Supplementary material for Alignment Tool Evaluation

Weichun Huang[1*], Joseph R Nevins, and Uwe Ohler[*]

Institute for Genome Sciences and Policy, Duke University, Durham, NC 27708
[1]Current address: Department of Biology, Boston College, Chestnut Hill, MA 02467

Evaluation on star tree simulation data

Based on the star phylogenetic tree, we used PSPE to generate benchmark promoter sequences at 15 different divergence distances. The data were simulated under the HKY85 nucleotide substitution model with Gamma and invariant rate (Γ+I) for modeling substitution rate heterogeneity. For each divergence distance, we generated 1,000 replicate homologous sets, each having four promoter sequences with the same divergence distance from their ancestral sequence. Each sequence contained exactly one functional binding site for each of the six transcription factors: Pax6, TP53, IRF2, PPARG, ROAZ, and YY1E2F. YY1E2F is a composite TFBS consisting of YY1 and E2F binding sites that reportedly interact with each other in cell cycle gene regulation [1]. Binding sites were subject to a set of functional constraints (Table 1) which were set to allow for turnover within a restricted distance, but keeping the overall order of the binding sites unchanged. Simulation allowed us to quantify the amount of turnover, how many non-aligned functional sites were due to turnover as compared to "simple" misalignments, and whether some tools would in fact be able to align functional sites despite turnover. We used this dataset to assess performances of five widely-used MSA tools: CLUSTALW [2], DIALIGN [3], AVID/MAVID [4, 5], LAGAN/MLAGAN [6], and MUSCLE [7]. The performance was measured as TFBS detection accuracy, defined as the proportion of nucleotides in functionally homologous TFBS which were correctly aligned. The detection accuracy reported here is the average value over 1,000 replicates at each divergence distance.

We compared the performance of the five tools in aligning sequences of two, three and four species, respectively. For two species (Figure 2), MUSCLE showed the highest overall detection accuracy (average over all of functional TFBS) across different divergence distances; LAGAN/MLAGAN performed better than AVID/MAVID, CLUSTALW, and DIALIGN for sequences of intermediate and large divergence distances; and CLUSTALW was slightly better than DIALIGN and AVID/MAVID for sequences of short and intermediate divergence distances. For three (Figure 2B) and four species (Figure 2C) alignments, MUSCLE still had the best overall performance, but DIALIGN gradually overtook the other three tools, whose relative performance order with respect to each other remained unchanged. The TFBS detection accuracy decreased as divergence distances increased for all tools.

For each tool, there were also significant differences in performance on different TFBS, and differences became more pronounced as sequence divergence increased. For example, in four species alignment, all tools were better at aligning YY1E2F and Pax6, which had low replacement turnover rates and short restricted distance for translocation, than for IRF2 and ROAZ, which had higher turnover rates and long restricted distances for translocation (Figure 3). Besides the restricted distance for translocation, other properties of TFBS, such as length, nucleotide composition and distance to neighboring TFBS, could have significant impact on its detection accuracy. For example, PPARG had a similar low turnover rate as TP53, but each tool had higher detection accuracy on TP53 than on PPARG. The degree of performance variation among TFBS was not always consistent among different tools; for instance, DIALIGN performed better on PPARG than MUSCLE, which had the highest detection accuracies for all other TFBS.

We also assessed the performance of each tool separately on aligning sequences of two, three and four species, respectively (Figure 4). Contrary to the belief that more distantly related species help to locate functional conserved sites, we found that the increase in number of species did not necessarily increase the TFBS detection accuracies of all tools. AVID/MAVID and LAGAN/MLAGAN showed a decrease in performance as the number of species increased, and the decrease was more significant with increasing divergence distance. CLUSTALW showed the same tendency, but difference in performance was less significant. Interestingly, MUSCLE had no significant difference in performance as the number of species increased, while DIALIGN improved its performance markedly across different divergence distances (Figure 4).

We made additional evaluations on three more promoter sequence datasets. The three datasets were simulated by PSPE using the same parameters except for one change each: the first one using a zero order model Markov model for background sequence simulation, the second without using Γ+I for rate heterogeneity, and the third using a different set of TFBS. The results were largely consistent with those reported above (see Supplementary Information [8]). Furthermore, we compared performances of the five tools in terms of their overall alignment sensitivity and TFBS sensitivity. We found that MUSCLE and CLUSTALW had slightly better overall alignment sensitivity than the other three, and the rank of TFBS sensitivities were in the same order as the detection accuracies (see Figure 5).

Evaluation on mammalian tree simulation data

The above evaluation on simulated orthologs of equal distance from the last common ancestor provided initial results about how different MSA tools perform as sequence divergence increases. In real applications, it is more common to observe species having different divergence distances from their last common ancestral sequence. It also generally assumed that an MSA tool should work better when aligning more closely related species at the beginning stage and adding more distantly related species in later stages, especially for those based on a progressive approach. Therefore, we additionally compared tool performance on simulated promoter sequences of five mammalian species in an attempt to arrive at a fair and more realistic assessment of the five MSA tools.

