International Journal of Foundations of Computer Science


# Identifying Potential Regulatory Elements by Transcription Factor Binding Site Alignment using Partial Order Graphs

Maryam Abdollahyan

*School of Electronic Engineering and Computer Science*
*Queen Mary University of London*
*Mile End Road, London E1 4NS, United Kingdom*
*m.abdollahyan@qmul.ac.uk*

Greg Elgar

*Regulatory Genomics Laboratory, The Francis Crick Institute*
*1 Midland Road, London NW1 1AT, United Kingdom*
*greg.elgar@crick.ac.uk*

Fabrizio Smeraldi

*School of Electronic Engineering and Computer Science*
*Queen Mary University of London*
*Mile End Road, London E1 4NS, United Kingdom*
*f.smeraldi@qmul.ac.uk*

Received
Accepted
Communicated by

Identification and functional characterization of regulatory elements in the human genome is a challenging task. A sequence feature commonly used to predict regulatory activity is the co-occurrence of transcription factor binding sites (TFBSs) in regulatory regions. In this work, we present a graph-based approach to detect frequently co-occurring TFBSs in evolutionarily conserved non-coding elements (CNEs). We introduce a graph representation of the sequence of TFBSs identified in a CNE that allows us to handle overlapping binding sites. We use a dynamic programming algorithm to align such graphs and determine the relative enrichment of short sequences of TFBSs in the alignments. We evaluate our approach on a set of functionally validated CNEs. Our findings include a regulatory signature composed of co-occurring Pbx-Hox and Meis binding motifs associated with hindbrain enhancer activity.

*Keywords*: regulatory element prediction; transcription factor binding site co-occurrence; partial order graphs; dynamic programming.

## 1. Introduction

Identification of regulatory elements in the human genome is a fundamental challenge in the field of genomics. Regulatory elements that coordinate the expression of genes act through the process of transcriptional regulation. Transcriptional reg-

ulation is mediated by transcription factors (TFs) that bind to specific motifs in the DNA in a cooperative manner. Thus, combinations of multiple transcription factor binding sites (TFBSs) co-occurring in close proximity are good predictors both of regulatory activity and to some extent of biological function [3, 21, 16]. This assumption, that TFBSs in clusters are more likely to act as regulatory elements than solitary binding sites, has formed the basis for the development of a number of algorithms. Examples of such algorithms are Cluster Buster [6], MSCAN [12], MCAST [2] and Ahab [18].

Motivated by the same assumption, we introduce a graph-based approach to detect over-represented co-occurring TFBSs in conserved regulatory regions. We do not directly use the DNA sequence of conserved non-coding elements (CNEs); instead, we consider the sequence of TFBSs identified in each CNE. The idea is to align such sequences of TFBSs and find the short subsequences of TFBSs that are frequently matched in the alignments, i.e. TFBSs that co-occur in the CNEs. This reduces the effect of spurious matches that are unlikely to occur in the same order in multiple sequences, while taking into consideration the spatial order of TFBSs. A similar approach was proposed in [10] which uses local alignments of TFBSs to predict regulatory elements. Analyzing the co-occurrence of TFBSs is complicated by the fact that binding sites may overlap. This rules out the use of classic alignment algorithms [17, 20] (that cannot handle overlapping subsequences) and $k$-mer-based methods (that count the occurrences of subsequences and would enumerate the overlapping subsequences indiscriminately). We use partial order graphs to handle the overlap of TFBSs.

We represent each sequence of TFBSs identified in a CNE as a directed acyclic graph (DAG). We then find the optimal alignment between two sequences of TFBSs by aligning their corresponding graphs using a modified dynamic programming-based alignment algorithm called the Partial Order-Partial Order (PO-PO) alignment algorithm [8], originally developed in the context of multiple sequence alignment. Finally, we measure the relative frequency of aligned TFBSs in the alignments with respect to a background distribution.

This article is organized as follows: in Section 2.1, we show how the partial order graph representing the sequence of TFBSs in a CNE is constructed. In Section 2.2, we overview the PO-PO alignment algorithm in a graph framework. Measuring the relative enrichment of co-occurring TFBSs in the alignments is discussed in Section 2.3. In Section 3, we present the results of testing our method on a set of functionally validated CNEs from the CONDOR database [22].

