Chapter 4

Evaluation

In the previous chapter, we described annotated alignments and the concerned parameter choice. Through random sequences, we additionally illustrated the characteristics of annotated alignments. In this chapter, we take a step further and focus on evaluating the annotated alignment approach – both in simulated and real data setting.

4.1 Design and implementation of Multi-step approach

First and foremost, our interest lies in studying how the simultaneous approach to predicting conserved TFBSs performs as opposed to a multi-step approach. Intuitively, it would seem reasonable to use an existing multi-step tool for such a comparison. However, most available multi-step tools are unsuitable for our setting:

- *either* the underlying alignment algorithms are different – rVista [113] and Siteblast [131] are based on a heuristic alignment program (BLASTZ [176]) while ConReal [18] and EEL [73] focus on globally aligning *only* the TFBS hits,
- *or* the TFBS annotation strategies are different – ConSite [109] uses a user-defined matrix similarity threshold ignoring statistical significance of hit scores, Cis-Ortho [21] considers the top N hits.

Another factor that complicates the usability of existing tools is the *definition* of a conserved TFBS hit. This is inter-linked with the strategy adopted by each method for dealing with gaps in the aligned hit regions. For example, Monkey [134] uses a heuristic to conservatively allow some gaps. On the other hand, ConSite [109] allows any number of gaps in the aligned regions of the hit pair as long as the pair lies in a region with high overall sequence similarity. For comparisons, it becomes difficult to find a universally applicable definition, particularly in light of the fact that SimAnn predicts gapless-ly aligned hits.

To enable a fair comparison in controlled setting and to restrict differences arising from such factors, we therefore developed two multi-step approaches for predicting conserved binding sites. Our main objective is to highlight the differences arising from the additional pair-profile states in the standard alignment algorithm.

Design: Both methods first align the two sequences using the Smith-Waterman (SW) algorithm with affine gap penalties (see Section 2.2.1). For TFBS hits, both sequences are scanned with the respective profile using the procedure described in Section 2.1.3. Note that here the choice of the score cutoff t influences the number of accepted hits. It can be used as the parameter to control the final balance between true and false positive predictions.

Next, the hits on each sequence are mapped onto the SW alignment as a basis for filtering out the conserved hit pairs. The two approaches differ with respect to this filtering. We distinguish between a **Relaxed** and a **Strict** filtering:

- Relaxed filter A hit pair is marked as conserved if the mapped hit on the first sequence overlaps positively with that on the second sequence in the alignment, irrespective of the number of gaps in the mapped regions of the alignment.

- Strict filter A pair is marked as conserved *only* if the mapped hits contain no gaps, and the hit on the first sequence is perfectly aligned with that on the second sequence.

By considering both the Relaxed and the Strict filters, we cover two extremes of the spectrum. While the Relaxed provides an over-estimate by allowing unlimited number of gaps in the aligned hits, the Strict provides a lower estimate with no leniency for alignment errors. Such a definition circumvents the necessity to define a conservation threshold and is similar to those adapted by other works ([57], [155]). Fig. 4.1 depicts the design of the multi-step approaches.

4.2 Evaluation of SimAnn – comparison with Multi-Step approach

Through comparisons with the multi-step approach, we also wish to evaluate the predictive performance of SimAnn on its own, addressing two issues: **a)** how the background sequence evolutionary distance and the quality of the profile to be searched for, influence predictions, and **b)** do the theoretically calculated profile penalties give reasonable predictions? That is, check the validity of the parameter choice. A brief outline of the procedure follows next, while details are discussed later (Section 4.2.1).

Brief Outline A large set of evolutionarily related sequence pairs is generated. Into each of these pairs, motifs sampled from a fixed profile P are implanted. The correct alignment and positions of the implanted motifs are stored for later evaluation. The sequence pairs are analyzed with SimAnn and the multi-step approach with the two filters, to detect conserved binding sites. All methods are provided with the profile P from which the implanted motifs have been sampled. For each method there is a single parameter which governs its true and false positive rates (`TPR` and `FPR`). We vary this parameter and plot the `TPR` versus `FPR`, yielding a receiver operator characteristic (ROC)-like curve (Appendix). Since for SimAnn this parameter is the profile penalty *pen*, this enables us to use the curves to assess the quality of the theoretically derived profile penalty choices. Finally, this analysis is carried out for different values of sequence relatedness (corresponding to the background sequence evolutionary distance) and different quality of profiles.

4.2.1 Simulation setting

Dataset Generation

For a fixed evolutionary distance and a fixed profile, we adopt the following strategy to generate a set of sequence pairs. We use the software program Rose version 1.3 [189]

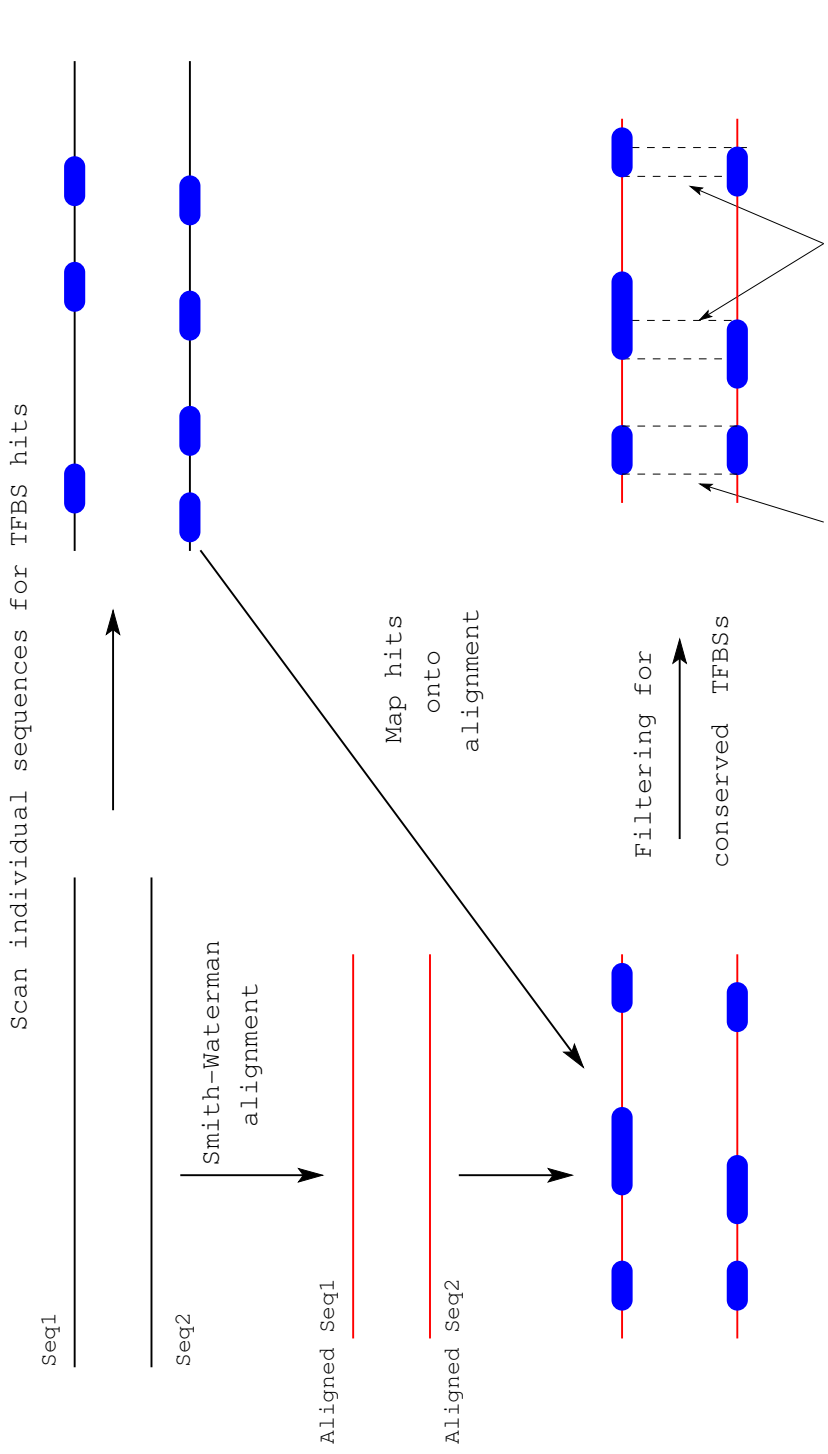


Figure 4.1: *Design of the multi-step approach.* In one step, each sequence is scanned individually for putative TFBS hits. In a separate step, the sequences are aligned using the Smith-Waterman alignment algorithm. The single-sequence hits are mapped onto the alignment to yield aligned hit locations. These may be longer than original sites, since they may contain gaps. Hits that are outside the alignment are not considered. Finally, two filtering strategies, Relaxed and Strict, are employed to extract respective conserved TFBS hits. While Relaxed predicts any overlapping pair as a conserved TFBS hit, the Strict requires a pair to be perfectly aligned (gap-less) to be predicted as conserved.

to simulate sequence pairs at specified evolutionary distance (called *Relatedness* in Rose) together with their true alignments.

