rearrangements have also been observed in NGS studies ([2]). While the ESP method allows for selection of relevant clones for capillary sequencing and thus for exact breakpoint localization, the short read technology involved in NGS methods makes the identification of complex rearrangements rather difficult (see [6] for a review). Most studies for automated structural variation (SV) discovery used a paired-end mapping strategy: both ends of short (200 - 1400 bp) fragments of DNA are sequenced and the reads are mapped to the reference genome. Abnormally mapping reads are compared to known signatures of simple insertions, deletions, inversions, insertions of longer segments of distant or novel sequence and translocations.

The detection of fingerprints of such simple SVs may be based on heuristics, possibly complemented with exact treatment of measurement error ([7]), or probabilistic (e.g. [8]) algorithms. While it is possible to manually detect complex rearrangements from discordantly mapping reads, neither the aforementioned methods nor any others I am aware of are capable of the automated description of a rearranged genomic landscape.

With short reads long enough to be uniquely aligned in part to the genome split-read alignment becomes possible. This technique can be used for direct detection of SV breakpoints captured by a single read. It is commonly used for identification of insertions and deletions, both as separate tools (e.g. [9]) or as a step of more comprehensive SV analyses. With high enough sequence coverage split-read mapping may also be used to map the breakpoints of other, perhaps more complex structural variations, including genomic rearrangements (e.g. [10]). These studies, however, only used these breakpoints to select genomic regions for further analysis by more precise experimental methods, and to the best of my knowledge, no split-read mapping approach for automated rearrangement discovery exists.

IV. The proposed framework

Our framework aims at the classification of individual tumor genomes based on their next-generation sequencing data. This process consists of three distinct steps, whose relation as well as their inputs and outputs are illustrated in Figure 1.



SNPs, indels, simple SVs
Next-generation sequencing data
Breakpoint mapping
Static genomic landscape (synteny block order)
Rearrangement analysis
Rearrangement events (most likely / distribution)
Classification
Diagnosis
Predictions
Database

Figure 1: Overview of the framework

The first step is the mapping of the rearranged genomic landscape from the short read data. The output of this preprocessing step is the relative order of conserved (so called synteny) blocks of the genome. Our method is based on split-read mapping, but unlike existing methods, it is designed for automatic reconstruction of the synteny block order. Reads which can be divided in two parts mapping to remote locations of the reference are referred to as *bridge reads*. Every such bridge read is considered as an evidence supporting an adjacency relation between the two genomic intervals ending at the breakpoint of the read.

The bridge reads are represented in a graph whose vertices belong to the interval ends, and edges represent the aforementioned adjacency relations. If we add an extra vertex for every interval (connected to its endpoints), symbolizing the passage through the interval, the problem of finding the order of synteny blocks reduces to finding a Hamiltonian path in this graph.

Obviously, one faces challenges applying this formal model to real-world data. First, next-generation sequencing data is inherently noisy: reads generated from repetitive or homologous regions of the genome map to several different locations, and introduce false adjacency edges to the graph. It is therefore very likely that multiple Hamiltonian paths exist in the graph, and our method must choose the one that is supported by the most reads. Second, the number of intervals and adjacencies in a real genome is rather high, which questions the feasibility of this NP-complete model. However, we found that the structure of the graphs obtained is surprisingly sparse, and can be further simplified by removing vertices with an unrealistically high number of putative neighbors (likely having resulted from repetitive regions). An integer linear programming formulation of the Hamiltonian path problem on such graphs was found to be soluble in a matter of seconds (simulated rearrangements on *E. coli* data).

Having determined the synteny block order, the rearrangement steps that could have lead to the given static landscape must be identified. We use the most general model which considers inversions, translocations, fusions, fissions, duplications and deletions. Due to the exponential complexity arising when one attempts to reconstruct the rearrangement process with such a general model, we employ a Markov-chain Monte Carlo method to sample from the distribution of the possible rearrangement pathways. In addition to its ability to provide a high-probability solution within reasonable time limits, this stochastic approach also allows for reporting not only the most parsimonious rearrangement scenario, but rather the probability distribution of several most likely ones. This makes our approach more robust in cases with not only a single reasonable rearrange pathway.

With the dynamics of rearrangements known the genome can be compared to known cancer genomes for classification. This last step is yet to be developed.

V. A simulator

Breakpoint detection from real next-generation data is a rather difficult task. Before having arrived to our current method we evaluated several candidate techniques and beyond any doubt we will keep doing so in our further attempts to refine our framework. The analysis of a given idea is much easier if one can easily implement it with the need to care for as few irrelevant details as possible. To facilitate this rapid development we designed and implemented a simulator program.

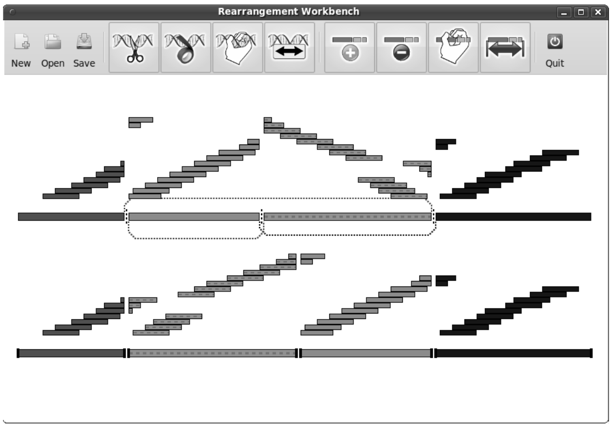


Figure 2: Screenshot of the simulator

The simulator has a graphical user interface (Figure 2) which allows the user to introduce inverted and non-inverted rearrangements to a hypothetical genome. The dual view allows for simultaneous inspection of the reference and the rearranged genome. One can map reads to the sample and see their mapping to the reference; various properties of the reads are customizable, and even mate-pair reads are supported.

The scenario set by the user can be exported to a file for the breakpoint detection software. However, the simulator allows for the implementation of such algorithms within the framework itself. This way, the input data along with the expected output is readily available for the algorithm which allows for short modification cycles, resulting in better interactive evaluation.

References

[1] M. R. Stratton et al., "The cancer genome.," *Nature*, 458(7239):719–24, 2009.

[2] E. D. Pleasance et al., "A small-cell lung cancer genome with complex signatures of tobacco exposure.," *Nature*, 463(7278):184–90, 2010.

[3] M. J. Clark et al., "U87MG decoded: the genomic sequence of a cytogenetically aberrant human cancer cell line.," *PLoS Genet.*, 6(1):e1000832, 2010.

[4] S. Volik et al., "End-sequence profiling: sequence-based analysis of aberrant genomes," *Proceedings of the National Academy of Sciences of the United States of America*, 100(13):7696–7701, 2003.

[5] B. J. Raphael and P. A. Pevzner, "Reconstructing tumor amplisomes.," *Bioinformatics*, 20 Suppl 1:i265–73, 2004.

[6] P. Medvedev et al., "Computational methods for discovering structural variation with next-generation sequencing.," *Nat. Methods*, 6(11 Suppl):S13–20, 2009.

[7] S. Sindi et al., "A geometric approach for classification and comparison of structural variants.," *Bioinformatics*, 25(12):i222–30, 2009.

[8] F. Hormozdiari et al., "Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes.," *Genome Res.*, 19(7):1270–8, 2009.

[9] K. Ye et al., "Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads.," *Bioinformatics*, 25(21):2865–71, 2009.

[10] R. Bueno et al., "Second generation sequencing of the mesothelioma tumor genome.," *PLoS ONE*, 5(5):e10612, 2010.