## 2. Methods

### 2.1. *Graph representation of a CNE*

Given a conserved non-coding sequence $S = s_1 s_2 \dots s_n$ over the alphabet $N = \{A, T, C, G\}$, its graph representation is constructed in the following steps: first, we assign a symbol to each TFBS identified in $S$ to obtain the partially ordered

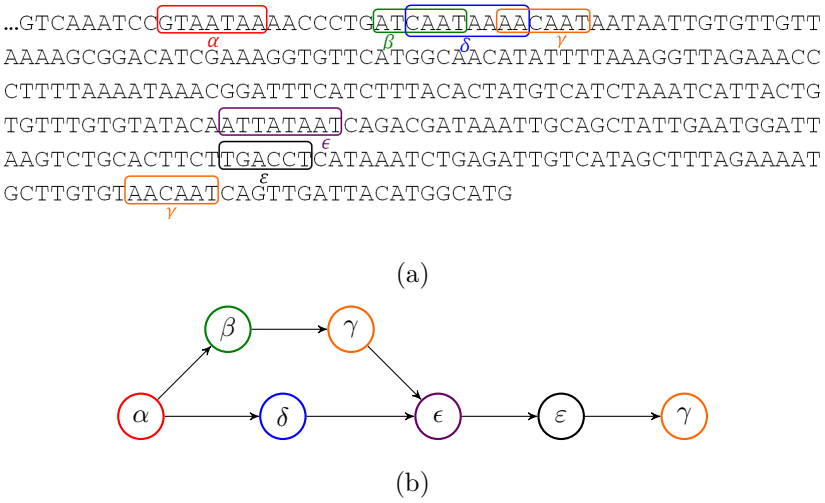



Fig. 1: (a) A CNE with seven identified TFBSs. (b) Graph representation of the CNE shown in (a). Vertices have the same color as the corresponding binding sites. In this example, TFBSs represented by vertices $\beta$ and $\gamma$ overlap with the TFBS represented by vertex $\delta$.

multiset $T = \{t_1, t_2, \ldots, t_m\}$. Elements in $T$ are in the same order as they are in $S$, i.e. $t_i < t_j$ $(i \neq j)$ if and only if in $S$, every nucleotide in $t_i$ comes before every nucleotide in $t_j$. Next, we transform this set into a directed acyclic graph (DAG) $G$. For each symbol in $T$, we create a vertex and label it with that symbol. In the case where the same TF binds to overlapping sites in a CNE, only a single vertex is created. We add an edge between two vertices if their corresponding symbols are consecutive in $T$. In this graph, each path from a source to a sink vertex (note that there may exist multiple source/sink vertices since $G$ can start/end with overlapping vertices) corresponds to a sequence of non-overlapping TFBSs that were identified in $S$. An example of the graph representation of a CNE is shown in Figure 1.

### 2.2. *PO-PO alignment of CNEs*

We use the Partial Order-Partial Order (PO-PO) alignment algorithm [8] for aligning a pair of CNEs. The PO-PO alignment algorithm is a generalization of the Partial Order Alignment (POA) algorithm [13], which was proposed as an approach to Multiple Sequence Alignment (MSA). In [13], linear representation of an MSA was replaced by a DAG called a Partial Order MSA (PO-MSA); classic dynamic programming-based alignment algorithms [17, 20] were modified to find the optimal alignment between a sequence and a PO-MSA. This involved adding the branches of the PO-MSA as additional surfaces to the dynamic programming matrix. The set of possible moves at each position in the matrix was extended accordingly to allow moves to any surface at junctions between the surfaces. The PO-PO align-