Rose: Beginning with a user-defined ancestral sequence or random sequence of user-defined length, Rose iteratively incorporates mutations following the branches of a mutation guide tree. This tree is either specified by the user, with individual branch lengths; or it is generated as a binary tree with branch lengths derived from a user-defined average distance (*Relatedness*). The nodes of the tree correspond to child sequences. Branch lengths govern the proportion of mutations – longer branch lengths imply more frequent mutations. The mutations themselves are derived using mutation matrices that are again branch-length dependent and are based on evolutionary models. Rose also allows the user to define variable mutation rates, yielding the possibility to introduce position-dependent degree of mutability.

For the selection and creation of indels, Rose uses an "inverted gap function" based on user-defined indel thresholds and an indel length function. The thresholds determine whether or not to introduce an indel; the indel length function decides the length of the indel. Rose accepts any quantized length function $l_g = (l_g^1, \ldots, l_g^{len})$ with $\sum_{i=1}^{len} l_g^i = 1$, such that the probability of selecting an indel ($g$ap) of length i in $(1, \ldots, len)$ is given by l_g^i. It should be stressed here that the indel distribution remains fixed across the evolutionary process, while the frequency of the indel length increases with branch length. Throughout the generation process, the true history of mutations is stored giving the advantage of knowing the *true* alignment of a pair of sequences. In the following, the term *Relatedness* stands for the corresponding parameter in Rose, and hence is associated with the evolutionary distance between the sequences. The term relatedness, which would be inversely proportional to the distance, is used in general when referring to sequence similarity levels.

Procedure: For our purpose, we describe the sequences to be at the leaves of a depth one binary tree with branch lengths proportional to the desired distance. The indel threshold is set to 0.002 for a better balance between substitutions and indels than with the default value. All other parameters are set to the default DNA settings – uniform background frequencies, Jukes-Cantor substitution model [90], a mean substitution rate corresponding to 1% mutations (0.013423) and finally both insertion and deletion functions given by $(.2, .2, .2, .1, .1, .1, .1)$. The final set consists of 50 sampled sequence pairs of an average length of 500.

The profile, given as a position-specific count matrix, is first converted to a regularized position-specific probability matrix (`PSPM`) as described in Section 2.1. For each sequence pair, two motifs are sampled independently from this `PSPM`. Finally, the true alignment of the sequence pair is cut at a random position and one of the sampled motifs is inserted into each sequence.

We repeat this construction for sequence relatedness values ranging from 10 to 50 at steps of 10 and for three profiles of differing quality, resulting in a total of 15 different data sets. As a measure of profile quality we use the *balanced quality* (Section 2.1.3) and retrieve matrices from TRANSFAC. In this example, we consider M00395 (poor quality, 0.199), M00690 (medium quality, 0.622) and M00360 (good quality, 0.967), although similar results are obtained for other matrices too.

Required Parameters: We can use the same parameters for the standard alignment part of SimAnn and the Smith-Waterman alignment algorithm underlying the two multi-step approaches. This ensures that the differences observed in the comparison of the three approaches can directly be attributed to those aspects of the methods which are added onto this basic alignment part. In the case of SimAnn this is the introduction of the pair profile states into the alignment algorithm and their special scoring.

To get the correct standard alignment parameters we first determine the substitution matrix that fits to the chosen evolutionary distance. This derivation is straightforward because by default ROSE uses a Jukes-Cantor substitution model and a uniform background letter distribution. Therefore, at desired sequence relatedness values we can use Equation (2.27) to formulate log-likelihood based substitution scores.

To find the appropriate gap penalty, we restricted ourselves to the set where the gap extension penalty is 1/10 of the gap open penalty. We estimate gap costs using the following simple simulation experiment. A set of sequence pairs at the fixed evolutionary distance is generated with ROSE, as described above. All generated sequence pairs are realigned under different gap open penalties and the proportion of gaps in the true and the recomputed alignments is compared to determine the optimal gap open penalty. The estimated gap costs for increasing relatedness values are: $80, 68, 59, 46, 38$ for *Relatedness* values of $10, 20, 30, 40$ and 50, respectively.

Calculation of the true and false positive rates

For the current evaluation, the true and false positive rates are calculated as follows. If the implanted motif pair is detected as a conserved TFBS hit, it is counted as a true positive (`TP`). So there can be at most one `TP` in each of the 50 sequence pairs. Since, in contrast to SimAnn, the multi-step approaches can predict *overlapping* conserved hits, we define the false positive rate as the relative amount of non-site sequence covered by predicted conserved pairs. Here, by *conserved* hit we simply mean gapless-ly aligned TFBS hits.

In SimAnn, the profile penalty *pen* aids in deciding when a pair of strings is annotated as a conserved TFBS hit. It is hence used as the varying parameter to generate the ROC-like curves. This additionally allows us to use the curves to validate the performance of our theoretically calculated profile penalties. Hence, we highlight three theoretical choices of the profile penalty – level 0.05 type I error, level 0.05 type II error and the balanced (Section 3.2) – on the curves for SimAnn. In the multi-step approaches the PSSM score cutoff t is varied over a wide range to determine their true and false positive rates.

It is worth mentioning here that although the parameter that is varied has been derived from the signal and background distribution in both the methods – the PSSM score cutoff is derived by comparing the background distribution π with the signal distribution P, and the profile penalty *pen* is calculated by comparing the background evolutionary distribution ϕ against the signal distribution P^2 – the resulting true and false positives are calculated by comparing against the known site locations. That is, the PSSM score cutoff only helps in deciding when a string is accepted as a TFBS hit, *not* when it is predicted as a conserved hit or when it is counted as a true positive. It is the filtering strategies that decide when it is predicted as a conserved hit, while the comparison with the known site decides when it is a true positive. Similarly in SimAnn, the profile penalty only aids in deciding if a pair of strings is more likely to be a pair-profile hit. The final prediction as a conserved hit, or

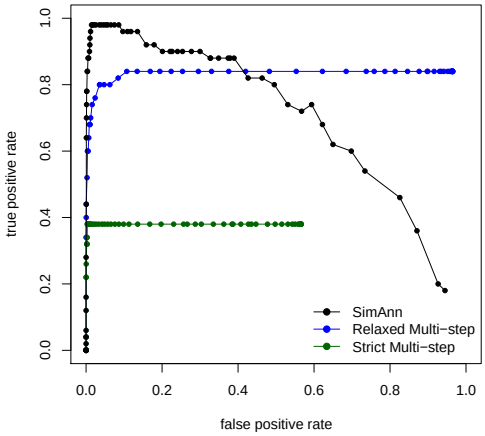


Figure 4.2: *Characteristics of the* TPR *versus* FPR *curves.* Color code: green – multi-step approach/Strict filter; blue – multi-step approach/Relaxed filter; black – SimAnn. The curves for SimAnn and Relaxed are comparable with a similar steep rise in the TPR as opposed to the FPR. At a desired FPR of, say 0.05, the TPR of SimAnn is much higher than that for multi-step with Relaxed filter. The TPR of multi-step with Strict filter does not rise as high as the others, while that of SimAnn decreases as the profile penalty decreases extremely. See text for details on observed behavior. Example is from medium quality profile M00690 at a distance of 30.

a pair-profile hit, is decided through the dynamic programming algorithm that has various other components. And whether it is a true positive is decided again by comparing with the known site locations. Yet, since both the parameters directly influence the number of accepted hits, thus affecting the TPR and FPR, using them to generate *ROC-like* curves seems justified.

4.2.2 Results and analysis

Characteristics of the curves

Varying the penalty/cutoff affects each method differently. To study the general trends of the TPR versus FPR curve of each method, we focus at one example with a fixed evolutionary distance (30) and profile (M00690, medium quality) (Fig. 4.2). In the following we use the term cutoff for both the pair-profile penalty and the PSSM score cutoff.

With decreasing cutoffs, the TPR and FPR of each method rises. In the multi-step approaches the curves (blue and green) level off at extremely low cutoffs. Counter-intuitively, in SimAnn (black) the TPR starts *decreasing* at extremely low values. This is because, in contrast to the multi-step approaches, SimAnn does not predict overlapping hits. At low cutoffs, it tries to induct maximum number of pair-profiles into the alignment – loosing the annotations predicted correctly at higher cutoffs. This explains the fall in the TPR with very low cutoffs and is characteristic of all curves corresponding to the annotated alignments (SimAnn as well as eSimAnn, as will be discussed later).

