Sequence-independent alignment of DNA shape at transcription factor binding sites (TFBSs)

Sivakanthan Kasinathan, Gabriel E. Zentner, Beibei Xin, Remo Rohs, and Steven Henikoff.

Conceptually similar to classical ungapped sequence alignment, pairwise alignment of DNA shape (**Fig. 1**) involves determining the optimal relative shift of shape features of one TFBS relative to shape features of another TFBS. Multiple alignment of shape features for a collection of TFBSs involves aligning sites to a common 'centroid' TFBS.

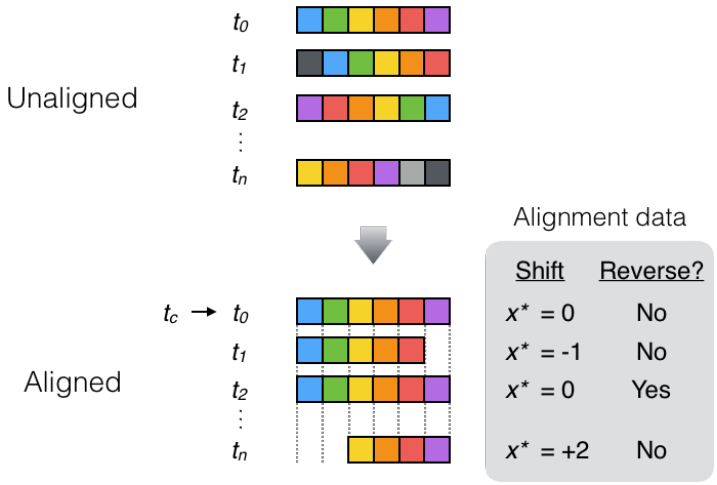

Figure 1 | Conceptual overview of sequence-independent shape alignment. A collection of one-dimensional unaligned vectors $\{t_1, \ldots, t_n\}$ is taken as input and the output is a set of aligned vectors and a designated centroid t_c against which all other vectors are aligned by shifting (left or right) or reversal.

Specifically, in order to align shape features independently of DNA sequence, windows of length L centered on the maximally cleaved position in each ChEC peak were generated. Values for shape parameters (helical twist, propeller twist, roll, and minor groove width) were computed within these windows as described previously (Zhou, *et al. Nucleic Acids Res.* 2013; PMID 23703209). Two $4 \times L$ matrices M_s and M_s' containing values for the four shape parameters at each position in the window are defined for each site s. M_s' represents a reversal of the ordering of columns of M_s (analogous to taking the reverse complement of a DNA sequence).

For a given pair of TFBSs a and b with associated shape matrices and horizontal shifts $x \in (-L, L)$ of one matrix with respect to another, we are interested in finding the shift $x_{a,b}^*$ that produces the optimal alignment of the two sites. Note that $x < 0$ and $x > 0$ correspond to left and right shifts of, *e.g.,* M_a relative to M_b, respectively. At each shift x, the 'goodness' of alignment of the shape matrices can be quantified by computing the cosine similarity S_x of the linearized matrices as follows; $M_a\,(i,j)$ and $M_b\,(i,j)$ represent the element at position (i,j) in the matrices being compared and $\|M_a\|_F$ and $\|M_b\|_F$ designate the Frobenius norms of the matrices.

$$S_x = \frac{\sum_{i=1}^{4}\sum_{j=1}^{L} M_a\,(i,j) M_b\,(i,j)}{\|M_a\|_F \|M_b\|_F}$$

Note that the numerator is equivalent to the tensor dot product of M_a and M_b. The quantity S'_x, which corresponds to the 'reversed' matrices M'_a and M'_b, is defined similarly. The shift corresponding to the best alignment is defined as follows.

$$x^*_{a,b} = \text{argmin}_x(S_x, S'_x)$$

The score for the optimal alignment is given by $S_{a,b} = \min(S_x, S'_x)$ and whether a reversal of the shape matrices produced the optimal shift is recorded.

For a set of TFBSs $T = \{t_1, \ldots, t_n\}$, all possible pairwise shape alignments are performed and the alignment scores and optimal shifts are retained in $n \times n$ matrices A and B, respectively. The distance from a site t_i to all other sites in T is given by the $\mathcal{L}^2$-norm of the column (or, equivalently, row) vector $A_{*,i}$, which is denoted $\left|A_{*,i}\right|$, of alignment scores. The centroid t_c, defined to have the lowest distance to all other sites in T, is given by $c = \text{argmin}_i\left(\left|A_{*,i}\right|\right)$. For $t_i \in T$, the optimal shift relative to t_c can be looked up in B at $b_{i,c}$. The final alignment of all sites is produced by shifting and/or reversing each site t_i with respect to t_c.

In practice, we used $L = 201$ to obtain shape data. We restricted cosine similarity computation to a 90 nt window centered on the peak midpoint and considered $x \in [-25,25]$ to minimize missing data in the aligned vectors (see, for example, aligned t_1 and t_n in **Fig. 1**). Further, given the well-documented MNase A/T preference, which occurs proximal to cleavage maxima, we excluded the data from the five central positions in windows corresponding to each site in the process of carrying out shape alignment. These sites were only excluded in the alignment procedure and were included in all subsequent analyses.