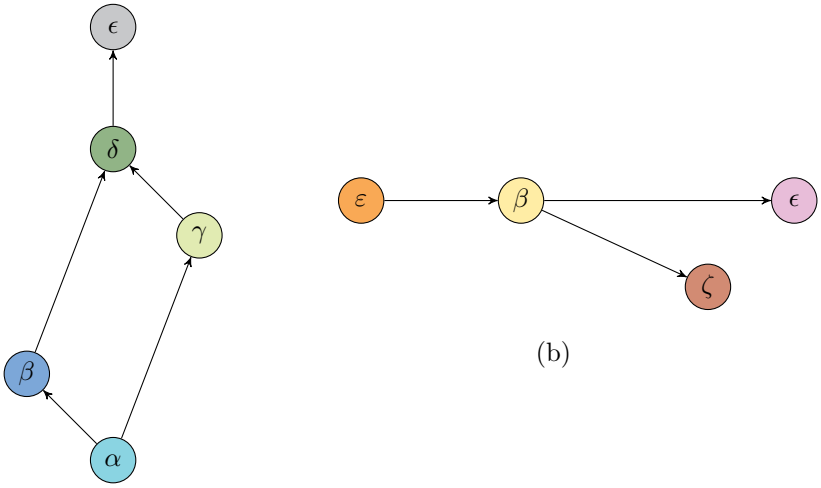
ment algorithm generalized the above approach to align two PO-MSAs. Here, we use this algorithm to find the optimal alignment between a pair of DAGs, each representing a CNE. An elegant derivation of this algorithm can be obtained using the graph-based approach that we introduced in [1] in the context of Mercer kernels for partial order graphs. In this framework, the PO-PO algorithm becomes a dynamic programming approach to finding the optimal path (corresponding to an optimal alignment) in the strong product graph of two DAGs.

We denote the vertex set and the edge set of a DAG $G$ by $V(G)$ and $E(G)$, respectively. A directed edge from vertex $u$ to vertex $v$ is written as $uv$. Given two DAGs $G_1$ and $G_2$, their strong product $G_1 \boxtimes G_2$ has vertex set $V(G_1) \times V(G_2)$, where vertices $(v_1, v_2)$ and $(u_1, u_2)$ are connected if and only if for $k \in \{1, 2\}$ either $v_k = u_k$ or $v_k u_k \in E(G_k)$. In this graph (which generalizes the dynamic programming matrix), each path corresponds to an alignment of a path in $G_1$ against a path in $G_2$. The objective is to find the path with the optimal alignment score in the set of all paths in $G_1 \boxtimes G_2$. This requires finding the move (incoming edge) with the optimal score at every vertex $(m, n)$ in $G_1 \boxtimes G_2$. Possible moves are aligning two symbols with substitution score $s(m, n)$ and indels (insertions or deletions) with gap penalty $g$. Vertices in $G_1$ and $G_2$ can have multiple predecessors. Hence when computing score $S(m, n)$ of a vertex $(m, n)$, all possible combinations of its incoming edges must be considered:

$$S(m,n) = \max \begin{cases} S(p,q) + s(m,n) & pm \in E(G_1) \text{ and } qn \in E(G_2) \\ S(m,q) + g & qn \in E(G_2) \\ S(p,n) + g & pm \in E(G_1) \end{cases} \tag{1}$$

In the case of sequences that do not contain overlapping TFBSs, the corresponding DAGs do not branch, and $m$ and $n$ can be thought of as simply positions in the sequences. Tracing the path that leads to the optimal alignment is done in the same way as in classic dynamic programming-based alignment algorithms. For global alignments, back-tracking starts from vertex $(m, n)$, where $m$ and $n$ are sink vertices in $G_1$ and $G_2$, respectively. For semi-global alignments, back-tracking starts from the highest scoring vertex $(m, n)$, where either $m$ or $n$ is a sink vertex. Starting from the chosen start node, the optimal alignment is traced back along the product graph to vertex $(s_1, s_2)$, where $s_i$ is a source vertex in $G_i$ for at least one $i \in \{1, 2\}$ (semi-global alignment) or both (global alignment). An example of an alignment is shown in Figure 2.

