BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–7

Statistics of local multiple alignments

*Amol Prakash**, *Martin Tompa*

Department of Computer Science and Engineering
Box 352350, University of Washington
Seattle, WA 98195-2350 U.S.A.
{amol,tompa}@cs.washington.edu

ABSTRACT

BLAST [Altschul et al., 1990] statistics have been shown to be extremely useful for searching for significant similarity hits, for amino acid and nucleotide sequences. While these statistics are well understood for pairwise comparisons, there has been little success developing statistical scores for multiple alignments. In particular, there is no score for multiple alignment that is well founded and treated as a standard. We extend the BLAST theory to multiple alignments. Following some simple assumptions, we present and justify a significance score for multiple segments of a local multiple alignment. We demonstrate its usefulness in distinguishing high and moderate quality multiple alignments from low quality ones, with supporting experiments on orthologous vertebrate promoter sequences.

1 INTRODUCTION

Sequence alignment is usually the first step when comparing multiple protein or DNA sequences, for comparative genomics and database similarity searches. While assessing significance of alignments is an important task, this becomes all the more important when aligning sequences with low similarity. Given any local alignment, it is very important to know how likely we are to see an equally good alignment on unrelated sequences.

Karlin and Altschul, 1990 employed the theory proposed by Gumbel, 1958 and presented the statistics for pairwise ungapped local alignments. Today, this has become the basis of the widely uses BLAST searches. The theory was also extended to assess the significance of multiple segments of a pairwise alignment [Karlin and Altschul, 1993]. Later, it was shown that the theory works for gapped local alignments [Altschul and Gish, 1996, Waterman and Vingron, 1994], though it is hard to theoretically estimate the Karlin-Altschul parameters (K, λ, H) [Karlin and Altschul, 1990] in this case. Recently Altschul et al., 2001 presented faster methods to estimate these parameters.

As more and more sequence data becomes available, multiple alignments have become increasingly important. In the last few years, many new multiple alignment tools have emerged, e.g. TBA [Blanchette et al., 2004], MLAGAN [Brudno et al., 2003], MAVID [Bray and Pachter, 2004], DIALIGN [Morgenstern, 1999], ClustalW [Chenna et al., 2003], etc. But on the statistical significance front, there has been little development. Ideas were presented to extend the statistics to 3-way alignments [Altschul and Lipman, 1990], but these ideas were not scalable. While it is believed that the theory extends to multiple alignments, it is a hard problem to solve. The main reasons for this are the lack of availability of a good scoring function, and an inherent problem in the null hypothesis (Stephen Altschul, *personal communication*). The commonly used sum-of-pairs scores are not well justified theoretically [Altschul, 1991] and exhibit a high entropy as the number of aligned sequences grows. The null hypothesis should be chosen so that, when it is rejected, all the sequences are related. This means that the null hypothesis must allow for arbitrary proper subsets of the sequences to be related, which is unwieldy.

Here we present a simple yet robust way to assess significance of multiple segments of a local multiple alignment. We expect these sequences to be *related* to each other, i.e. they have a high local sequence similarity (e.g. orthologous sequences). First we develop a null hypothesis that includes the possibility of only a subset of sequences being related. Making some strong yet reasonable assumptions we keep this set of possibilities small, which makes the whole approach scalable. Secondly, we develop a log-likelihood based scoring function that is consistent with the Karlin-Altschul statistics. Using these two ideas, we extend the BLAST theory to develop methods for assessing significance of multiple local alignments. With some simple assumptions we also handle gapped alignments and extend the analysis to multiple segments.