In general, the Strict multi-step performs worst while the Relaxed multi-step and SimAnn perform better and comparable. At desired levels of type I and type II errors (top left), SimAnn outperforms the Relaxed multi-step approach with a higher TPR at the same FPR. Errors in the underlying alignment, in this case indels in the aligned hit regions, underlie the poor performance of the multi-step methods, especially the Strict approach which penalizes

indels severely. On the other hand, in SimAnn local rearrangements tend to protect the known conserved pair and such effects are dampened.

A point to stress here is that although the Relaxed filter performs comparably to SimAnn, the predicted conserved pairs are not necessarily perfectly aligned in the optimal alignment. They can be interrupted by any number of gaps making them difficult to stand out as a conserved binding site. Contrarily, the predictions from SimAnn are perfectly aligned, gapless pairs of profiles and the conserved binding site is clearly identifiable.

Effects of distance and profile

In Figures 4.6, 4.7 and 4.8, we show the results for each method with profiles of good, medium and poor quality, respectively and under *Relatedness* values corresponding to low (10), medium (40) and high (50) evolutionary distances. As would be expected, for each profile as distance increases the performance of each method deteriorates. Similarly, at the same distance, each method performs worse as profile quality worsens. This is because, the poorer profile yields higher proportion of false predictions and allows more ambiguity in the aligned hit regions. We shall illustrate this with examples in more details later.

Validity of the profile penalty

In each of Figures 4.6, 4.7 and 4.8, the theoretically calculated profile penalties are highlighted. The balanced profile penalty (red) and the type-II error penalty at level 0.05 (cyan) both fall into a region where true and false positive rates show reasonable combinations. Thus, a balanced profile penalty can be chosen when high sensitivity is required while the type-II error penalty at level 0.05 can be chosen for needs of high specificity.

Discussion and conclusion

The multi-step approaches rely on a pre-determined optimal alignment and single sequence `PSSM` hits. At greater evolutionary distances, the chances of aligned true hit locations containing more indels increase. To combine the results from individual steps for extracting conserved hits, one needs to adopt additional strategies. In a simultaneous approach like that of SimAnn, such a necessity does not arise. Especially, when searching for perfectly aligned hits (such as those predicted by SimAnn and Strict multi-step), a simultaneous approach proves to be advantageous.

Consider the example shown in Fig. 4.3 depicting a combination of good profile and high evolutionary distance (high *Relatedness* parameter in Rose). The implanted motifs are highlighted in red. Since the profile is specific, single sequence scanning in the annotation step correctly predicts the true motif in each sequence. But since the sequences are far apart, in the alignment step the SW alignment contains indels in the aligned true motif locations. On combining the results of the two steps the Strict multi-step approach hence fails to predict the true motif pair as conserved. On the other hand, the Relaxed multi-step predicts the true pair but as poorly aligned. In the SimAnn alignment, the combined score of both the motifs is high since the profile is good and hence enables the correct prediction by bringing together the nucleotide pairs in one pair-profile hit.

This effect of indels interrupting true motif locations in the alignment is more pronounced as the profile quality deteriorates. The reason is that the sampled motifs are more degenerate

```
Seq 1: 319  TCTG--TAGCGCGAGTTTT--AATCGTCCCACTCTAACGACTTTCTCTCT
              ||  |X|| |||   ||  |X|||||        |||  |||XXX|||
Seq 2: 295  --TGATTCGC-CGA---TTGGACTCGTC--------ACG--TTTAAATCT
```

(a) Smith-Waterman alignment

```
Seq 1: 319  TCTG--TAGCGCGAGTTTT--AATCGTCCCACTCTAACGACTTTCTCTCT
              ||  |X|| |||   ||  |XPPPPPPPPPPPPP|          |||
Seq 2: 295  --TGATTCGC-CGA---TTGGACTCGTCACGTTTAAA----------TCT
```

(b) SimAnn alignment

Figure 4.3: *Good profile, high distance:* Portion of the alignments under the multi-step approach and SimAnn for an example case of the good profile (M00360) at a high distance (50). In the Smith-Waterman alignment of the multi-step approach, the true hits (red) are interspersed with indels, which are pushed out in the case of SimAnn (blue).

```
Seq 1: AAAGGCTAT---TGACGGTCCGCAATCGT--CTCT-------GGGA--TTGCTC
       |||   |X|   |||XX||     ||| |  ||||       ||||  |||X|X
Seq 2: AAA---TCTGCCTGATAGT-----ATC-TAGCTCTGCGAGCAGGGATCTTGGTT
```

(a) Smith-Waterman alignment

```
Seq 1: AAA---GGCTATTGACGGTCCGCAATCGT--CTCT-------GGGA--TTGCTC
       |||   |PPPPPPPPP        ||| |  ||||       ||||  |||X|X
Seq 2: AAATCTGCCTGATAGT--------ATC-TAGCTCTGCGAGCAGGGATCTTGGTT
```

(b) SimAnn alignment

Figure 4.4: *Poor profile, high distance:* Portion of the alignments under the multi-step approach and SimAnn for an example case of the poor profile (M00395) at a high distance (50). True hit marked in red, the prediction by SimAnn is marked in blue. Both approaches fail since the combined effect of high distance and poor profile quality yields an alignment which has no overlap between the true hit pair.

and hence less alike. Each method suffers, especially when the alignment generated is distorted enough to allow no overlap between the aligned true motif locations. Fig. 4.4 illustrates this where each method fails to predict the correct motifs. SimAnn, in fact shifts the gaps to predict another pair as a putative conserved hit. Overall, both the Relaxed and SimAnn produce a higher number of false positives – an almost random performance in the case of Relaxed. While the Strict has both low `FPR` as well as `TPR` because of its intolerance to indels.

With regard to the profile penalty, an interesting feature is that the type II error penalty is lower than the type I error penalty. Two factors contribute to this: First, the signal score distribution curve is more separated from the background score distribution curve – for the same pvalue this implies a lower type I penalty. And second, at lower penalties the `TPR` in SimAnn starts deteriorating. And hence for a higher `TPR`, we need a higher cutoff leading to a higher type II penalty.

Having the additional choice of jumping to a pair-profile state protects SimAnn from the pitfalls of possible mis-alignments. On the other hand, scoring two motifs independently yields more extreme scores for both good and bad motifs. Non-consensus nucleotide pairs at a position get extremely negative scores and consensus pairs get highly positive scores.

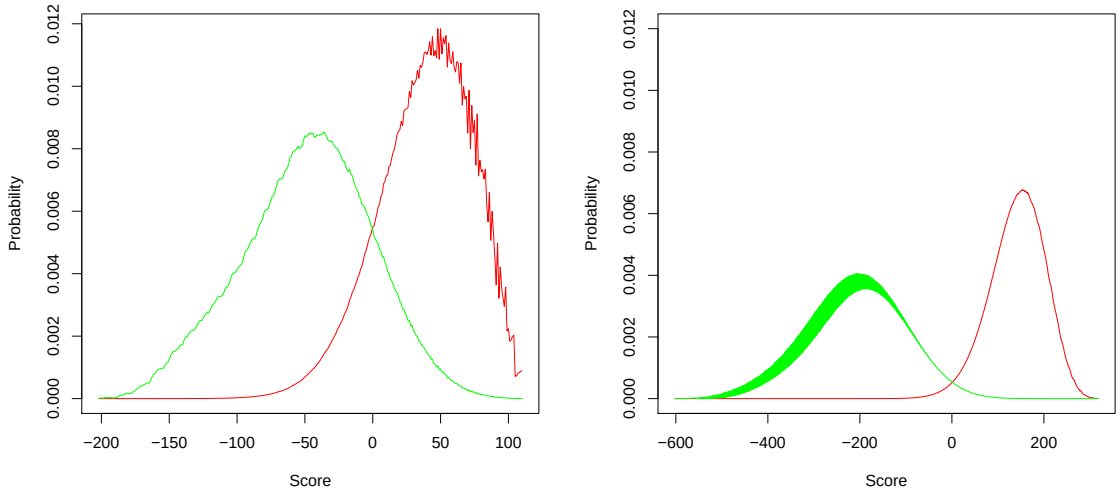


Figure 4.5: Score distributions under the signal (red) and the background (green) models in the case of `PSSM` (left) and the basic `PSA` (right). The latter has a lower overlap between the signal and background highlighting the better signal-to-noise ratio here. The figures are for poor profile (M00395) at a high distance (50).

While this has the direct consequence of a better signal-to-noise ratio in the case of SimAnn (Figure 4.5), it also increases the chances of missing a weak true motif pair in the light of a nearby strong false motif pair.

The simulation experiments discussed here contrast the simultaneous approach against the multi-step one. They provide a proof of principle for the theoretical motivations of the respective parameters in the basic algorithm. We now turn to a similar analysis for the case of the extension eSimAnn, where TFBS-specific evolutionary constraints are utilized.