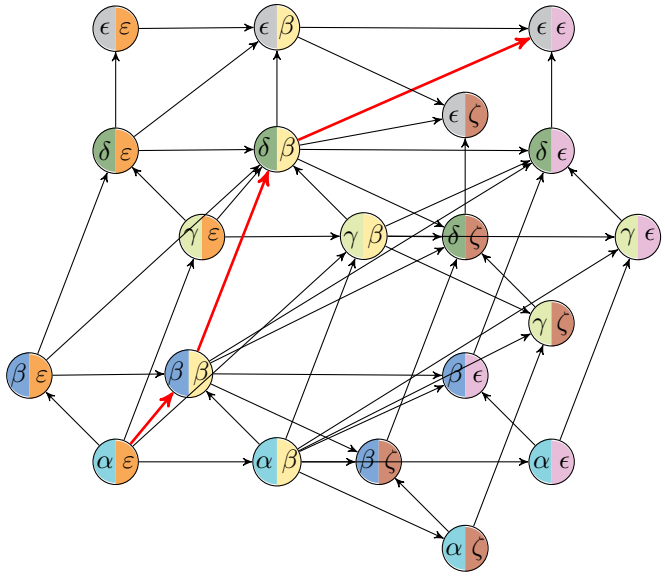
We note that using the Needleman-Wunsch [17] alignment algorithm, finding the optimal alignment between two sequences with overlapping subsequences requires aligning all possible pairs of sequences (without overlapping subsequences) corresponding to the alternative paths in their DAGs, which will result in an exponential complexity as the number of overlaps increases. In contrast, the above algorithm finds the optimal alignment between two DAGs efficiently, with a time complexity that is quadratic with respect to the number of vertices in each DAG.




(a)




(b)




(c)

$\alpha\beta\,\delta\,\epsilon$
$\varepsilon\,\beta{-}\epsilon$

(d)

Fig. 2: (c) Strong product graph of DAGs shown in (a) and (b). The path corresponding to the optimal global alignment shown in (d) is coloured in red. (d) Optimal global alignment between the two sequences represented by DAGs shown in (a) and (b).

### 2.3. *Measuring the frequency of co-occurring TFBSs*

We find the optimal alignment between all pairs of CNEs in the dataset. For each pair of CNEs, we obtain two alignments, one for each of the two possible relative orientations of the sequences. The optimal alignment between the CNEs is chosen as the one with the highest score.

We search the alignments for words composed of up to four aligned symbols, i.e. co-occurring TFBSs. We then compute the relative frequency of each word with respect to a background distribution (see Section 3.1 for details) as follows: let $n_C(w)$ be the number of occurrences of word $w$ in the alignments of sequences in the main dataset (denoted by $C$), and let $n_B(w)$ be the number of occurrences of $w$ in the alignments of sequences in the background distribution (denoted by $B$) using the same type of alignment (global or semi-global). We denote the length of $w$ by $|w|$. Not all words occurring in $C$ are present in $B$, and vice versa, i.e. $n_B(w) = 0$ or $n_C(w) = 0$ for some $w$. To account for unseen words, we apply Laplace smoothing by adding the constant $\lambda$ to all counts of $w$. The probability of occurrence of $w$ in the alignment of main sequences is computed as follows:

$$P_C(w) = \frac{n_C(w) + \lambda}{\sum\limits_{\substack{w' \in C \cup B \\ |w'|=|w|}} (n_C(w') + \lambda)} \tag{2}$$

Note that in computing $P_C(w)$, only words of the same length as $w$ are considered. The probability of occurrence of $w$ in the alignment of background sequences, $P_B(w)$, is computed in the same way. The relative frequency of $w$ is computed as follows:

$$R_{CB}(w) = \frac{P_C(w)}{P_B(w)} \tag{3}$$

## 3. Evaluation

We tested our approach on a set of CNEs downloaded from the CONDOR database [22]. Many of these CNEs have been functionally validated previously [9]. We found both the optimal global and semi-global alignments of CNEs in this dataset, which we will refer to as the *global* and *semi-global* sets, respectively. We then extracted the words of length two, three and four in the alignments and computed their relative frequency with respect to a background distribution (see below). Finally, we selected the over-represented words and compared them to the results of functional assays.