In the Results section, we apply this new theory to thousands of sets of orthologous promoter regions from a genome-wide study in the vertebrates. We demonstrate that the new statistics are capable of distinguishing between multiple alignments of truly orthologous sequences and those of "nearly orthologous"

*Corresponding author

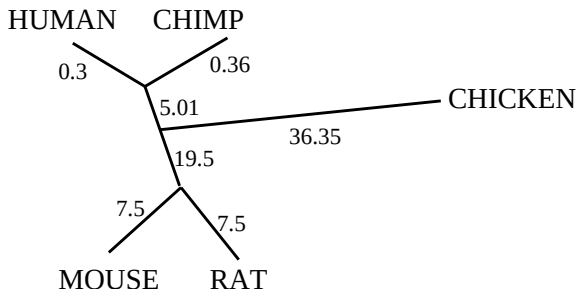

Fig. 1. Phylogenetic tree relating the vertebrate species used in this study. The branch lengths are estimates from Blanchette et al. (*in preparation*).

sequences. Although we apply this theory only to promoter sequences here, it applies equally to many other multiple alignment situations.

2 METHOD

2.1 Null Hypothesis

If we build the multiple alignment statistics on the same ideas as the BLAST statistics, the null hypothesis would be that all of the sequences are completely unrelated. This is a very weak null hypothesis because in most cases of multiple alignments, a subset of the sequences will be related to each other and unrelated to the rest. We want to distinguish these cases from those where all the sequences are related to each other. So we want our null hypothesis to consider the cases where any proper subset of the sequences might be related to each other but unrelated to others, so that when we reject it, we can claim that all the sequences in the multiple alignment are related to each other.

Listing all the possible cases of the partition of S sequences being related will result in a number super-exponential in S, thus resulting in a nonscalable solution. To manage this we make an assumption. We assume that there is an unrooted phylogenetic tree relating the sequences, and for ease of discussion will think of the sequences as coming from different species. The null hypothesis will be restricted to those cases where a single branch (whose removal disconnects the species into two subsets) is exhibiting unrelated behavior, i.e. the two subsets are related amongst themselves, but unrelated to each other.

Suppose we have a sequence from each of the species from the tree in Figure 1. Thus the null hypothesis covers seven cases, corresponding to the seven branches of Figure 1. As an example, the branch with length 5.01 corresponds to the case when the human and chimp sequences are related, and the mouse, rat and chicken sequences are related, but these two subsets of sequences are unrelated to each other. The assumption we make is that when we reject all the cases of the null hypothesis, we can infer that all the sequences are related.

What motivates this is that, in practice, the greatest difficulty for the scoring function is to distinguish between the case when all sequences are related and the case when there is unrelated behavior only on a single branch. If the scoring function can distinguish these, it will also distinguish the cases that show even more unrelatedness. Our experiments on vertebrate promoter sequences confirm this belief (see Figures 2 and 3). As an example, consider the case when the chicken and human sequences are related and the rat, mouse and chimp sequences are related, but these two subsets are unrelated to each other. This would be approximated by the case in the null hypothesis where the human sequence is unrelated to the other four sequences, because a mutation on the long branch to chicken is not unlikely.

Even when a phylogenetic tree is not readily available, most multiple alignment tools still require a tree relating the sequences. They either generate this tree themselves (ClustalW) or the user uses some phylogeny building tool to output a tree. This same tree can be used to list all the cases covered by the null hypothesis.

2.2 Scoring function

As suggested by Karlin and Altschul, 1990, the ideal scoring function for their theory is the log likelihood score. For pairwise comparison, the score for residues i and j is $sc(i,j) = log(\frac{q_{ij}}{p_i p_j})$ where q_{ij} is the target frequency of seeing i and j aligned and p_i (p_j) is the background frequency of i (j). We built our scoring function on the same ideas.

