Poster J-7

Statistical Analysis of Retroviral Insertional Mutagenesis Data

Authors:
Jeroen de Ridder (*Information and Communication Theory Group, Delft University of Technology*)
Jaap Kool (*Division of Molecular Genetics, The Netherlands Cancer Institute*)
Anthony Uren (*Division of Molecular Genetics, The Netherlands Cancer Institute*)
Lodewyk Wessels (*Division of Molecular Biology, The Netherlands Cancer Institute*)
Marcel Reinders (*Information and Communication Theory Group, Delft University of Technology*)

Short Abstract: The presence of multiple mutations within one tumor hints toward cooperation between the mutated genes. We propose a two-dimensional Gaussian Kernel Convolution method, a computational technique that identifies the cooperating mutations in mutagenesis data. This resulted in the discovery of new regions in the genome that may be involved in tumorigenesis.

Long Abstract:

Introduction

In retroviral insertional mutagenesis experiments, genes involved in the development of cancer are identified by determining the loci of viral insertions from tumors induced by retroviruses in mice. After infecting a host cell, the retrovirus inserts its own DNA into the host cell's genome, mutating the host cell's DNA in the process. The mutation may cause alteration in expression of genes in the vicinity of the insertion or, when inserted within a gene, alteration of the gene product. When the affected gene is a cancer gene, activation of a proto-oncogene or inactivation of a tumor-suppressor gene can cause uncontrolled proliferation (cell division) of cells. Eventually this may give rise to tumors.

From a statistical point of view the challenge is to find the regions in the genome that carry insertions in multiple independent tumors significantly more frequently than expected by chance. Such a region is called a Common Insertion Site (CIS), and its location is highly correlated with the location of genes involved in tumor development. In most cases, however, the oncogenic state is reached by an accumulation of multiple independent mutations, rather than a single insertional mutation inducing proliferation. Therefore, cooperation between virally targeted genes, also play an important role in tumor formation.

Over the last few years an extensive amount of insertional mutagenesis data has been published, many of which are compiled in the Retroviral Tagged Cancer Gene Database (RTCGD) (URL: http://RTCGD.ncifcrf.gov). At this moment the database contains approximately 4000 insertions from 1076 tumors. The vast majority of these insertions have been acquired in twenty different screens, using different genetic backgrounds (different tumor types, knock-outs, etc). In this study we analyze the combined data from all the screens in the RTCGD, irrespective of the genetic background or cancer predisposition of the mice used in the screens.

Because the data from the RTCGD is far from saturated (there are relatively few tumors), not all oncogenic regions are hit frequently enough to be labeled a CIS. Still these insertions may play a role in the tumor development, for instance by cooperating with another insertionally targeted gene. It is therefore important that in co-mutation analysis not only CISs are considered, because the insertions that did not contribute to a CIS may be, in combination with another insertion, observed significantly more frequently than expected by chance. Therefore, we propose to analyze the insertion data in the co-occurrence space. We define an Insertion Co-occurrence (IC) as a unique combination of insertions within one tumor, and the Common Co-occurrence of Insertions (CCI) as observing the combination of two insertions significantly more frequently than expected by chance across multiple tumors.

Methods

We propose a two-dimensional Gaussian Kernel Convolution method (2DGKC), a computational technique that identifies the cooperating mutations in insertional mutagenesis data. Similar to the Parzen window approach, the method positions a two-dimensional Gaussian kernel function at every IC in the cooccurrence space and sums the result. For the regions with a high density of ICs the kernel functions will overlap and therefore result in high peaks. Now, the regions with a significantly increased number of ICs are determined by evaluating the statistical significance of observing a certain peak height. For this purpose a null distribution is computed on the collection of peaks resulting from the 2DGKC applied to a randomly permuted version of the insertion data. Peaks exceeding the alpha-threshold can now be labeled CCIs, signifying the co-mutations that are statistically significant across all different screens in the RTCGD. We correct for the increased probability of detecting false CCIs due to large number of ICs.

Results and Conclusions

Until now, the main focus of analysis on insertional mutagenesis data has been one-dimensional, that is, discovering regions in the genome that are causal for tumor development, the CISs. The multidimensional analysis of the insertion data, carried out in this study, results in the discovery of statistically significant co-mutations, indicating the presence of cooperating oncogenes that play a role in tumor development and would have gone undetected in a one-dimensional analysis. We found 202 statistically significant Common Co-occurrences of Insertions (CCIs). Apart from known interactions such as the interaction between Myc and Pim1, also unknown interactions are discovered.

The methods presented are especially beneficial for data from high throughput screens with many insertional mutations per tumor. Therefore, the methods may be applied to other types of genome wide mutagenesis data as well, for example data from transposon screens. As the amount of data increases, extensions to higher dimensions become very interesting. In the framework we created, these extensions are fairly straightforward.