Transcription Factor Binding Site-based Alignment of Conserved Non-coding Elements

Maryam Abdollahyan[1], Fabrizio Smeraldi[1], Boris Noyvert[2] and Greg Elgar[2]

1.School of Electronic Engineering and Computer Science, Queen Mary University of London, UK 2.The Francis Crick Institute, Mill Hill Laboratory, London, UK

m.abdollahyan@qmul.ac.uk

Introduction

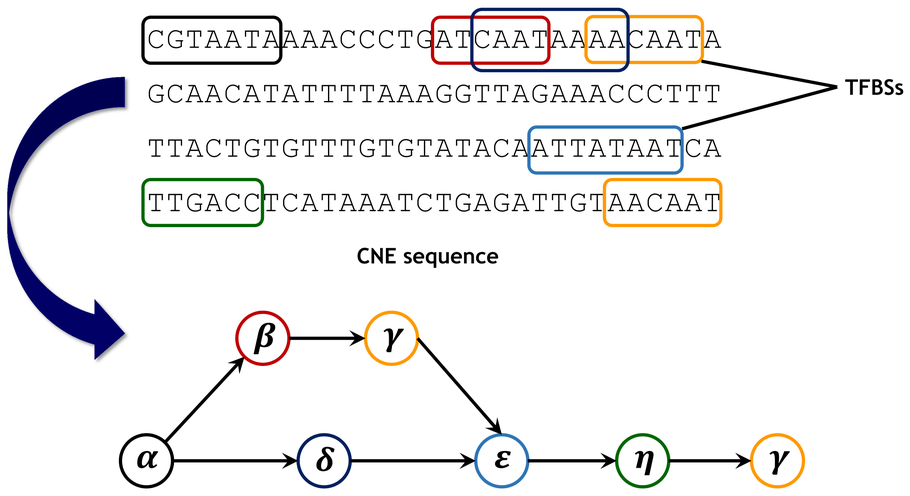
Identification and functional characterisation of regulatory modules in the human genome is a challenging task. Regulatory regions that control gene expression act through the sequence-specific binding of transcription factors, proteins that generally bind to short and often redundant motifs. Given the limited DNA alphabet, putative transcription factor binding sites (TFBSs) occur frequently even in relatively short stretches of the genome, yet only a tiny fraction of them are functional in any tissue at any given time. Thus, the presence of single TFBSs is a poor predictor of function; however, combinations of multiple TFBSs co-occurring in close proximity have been found to be good predictors both of regulatory activity and of biological function. Analysis of the co-occurrence of TFBSs is complicated by the fact that binding sites overlap.

In this work, we looked at the enrichment of co-occurring binding sites of 31 transcription factors associated with developmental patterning in a set of over 5000 conserved non-coding elements (CNEs). We used a a graph-based approach which allows us to handle overlapping TFBSs efficiently.

Graph Representation of CNEs

Given a conserved non-coding sequence $\boldsymbol{S = s_1, s_2, \dots s_n}$ over the alphabet $\boldsymbol{N = \{A, C, G, T\}}$, its graph representation is constructed in the following steps:

- Assign a symbol to each TFBS identified in $\boldsymbol{S}$ to obtain the partially ordered multiset $\boldsymbol{T = \{t_1, t_2, \dots t_m\}}$
- For each symbol in $\boldsymbol{T}$, create a vertex and label it with that symbol
- Add an edge between two vertices if their corresponding symbols are consecutive in $\boldsymbol{T}$



