Cost functions for finding regulatory elements with pairwise alignment programs

When searching for motifs in multiple alignments, a primary level of conservation already exists, embedded in the multiple alignment, and the tools only detect regions of maximum local conservation. sim addresses a different category of problems, detecting good local similarities between two unaligned sequences. Although not intrinsically a motif finder, it can be used to extract candidate regulatory regions by comparing non-coding sequences from species that are relatively distant but which share similar regulatory apparatus. Sequence comparisons between human and mouse could be particularly informative, these being the two most extensively studied organisms that conform to our criteria.

In an ideal setting, only functional sites common to the two species would be conserved, and each would be supported with thorough experimental evidence. In practice, however, comparing promoter regions for homologous genes reveals regions of good conservation with no documented regulatory function, and whether these are functional sites which haven't been investigated yet, or simply regions of lower mutation rate, is an open question. With the fine-grained, nucleotide-level cost function of the previous section these positions will appear as false positives, and thus be penalized in the assessment. Ultimately, if the total length of the functional regions is relatively small in comparison to the region's length, the best FN+FP count could be obtained for the trivial case of the empty alignment. Hence the scoring model of the previous sections is no longer appropriate. The fact that the reference site sequences compiled using TRRD, EPD, and the literature often extend fairly far beyond their agreed consensus regions (e.g., the binding site sequence for TBP within the HSP70 human promoter is listed as tgacgacTTATAAAgcccagggg) further complicates the problem. Based on these observations we suggest two alternative scoring schemes, open to investigation.

The first is to use a coarser cost function that does not incorporate false positives. For example, "hit-and-miss" scoring schemes based on partial matches between the reference sites and the regions reported by the program meet the above criteria. With such an approach, a hit can be declared if, for instance, the fraction of the landmark covered by the match exceeds a fixed threshold. To compensate for a scoring scheme biased toward a "principal" sequence, a unique global score that takes both sequences into account can be computed as:

$$\text{TP\_hit\_ratio} = \frac{\text{N\_com\_sites\_found(human)} + \text{N\_com\_sites\_found(mouse)}}{\text{N\_com\_sites\_total(human)} + \text{N\_com\_sites\_total(mouse)}} ,$$

where N_com_sites_total(x) denotes the number of sites known to be common to both species and annotated for species x, while N_com_sites_found(x) is the number of such sites that were found by the program. Note that sites found by the program which are not in the reference set, i.e., false positives, are ignored.

In general, such approaches capture the sensitivity but not the specificity of the programs evaluated and favor parameter settings that produce long alignments with low similarity scores. To prevent this, a complementary measure for specificity should be computed using the sites that are known to be organism-specific and thus should not be reported by the program, e.g.:

$$\text{TN\_hit\_ratio} = \frac{\text{N\_spec\_sites\_notfound(human)} + \text{N\_spec\_sites\_notfound(mouse)}}{\text{N\_spec\_sites\_total(human)} + \text{N\_spec\_sites\_total(mouse)}} .$$

However, this approach is limited to the regions where data on both types of regulatory sites, common and specific, are available.

A second alternative is to maintain a fine-grained scoring function that would still satisfy the requirement for a small false positives plus false negatives count while seeking a balance between the two components. Figure .n. shows how these two indicators, calculated as fractions of the maximum attainable values, vary with the parameter set for sim, for the CEBPA promoter region. During the experiments, the score for a nucleotide match was kept constant, while all penalties were varied linearly by a factor of λ. Two types of measurements were performed, namely, with respect to all functional sites in each species, and with respect to the common sites only. In both cases, the errors were measured separately in each organism: in Figure .n., for instance, a false positive corresponds to a position in the mouse sequence that was included in the alignment reported, but which did not belong to any common (Fig .n.A) or non-differentiated (i.e. either common or organism-specific, Fig .n.B) reference sites in mouse. Analogous representations can be obtained for human.

The Correlation Coefficient (CC) corresponds to the Pearson product-moment correlation coefficient in the particular case of two binary variables, "prediction" (with the values "predicted as functional" and "predicted as non-functional"), and "reality" (with the values "functional" and "non-functional") in this context. Its long recognized virtues as a good global measure for overall accuracy recommended it as a preliminary quality indicator for our assessments.

The experiments reinforced our position that the results could be highly dependent on the species used for reference. On the graphs of Figure .n+1., different sets of λ values produced the maximum values of the CC in the two organisms: λ=2.0 for human (panel A), and λ=1.4,1.5 for mouse (panel B), regardless of the set of reference sites used ("common", "all"). The exact values of the maxima are included in the table of Figure .n+1.C. Interestingly, while the set of program-parameter values for the human-"best" lambda (2.0) still scores very well in measurements against functional sites in mouse (panel B) when only the common sites are used (CC=0.27), it performs very poorly when all the sites in the mouse are considered (CC=0.05). Furthermore, a wide range of λ-values which produce CC maxima in mouse-"common"-sites measurements are far from optimal when all sites are considered (λ=1.6, 1.7, 1.8). Intuitively, there is a good correlation between the "common" sites in human and mouse, on one hand, and between the "common" and non-differentiated sites in human, on the other hand, but a relatively poor correlation between the "common" and non-differentiated sites in mouse. Hence, one would expect to find functional landmarks in mouse that are clearly divergent from their homologous sequences in human. Conversely, the CC values corresponding to the mouse-"best" λ=1,4, 1.5, measured against all of the sites in human (panel A), drop below 0. Consequently, the choice of organism and the set of reference sites, if not carefully selected, could bias the conclusions of optimization. In the experiments above, a value that produces optimum or near-optimum results in both human and mouse measurements when only the common sites are used is λ=2.0 and can be chosen with high assurance to tune the program's parameters.

Type of experiment	λ_{opt}	CC_{opt}
common_human	2.0	0.22
all_human	0.8 – 1.3, 2.0	0.06
common_mouse	1.4 – 1.8	0.28
all_mouse	1.4, 1.5	0.17

Figure 1: Optimum lambda values