We applied the same evolution models and transcription factors as above to simulate promoter sequences, but used a phylogenetic tree of five mammalian species (Figure 1B). We scaled this mammalian tree at 10 different levels from 0.25 to 10, relative to the distances shown, and generated a sequence data set at each scale level (defined as divergence scale coefficient), where each dataset contained 1,000 replicates of orthologous promoter sequences of the five species. We used each tool to align the sequences and calculated its TFBS detection accuracy, and report the average detection accuracy over 1,000 replicates at each scale level.

For the two species (human and baboon) alignment, all five tools showed high detection accuracies of TFBS with no significant difference between each other (Figure 6A). When adding more distant species such as mouse to the alignment, we found that TFBS detection accuracies of all tools were dramatically decreased, especially those of MAVID and CLUSTALW (Figure 6B, C, D). Again, we observed marked differences in performance between different tools for three or more species alignments. Overall, MUSCLE had the highest detection accuracy among all tools across all divergence scale coefficients; MAVID had a slightly worse performance than all others; and CLUSTALW, DIALIGN and MLAGAN showed similar performance, although their relative order in performance varied with the number of species or a change of the divergence scale coefficient. As expected, the TFBS detection accuracy decreased for all tools as divergence scale coefficient increased.

The ability of a tool to detect the presence of a common TFBS varied among different TFBS, depending on TFBS base composition, length, and restricted translocation distance, as well as the divergence scale coefficient of the phylogenetic tree. For example, Figure 7 shows that detection accuracies differed significantly among TFBS in the alignments of the five species. In addition, the same figure shows that all tools had higher detection accuracies for TFBS with low replacement turnover rates, such as YY1E2F and Pax6, than those with high replacement turnover rates, such as IRF2 and ROAZ. While MUSCLE showed a better performance than all other tools, CLUSTALW as the oldest tool performed slightly better than DIALIGN, MAVID, and MLAGAN in at least some cases (YY1E2F and ROAZ). Additionally, for YY1E2F, Pax6 and TP53, MUSCLE showed higher TFBS detection accuracies than the baseline of SimuALN, suggesting its capability of correctly aligning at least some TFBS subject to turnover, *i.e.* homologous only at the functional level. At large divergence scale coefficients, however, no tool seemed to perform well in detecting ROAZ.

When looking at the performance of each tool individually (Figure 8), we found that the TFBS detection accuracies of all tools decreased when adding one or more distant species to the human/baboon alignment. For alignments from three to five species, the TFBS detection accuracies of DIALIGN and MUSCLE showed little change, those of CLUSTALW and MLAGAN had a noticeable change and that of MAVID markedly decreased, especially at large divergence scale coefficients. Interestingly, MLAGAN showed better performance in detecting TFBS for five species alignments than for four species alignments, which we did not observe in our evaluation on the star tree simulation data. We also compared tool performance again with respect to overall alignment sensitivity and TFBS sensitivity. We found that MUSCLE and CLUSTALW had slightly better overall alignment sensitivity than the other three, and the rank of TFBS sensitivities were in the same order as those of their detection accuracies (Figure 9).

References

1. ***Schlisio S, Halperin T, Vidal M, Nevins JR: Interaction of YY1 with E2Fs, mediated by RYBP, provides a mechanism for specificity of E2F function. Embo J 2002, 21(21):5775-5786.***

2. ***Thompson JD, Higgins DG, Gibson TJ: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994, 22(22):4673-4680.***

3. ***Morgenstern B: DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics 1999, 15(3):211-218.***

4. ***Bray N, Dubchak I, Pachter L: AVID: A global alignment program. Genome Res 2003, 13(1):97-102.***

5. ***Bray N, Pachter L: MAVID multiple alignment server. Nucleic Acids Res 2003, 31(13):3525-3526.***

6. ***Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S: LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 2003, 13(4):721-731.***

7. ***Edgar RC: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004, 32(5):1792-1797.***

8. ***Huang W: PSPE Online Supplementary Materials at http://biomedempire.org. In.; 2006.***

9. ***Vlieghe D, Sandelin A, De Bleser PJ, Vleminckx K, Wasserman WW, van Roy F, Lenhard B: A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. Nucleic Acids Res 2006, 34(Database issue):D95-97.***

10. ***Siepel A, Haussler D: Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol Biol Evol 2004, 21(3):468-488.***

Tables

Table 1: Functional TFBS constraints used in the promoter simulation. The accession numbers in the second column are from the JASPAR database [9]. "Location" refers to the restriction on the upstream minimum and maximum distances to transcription start site. YY1E2F is a composite TFBS created by joining the YY1 and E2F sites.

Name	**Accession #**	**Len**	**Strand**	**Location (min, max)**	**Copy # (min, max)**	**Cutoff**
YY1E2F	MA0095 (YY1) MA0024 (E2F)	13	+	(20, 30)	(1, 1)	0.90
Pax6	MA0069	14	+	(50, 70)	(1, 1)	0.90
TP53	MA0106	20	+	(360, 400)	(1, 1)	0.90
IRF2	MA0051	18	+	(420, 480)	(1, 1)	0.90
PPARG	MA0066	20	+	(2000, 2080)	(1, 1)	0.90
ROAZ	MA0116	15	+	(2100, 2200)	(1, 1)	0.90

Table 2: Simulation parameters used by PSPE for generating benchmark promoter sequences.