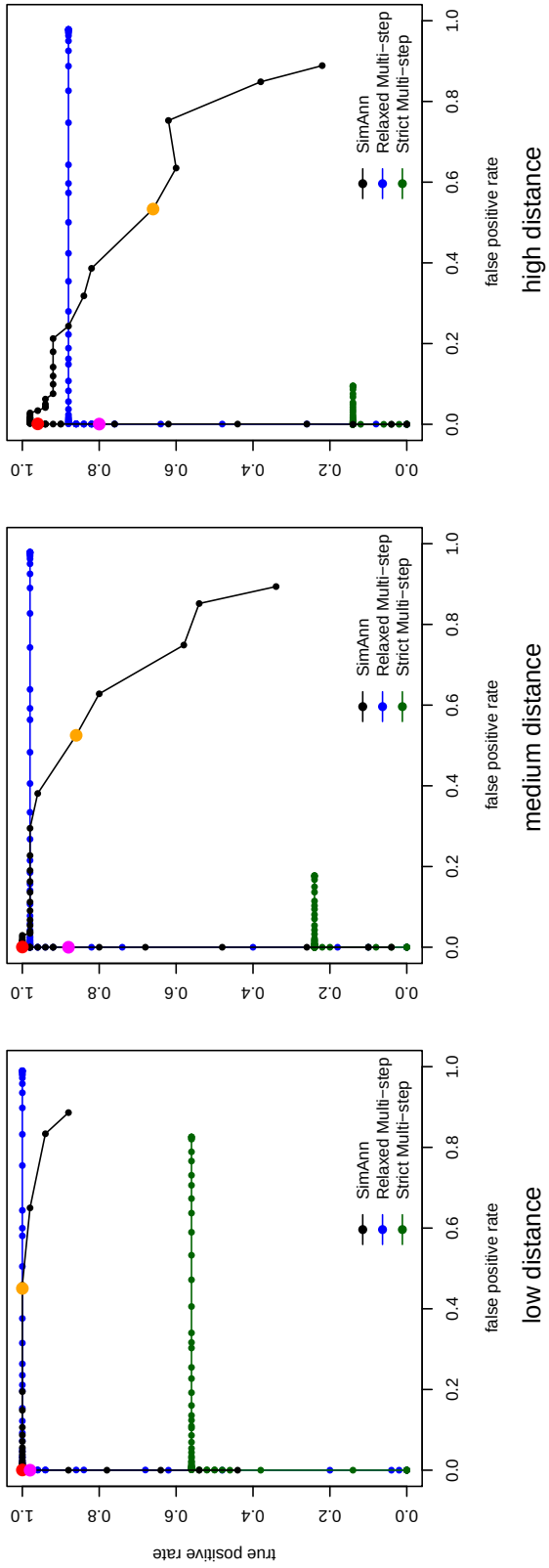


Figure 4.6: Performance comparison of SimAnn and the two multi-step approaches at three evolutionary distances and for the good profile M00360. The curves illustrate the interplay of false and true positive rates on the simulated test sets at varying penalty/cutoff parameters. Color code: green – multi-step approach/Strict filter; blue – multi-step approach/Relaxed filter; black – SimAnn. On the SimAnn curve the statistically motivated profile penalty choices are highlighted. Color code: orange – type-I error penalty at level 0.05; red – balanced penalty; cyan – type-II error penalty at level 0.05.

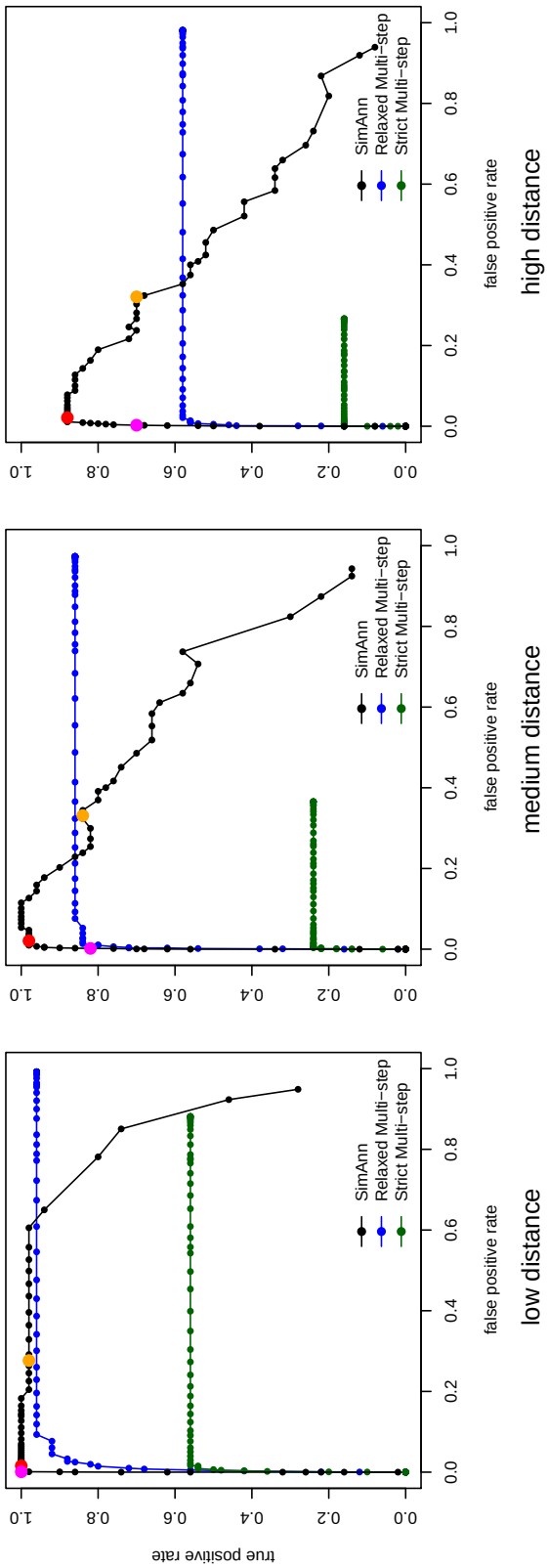


Figure 4.7: Performance comparison of SimAnn and the two multi-step approaches at three evolutionary distances and for the medium profile M00690. The curves illustrate the interplay of false and true positive rates on the simulated test sets at varying penalty/cutoff parameters. Color code: green – multi-step approach/Strict filter; blue – multi-step approach/Relaxed filter; black – SimAnn. On the SimAnn curve the statistically motivated profile penalty choices are highlighted. Color code: orange – type-I error penalty at level 0.05; red – balanced penalty; cyan – type-II error penalty at level 0.05.

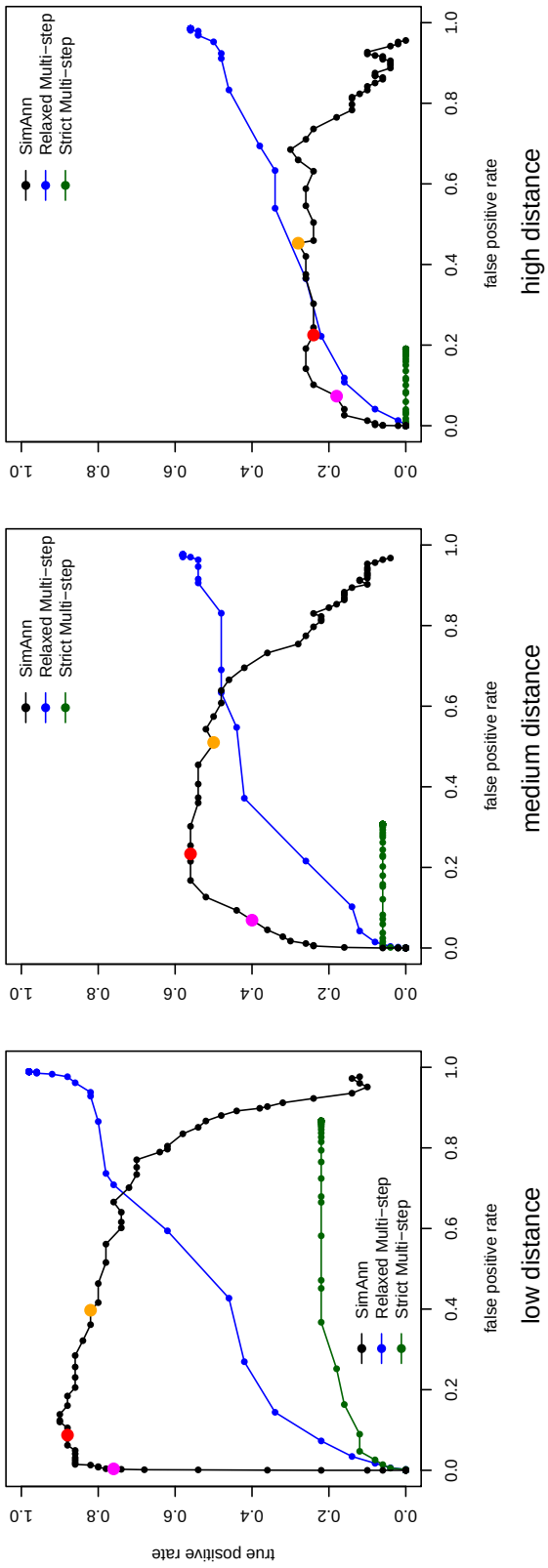


Figure 4.8: Performance comparison of SimAnn and the two multi-step approaches at three evolutionary distances and for the poor profile M00395. The curves illustrate the interplay of false and true positive rates on the simulated test sets at varying penalty/cutoff parameters. Color code: green – multi-step approach/Strict filter; blue – multi-step approach/Relaxed filter; black – SimAnn. On the SimAnn curve the statistically motivated profile penalty choices are highlighted. Color code: orange – type-I error penalty at level 0.05; red – balanced penalty; cyan – type-II error penalty at level 0.05.

