How to know transcription factors by the company they keep

The ENCODE project is a massive data-collection effort set out to understand the function of the human genome. The collection comprises many types of genomic data, including the localization of transcription factors onto DNA. Transcription factors are proteins that bind to specific patterns of DNA (called recognition motifs) to control how genes are turned on or off. The way this function is achieved is still unknown. In order to gain insights into this mechanism, we analyzed ENCODE data to study how transcription factors interact together with specific DNA regions.

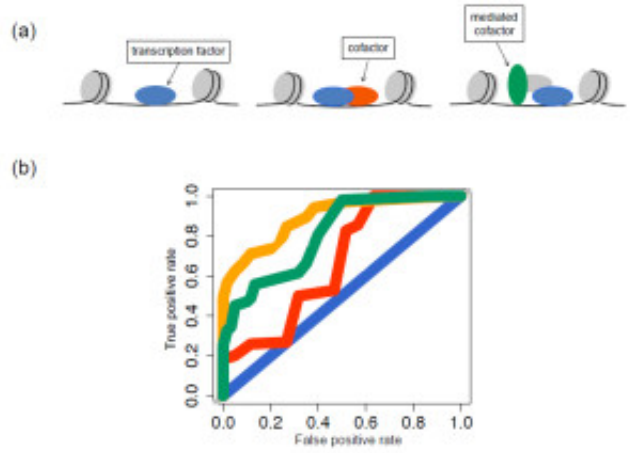

Fig. 1. (a) Graphical representation of transcription factors partners and binding sites onto DNA. In addition to transcription factors (in blue), partners (cofactors, in orange; mediated cofactors, in green) can associate with them onto specific DNA regions. (b) Performances of PAnDA method on ENCODE data. The integration of the network for the prediction improves the performances (from blue to yellow, increasing performances towards more correctly identified binding sites [True positive rate] and less incorrectly identified ones [False positive rate]).

We found that the association of multiple transcription factors (called network) is a fundamental feature to explain their localization onto DNA. We developed a computational method that uses this feature to predict where a transcription factor will localize in the genome. This tool is called PAnDA (Protein And DNA Associations).

The PAnDA method predicts the ability of human transcription factors to localize onto DNA. The key ingredients of PAnDA predictions are three: the recognition motifs of transcription factors, their abundance in the cell, and their partners (called cofactors and mediated cofactors, as shown in the Fig. 1).

Thanks to PAnDA we found that not only transcription factors but also their partners recognize specific DNA patterns. This observation led to the idea of including the network information in the computational model. We analyzed more than 400 ENCODE datasets and we reached a

performance of more than 80% in localizing transcription factors binding sites. This result shows that our observations are statistically robust and well described by a theoretical framework.

We also tested PAnDA method on ENCODE datasets of transcription factors without known recognition motifs. Remarkably, PAnDA reaches a performance of 82%. Overall, PAnDA approach highlights that the network itself contains enough information to localize transcription factors on DNA even in absence of known recognition motifs.

The most innovative aspect of our work is that it introduces a cell-specific view of transcription factors networks, which opens up the way for efficient and effective manipulation of cellular processes. Our findings are of great practical relevance to a number of research lines, from engineering expression of genes to somatic stem cell reprogramming. PAnDA tool will raise new fundamental questions in the field and will inspire future research on topics like the evolution of regulatory networks and the formation of macromolecular complexes. In summary, we better understand transcription factors by taking into account the company they keep.

Publication

By the company they keep: interaction networks define the binding ability of transcription factors.
Cirillo D, Botta-Orfila T, Tartaglia GG.
Nucleic Acids Res. 2015 Oct 30