Directed acyclic graph representing the sequence of TFBSs identified in a CNE

Aligning Partial Order Graphs

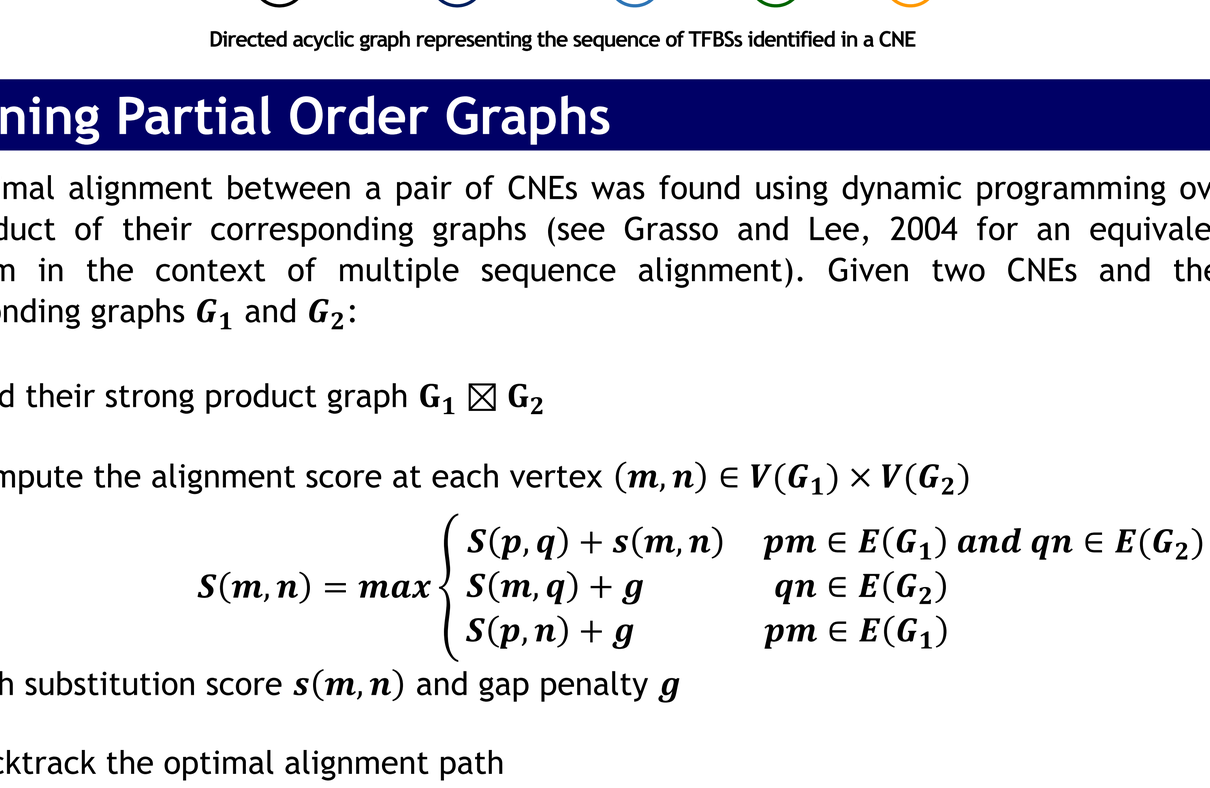
The optimal alignment between a pair of CNEs was found using dynamic programming over the product of their corresponding graphs (see Grasso and Lee, 2004 for an equivalent algorithm in the context of multiple sequence alignment). Given two CNEs and their corresponding graphs $\boldsymbol{G_1}$ and $\boldsymbol{G_2}$:

- Find their strong product graph $\mathbf{G_1} \boxtimes \mathbf{G_2}$
- Compute the alignment score at each vertex $(\boldsymbol{m, n}) \in \boldsymbol{V(G_1) \times V(G_2)}$

$$\boldsymbol{S(m,n) = max} \begin{cases} \boldsymbol{S(p,q) + s(m,n)} & \boldsymbol{pm \in E(G_1)\ and\ qn \in E(G_2)} \\ \boldsymbol{S(m,q) + g} & \boldsymbol{qn \in E(G_2)} \\ \boldsymbol{S(p,n) + g} & \boldsymbol{pm \in E(G_1)} \end{cases}$$

  with substitution score $\boldsymbol{s(m,n)}$ and gap penalty $\boldsymbol{g}$

- Backtrack the optimal alignment path



Two alignment scores, one for each relative orientation of graphs were computed. The higher score was chosen as the final score.