Evolution distance per step	0.05 substitution per site
Length of root sequences	3000 *bps*
Background sequence model	Markov Order of Third
Base frequencies	A=0.258, C=0.242, G=0.242, T=0.258
Substitution Model	HKY85
Transition/Transversion Ratio	20:1
Rate heterogeneity	Gamma (1.0) + Iota (0.1)
Range of GC content	(0.45, 0.55)
Gap model	Negative Binomial Distribution (1, 0.5)

Figures

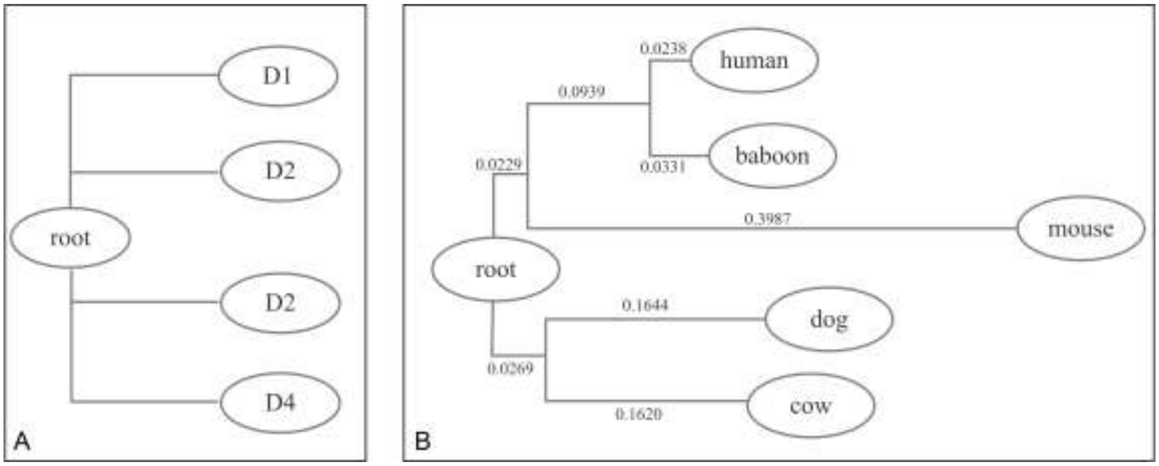

Figure 1: The two phylogenetic trees for promoter sequence simulation. (A) The star topology. In the star tree, four descendants (node D1 to D4) are evolved independently from the root sequence, and they have the same divergence distance from the root. We used D1 and D2 for two species alignments, and D1, D2 and D3 for three species alignment. (B) The phylogenetic tree of five mammals. The evolutionary distances shown in the tree were recently inferred from the coding region of orthologous genes [10]. In our simulation, we used the tree scaled at 10 different levels relative to the evolutionary distances shown.

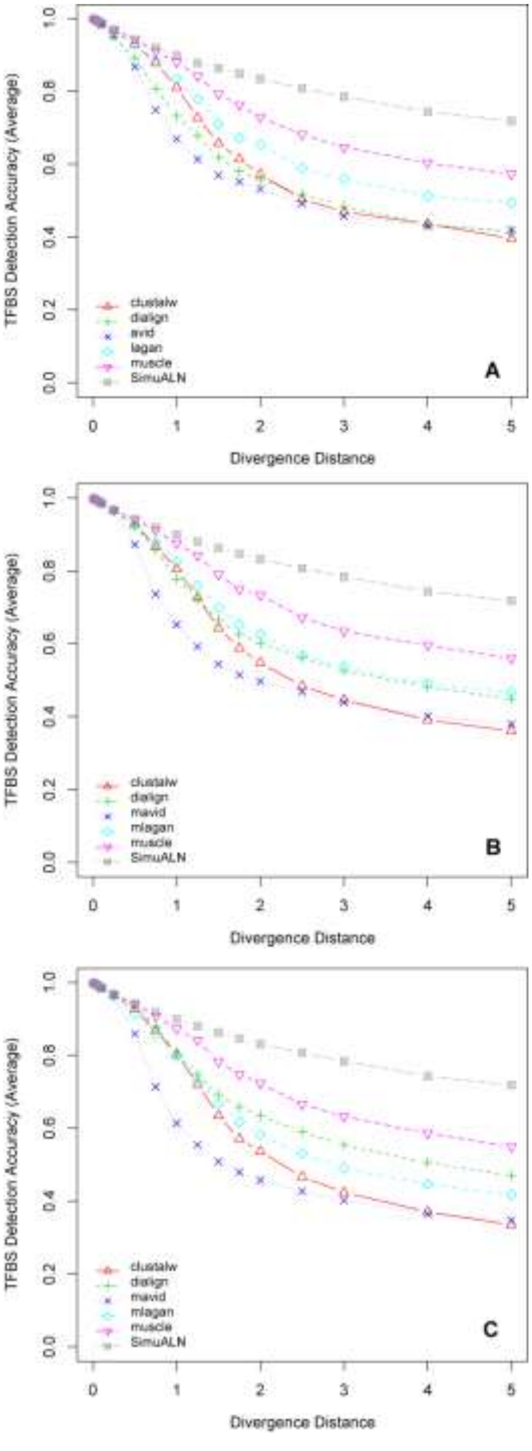

Figure 2: Performance comparison of alignment tools for TFBS detection accuracy. The Y-axis is the TFBS detection accuracy, the X-axis is the divergence distance measured by the number of substitutions per site. The SimuALN stands for the simulated alignment and its measure indicates the proportion of TFBS not subject to replacement turnover in the descendent sequences, and thus aligned in simulated alignments. (A) two species alignments, (B) three species alignments, and (C) four species alignments.

