32. ELEANALYZER:an automated method for analysis of retrotransposons in genomes and prediction of their insertion sites

Kamal Rawal [1], Alok Bhattacharya1 [3] Sudha Bhattacharya [2] and Ram Ramaswamy [1, 4]

[1] Center for Computational Biology and Bioinformatics, School of Information Technology, Jawaharlal Nehru University, New Delhi, India
[2] School of Environmental Sciences, Jawaharlal Nehru University, New Delhi, India.
[3] School of Life Sciences, Jawaharlal Nehru University, New Delhi, India.
[4] School of Physical Sciences, Jawaharlal Nehru University, New Delhi, India.

ELEANALYSER identifies unoccupied insertion sites, characterizes occupied insertion sites and finds their distribution. The algorithm incorporates Bayesian scoring, support vector machine and boosting. A webserver has been developed for automated analysis of truncation hotspots. It also allows examination of the occurrence of insertion elements in the vicinity of genes.

Retrotransposons are mobile genetic elements which can amplify from one location of genome to another using an RNA intermediate. They are present in varying extent in eukaryotic genomes including parasite called E. histolytica. This parasite is the causative agent of amoebiasis, a highly prevalent disease in developing countries. Autonomous Non-LTR elements encoding their own retrotransposition machinery are called as Long Interspersed Elements (LINEs). Short nonautonomous elements that borrow this machinery for propagation are called Short Interspersed Elements (SINES). They plays important role in genome by producing variety of effects like genome expansion, insertional mutagenesis of genes leading to inherited disorders, unequal homologous recombination events and can lead to transduction of sequences downstream to them leading to exon reshuffling. They have also been shown to influence the expression of genes flanking their sites of insertion and SINEs have been implicated to work as stress sensors in the cell giving a function to otherwise known as “junk” DNA.

Here we report a computational method called as ELEANALYSER pipeline for automated analysis of retrotransposon elements in genome which attempts to predict unoccupied insertion sites, identify and characterizes occupied site of an element, finds their distribution, predicts truncation hot spots, finds out relationship of element with genes and other instances of elements, and examines disruption of genes by them.

ELEANALYSER can be considered as series of programs written in Perl/Bioperl along with local version of Blast running in a sequential mode which attempts to answer a biological question at each step. The pipeline also incorporates number of genomes such as Human and E. histolytica along with elements such as L1 and EhLINEs for searching. The program filters redundant hits by comparing 5’ flanking sequences via pairwise global alignment and pairs with greater than 90% identity are deemed redundant. The program inspect flanking region for the occurrence of another instance of element and it also examines for presence of gene through BlastX search against nr database.

Applying this program on recently sequenced E. histolytica genome we found that there are 3 classes of LINEs (EhLINE 1, 2, 3) and two classes of SINEs (EhSINE1 and 2) with a third class of EhSINE that appears to be present in a single copy (1). The EhLINE1 is present in most copies (409) in genome whereas EhLINE3 have least number of copies (52). It might be possible that EhLINE3 is on its way out of genome because of most of its copies have accumulated too many mutations. Most of the copies of EhLINEs and EhSINEs are truncated. A novel feature was the presence of two types of members- some with a single long ORF (less frequent 20%) and some with two ORFs (more frequent 80%) in both EhLINE1 and 2. The two ORFs were generated by duplication of 5 nt (AAGCA) leading to stop codon which was confirmed subsequently by RTPCR experiments. The EhLINEs/SINEs together account for 6 % of the E. histolytica genome. Computational analyses and Southern hybridization studies of PFGE separated chromosomes of E. histolytica with EhLINE1 and EhSINE1 probes showed that these elements reside on all chromosomal bands, do not seem to be telomeric, and might be dispersed in the E. histolytica genome.

However, all the elements seemed to insert in AT-rich sequences, with a clear preponderance of T-residues in a 50-nt stretch upstream of the site of insertion of each element. In most (80%) of the cases, the EhLINEs/EhSINEs were associated with either genes or another instance of elements. In general, 50% of the time, when a gene was found near an element, it was present within the first 0.5 kb, whereas an element was found within the first 0.1 kb. Such close proximity of elements to one another could be due to clustering of sequences that serve as favorable target sites for insertion. Amongst the genes present near the elements, most (about 63%) were found to be hypothetical. Other genes found in the vicinity were Kinases, GTPases, and heat shock proteins but house keeping genes were rarely found. No instance was encountered of an element inserted within a gene. Presence of hundreds of copies of EhSINE3 (1 copy in E. histolytica) in non pathogenic type called Entamoeba. dispar and elements' insertion near the coding region raises the possibility that these elements may influence the phenotype of E. histolytica and the pathogenesis of amoebiasis.

Using a particular module called Insertion site finder (ISF), we found that insertion sites contains complex patterns in terms of physical properties of DNA such as local bendability, A philicity, & protein induced deformability and thermodynamic properties such as stacking energy, & DNA denaturation. We report presence of combination of these patterns serve as signals for insertion. We have developed a scoring system which incorporates diverse signals using Bayes' rule to compute the posterior probabilities of these signals from the training set consisting of already occupied sites and negative set which contained set of randomized sequences. The scores for different signals were then used to develop a composite signal by combining them with different coefficients using modified Adaboost algorithm. The support vector machines were used an additional machine learning system for classification of insertion sites.

The sensitivity and specificity of ISF was calculated using two independent test sets containing positive and negative examples respectively. The accuracy level of greater than 91 % was obtained during testing of ISF. ISF was used to screen unoccupied sites in the E. histolytica genome. Computationally predicted sites were verified through in vitro experiments where endonuclease seems to nick predicted sites preferably. Similar signals were seen in L1 element of human raising the possibility that these signals are present in other genomes also.

We are characterizing more elements in other genomes. A user friendly web server is also being developed.

Reference:

Bakre A. & Rawal K. et al, The LINEs and SINEs of Entamoeba histolytica: Comparative analysis and genomic distribution. Experimental Parasitology. 2005 Jul;110 (3):207-13.