Hypothesis

Supplementary material

| Tools         | Nucleotide level | Binding site level | Motif level |
|--------------|-----------------|--------------------|-------------|
|              | nPC  | nSn  | nSp  | sPC  | sSn  | sSp  | mSr  | sSr  |
| AlignACE     | 0.17 | 0.24 | 0.198 | 0.276 | 0.399 | 0.379 | 0.91  | 0.54  |
| MotifSampler | 0.19 | 0.225 | 0.29 | 0.31 | 0.442 | 0.392 | 0.91  | 0.61  |
| Consensus    | 0.25 | 0.282 | 0.335 | 0.341 | 0.431 | 0.458 | 0.93  | 0.61  |
| MEME         | 0.26 | 0.365 | 0.305 | 0.397 | 0.565 | 0.54  | 0.94  | 0.63  |
| PossumSearch | 0.32 | 0.348 | 0.401 | 0.469 | 0.557 | 0.584 | 0.95  | 0.65  |
Average       | 0.238 | 0.292 | 0.308 | 0.3586 | 0.4788 | 0.4706 | 0.928 | 0.608 |

Table 1: Prediction accuracy on the E. coli intergenic region dataset is given.

| Algorithm        | Best | Worst | Mean | Standard deviation |
|------------------|------|-------|------|--------------------|
| AlignACE         | 0.17 | 0.02  | 0.10 | 0.07               |
| MotifSampler     | 0.19 | 0.03  | 0.11 | 0.08               |
| Consensus        | 0.25 | 0.07  | 0.15 | 0.09               |
| MEME             | 0.26 | 0.05  | 0.15 | 0.10               |
| PossumSearch     | 0.32 | 0.06  | 0.16 | 0.13               |

Table 2: The statistics of the top five predictions in terms of nPC on ECOLI_ACTIVATORS_RDB_40A set.

| Algorithm | Motif width |
|-----------|-------------|
|           | 5  | 10 | 15 | 20 |
| AlignACE  | 0.068 | 0.241 | 0.194 | 0.131 |
| MotifSampler | 0.031  | 0.240 | 0.196 | 0.136 |
| Consensus | 0.162 | 0.248 | 0.282 | 0.148 |
| MEME      | 0.149 | 0.246 | 0.254 | 0.136 |
| PossumSearch | 0.177 | 0.251 | 0.292 | 0.158 |

Table 3: Influence of estimated motif width on the nucleotide level prediction accuracy (nPC)

Nucleotide level accuracy
First, for each target binding site with overlapping predicted binding sites in an input sequence, we define the following values for calculating accuracy metrics at the nucleotide level: nTP (true positive), the number of target binding site positions predicted as binding site positions; nTN (true negative), the number of non-target binding site positions predicted as non-binding site positions; nFP (false positive), the number of non-target binding site positions predicted as binding site positions; nFN (false negative), the number of target binding site positions predicted as non-binding site positions. The sensitivity, specificity and performance coefficient over a pair of target/predicted binding sites is defined as:

\[ nS_n = \frac{nTP}{nTP + nFN}, \quad nS_p = \frac{nTP}{nTP + nFP}, \quad nPC = \frac{nTP}{nTP + nFP + nFN} \]

According to this definition, the nPC value ranges over (0, 1) with the perfect prediction being the value of 1. Compared with the correlation coefficient (CC) [19, 20], nPC has several benefits: it is straightforward to interpret and practically, it also tells the experimental biologists the probable ranges that the true binding sites are located around the predicted positions.

Binding site level accuracy
The binding site level accuracy indicates whether predicted binding sites overlap with true binding sites by one or more nucleotide position. We define, sTP, sFP and sFN as follows: sTP, the number of predicted binding sites which overlaps with the true binding sites by at least 1 nt; sFP, the number of predicted binding sites which have no overlaps with the true binding sites; sFN, the number of true binding sites that have no overlaps with any predicted binding sites. For each input sequence, we define the following accuracy metrics at the binding site level:
The binding site level accuracy score of an input sequence set (e.g. ArcA) is the average of the scores over all its sequences. The binding site level accuracy score of the entire benchmark dataset is the average of the scores for all input sequence sets.

**Sequence motif level accuracy**

To evaluate the capability to find at least one binding site in an input sequence, we define the sequence level success rate as the number of sequences $N_s$ that have at least one correctly predicted motif divided by the total number $N$ of sequences in an input sequence set:

$$sSr = \frac{N_s}{N}$$

The overall sequence success rate of an algorithm is thus the average of $sSr$ over all the input sequence sets. We introduce the motif level success rate score $mSr$, a sensitivity measure, to evaluate the adaptability of an algorithm to different types of motifs, is defined as the number of target motif groups $N_p$, which have at least one correctly predicted binding site divided by the total number of target motifs ($M = 45$). A prediction is regarded as correct when the predicted motif overlaps with the target motif by at least 1 nt.

$$mSr = \frac{N_p}{M}.$$