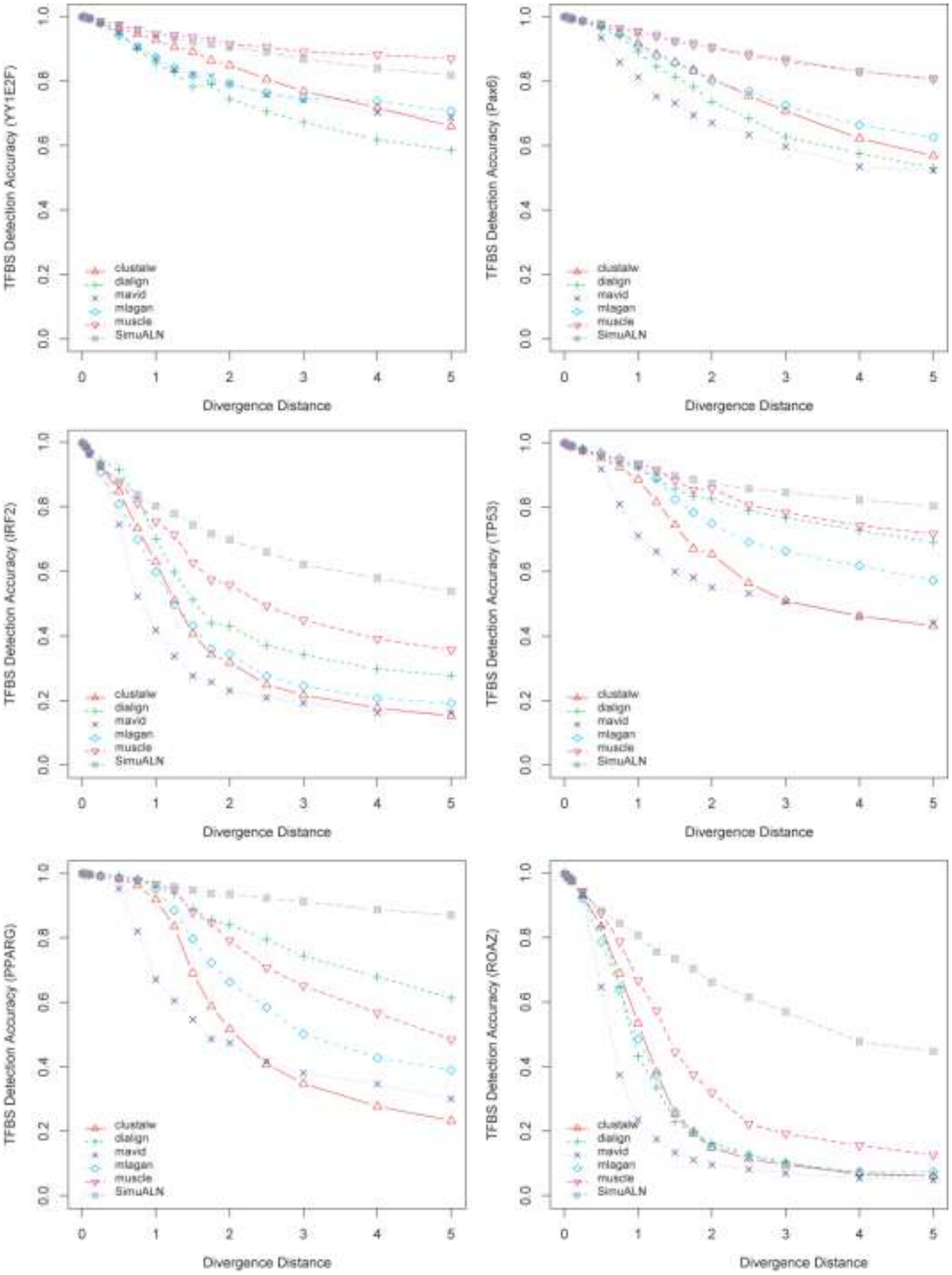

Figure 3: The detection accuracy on individual TFBS in four species alignments. All five tools perform better on detecting TFBS YY1E2F and Pax6, which have low replacement turnover rates and a short restricted distance for translocation, than on detecting IRF2 and ROAZ, which have high turnover rates and long restricted distances for translocation. Overall, MUSCLE performs superior to other four tools, while DIALIGN shows good performance on detecting TP53 and PPARG, which have long restricted translocation distances but relatively low replacement turnover rates.