Measuring the Enrichment of TFBSs

Relative enrichment of short (2-4 binding sites) sequences of aligned TFBSs in the alignments of CNEs (denoted by $\boldsymbol{C}$) was computed with respect to those found in the alignments of a background distribution (denoted by $\boldsymbol{B}$) containing TFBSs identified in shuffled conserved non-coding sequences.

The relative enrichment of a TFBS alignment $\boldsymbol{w}$ is:

$$\boldsymbol{R_{CB}(w) = \frac{P_C(w)}{P_B(w)}} \qquad \boldsymbol{P_C(w) = \frac{n_C(w) + \lambda}{\sum_{\substack{w' \epsilon C \cup B \\ |w'|=|w|}} (n_C(w') + \lambda)}}$$

where $\boldsymbol{n_C(w)}$ is the number of occurrences of $\boldsymbol{w}$ in $\boldsymbol{C}$. $\boldsymbol{P_B(w)}$ is computed in a similar way to $\boldsymbol{P_C(w)}$. To account for unseen TFBS alignments, constant $\boldsymbol{\lambda}$ was added to all counts.

Results

Aligned Symbols	Motif(s)	Relative Enrichment
υβ	***Zic, Meis/Tgif/Pknox***	**34.75**
ββ	*Meis/Tgif/Pknox*	32.41
θβ	***Pbx-Hox, Meis/Tgif/Pknox***	**19.97**
βν	*Meis/Tgif/Pknox, Hoxd10,d13*	18.34
βγ	*Meis/Tgif/Pknox, Pou/Oct*	18.31
ττ	*Homeodomain*	17.85
γθ	*Pou/Oct, Pbx-Hox*	15.04
υα	*Zic, Maf*	13.21
τγ	*Homeodomain, Pou/Oct*	13.20
γα	*Pou/Oct, Maf*	11.90

Top 10 co-occurrences of TFBSs with the highest relative enrichment in globally aligned CNEs

- The shared sequence signature (**θβ**) composed of spatially co-occurring *Pbx-Hox* and *Meis* has been functionally validated in this set of CNEs and is associated with coordination of gene expression in the developing hindbrain (Grice et al., 2015).
- *Meis* and *Zic* (**υβ**) are involved in the patterning of both the brain and spinal cord, and are likely to be co-expressed spatially and temporally in the embryo (Biemar et al. 2001 and Nagai et al. 1997).
- The obtained relative enrichments of TFBS alignments are stable (in terms of being over-represented or under-represented) irrespective of the choice of alignment type (global or semi-global).
- TFBS alignments of length 3 and 4 are rare ($< \mathbf{4}\%$).

References

Grasso, C. and Lee, C., 2004. Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. *Bioinformatics, 20*(10), pp.1546-1556.

Grice, J., Noyvert, B., Doglio, L. and Elgar, G., 2015. A simple predictive enhancer syntax for hindbrain patterning is conserved in vertebrate genomes. *PloS one, 10*(7), p.e0130413.

Biemar, F., Devos, N., Martial, J.A., Driever, W. and Peers, B., 2001. Cloning and expression of the TALE superclass homeobox Meis2 gene during zebrafish embryonic development. *Mechanisms of development, 109*(2), pp.427-431.

Nagai, T., Aruga, J., Takada, S., Günther, T., Spörle, R., Schughart, K. and Mikoshiba, K., 1997. The Expression of the MouseZic1, Zic2, andZic3Gene Suggests an Essential Role for ZicGenes in Body Pattern Formation. *Developmental biology, 182*(2), pp.299-313.

We gratefully acknowledge the travel fellowship awarded by the ECCB 2016