### 3.1. *Data*

The main dataset consists of four orthologous sets of human, mouse, rat and fugu CNEs retrieved from the CONDOR database. We chose a set of 31 binding sites of representative family members of TFs known to play a role in developmental

Table 1: Names of TF families from which the representative TFs were chosen.

| Transcription Factor Families | | |
|---|---|---|
| Cdx | Meis | Runx |
| Ets | Nkx | Six |
| Forkhead | Nrf | Sox |
| Gata | Pax | Tcf |
| Hmx | Pbx | Tfap |
| Hox | Pitx | Zic |
| Irx | Pou | |
| Maf | Rfx | |

patterning (Table 1). The binding preferences for these TFs were extracted from the UniPROBE [19] and JASPAR [14] databases. We scanned the CNEs for the occurrence of all TFBSs using FIMO [7]. For more details on the data see [9].

The relative frequency of each word in the PO-PO alignments of the above dataset was computed with respect to a background distribution obtained by shuffling the sequences in the main dataset. Each sequence in the background distribution was generated by randomly shuffling a CNE in the main dataset. We repeated this process ten times to generate ten sets of shuffled sequences. The number of occurrences of a word in the alignments of background sequences was averaged over the ten sets.

### 3.2. *Parameters*

The alignment parameters and the Laplace smoothing constant were set as follows: the matching score between two TFBSs was defined in a way that takes the count of each TFBS in the main dataset into consideration. Let $n(t)$ be the number of times that TFBS $t$ has been detected in CNEs in the main dataset and $N$ be the total number of detected TFBSs in the main dataset. The matching score $s(t, r)$ is defined as:

$$s(t,r) = \begin{cases} \log \dfrac{1}{P^2(t)} & \text{if } r = t \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $P(t) = n(t)/N$. The linear gap penalty is -1. The smoothing constant ($\lambda$) was set to 1.

### 3.3. *Results*

The global and semi-global sets contain 229 and 270 words, respectively. The number of words of length three and four is low and collectively, they constitute less than

Table 2: Top five Over-represented words of length two in the global set and their relative frequency (symbols used to represent TFBSs are arbitrary and are only included to allow a rapid assessment of the similarity between the words).

| Word | TFs | Relative Frequency |
|---|---|---|
| $\gamma\delta$ | Meis, Pbx-Hox | 20.3 |
| $\delta\epsilon$ | Pbx-Hox, Zic | 5.4 |
| $\gamma\epsilon$ | Meis, Zic | 2.3 |
| $\alpha\gamma$ | Cdx-2, Meis | 1.8 |
| $\beta\gamma$ | Hoxd10-Hoxd13, Meis | 1.5 |

15% of the words in the sets. The two sets have 207 words in common. In 99% of cases, the words in the global set that are over-represented with respect to the background distribution are also over-represented in the semi-global set, and vice versa. Hence, the results are stable irrespective of the type of alignment. The top five co-occurring TFBSs with the highest relative frequency are listed in Table 2.

The words '$\delta\epsilon$' and '$\gamma\epsilon$' are of note since Zic has been shown to regulate retinoic acid (RA) signaling during the early development of the embryo, which affects the expression levels of Hox and Meis during the hindbrain patterning [5]. Moreover, both Meis and Zic are involved in the patterning of the brain and the spinal cord, and as such are likely to be co-expressed spatially and temporally in the embryo [4, 15]. The word '$\beta\gamma$' represents the known interaction of Meis with Hox [11].

The regulatory activity driven by the highest ranked word '$\gamma\delta$', composed of co-occurring Meis and Pbx-Hox binding motifs, has been previously functionally validated in our dataset [9]. In [9], this syntax was identified in a set of conserved vertebrate hindbrain enhancers. The authors showed that Meis TFBSs are frequently proximal (within 100bp) to Pbx-Hox TFBSs, and that both TFBSs are required for hindbrain enhancer function. They then used this syntax to accurately predict hindbrain enhancers in 89% of cases from our dataset. Furthermore, they refined and used this syntax to predict over 3,000 hindbrain enhancers across the human genome, demonstrating the predictive power of this approach.

