Integrating current knowledge to predict mutations effect on splicing.

ID - 227

Cereda Matteo[1], Sironi Manuela[1], Comi Giacomo P[2], Pozzoli Uberto[1]

[1]Bioinformatics, Scientific Institute I.R.C.C.S. E.Medea

[2]Dino Ferrari Centre, Department of Neurological Sciences, University of Milan, IRCCS Ospedale Maggiore Policlinico, Mangiagalli and Regina Elena Foundation , 20100 Milan, Italy

Motivation

The production of functional messenger RNAs in metazoans is critically dependent upon the accuracy of pre-mRNA splicing, a highly regulated process assuring that introns are removed and ordered array of exons is maintained in mature transcripts. The process requires an accurate recognition of exons, in particular a precise determination and pairing of 5' and 3' splice sites by the splicing machinery. Over the past few years several approaches have been explored to allow for the exact detection of splice sites and an increasing number of sequence elements have been identified that are involved in the splicing mechanism and in its regulation.

In this work we wished to test whether a tool can be built that uses the known splicing determinants to predict the effect of mutations/variations on splicing. A similar tool can be very effective when screening for mutations in non coding regions by a priori selecting region/positions that are predicted, if mutated, to yield splicing aberrations; moreover, it can give useful information when missense mutations have already been identified and, more in general, can be used to predict the possible effects of SNPs.

Methods

To create the exons database, protein coding genes annotated in GenBank were selected considering a single id for each transcription cluster. The data set was composed by 134623 canonical exons (AG-GT). A great number of known elements that take part in splicing regulation was identified in the exon set. In particular, we evaluated: consensus values, exon length (EL), Branch Point position (BPP), Exonic Splicing Enhancers motifs (ESE), Exonic Splicing Silencers (ESS), UAGG and GGGG motifs.

Frequency distributions were evaluated for exon size, branch point distance and relative element abundance. A number of intronic sub-sequences (pseudo exons) were selected whose flanking regions resemble splice sites (a threshold of 0.6 on consensus values was considered) and whose length was comprised between the 0.005th and 0.995th quantiles of exon length distribution. For both exons and pseudo-exons EL, BPP, ESE, RESE and ESS were scored according to frequency distributions in the real exons set. A total of 8 independent set of data were created each containing scores from 10000 real and 10000 pseudo exons. To extract the most relevant features of this set we choose a Recursive Feature Elimination (RFE) method built upon a Support Vector Machine technique with a linear kernel. RFEs were performed with Leave-One-Out cross-classification approach on 6 independent sets. Once determined the most relevant variables, a SVM with radial kernel and 500-fold cross-validation was trained on an set different from the ones used in RFE to identify the model. Another set was used as test set to measure performance in terms of sensitivity, specificity, accuracy and Matthews correlation coefficient (MCC). A preliminary set of experimentally well studied splicing mutations was then analysed comparing the predictions of this SVM model on wild type/mutated sequences in order to verify if the predictions of our model correspond to the described splicing aberrations.

Results

The RFE technique applied to the 6 independent datasets allowed the identification of 7 features that resulted most important in each set. These features were then used to train the SVM with radial kernel obtaining a final model (sensitivity= 0.96, specificity=0.93, accuracy=0.95 and MCC=0.89). The model was then used to predict mutation effects on splicing (see table 1): out of 15 studied mutations 10 were completely predicted, in one case the effect was only partially predicted, in 2 cases the predicted effect was different from the described one and in 2 cases no effect was predicted. The preliminary results are quite encouraging and deserve further investigations. A greater number of cases should be collected to better test the effectiveness of the model and different strategies in features selection should be investigated.

Availability: http://bioinformatics.emedea.it

Image: http://bioinformatics.emedea.it/images/tab1.png

Email: matteo.cereda@bp.lnf.it