4.3 Evaluation of eSimAnn – comparison with SimAnn

To study how incorporating evolutionary constraints on TFBSs influences the alignment as well as pair-profile predictions, we adopt a simulation strategy similar to that in the previous section. The main difference is that we now consider *evolutionarily related* motif samples. The motifs to be embedded are generated by evolving a sampled instance according to the Felsenstein 1981 (F81) or the Halpern-Bruno (HB) model (Section 2.2.3) to yield the second motif in a pair.

4.3.1 Simulation setting

Dataset Generation

For simulating sequence pairs with TFBS evolution modeled according to the F81 model, we used the the software program Rose v1.3 (described in Section 4.2.1). For those with TFBS evolution modeled by the HB model, we used the software program CisEvolver [155], described briefly later. In both cases, we generate 50 pairs of evolutionarily related sequences with average lengths of 500 and Jukes-Cantor as the background evolutionary model. Again, we use the count matrices from TRANSFAC [123] – M00360 (good quality, 0.967), M00690 (medium quality, 0.622) and M00395 (poor quality, 0.199). Gaps are not allowed in the motif locations. In both approaches, the sequence pairs were taken to be at the leaves of a simple depth one binary tree, with branch lengths proportional to the distance. Background frequencies were set to uniform.

CisEvolver: Similar to Rose, CisEvolver takes an ancestral sequence and evolves it along a mutation guide tree, storing the true alignment of the generated sequences. However, distances are absolute here and not multiples of μ as in Rose. By default, binding site locations are evolved according to the Halpern-Bruno model and background sequences according to the Hasegawa Kishino Yano 1985 (HKY85 [76]), although other models can also be employed. Indel events are modeled according to a Poisson distribution and an empirical indel length frequency distribution.

Procedure for CisEvolver: In CisEvolver, an ancestor sequence with a single binding site implanted at a random position is evolved to a desired branch length, yielding a pair of evolutionarily related sequences. As mentioned earlier, the implanted motif evolves according to the HB model and we store the true locations of the evolved motifs provided by CisEvolver. For indel frequencies, we use the example file provided in the CisEvolver package, but set the relative indel rate to 0.05 for a reasonable proportion of gaps in the true alignments. Two distance settings of 0.1, 0.5 are used.

Procedure for Rose: In Rose, for background sequence evolution we use the default DNA parameters with the indel thresholds at 0.002. Motifs are evolved according to the F81 model – at each position the stationary distribution is set to the corresponding position-specific letter distribution of the profile. A random position in a sequence is chosen and a motif sampled from the respective profile is implanted. At the equivalent position in the true alignment the evolved motif is implanted in the second sequence. The distances corresponding to close and far are 10 and 50, respectively.

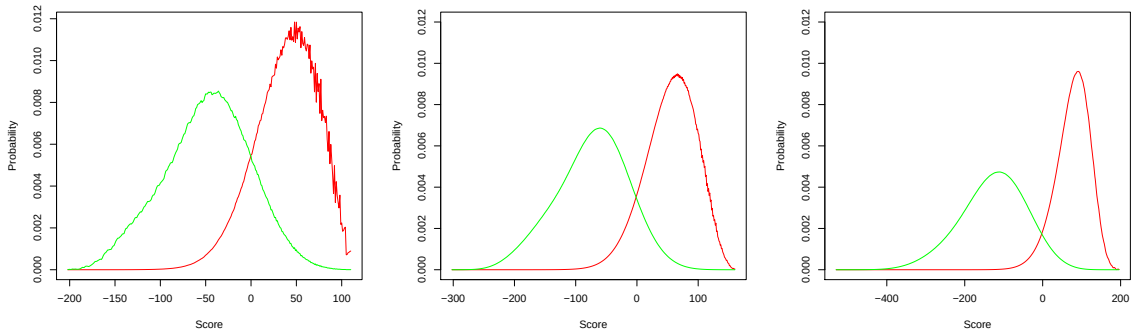


Figure 4.9: Score distributions under the signal (red) and the background (green) models in the case of PSSM (left), PSA formulated using the Halpern-Bruno model (middle) and the basic PSA (right). Note that the distance here is in terms of the branch length in Cis-Evolver, yielding a different substitution scoring matrix. While the HB-based PSA provides a better segregation than the single sequence case, the curves corresponding to the basic PSA are most separated. The figures are for poor profile (M00395) at distance 0.5.

Required Parameters: To run SimAnn and eSimAnn, we derived the respective substitution scoring matrix for each distance under the Jukes-Cantor model and uniform background frequencies. The gap costs are estimated as before (Section 4.2.1). The estimated gap costs for increasing branch length values are: $89, 40$ for branch lengths of 0.1 and 0.5, respectively.

For the pair-profile parameters, SimAnn is run with PSA calculated using the independence assumption, while eSimAnn is run with that calculated using the corresponding evolutionary model. Both SimAnn and eSimAnn, are run on each pair using a wide range of profile penalties. At each penalty, the predicted hit pairs are compared with the true locations to retrieve true and false positives (TPs and FPs). A prediction is called a true positive only if it overlaps exactly with the true site locations on both the sequences. Finally, the true and false positive rates are plotted at varying cutoff/penalty.

4.3.2 Results and analysis

Figures 4.10(a) and 4.10(b) show the results for eSimAnn (black) and SimAnn (blue) at two distances (0.1 and 0.5) for medium and poor quality profiles (TRANSFAC Ids M00690 and M00395, respectively) using the Halpern-Bruno model. Curves for the simulation using the Felsenstein model are attached at the end of the chapter (Figures 4.13(a) and 4.13(b)).

Discussion and conclusion

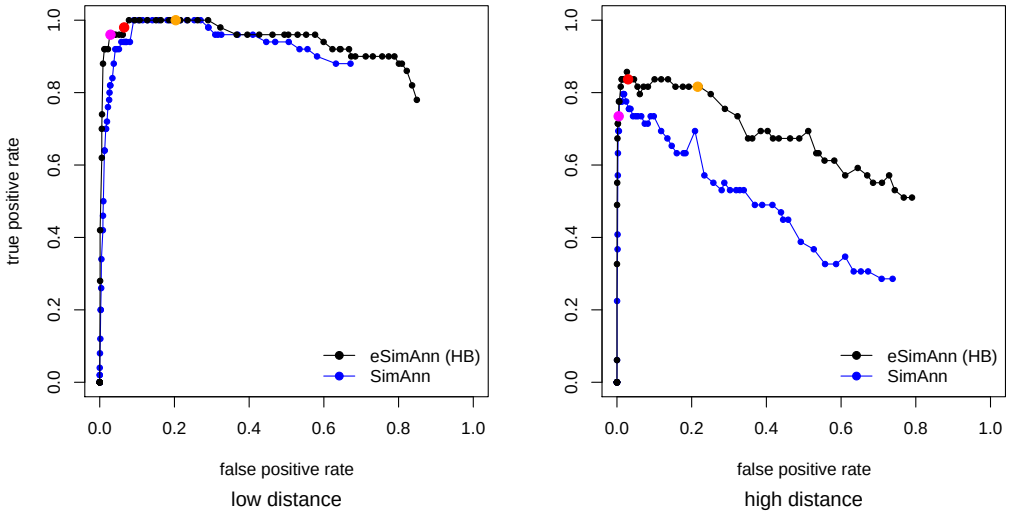
Both eSimAnn and SimAnn are founded on the same algorithmic premises; it is how a pair of strings is scored that yields the difference in performance. For modeling real binding site evolution, both the F81 as well as the HB model have been shown to be reasonable [133]. Incorporating information about binding site evolution, as provided through either model seems a logical step towards a biologically more meaningful scoring scheme. This is how eSimAnn works.

In eSimAnn, the PSA score as well as the profile penalty are calculated based on four main constituents – binding site description (given by the profile P), corresponding binding site evolutionary distribution ($\tilde{\phi}$), background (BG) sequence distribution (π) and background evolutionary distribution (ϕ) (see Section 3.2.2). In case of both the evolutionary models, the binding site evolutionary distribution relies on the information of the background sequence evolution (see Equations 3.6 and 3.5 for the general case and 3.7, 3.8 for the individual models). Since, in eSimAnn the signal is given by the binding site evolution, the corresponding signal score distribution hence depends on the background evolutionary distribution. Clearly, this introduces dependency on the background in eSimAnn, yielding a greater overlap between the signal and background score distribution curves. In SimAnn, the signal score distribution is calculated solely using the profile, resulting in a greater separation between the signal and background curves.