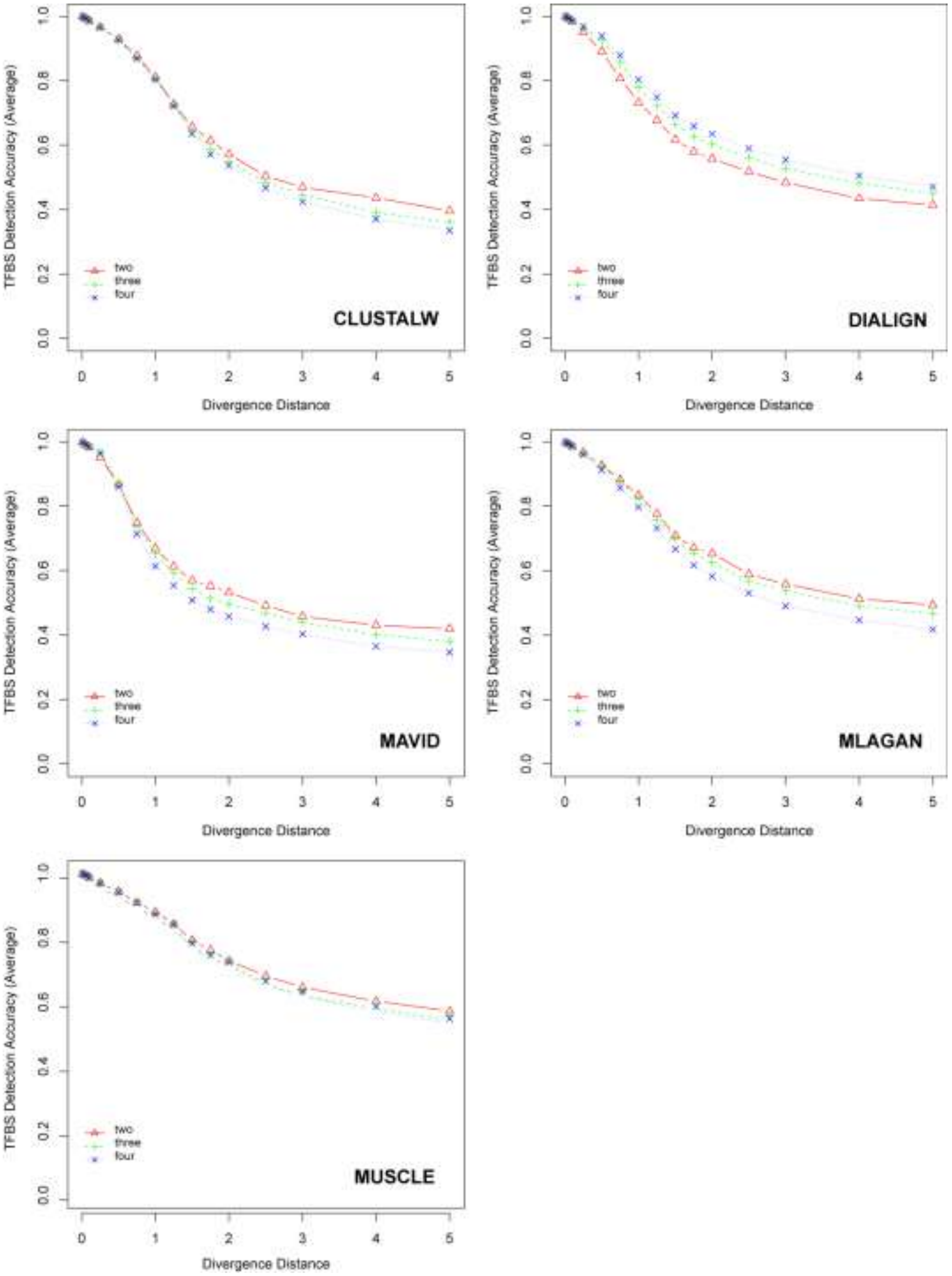

Figure 4: The effects on TFBS detection accuracy of five alignment tools as the number of species increases. Each subfigure shows a comparison of TFBS detection accuracy of the tools on aligning promoter sequences of two, three, and four species, respectively. The figure shows that the performances of CLUSTALW, AVID/MAVID and LAGAN/MLAGAN decrease as the number of species increases, especially at large divergence distances. The performance of MUSCLE is relatively unaffected; only DIALIGN shows improvement.

