Correlation between structural properties of promoters and physico-chemical features of Transcription Factors

ID - 240

Brilli Matteo[1], Fani Renato[1], Lio' Pietro[2]
[1]Dept. of Animal Biology and Genetics, University of Florence
[2]Computer laboratory, University of Cambridge

Motivation

Biological mechanisms underlying the regulation of gene expression are not completely understood. It is known that they involve binding of transcription factors (TF) to regulatory elements (motifs) on gene promoters. However, attempts to computationally predict such elements in promoters typically yield an excess of false positives, because regulatory motifs are short, degenerated and embedded into large regions of non--coding DNA. They may be shared by co--regulated genes thus modulating in a similar way the expression patterns of these genes.

The analysis of expression data allows the identification of co--regulated genes, likely controlled by common regulatory mechanisms.

Several authors have shown that curved regions are often found near functionally important sites such as promoters and origins of replication (Kanhere et al., 2004, Bolshoy et al., 2000, Kozobay et al., 2006). Detailed studies of hydrogen bonding taking place between a transcription factor and its targets have notably shown that, while specific side chain-base interactions certainly exist, the majority of hydrogen bonds involve the DNA backbone, underlining the importance of the overall DNA conformation within a complex and implying that sequence-dependent changes in DNA structure, or in its mechanical or dynamic properties, might also play a role in recognition. The distribution of curved DNA in promoter regions is evolutionarily preserved, and it is mainly determined by temperature of habitat (i.e Bolshoy et al., 2000). Experimental evidence has demonstrated contributions of DNA curvature in regulating the transcription of several genes and have led to the idea that recognition generally involves both direct (hydrogen bonding, steric fit, etc.) and indirect (DNA structural adaptation) components. DNA is negatively charged and those residues in a TF which contribute to its overall net charge or dipole moment might also be important for target motifs recognition.

The dipole is the first moment of the charge distribution around the rotational centre of drag which is estimated as the geometric centre of all atoms weighted by their van der Waals radius.

We show a study aimed at characterizing the relationships existing between DNA curvature and dipole moment of TFs.

Methods

Structural analysis of promoters:

Here we present an implementation of MotifScorer (Brilli et al., 2006) with built-in curvature tables which are used to characterize the structural properties of promoter regions. These structural data can be used to characterize the regions surrounding regulatory motifs identified using motif finding algorithms and Partial Least Squares regression against a compendium of expression data.

Dipole moment calculation:

A plausible net charge of proteins can be calculated summing up the number of charged residues: +1 for Lys, Arg and His; -1 for Asp and Glu which can be taken as the mean charges at pH=7. A much more accurate statistics can be obtained if atomic coordinates of a protein are available, allowing the computation of the dipole moment of a protein. We have implemented an algorithm written in Java that uses Matlab and Colt library which computes the protein radius and the pH dependent partial charges of all polarised atoms and ionised groups. Missing hydrogen atoms are constructed according to the equivalent bond lengths in small organic molecules.

Results

Using this refined algorithm we have found a bivariate curve of the product of dipole and charge i.e. most of the transcription factors and other DNA binding proteins have either a large positive charge or a statistically larger dipole, or the product of the two with respect to a data set of non transcription binding proteins.

The analyses we have carried out using the human proteome shows that the average of net charges of all proteins follows a gaussian distribution with mean +11.

Interestingly charge distribution of eukaryotic proteins was broader than bacteria (Escherichia coli + Haemophylus influenzae) where the distribution was much narrowed (from -10.5 to +16) with the mean value of about 5. The reason for such differences is that there are many more gene duplications in Eukarya which have a large number of families of DNA protein binding sites.

As expected, the net charge of a protein strongly depends on pH. At low pH many DNA binding proteins are highly charged while at bacterial intracellular physiological pH they maintain a small positive charge. Conversely, the dipole moment is less sensitive on pH, suggesting that it might be more important than charge, particularly in orienting the protein with respect to DNA.

Moreover, we discuss the correlation existing between curvature and the dipole moment of transcription factors in Eucaryotes and Prokaryotes.

Email: matteo.brilli@dbag.unifi.it