First we need to build an alphabet. Suppose we have a multiple alignment of DNA sequences. (The ideas apply equally well to protein sequences.) In a multiple alignment it is possible that for some regions, some sequences are too dissimilar to align. Obviously, we would like to know the evolutionary history of every residue of every species, but our current understanding of evolution is insufficient to do this reliably. Thus for understanding the history of such dissimilar regions, aligning to segments that are most similar by simple sequence similarity or treating these regions as insertions/deletions may be unrealistic and may penalize the overall alignment heavily. It is quite likely that these regions have mutated to the extent that they cannot be aligned. The multiple alignment program TBA [Blanchette et al., 2004] is built on these premises, and thus it outputs threaded local alignments, each of which may contain only a subset of species. To handle such cases, we will introduce a special character ϵ. Thus an ϵ-added alignment states that any sequence containing ϵ in these regions looks no different than an unrelated sequence. This problem does not come up in a pairwise alignment. Table 1 gives an example of a TBA alignment and the corresponding ϵ-added alignment for our purposes.

Table 1. Table showing the TBA alignment and the corresponding ϵ-added alignment for a handcrafted example containing sequences from human, chimp, mouse and rat. The first row shows two blocks of the TBA alignment and the second row shows the conversion of each block to an ϵ-added alignment containing all species. The species having no data in the block are assigned a sequence consisting of ϵ.

TBA alignment	HUMAN	CCCGTGTGT
	CHIMP	CCCGTTTGT
	MOUSE	CCGCTTTGT
	MOUSE	GTACGT—-TTTGTGTGTAAAAAC
	RAT	GGACGAGGATTTGGGGGTAAAAC
ϵ-added alignment	HUMAN	CCCGTGTGT
	CHIMP	CCCGTTTGT
	MOUSE	CCGCTTTGT
	RAT	$\epsilon\epsilon\epsilon\epsilon\epsilon\epsilon\epsilon\epsilon\epsilon$
	HUMAN	ϵ
	CHIMP	ϵ
	MOUSE	GTACGT—-TTTGTGTGTAAAAAC
	RAT	GGACGAGGATTTGGGGGTAAAAC

We would also like to handle gaps in our alignments. It has been shown that the BLAST theory extends to affine gap penalties [Altschul and Gish, 1996, Waterman and Vingron, 1994, Altschul et al., 1997]. Building affine gap penalties is a much more difficult problem in multiple alignments than in pairwise alignments. Thus we make the standard simplifying assumption here. We will treat gaps as individual characters. This strategy has been followed by Cooper et al., 2004 to compute evolution matrices using multiple alignments of human, mouse and rat sequences. While this assumption is surely unrealistic as it penalizes long gaps very heavily, we chose to use it as it gives us a first handle on gaps for multiple sequences.

We will assume a model of evolution M for the residues. McGuire et al., 2001 suggested a version of the F81 model [Felsenstein, 1981] that included gaps. Cooper et al., 2004 used the same gapped version of the Jukes-Cantor model [Jukes and Cantor, 1969] (a simplified version of the F81 model) to contrast the evolutionary rates in the primate and rodent lineages. The model treats the gaps in the same way as it treats nucleotides. The theory that we present below can be extended for any evolutionary model. We use the F81 model for simplicity.

Thus our character set is {A,C,T,G,ϵ,-}. Now, suppose we have S sequences of length N related by a phylogenetic tree T (having branch lengths). Let $\alpha_1, \alpha_2, ...\alpha_S$ be the characters observed at a particular column of the multiple alignment. Suppose we need to reject the hypothesis that there is an unrelated behavior at branch k. Suppose branch k joins subtrees t_1 (having characters $\beta_1, \beta_2, ...\beta_i$ at the leaves) and t_2 (having characters $\beta_{i+1}, \beta_{i+2}, ...\beta_S$ at the leaves). Then, analogous to the Karlin-Altschul log likelihood score, the score for observing this column of multiple alignment is as follows:

$$sc_k(\alpha_1, ...\alpha_S|T, M) = log(\frac{Pr(\alpha_1, ...\alpha_S|T, M)}{Pr(\beta_1, ...\beta_i|t_1, M)Pr(\beta_{i+1}, ...\beta_S|t_2, M)}) \quad (1)$$

We assume a background distribution of the characters {A,C,T,G,-} and assume ϵ to be sampled from that distribution. We tried different values for the background probability of gaps and found the Karlin-Altschul parameters to be similar in the various cases. Thus for our simulations, we assumed the background probability of the gap character to be 0.01, and the various nucletotides {A,C,T,G} to be uniformly distributed. Using these probabilities as priors, we compute the probability of observing specific characters at the leaves of a tree (the numerator and denominator of Equation 1), using a dynamic programming algorithm very similar to the one used by Felsenstein, 1981. This scoring function behaves in the way that we would want it to. A column in which all sequences in the multiple alignment have the same residue scores higher than a column in which some sequences have ϵ.