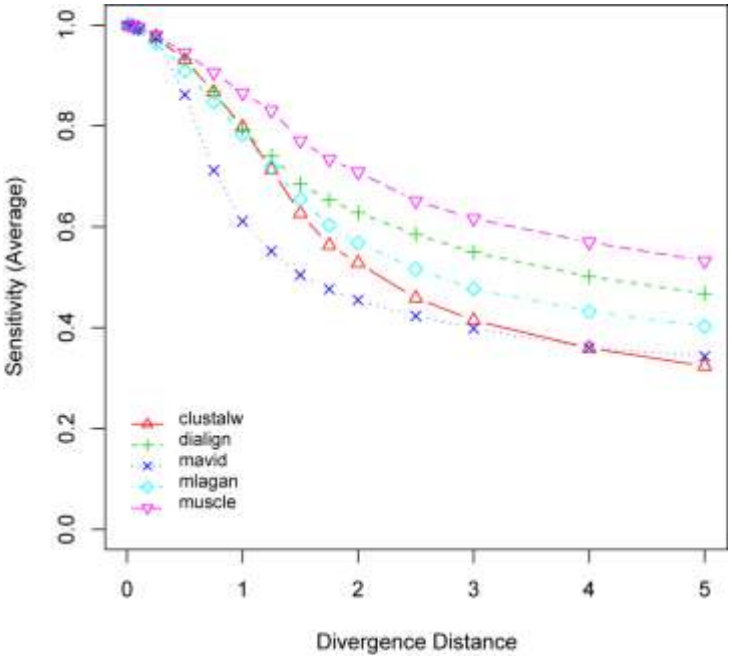

Figure 5: The average alignment sensitivity of TFBS on four species alignment. The relative order of TFBS sensitivity is almost identical to the order of TFBS detection accuracy (see Figure 2C).

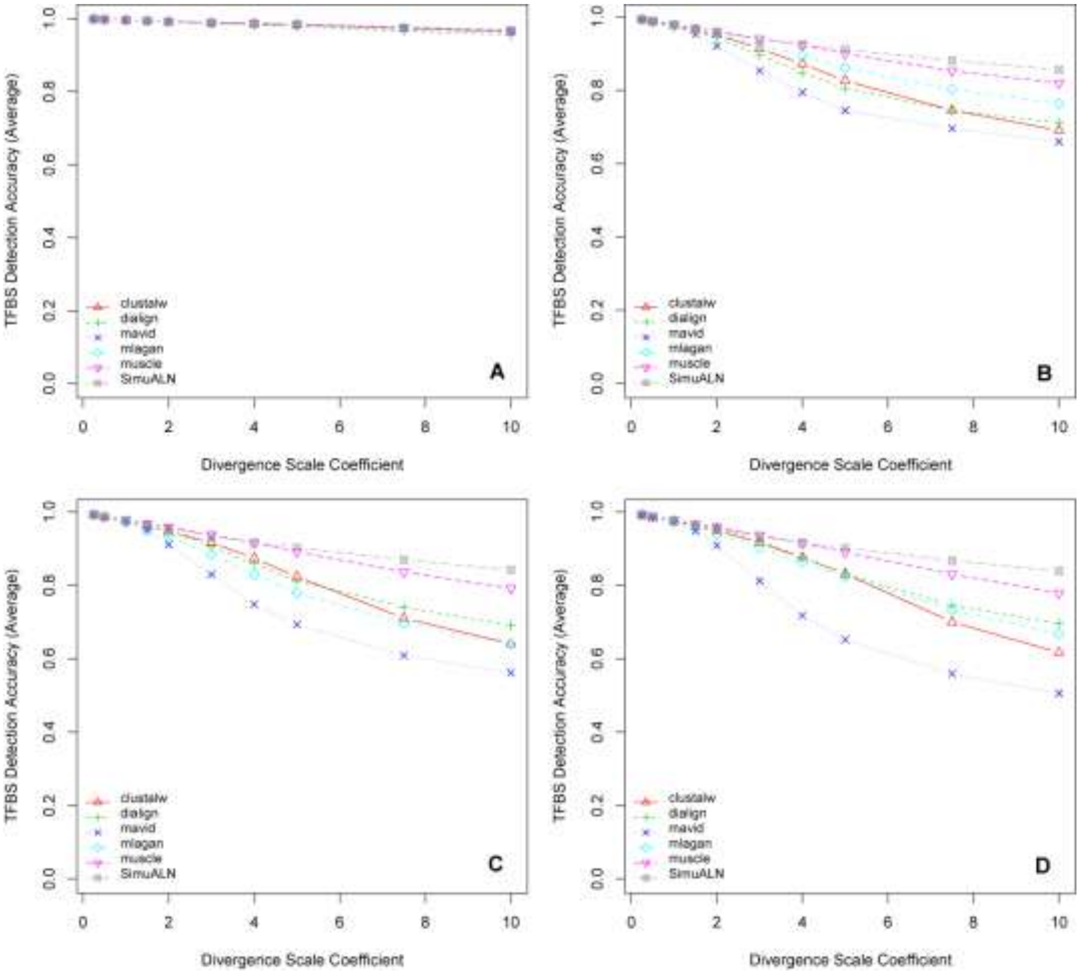

Figure 6: The average TFBS detection accuracy of five tools for mammalian sequence alignment. The Y-axis is the TFBS detection accuracy average on six TFBS, and the X-axis is the divergence scale coefficient of the mammalian phylogenetic tree (Figure 1B). The SimuALN stands for the simulated alignment and its measure indicates the proportion of TFBS not subject to replacement turnover in descendent sequences, and thus aligned in simulated alignments. (A) Two species alignments of human and baboon. (B) Three species alignments of human, baboon and mouse. (C) Four species alignments of human, baboon, mouse, and dog. (D) Five species alignment.

