Statistical method for filtering sequencing error from minor clonal mutation in sequencing data and implementation

Vojtěch Kulvait [1,3] Kateřina Machová Poláková [2] Tomáš Stopka [1,3]

[1]First Faculty of Medicine, Charles University in Prague [2]The Institute of Hematology and Blood Transfusion, Czech Republic [3]PersMed s.r.o., Czech Republic

Introduction

Clonal disorders (cancer, leukemia, CML)

- Patients may develop multiple clones with different aggressiveness and response to the treatment.
- Presence of particular clones, particularly in CML, indicates a need for therapeutic intervention.
- Different therapeutics target different clones.
- Clones may be present in low abundances in samples.
- There is a need for early detection of clone development.

Typical sequencing data

- Data are sequenced with high coverage (500x or more).
- Only particular regions of interest are sequenced, amplicon sequencing.
- It is hard to distinguish between an error and actual presence of a subclone.
- PCR based preparation of samples possesses errors that do not respect Poisson distribution.
- Some error types (AG) are more frequent than others (AC).

Objectives

Filtering relevant mutations

- Algorithm should detect SNS errors.
- Current approaches to mutation detection are based on heterozygosity (50% abundance) or homozygosity (100% abundance) assumption.
- Final method have to be sensitive enough not to discard relevant information.

Modeling error occurrence

For given sequencing technology, sample processing and data analysis pipeline, we model probability of an SNS error in data as a function $p = p(r_n, m_n, N, N_m)$, where:

- r_n ... reference nucleotide at given position,
- m_n ... mutated nucleotide at given position,
- N ... number of reads,
- N_m ... number of reads with given mutation.

For given number of reads N per base we then approximate function p by fitting negative binomial distribution. Fitting is done using control samples for which no mutations are expected to be present. Control samples should be processed the same way as the samples of interest.

Testing data

Control samples and CML samples from Roche 454 amplicon sequencing of BCR-ABL gene.

SNS type	Mean	Variance
AG & TC	**9.169**	**19.626**
CT & GA	**2.460**	**4.717**
AT & TA	0.457	0.886
AC & TG	0.245	0.375
CG & GC	0.147	1.593
CA & GT	0.113	0.191

Table 1: Errors per 3000 reads in control samples.

Filtering algorithm

For a certain candidate mutation, read depth N, N_m reads supporting mutation and p-value level P we call a mutation significant if

$$p(r_n, m_n, N, N_m) < P. \quad (1)$$

Multiple hypothesis testing

Since the data are acquired from amplicon sequencing, it is natural to perform multiple hypothesis testing adjustment.

Implementation

Algorithm has been implemented in Java. It has been implemented as part of software to run the whole processing pipeline. This software has XML configurable front end to run particular commands from command line.
We are discussing possible implementation of the software or this particular algorithm as part of Galaxy framework.

Algorithm validation

In a recent publication [1] we discuss the application of the algorithm on the data from NGS sequencing of BCR-ABL kinase domain in CML samples. The work describes impact of mutations in BCR-ABL in CML on a drug selection as certain BCR-ABL kinase inhibitors fail to be effective for particular mutation subclones.
In the recent validation study samples from CML patients independently evaluated in two distant sites at Czech Republic and Italy were used. Control samples were also independently prepared. Data processing and also the application of the described algorithm were then performed for both data sets in the same way. From 22 samples and 24 total mutations there was 92% agreement in mutation detection and 90% agreement in mutation profile detected from the sample. Consistent results were obtained in a recent extension of the study where the data obtained in another site in Germany were added.

Outlook

We are working on these possible extensions:

- Test the algorithm on Illumina data, extend testing datasets.
- Use mutational profile of whole reads to distinguish more complex subclones.
- Implement the algorithm using Galaxy framework.

References

[1] Katerina Machova Polakova, Vojtech Kulvait, Adela Benesova, Jana Linhartova, Hana Klamova, Monika Jaruskova, Caterina de Benedittis, Torsten Haferlach, Michele Baccarani, Giovanni Martinelli, Tomas Stopka, Thomas Ernst, Andreas Hochhaus, Alexander Kohlmann, and Simona Soverini.
Next-generation deep sequencing improves detection of BCR-ABL1 kinase domain mutations emerging under tyrosine kinase inhibitor treatment of chronic myeloid leukemia patients in chronic phase.
Journal of Cancer Research and Clinical Oncology, 141(5):887–899, November 2014.

Acknowledgement

This work was supported by BIOCEV - Biotechnology and Biomedicine Centre of Academy of Sciences and Charles University in Vestec, project supported by the European Regional Development Fund.

E-mail

- kulvait@gmail.com