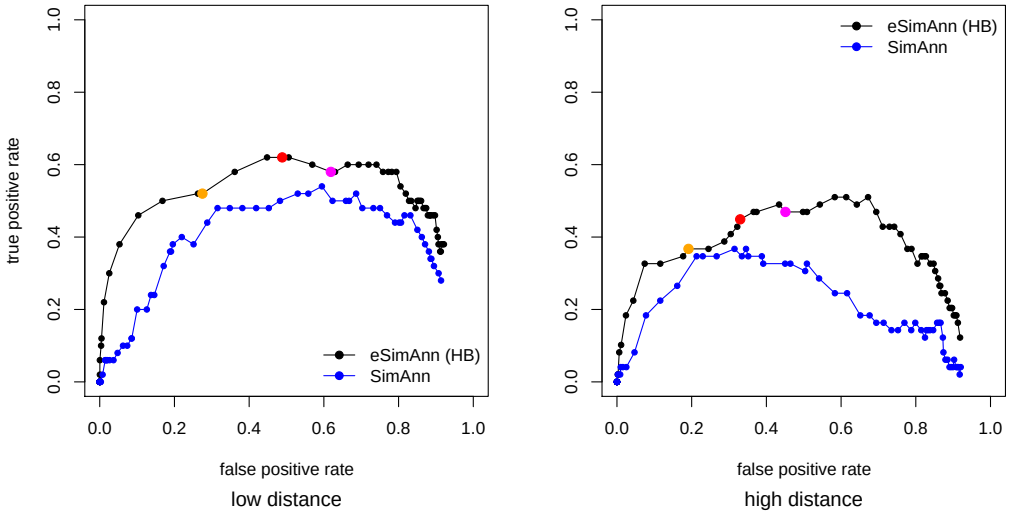
In Figure 4.9, the score distributions under signal and background are shown for the three cases: single sequence PSSM, HB-based PSA and the basic independent PSA. We can see that the curves for the HB-based PSA are more segregated than those for the PSSM, but less segregated than those for the basic PSA. Additionally, with increasing evolutionary distance, the separation increases (data not shown), approaching the separation of the basic PSA score distributions at large distances. This is in line with our discussion in Section 3.2.2, where we saw that as distance increases (ie. μt increases) and approaches infinity, the equations of eSimAnn approach those of SimAnn.

Despite the greater overlap between the signal and background in the eSimAnn score distributions, the TPR versus FPR curves show a significantly better performance in eSimAnn as compared to SimAnn (Figures 4.10 and 4.13). On the one hand, this indicates the inability of the theoretically derived penalties to be completely interpretable as the parameter that controls the final true and false predictions in the annotated alignment approach. On the other hand, the locations of the theoretically derived penalties (highlighted in the eSimAnn curves) indicate a reasonable proportion of the type I and type II errors.

In a simulation study, where the artificial sites are generated using the relevant evolutionary model, eSimAnn scores the true site pair more reasonably as compared to SimAnn. Since the setting is more in line with biological knowledge, eSimAnn seems to provide a better performing approach, while still being simultaneous.



(a) Medium quality profile (M00690)



(b) Poor quality profile (M00395)

Figure 4.10: *Simulation results for Halpern-Bruno model.* Comparison of eSimAnn (black) with SimAnn (blue) at low (0.1) and high (0.5) distances using profiles of medium (top) and poor (bottom) quality, respectively. The curves show the true versus false positive rate calculated at varying profile penalties. On the eSimAnn curves the theoretically derived profile penalties are highlighted. (orange – type-I error penalty at level 0.05; red – balanced penalty; cyan – type-II error penalty at level 0.05.)

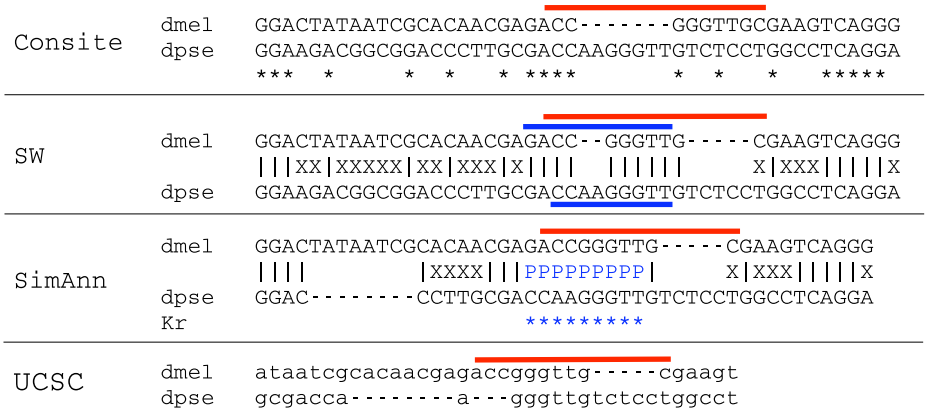


Figure 4.11: *Alignment region of the Krüppel 4 site.* Red lines indicate the true location, while the blue lines mark predictions made by respective methods. SW stands for the Smith-Waterman alignment used for the multi-step approaches.

4.4 Applications on real data

Having seen the performance of the simultaneous approach (both with and without evolution of TFBSs) in a controlled setting, we now proceed to illustrations via real data.

4.4.1 Extracting conserved binding sites in Drosophila: a case study

Even-skipped stripe 2 enhancer The even-skipped stripe 2 region in *D. melanogaster* and *D. pseudoobscura* is a well-characterized cis-regulatory module [187, 116, 115] containing multiple binding sites for at least four transcription factors: Bicoid, Hunchback, Giant and Krüppel. There is a total of 17 experimentally verified sites for these factors in this region and the corresponding count matrices are available [159]. We retrieved the orthologous enhancer sequences using the Genbank identifiers provided in [115]. The lengths of the individual sequences amount to 799 in *D. melanogaster* and 1028 in *D. pseudoobscura*.

We compared SimAnn with the multi-step approaches and a third tool called ConSite available online [170]. We consider ConSite because it is also a multi-step approach where first alignments are generated and conserved regions extracted. Then, sequences are scanned for putative hits using a score cutoff which does not consider the background letter distribution. Finally, only those hits that are situated in conserved regions and lie at equivalent positions in the alignment are output as conserved pairs.

Both SimAnn and our multi-step approaches are run with the standard HOXD70 substitution scoring matrix with gap open cost of 400 and extension cost of 30. The count matrices describing the relevant factors are preprocessed as described in Section 3.2.1 to calculate the basic profile related parameters for SimAnn. For sequence scanning within the multi-step approaches, count matrices are converted into scoring matrices and score cutoffs are determined along the lines of [158]. For ConSite, we use two main parameter settings, the default and with conservation and matrix score cutoff of 70%. Count matrices are same as above. In all methods we count a prediction correct if it overlaps with the known binding

site by more than quarter of the length of the PSSM. Overlapping predictions of the same PSSM are counted only once.

Out of 17 sites the Relaxed multi-step approach predicts 10 while the Strict predicts 9 sites correctly, with no false positives. The one site that is missed out by the Strict filter due to gaps in the alignment is the Krüppel 4 site. With ConSite, the default settings yield much fewer predictions, namely 5 out of 17. When the matrix score cutoff is lowered to 70%, this number increases to 10, at the cost of predicting additional 10 false positives.

When SimAnn is run with all four profiles together it predicts 9 sites correctly. We also run SimAnn supplying each profile individually to check whether overlapping binding sites pose a problem and this raised the number of true positives to 11. In first case we obtained 3 false positives, while in the second we obtained 4.

Overall, our multi-step approaches and SimAnn perform very similar, which is expected since they are based on the same premises algorithmically and parametrically. But ConSite has a slightly poorer performance since the gain in sensitivity by lowering cutoffs results in a drastic increase in the number of false positives, too.

It is worth looking in more detail at the Krüppel 4 site mentioned above because it is the only site which resides in a region of ambiguous alignment. The results of all methods are shown in Figure 4.11. The Strict multi-step approach and ConSite fail to predict the site because of gaps in the underlying alignment. ConSite predicts it, but only after the matrix score cutoff and the conservation cutoff are reduced to 60% and 40% respectively. The Relaxed multi-step approach and SimAnn successfully predict the site. However, with SimAnn the nice feature is that the binding site stands out more clearly. Through the UCSC alignment of the site, also shown in Figure 4.11, one can see that there is no clear *correct* alignment — the UCSC alignment differs considerably from the rest.