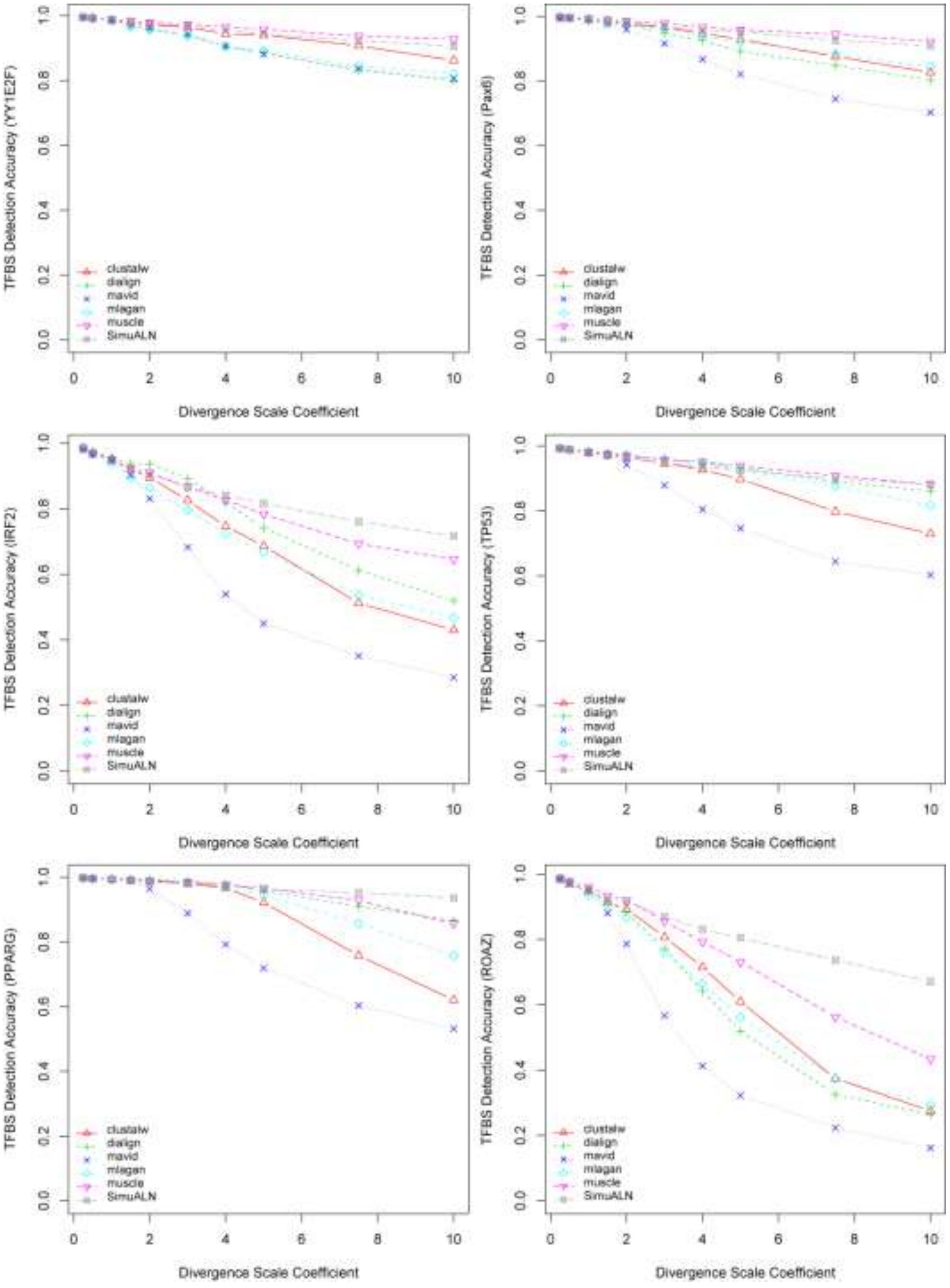

Figure 7: The detection accuracy of individual TFBS on five-way mammalian alignments. All five tools perform better on detecting YY1E2F and Pax6 which have low replacement turnover rates and short restricted distance for translocation than on detecting IRF2 and ROAZ which have high turnover rate and long restricted distance for translocation. MUSCLE shows an overall better performance than the other four tools. MLAGAN performs better than DIALIGN on YY1E2F, PAX6, PPARG and ROZA, while DIALIGN shows a better performance than MLAGAN on TP53 and PPARG, which have a long restricted distance for translocation but a relatively low replacement turnover rate.

