Identification of couples and triples of co-occurring transcription factor binding sites sequences from relative distance and GO filter.

ID - 114

Pesando Igor[1,2]
[1]Department of Theoretical Physics, University of Torino, Italy
[2]INFN, sezione di Torino, Italy

Motivation

Transcription regulation in eukariotes is known to occur through the coordinated action of multiples transcription factors (TFs).

The problem is then both to find the TF binding sites (TFBSs) and how TFs interact. The interaction can be either homotypic, involving binding of only one TF to multiple sites in an ORF upstream, or heterotypic, involving of more than one TF.

Despite many cases are heterotypic most of the methods developed are targeting the homotypic case.

In this work we present a novel computational method for the identification of both TFBSs and the search for heteotypic cooperation between them in yeast genome.

Methods

We propose the following approach to cope with the previous problems.

We have developed a distance distribution between TFBSs which takes also care of the multiple occurrences of the same TFBS in one single ORF upstream.

We have then verified the accuracy of this distance distribution on randomized data.

Since this distribution is entropyc in nature it works well when either we consider a word, i.e. a sequence of basis, and its reverse complement to be or not to be the same BS.

Using this distance distribution we have selected non overlapping words couples of given length as well as non overlapping words triples (of more limited length, say 5/7 bp per word because of the computation time) with a FDR of about or less 0.01 with respect to the null model made of randomized upstreams.

To these couples and triples we could associate sets of ORFs chosen accordingly to where the given couples/triples appear.

These sets have been filtered by means of their GO annotation using an algorithm by Cora' et al. and only those with FDR less than 0.01 were kept. The FDR has been established by randomly generating samples where the fake sets have the same numbers of ORFs as the real data.

In this way we could associate some putative GO terms to couples/triples.

We will now explain in more details the previous steps.

The selection operated using the distance distribution cannot be done considering directly the probability of a configuration of words since it depends on how many 'interactions' there are between these words, i.e. on how many ORF upstreams they can be found at the same time, on how many occurrences the words have on each ORF upstream and how distant they are.

We have then decided to consider as biological significant the number of interactions and therefore we have numerically determined which value of the probability corresponds to a FDR of about 0.01; this value has been chosen to have a quite reliable value for the probability cutoff.

Afterwards we have verified that restricting the couples/triples to those with lesser distance probability the corresponding FDR for the ORFs sets passing the GO filter improves.

We have also verified that better FDR are obtained when restricting our attention to more specific GO terms.

We have then chosen the couples/triples so that the FDR for ORFs set was less than 0.01.

Results

We have applied the previous approach to the yeast Saccharomyces cerevisiae.

Using method describe before based on the distance distribution we have selected non overlapping words couples of length 6 and 7 bp as well non overlapping words triples of length of 5,6 and 7 bp with a FDR of about or less 0.01 from S, cerevisiae 500bp ORF upstreams.

In the case of triples of words of 7 bp the space of possibilities is more than 5*10**10. The time required for the complete analysis of couples is of the order of some hours while it is of days for the 6 bp words triples. To assess our method we have compared the putative couples and triples with Lee's ChIP-CHIP data and the cooperativity analysis already performed in literature.

In the case of triples of 7bp length words the triples passing the distance filter with a FDR of 0.01 are 363088. To obtain a FDR of 0.01 in the GO filter we selected the 290049 best triples corresponding to the 80% of triples chosen according to the best distance probability.

Of these triples 599 could be assigned to at least one GO term whose size was less than the 10% of ORFs and with a Pvalue of bigger than 10**(-6).

Most of these triples are associated with rRNA processing, nucleolus and similar terms.

Accordingly the only 22 out of 599 triples are associated with a Bonferroni corrected Pvalue less than 0.05 with TFs involved in the cell cycle using ChIP-CHIP data.

By far more interesting results are obtained considering triples with 5bp words and couples with 6bp words.

In the latter case starting with 4766 couples passing the distance filter we find 294 couples after the GO filter out of which 85 are associated to at least one TF using ChIP-CHIP data.

The strongest signals of our analysis are associated with the TFs RAP1-FHL1, the cell cycle TFs MBP1-SWI6-SWI4 and HIR1-HIR2.

Nevertheless we have associated 30 out of 107 TFs using the Bonferroni correction and about 60 using the Benjamini-Yekutieli one (which is stricter than the Benjamini-Hochberg).

Availability: http://www.to.infn.it/~ipesando/genomics/nples_yeast/supplementary.html

Email: ipesando@to.infn.it