Comments Examples of known binding sites present in regions that have ambiguous alignments are difficult to come by. Usually, experimental knowledge is biased by information about conservation across species, thus yielding verified sites that are usually perfectly align-able. In such cases, the simultaneous approach provides only a marginal edge over the multi-step approaches. Still, there are examples (especially in Drosophila) where known binding sites have been found to lie in poorly conserved regions and using our proposed approach there might prove beneficial. Although a case-by-case analysis of such examples is beyond the scope of the present work, studying the differences in the alignments of the binding sites (simultaneous versus multi-step) would provide an insight into their evolutionary properties. A by-product of such an endeavour would be the compilation and analysis of a dataset of known sites that lie in regions with ambiguous alignments.

4.4.2 Evaluation on a human-mouse testset

Above, we illustrated how a detailed analsis of a cis-regulatory region can be performed with our approach. Our focus now is to highlight its applicability on a larger testset, in this case, of human-mouse sequences. In the following, we begin with an evaluation of SimAnn in context of two existing multi-step tools, ConSite [109] and SITEBLAST [131]. In the latter half of the section, we turn our attention to the evaluation of eSimAnn in comparison with

the state-of-the-art in multi-step tools, Monkey [134], that explicitly considers evolution of binding sites.

The setting consists of a testset of 110 uniquely mapped TFBSs in 57 promoters of human-mouse orthologous gene pairs, compiled by Lenhard *et al.* for the analysis of their method ConSite [109]. Removing those examples that contain ambiguous characters (Ns) results in a set of 98 experimentally verified binding sites in orthologous human-mouse sequence pairs, with high quality position-specific count matrices [168] provided.

Evaluation of SimAnn – ConSite and SITEBLAST

We briefly introduced the two multi-step methods ConSite [109] and SITEBLAST [131] in Section 1.5. We now provide a more detailed overview of the methods and present the respective results on the human-mouse testset.

ConSite: As noted earlier, ConSite is the prototypical conserved TFBS identification method with separate alignment and TFBS annotation steps. For generating alignments, the ConSite article mentions the use of the alignment algorithm DPB (unpublished) which optimizes for global alignment of long sequences containing short, colinear segments of similarity. However, the online tool itself uses a related algorithm ORCA (described in a later work [191]). In ORCA, short stretches are identified using BLASTN and the intermediate sequences are aligned using the Needleman-Wunsch algorithm. A user-defined conservation cutoff or that corresponding to the top 10% conserved windows in a scan of the alignment are used to extract conserved regions. For TFBS scanning, it uses highly curated PSSMs from the JASPAR database [168]. A substring is labeled a hit if it exceeds a user-defined matrix score threshold (or the default of 80%). The matrix score threshold itself is calculated without taking background sequence into account and hence is not guaranteed to be statistically significant. Finally, a conserved binding site is defined as one which has a TFBS hit on each sequence in a conserved window at equivalent positions in the alignment.

Since ConSite, as presented in the respective article, is only available online, we chose to use the analysis results of the authors on the initial dataset of 110 known sites as a coarse estimate of performance. All analyses were performed at a conservation cutoff of 70% and a window size of 50 base pairs. Two measures were used for performance evaluation – sensitivity and "predictive selectivity". The former is defined as the fraction of known sites detected, and the latter as the average number of predicted TFBSs per 100 bp of promoter sequence when scanning with *all* the `PSSM`s. A prediction is called a true positive when it overlaps with a known site by more than half the length of the shorter of the count matrix or the known site. Calculations are performed at varying matrix score thresholds and results presented as figures.

From the figures in the ConSite article, one can see that the sensitivity increases with decreasing matrix score thresholds: from $\sim 15\%$ at matrix score threshold of 90% to $\sim 70\%$ at a threshold of 65%. The "predictive selectivity" values increase from ~ 3 to ~ 100 predictions per 100 base pairs in the same threshold range. Clearly, the increase in sensitivity with decreasing matrix thresholds comes at a cost of increase in the proportion of false predictions. However, since the aim of the authors was to highlight the advantage of using orthologous sequence information as opposed to single sequence analysis, no measurements of the standard false positive rates (or specificity) have been discussed.

SITEBLAST: The tool SITEBLAST is a heuristic alignment algorithm where putative TFBS hits on individual sequences are used as seeds to generate the final alignment. Either consensus sequences or binding site profiles can be provided which are then scanned for matches against each sequence. For the latter, position-specific scoring matrices tailored to the GC-content of each sequence are formulated and the score threshold computed according to desired p-value (restricted to the range of 0 to 0.1) or power settings. The computations here are performed using the tool PATSER [80]. Once the seeds are found, it uses BLASTZ [176] to compute the alignment whereby each hit on one sequence is paired with each equivalent hit in the second sequence, respectively.

SITEBLAST is available as a downloadable stand-alone program and, as the first step, we ran SITEBLAST on the testset of 98 sites using the default settings, which implies a p-value cutoff of 0.001 and the HOXD70 DNA substitution scoring matrix. As is common with multi-step methods, SITEBLAST predicts overlapping hits and therefore, we calculated the true and false positive predictions at nucleotide level. A prediction is called a true positive if it overlaps by more than half the length of the shorter of either the count matrix or the known site (similar to [109] above). Multiple overlapping true hits are ignored and the number of correctly predicted bases is limited to the length of the respective matrix. It is important to mention here that SITEBLAST corrects for multiple testing and hence the resulting predictions are on the conservative side.

At default settings, SITEBLAST predicts only 2 out of the 98 known sites, yielding a true positive rate (`TPR`) of 0.0089 (0.8%) and no false predictions. Even on considering the maximum allowed p-value cutoff (0.099), the `TPR` of SITEBLAST only increases to 0.10149 (10%) and the `FPR` to 0.000835 (0.08%). As mentioned above, SITEBLAST adjusts for multiple testing – user-defined p-value is divided by the length of the scanned sequence. Clearly, this results in an overall low proportion of predictions, consequently leading to a low true as well as false positive rate. We then ran SITEBLAST with modifications to prevent multiple testing corrections. At the default p-value of 0.001 it resulted in a `TPR` of 0.1985 ($\sim$ 20%) and an `FPR` of 0.006835 (0.6%). At the other extreme of p-value at 0.099, the corresponding values increased to 0.437 ($\sim$ 40%) and 0.123 ($\sim$ 12%).

SimAnn results: For input to SimAnn, we calculated the independent `PSA` and the default balanced profile penalty for each profile. For the DNA substitution scoring matrix, we used the standard HOXD70 matrix, with a gap open cost of 400 and an extension cost of 30. Besides being shown to be suitable for human-mouse comparisons [42], it is also the default substitution matrix used by SITEBLAST. At these settings, SimAnn yields a `TPR` of $\sim$ 30% and `FPR` of $\sim$ 5%.

The default run of SimAnn yields a sensitivity ($\sim$ 30%) that is slightly greater than that of ConSite ($\sim$ 15%) at the higher end of the matrix score threshold range but much lower than at the higher end of the score threshold ($\sim$ 70%). On a separate analysis with 14 examples, the authors report that ConSite correctly predicted $\sim$ 80% of the true sites at the matrix score threshold of 60%. Our experience with the Drosophila case study on ConSite shows, as is also noted elsewhere [36], that the increased sensitivity at such low thresholds is accompanied by an increase in the proportion of false positives and a full-fledged analysis of the specificity of the method is missing in [109].

Comparison with SITEBLAST is not so straightforward. As we saw, the overall predictions of SimAnn far exceed those of SITEBLAST, with both the true and false positive rates greater than those of SITEBLAST – $\sim 30\%$ and 5% as opposed to 10% and 0.08% at the highest allowed p-value in SITEBLAST. Varying the p-value between the allowed range of $0 - 0.1$ also does not yield comparable results with the corresponding variation of the type I cutoff in SimAnn (data not shown). Modifying SITEBLAST to not correct for multiple testing, results in an improvement in the total number of predictions, with the `TPR` $\sim 20\%$ and `FPR` $\sim 0.6\%$ at the default p-value of 0.001. When SimAnn is run with a level 0.001 type I cutoff, the corresponding true and false rates are 27% and and 3%, indicating the slightly better performance of SITEBLAST. This difference in the amount of false predictions rises due to the fact that the dataset consists of well-conserved regions which are comparably straightforward to align. Lowering the profile penalty simply encourages the SimAnn algorithm to annotate aligned substrings as pair-profile hits.

Evaluation of eSimAnn – Monkey

We now turn to the evaluation of the extended version eSimAnn which allows for the possibility to explicitly incorporate evolution of binding sites. To this end, we compare eSimAnn with an existing method Monkey.

Monkey is the state-of-art multi-step method that also employs TFBS-evolution explicitly. Based on putative TFBS hits on a reference sequence, Monkey scans corresponding regions of other sequences in a multiple alignment to identify putative conserved TFBS hits. For each such hit, Monkey outputs a p-value estimate based on the Halpern-Bruno model. As with other multi-step methods, Monkey needs to deal with gaps in the aligned hit locations. To this end, it uses a heuristic that conservatively allows some gaps while disallowing too many.