## 4. Conclusion

We presented an approach to identify over-represented combinations of TFBSs in conserved regulatory regions based on the alignment of sequences of TFBSs appearing in CNEs. We showed how the overlap of TFBSs can be handled by representing the sequences as partial order graphs and described a modified dynamic programming algorithm to align these graphs. Moreover, we discussed a way to measure the relative frequency of subsequences of TFBSs in the pairwise alignments in order to detect frequently co-occurring binding motifs. Comparison between the results obtained using our approach and those obtained using functional assays showed that

our approach can be employed to computationally identify combinations of TFBSs which can then be prioritized for functional validation.

## References


[1] M. Abdollahyan and F. Smeraldi, POKer: a partial order kernel for comparing strings with alternative substrings, *Proceedings of the 25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, (2017), pp. 263–268.

[2] T. L. Bailey and W. S. Noble, Searching for statistically significant regulatory modules, *Bioinformatics* **19**(suppl_2) (2003) ii16–ii25.

[3] B. P. Berman, Y. Nibu, B. D. Pfeiffer, P. Tomancak, S. E. Celniker, M. Levine, G. M. Rubin and M. B. Eisen, Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome, *Proceedings of the National Academy of Sciences* **99**(2) (2002) 757–762.

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

[5] D. L. Drummond, C. S. Cheng, L. G. Selland, J. C. Hocking, L. B. Prichard and A. J. Waskiewicz, The role of Zic transcription factors in regulating hindbrain retinoic acid signaling, *BMC developmental biology* **13**(1) (2013) p. 31.

[6] M. C. Frith, M. C. Li and Z. Weng, Cluster-Buster: Finding dense clusters of motifs in DNA sequences, *Nucleic acids research* **31**(13) (2003) 3666–3668.

[7] C. E. Grant, T. L. Bailey and W. S. Noble, FIMO: scanning for occurrences of a given motif, *Bioinformatics* **27**(7) (2011) 1017–1018.

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

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

[10] O. Hallikas, K. Palin, N. Sinjushina, R. Rautiainen, J. Partanen, E. Ukkonen and J. Taipale, Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity, *Cell* **124**(1) (2006) 47–59.

[11] Y. Jacobs, C. A. Schnabel and M. L. Cleary, Trimeric association of Hox and TALE homeodomain proteins mediates Hoxb2 hindbrain enhancer activity, *Molecular and Cellular Biology* **19**(7) (1999) 5134–5142.

[12] Ö. Johansson, W. Alkema, W. W. Wasserman and J. Lagergren, Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm, *Bioinformatics* **19**(suppl_1) (2003) i169–i176.

[13] C. Lee, C. Grasso and M. F. Sharlow, Multiple sequence alignment using partial order graphs, *Bioinformatics* **18**(3) (2002) 452–464.

[14] A. Mathelier, X. Zhao, A. W. Zhang, F. Parcy, R. Worsley-Hunt, D. J. Arenillas, S. Buchman, C.-y. Chen, A. Chou, H. Ienasescu *et al.*, JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles, *Nucleic acids research* **42**(D1) (2013) D142–D147.

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

[16] S. Nandi, A. Blais and I. Ioshikhes, Identification of cis-regulatory modules in promoters of human genes exploiting mutual positioning of transcription factors, *Nucleic acids research* **41**(19) (2013) 8822–8841.
[17] S. B. Needleman and C. D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, *Journal of molecular biology* **48**(3) (1970) 443–453.
[18] N. Rajewsky, M. Vergassola, U. Gaul and E. D. Siggia, Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo, *BMC bioinformatics* **3**(1) (2002) p. 30.
[19] K. Robasky and M. L. Bulyk, UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein–DNA interactions, *Nucleic acids research* **39**(suppl_1) (2010) D124–D128.
[20] T. F. Smith and M. S. Waterman, Identification of common molecular subsequences, *Journal of molecular biology* **147**(1) (1981) 195–197.
[21] A. Vandenbon, Y. Kumagai, S. Akira and D. M. Standley, A novel unbiased measure for motif co-occurrence predicts combinatorial regulation of transcription, *BMC genomics* **13**(7) (2012) p. S11.
[22] A. Woolfe, D. K. Goode, J. Cooke, H. Callaway, S. Smith, P. Snell, G. K. McEwen and G. Elgar, CONDOR: a database resource of developmentally associated conserved non-coding elements, *BMC developmental biology* **7**(1) (2007) p. 100.