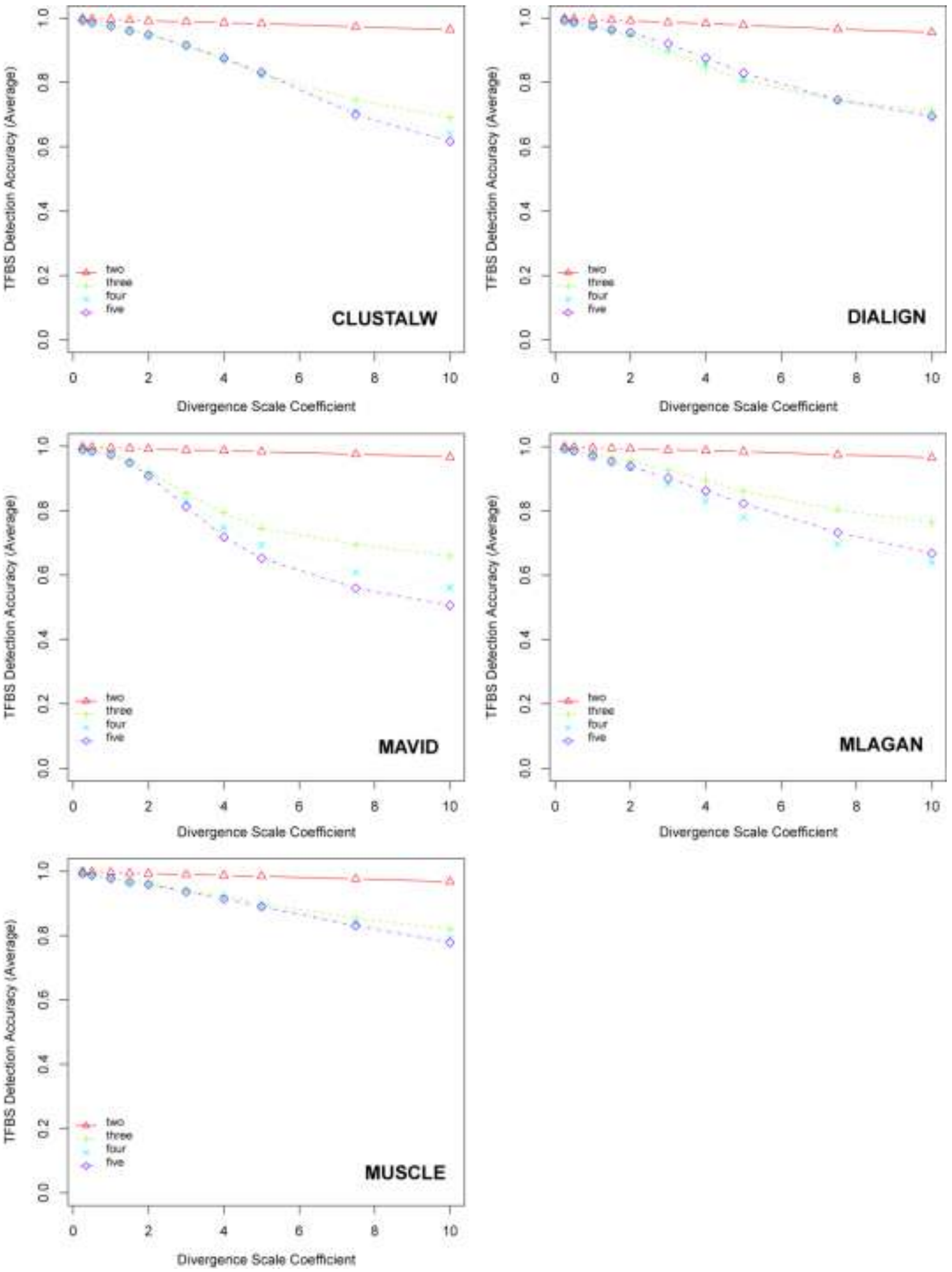

Figure 8: The effects of the number of aligned mammalian species on the TFBS detection accuracy. Each subfigure shows the performance of a tool in aligning a different number of species. Human and baboon were used for two species alignment, mouse was added for three species alignment, all five species but cow were used for four species alignment. While all tools have almost the same performance for aligning the two closely related species human and mouse, MUSCLE and DIALIGN perform better than other tools in maintaining or improving performance when adding more species to the alignment.

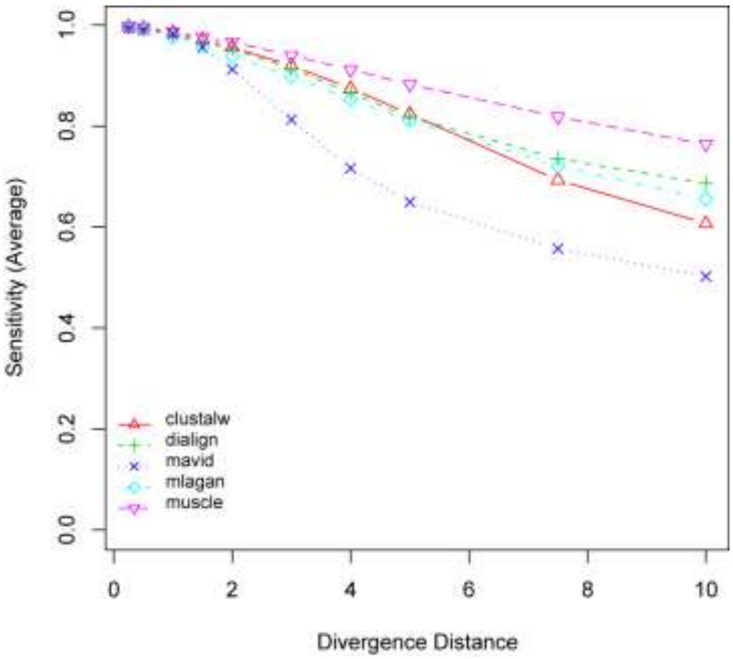

Figure 9: The average TFBS sensitivity of five tools on aligning TFBS in five mammalian species. The relative order on TFBS sensitivity of five tools is almost the same as the order on their TFBS detection accuracy (see Figure 6D).