To run eSimAnn, we again used the HOXD70 substitution scoring matrix. Since Monkey uses the HB model for binding site evolution, the profile-related parameters in eSimAnn (`PSA` and *pen*) are calculated using the same. For modelling background evolution, we simply used the Jukes-Cantor model with uniform background frequencies. We ran Monkey on alignments generated using ClustalW with the default parameters. The same distance as estimated from the HOXD70 matrix and Jukes Cantor model was used and the background frequencies set to uniform. Similar to SITEBLAST, Monkey outputs overlapping hits and the definitions of true andd false predictions remain the same.

Monkey outputs a list of putative hits with the associated p-values making it non-trivial to decide on the appropriate pvalue threshold. To deal with this, we consider a range of pvalue thresholds. We ran eSimAnn with profile penalties calculated for each of these pvalue thresholds and plotted the resulting proportion of true and false predictions under both methods (Figure 4.12). As can be seen, eSimAnn performs comparably to Monkey at reasonable false positive levels (below 0.05). However, at extremely low pvalue thresholds, the eSimAnn curve is lower than that of Monkey. This is because, in Monkey, the pvalue threshold is applied only on the candidate individual sequence TFBS hits, thus limiting the false positives. In eSimAnn, on the other hand, the algorithm fits in maximum pair-profile hits into the alignment at drastically low profile penalties. Thus, for reasonable false positive rates, eSimAnn performs well as compared to Monkey, while giving the benefit of a simultaneous approach that predicts perfectly aligned TFBSs.

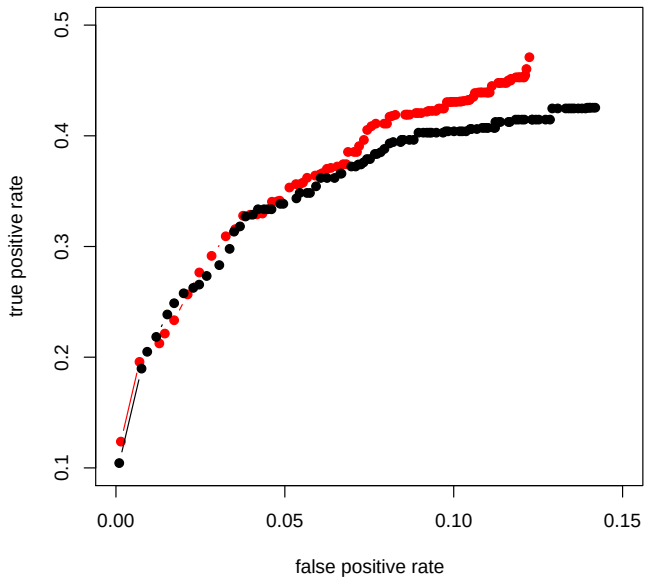


Figure 4.12: *Real-data analysis.* True versus False positive rates on real testset of human-mouse sequences with experimentally verified binding sites for eSimAnn (black) and Monkey (red).

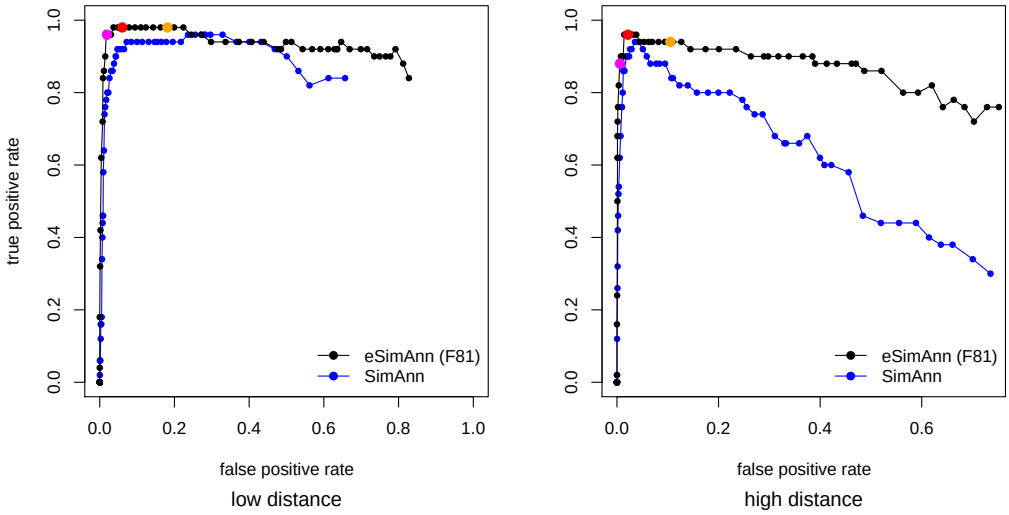
4.5 Summary and Discussion

In this chapter, we demonstrated the advantages gained by a simultaneous alignment and annotation approach. Through simulations, we showed how parameters can be estimated for any profile of interest without the need of training data. Moreover, perfectly aligned binding sites are predicted in one step obviating the necessity to use post-processing strategies for handling mis-alignments. And finally, local rearrangments in the alignment help in bringing forth putative conserved binding sites. Especially in the case of poor sequence conservation or profile quality, the approach performs better than a multi-step strategy.

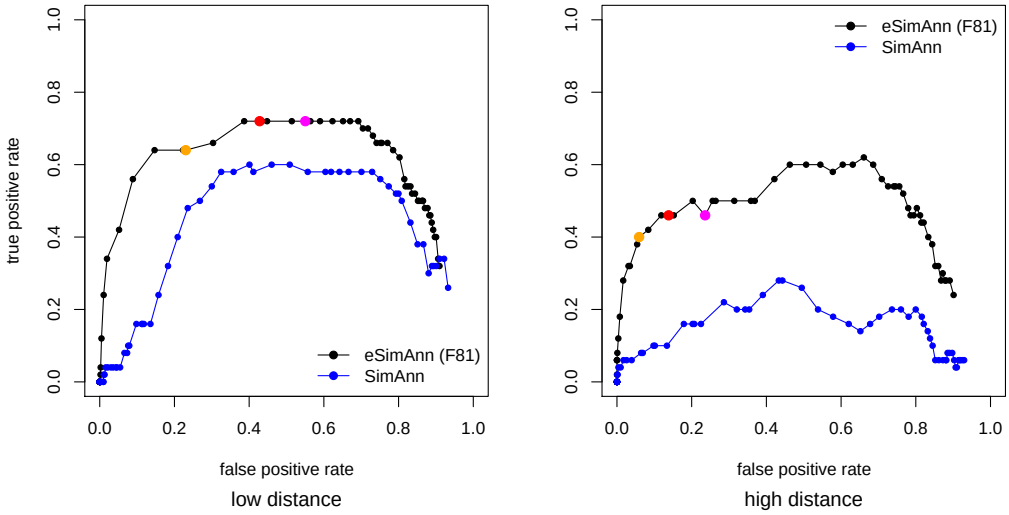
In real-world setting, the annotated alignment approach is best suited for detailed analysis of a regulatory region known to be conserved between two species with available information of certain essential transcription factors. When conservation is weak and it is difficult to identify conserved binding sites, the approach can assist a lot in understanding the potential regulatory mechanisms in the region. As an illustrative case study, we used the Drosophila eve stripe 2 enhancer. However, for analyzing arbitrarily big conserved regions with a large number of profiles, the approach is not particularly suited. The resulting multiple testing problems and the increased complexity of the extended alignment model could hinder performance and well-established multi-step approaches may be more preferable.

With regards to larger testsets, most compiled datasets of experimentally verified binding sites consist of well-studied instances, which are usually well-conserved too. The performance of the annotated alignment approach is then comparable to multi-step tools, as we saw in the analysis with human-mouse data. In a recent development, Moses *et al.* [135] focussed on the large-scale characterization of the evolutionary properties of the binding

sites of the transcription factor Zeste in Drosophila. Using an updated version of the tool Monkey [134] and adopting strategies to deal with errors in the pre-computed alignments, they hypothesized about the large scale evolutionary loss or gain of the binding sites. The fact that analytical measures are needed to handle mis-alignments stresses on the importance of the underlying alignments. Using a simultaneous approach that combines binding site characteristics with background sequence conservation properties may be advantageous in such a setting.



(a) Medium quality profile (M00690)



(b) Poor quality profile (M00395)

Figure 4.13: *Felsenstein 1981 model.* Comparison of eSimAnn (black) with SimAnn (blue) at low (10) and high (50) distances using profiles of medium (top) and poor (bottom)quality, respectively. The curves show the true versus false positive rate calculated at varying profile penalties. On the eSimAnn curves the theoretically derived profile penalties are highlighted. (orange – type-I error penalty at level 0.05; red – balanced penalty; cyan – type-II error penalty at level 0.05.)