2.3 Computing the significance

Again following the BLAST statistics, we write down the equation to compute the p-value of the scoring function. While doing this, we make another assumption. In the null hypothesis, our cases assumed that the sequences in each of the subtrees joined by branch k are completely related to each other. We will assume the number of possible choices of aligned segments to be quadratic in the length of the sequence (N choices from each of the two subtrees). This assumes that whenever we choose a position in a particular sequence, this determines the corresponding position in the other sequences of the related subtree.

We will also assume the Karlin-Altschul parameters [Karlin and Altschul, 1990] to be the average over all the null hypothesis cases (in our experiments, we find these parameters to be fairly similar over the various cases of the null hypothesis). We convert the score of Equation 1 to the normalized score $sc'_k = \lambda sc_k - ln(KN^2)$, where K and λ are the Karlin-Altschul parameters.

Thus the p-value for the null hypothesis case k (unrelated behavior on branch k) [Karlin and Altschul, 1990] is as follows:

$$p\text{-}value(x_k|k) = Pr(sc'_k \geq x_k|k) = 1 - exp(-e^{-x_k}) \quad (2)$$

Letting E be the number of branches of the tree, the p-value for the full null hypothesis is as follows:

$$\begin{aligned} &p\text{-}value(x_1, ..., x_E | one\ branch\ has\ unrelated\ behavior) \\ &= \sum_{k=1}^{E} Pr(k | one\ branch\ has\ unrelated\ behavior) \\ &\quad \times p\text{-}value(x_k|k) \\ &= \frac{1}{E} \times \sum_{k=1}^{E} p\text{-}value(x_k|k) \end{aligned} \tag{3}$$

Here we have assumed all branches to be contributing equally. We can instead alter it to make the contribution weighted by the branch lengths, so the longer branch lengths contribute more.

When we have multiple segments, we will use the sum statistics proposed by Karlin and Altschul, 1993. For the null hypothesis case k, let $sc'_{k,1}, sc'_{k,2}, ..., sc'_{k,r}$ be the normalized scores of the best r nonoverlapping segments. Let us define $total_{k,r}$ as : $total_{k,r} = \sum_{i=1}^{r} sc'_{k,i} - ln(r!)$. Thus the p-value is as follows:

$$\begin{aligned} &p\text{-}value(z_{k,r}|k) = Pr(total_{k,r} \geq z_{k,r}|k) \\ &= \int_{z_{k,r}}^{\infty} \frac{e^{-t}}{r!(r-2)!} \left(\int_0^{\infty} y^{r-2} exp(-e^{(y-t)/r}) dy\right) dt \end{aligned} \tag{4}$$

and the p-value for the full null hypothesis is as follows:

$$\begin{aligned} &p\text{-}value(z_{1,r}, ..., z_{E,r} | one\ branch\ has\ unrelated\ behavior) \\ &= \frac{1}{E} \times \sum_{k=1}^{E} p\text{-}value(z_{k,r}|k) \end{aligned} \tag{5}$$

When deciding the best value of r, we can choose the value that results in the best p-value.

3 ESTIMATING THE PARAMETERS

Figure 1 shows all the species that we have used in our analysis. The first step in this analysis is the estimation of the Karlin-Altschul parameters. This is done using simulations.

We simulated evolution over the phylogeny shown in Figure 1 using a tool by Blanchette et al. (*in preparation*). Using models for substitutions, insertions, deletions and inversions, this tool generates sequences for the leaves of the phylogeny (the various species). We created 1000 such data sets, where each data set contained a sequence for each of the species shown in Figure 1. From these, 5000 data sets were generated for each branch k of the tree, such that the branch k exhibited unrelated behavior. This was simulated by choosing orthologous sequences from the two different sets of species (the two subtrees joined by k), for example picking orthologous sequences from human and chimp, and another orthologous set from mouse and rat for the branch joining rodents and primates. In this way we generated random data sets from the null hypothesis. Now we need to compute the highest scoring local alignments for these data sets.

In another study [Prakash and Tompa, 2005], we compared many multiple alignment tools and showed TBA [Blanchette et al., 2004] to perform the best for our specific purpose of aligning regulatory elements. Ideally, we would like to report the highest scoring local alignment, but this is an NP-hard problem. As a proxy, we parse the TBA alignment for the highest scoring segment using the scoring function of Equation 1. Another advantage of using TBA alignments is that the segments that it reports as aligned may contain only a subset of species (i.e. TBA doesn't align a species in a region if it does not look similar enough). We take the unaligned segments to be made up of ϵ. A handcrafted example was shown in Table 1.

As described above, we created random data sets following our null hypothesis. We created 5000 data sets for each branch exhibiting unrelated behavior and let TBA report the highest scoring local alignment on each.

Equation 2 can be rewritten as follows:

$$ln(KN^2) - \lambda x_k = ln(-ln(Pr(sc_k < x_k|k))) \tag{6}$$

Doing a least square linear fit using Equation 6, we estimate the Karlin-Altschul parameters K and λ. The other Karlin-Altschul parameter (H) that accounts for edge effects and better estimates of effective sequence lengths is not important in these cases as the sequence lengths are much longer than the aligned segment lengths.

Table 2 shows the values of K and λ estimated for human/chimp/mouse/rat and human/chimp/mouse/rat/chicken data sets. Experiments were done for different values of N (sequence length). The RMS error of the least square linear fit of K and λ is also shown.

There are a few observations that we can make looking at Table 2. The first is the small values of the RMS error, thus giving us confidence in the linear relationship. Secondly, the values of K and λ seem to be dependent on the sequence length, perhaps because the main assumption made while inferring Equation 2 (number of choices of aligned segments being quadratic in N) may not be entirely accurate. Thirdly, the values of K and λ are comparable whether we include chicken or not. We need to do additional experiments with more species, but we believe that the values of these parameters will be robust to small changes in the tree. We also observe a variation of less than 5% in the values of these parameters when we repeat the experiments. While this is an acceptable variance for K, we would like to improve this estimate for λ as Equation 2 is doubly exponential in it.

Table 2. Table showing the values of the Karlin-Altschul parameters for various choices of sequence lengths. Experiments are done for two types of data sets : human/chimp/mouse/rat and human/chimp/mouse/rat/chicken. For each experiment, 5000 data sets are generated and for each, every branch is forced to induce unrelated behavior. The parameters are calculated using a least square linear fitting and the RMS error of the fit is also reported.

Sequence Length (N)	Chicken included ?	K	λ	RMS Error
800	n	4.4×10^{-5}	0.35	0.074
1000	n	2.9×10^{-5}	0.33	0.055
1200	n	2.5×10^{-5}	0.34	0.046
1500	n	2.3×10^{-5}	0.35	0.062
800	y	6.5×10^{-5}	0.37	0.038
1000	y	6.2×10^{-5}	0.38	0.049
1200	y	5.0×10^{-5}	0.38	0.058
1500	y	3.7×10^{-5}	0.38	0.060

4 RESULTS

Once we have estimated values of K and λ, we can use Equations 4 and 5 to compute p-values of multiple alignments. In this section, we exhibit the usefulness of this work.

In another study [Prakash and Tompa, 2005], we collected large sets of high confidence orthologous promoter sequences from human, chimp, mouse, rat and chicken. This was done by collecting orthologous genes and filtering out those that did not have orthologous transcription start sites. The upstream sequences were masked for repeats using RepeatMasker [Smit et al., 1996-2004] and DUST [Tatusov and Lipman]. This left us with 4215 data sets of orthologous genes from human, chimp, mouse and rat and 777 data sets from human, chimp, mouse, rat and chicken. For each of these, we align the length 1000 upstream sequences using TBA and then compute the p-value of each. Figure 2 plots the cumulative distribution function of this p-value for the human/chimp/mouse/rat data sets. Using these data sets, we also create 5000 data sets having one branch exhibiting unrelated behavior and 5000 data sets having multiple branches exhibiting unrelated behavior. The cumulative distribution function of the p-value of the alignments of each of these two sets is also plotted in Figure 2.

Using the parameters trained on the simulated data sets having chicken, we repeated the above experiment for human, chimp, mouse, rat, chicken using all of the 777 high confidence upstream regions. The results are shown in Figure 3.

Using Figures 2 and 3, we can decide to use a p-value threshold that makes most of the multiple alignments of single branch random data sets insignificant (say, 10^{-4}, plotted as a vertical line in the two figures). Thus, we can conclude that we have high confidence multiple alignments on the orthologous data sets for approximately 90% of human/chimp/mouse/rat data sets and 40% of the human/chimp/mouse/rat/chicken data sets. Using such computations, we can differentiate high quality multiple alignments from lower quality ones.

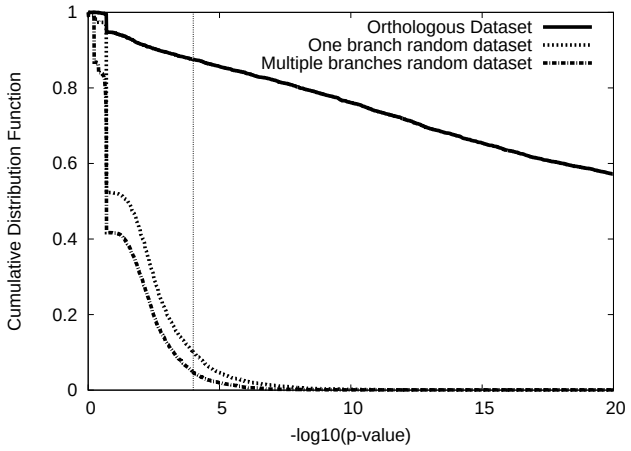

Fig. 2. Cumulative distribution function of the p-value computed for many human/chimp/mouse/rat data sets. Graphs are plotted for the 4215 orthologous data sets, 5000 random data sets having unrelated behavior on a single branch and 5000 random data sets having unrelated behavior on multiple branches.

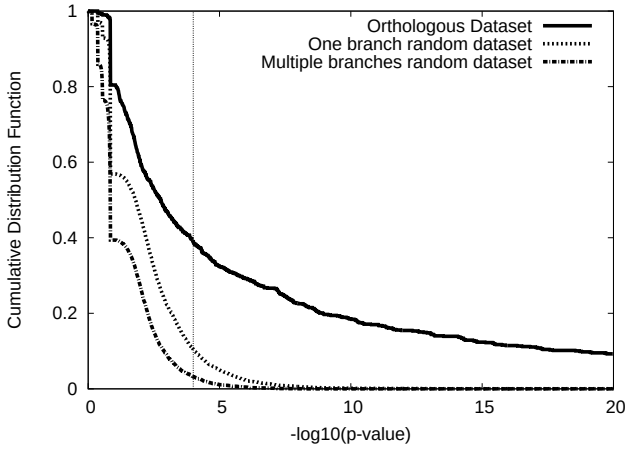

Fig. 3. Cumulative distribution function of the p-value computed for many human/chimp/mouse/rat/chicken data sets. Graphs are plotted for 777 orthologous data sets, 7000 random data sets having unrelated behavior on a single branch and 5000 random data sets having unrelated behavior on multiple branches.

5 DISCUSSION

We have presented a way to extend the BLAST statistics to multiple segments of a multiple local alignment. While the method is itself generic, we show a particular application for promoter regions of vertebrates. The results show its usefulness in distinguishing high (human/chimp/mouse/rat) and moderate (human/chimp/mouse/rat/chicken) quality alignments from low quality ones. Also, instead of averaging the

p-values of the individual null hypothesis cases in Equation 5, we can report the case with maximum p-value, and thus predict the most likely branch that is exhibiting unrelated behavior.

In this work we have made certain assumptions. Here we discuss their validity and the possibilities to remove these in the future.

- We assumed that there is a phylogenetic tree relating the sequences (along with the branch lengths). This is particularly hard when the sequences being aligned may not be orthologous, but as most progressive multiple alignment tools follow a phylogeny to create a multiple alignment, the same tree can be used for computing the significance.

- We assumed an evolutionary model for the purposes of computing the scoring matrices. While we used the F81 model, any other model can be substituted in this analysis. For protein sequences, we can use varying BLOSUM [Henikoff and Henikoff, 1992] matrices (corresponding to the evolutionary distances) or use a single evolutionary matrix as suggested by Altschul, 1993.

- To restrict the number of choices for subsets of species being related to each other and unrelated to others, we assumed that the unrelated behavior can happen only on one branch of the phylogenetic tree. In the Results section we showed that the performance is similar even when the unrelated behavior happens on multiple branches. When we reject the hypothesis that a single branch of the tree exhibits unrelated behavior, our assumption allows us to infer that all the sequences are related. Thus this method scales very nicely with the number of sequences in the analysis.

- For managing gaps, we treated a single gap as a character. While this is surely unrealistic, there are no better methods currently known for scoring gaps in multiple alignments. When better understanding of the scoring of gaps develops, we can modify our analysis to include it.

- To estimate the Karlin-Altschul parameters, we need to find the highest scoring multiple local alignment. This is an NP-hard problem. Thus we assumed that if the score of the highest scoring segment is high enough TBA would find it. While we could use any other tool for this purpose, the reason we chose to use TBA was because in another analysis [Prakash and Tompa, 2005] we showed TBA to be very sensitive and specific for similar purposes.

- We assumed the Karlin-Altschul parameters to be independent of the tree and the branch displaying nonorthologous behavior. This is not entirely correct, as we show small variability in their values. Some initial analysis shows these values to change by 10%-20% when we go to very different phylogenies. This shows that the assumption may not hold beyond small variations in the phylogeny. So if we are given an entirely new set of species with a very different phylogeny, it seems best to re-estimate the values of these parameters using simulations. Currently we are working on understanding the relationship between the Karlin-Altschul parameters and the phylogeny and hope to be able to remove this bottleneck in future.

We need faster methods to compute the Karlin-Altschul parameters. The ideas presented by Altschul et al., 2001 are not directly applicable, as they require us to find all high scoring local multiple alignments for a given set of sequences, which is a hard problem. Currently this is the computational bottleneck for our analysis. Exact and fast computation of these parameters is necessary to understand their relationship to the phylogeny, unrelated branch, background distribution, and evolutionary model. This can enhance the analysis and help us differentiate higher quality multiple alignments from lower ones at improved resolution.

When we include chicken in our analysis, a smaller fraction of data sets show significant alignments (40% vs. 90%). This should not be viewed as a negative result: it simply suggests that it is harder to reliably align a chicken promoter sequence to primates and rodents than it is to align primates with rodents. A similar observation was made by two recent studies [Margulies et al., 2003, International Chicken Genome Sequencing Consortium, 2004] which reported a very small fraction of regulatory elements to be conserved between human and chicken.

Simultaneously, we would like to work on showing the methods' applicability to other areas, for example assessing significance of multiple alignments of functionally related protein sequences. We also plan to work on extending the ideas to help build a better multiple alignment tool. Also, the sensitivity of BLAST searches may be improved significantly by aligning the good hits and assessing the significance of that multiple alignment.

ACKNOWLEDGMENTS

We would like to thank Steven Altschul, Webb Miller, Michal Linial, Larry Ruzzo, Mathieu Blanchette, Zizhen Yao and Zasha Wienberg for some insightful suggestions. This material is based upon work supported in part by the National Science Foundation under grant DBI-0218798 and by the National Institutes of Health under grant R01HG02602.

REFERENCES

S. Altschul and D. Lipman. Protein database searches for multiple alignments. *Proceedings of the National Academy of Science USA*, 87:5509–5513, 1990.

S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215: 403–410, 1990.

S. Altschul, T. Madden, A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Statistical Scince*, 25 (17):3389–3402, 1997.

S. F. Altschul. A protein alignment scoring system sensitive at all evolutionary distances. *Journal of Molecular Evolution*, 36(3): 290–300, Mar. 1993.

S. F. Altschul. Amino acid substitution matrices from an information theoretic perspective. *Journal of Molecular Biology*, 219(3):555–565, June 1991.

S. F. Altschul and W. Gish. Local alignment statistics. *Methods in Enzymology*, 266:460–480, 1996.

S. F. Altschul, R. Bundschuh, R. Olsen, and T. Hwa. The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Research*, 29(2):351–361, 2001.

M. Blanchette, E. Green, W. Miller, and D. Haussler. *In preparation*.

M. Blanchette, W. Kent, C. Riemer, L. Elnitski, A. Smit, K. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E. Green, D. Haussler, and W. Miller. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research*, 14(4):708–715, Apr. 2004.

N. Bray and L. Pachter. MAVID: Constrained ancestral alignment of multiple sequences. *Genome Research*, 14:693–699, 2004.

M. Brudno, C. Do, G. Cooper, M. F. Kim, E. Davydov, E. D. Green, A. Sidow, and S. Batzoglou. LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Research*, 13(4):721–731, 2003.

R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research*, 31:3497–3500, 2003.

G. M. Cooper, M. Brudno, E. A. Stone, I. Dubchak, S. Batzoglou, and A. Sidow. Characterization of Evolutionary Rates and Constraints in Three Mammalian Genomes. *Genome Research*, 14(4):539–548, 2004.

J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

E. J. Gumbel. *Statistics of Extremes*. Columbia University Press, New York, 1958.

S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Science USA*, 89(22):10915–10919, Nov. 1992.

International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Genome Biology*, 432:695–716, Dec. 2004.

T. Jukes and C. Cantor. Evolution of protein molecules. In H. N. Munro, editor, *Mammalian protein metabolism III*, pages 21–132. Academic Press, New York, 1969.

S. Karlin and S. F. Altschul. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proceedings of the National Academy of Science USA*, 90:5873–5877, June 1993.

S. Karlin and S. F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Science USA*, 87(6):2264–2268, Mar. 1990.

E. Margulies, M. Blanchette, NISC Comparative Sequencing Program, D. Haussler, and E. Green. Identification and characterization of multi-species conserved sequences. *Genome Research*, 13 (12):2507–2518, 2003.

G. McGuire, M. Denham, and D. Balding. Models of Sequence Evolution for DNA Sequences Containing Gaps. *Molecular Biology and Evolution*, 18(4):481–490, 2001.

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

A. Prakash and M. Tompa. Comparative genomics for regulatory element discovery in vertebrates. *Submitted*, 2005.

A. F. A. Smit, R. Hubley, and P. Green. Repeatmasker Open-3.0, 1996-2004. `http://www.repeatmasker.org`.

R. Tatusov and D. Lipman. *In preparation*.

M. Waterman and M. Vingron. Sequence Comparison Significance and Poisson Approximation. *Statistical Scince*, 9(3):367–381, 1994.