Pacific Symposium on Biocomputing 4:150-161 (1999)

A PROBABILISTIC APPROACH TO CONSENSUS MULTIPLE ALIGNMENT

B. LAZAREVA-ULITSKY and D. HAUSSLER
Department of Computer Science, University of California at Santa Cruz,
Santa Cruz, CA 95064, USA

We consider the problem of obtaining the maximum *a posteriori* probability (MAP) estimate of a consensus ancestral sequence for a set of DNA sequences. Our maximization method, called ASA (dnA Sequence Alignment), can be applied to the refinement of noisy regions of a DNA assembly, to the alignment of genomic functional sites, or to the alignment of any set of DNA sequences related by a star-like phylogeny. Along with the optimal consensus, ASA finds suboptimal solutions together with their relative probabilities. The probabilistic approach makes it possible to establish the limits to which an ancestor can in principle be recovered from diverged sequences. In simulations on rather short synthetic sequences (of length up to 80) with different coverage and error rates ranging from 5% to 30%, ASA restored the consensus from noisy observations essentially as best as is theoretically possible for the given error rates. We also illustrate the performance of ASA on the alignment of E.Coli promoters and the Alu-Sb subfamily of human repeat sequences. Since our model is a special case of a profile HMM, we give a comparison between these two approaches, as well as with other DNA alignment methods.

1 Introduction

This paper deals with a probabilistic model of *consensus multiple alignment.* Suppose that N observed sequences $\mathbf{S}_1, \ldots, \mathbf{S}_N$ evolved independently from a common unknown ancestor $\mathbf{C}$. The evolutionary process introduces insertions, deletions, and substitutions in the sequences with probabilities λ, μ, and γ, respectively. The problem we consider is to restore the original sequence $\mathbf{C}$, or in other words, to infer the consensus.

When sequences are related via an evolutionary tree it is common to simultaneously infer the correspondence between letters and the tree labels using high-dimensional dynamic programming[1]. In our case of consensus alignment for the star topology[2], we are interested primarily in the ancestor and consider all possible alignments that contribute to it. After the consensus is inferred, the multiple alignment can be obtained trivially by aligning every sequence to the consensus. Our approach is close in spirit to the treatment of probabilistic pairwise alignment in ([3,4]) and can also be considered to be a special kind multiple sequence alignment via a profile HMM[5,6,4]. Indeed, every pair of $\mathbf{C} \equiv c_1, \ldots, c_L$ and error parameter $\mathcal{E} = (\lambda, \mu, \gamma)$ corresponds to a profile HMM with L Match states and transition probabilities into Insertion and Deletion states equal to λ and μ, respectively. The output probabilities in state i

are determined by c_i and substitution rate γ. Our methods for estimating the parameters of this HMM are different from those normally used, however.

Let us assume an *a priori* distribution $P_0(\mathbf{C})$ on possible ancestor sequences $\mathcal{C}$. Then the most plausible ancestor sequence would be the one that maximizes $P(\mathbf{C}|\mathbf{S}_1, \ldots, \mathbf{S}_N, \mathcal{E})$, the posterior probability of ancestor given the set of observations. We show in section 2 that this probability can be effectively evaluated up to a constant. Therefore the problem of consensus restoration is an optimization problem on the discrete set $\mathcal{C}$ of possible consensus sequences

$$\mathbf{C}^* = \mathrm{argmax}_{\mathbf{C}\in\mathcal{C}} P(\mathbf{C}|\mathbf{S}_1, \ldots, \mathbf{S}_N, \mathcal{E}). \tag{1}$$

It follows from the HMM interpretation of our model, that consensus reconstruction can be viewed as the problem of optimizing the architecture of the HMM embodied in the discrete HMM “parameter” $\mathbf{C}$. This boils down to estimating the size of the HMM model (the number of match states) and the choice of the output probability distribution (from four possibilities, each determined by the fixed substitution rate γ) in every match state. The problem of architecture choice has been known to be a serious issue in HMM training [7] and for profile HMMs has been treated heuristically [5,4]. Though these heuristics work well for protein sequence alignment, especially when there are relatively conserved regions almost free from indels, they are not very reliable in predicting the consensus for noisy DNA sequences, as illustrated in Section 3.1. Our approach is complimentary to the classical HMM training: we optimize architecture, but heuristically fix all continuous parameters.

We approach the optimization problem (1) by combining deterministic optimization with a Markov Chain Monte Carlo (MCMC) method that samples from $P(\mathbf{C}|\mathbf{S}_1, \ldots, \mathbf{S}_N, \mathcal{E})$. We initialize the deterministic optimization with one of the sequences in the data set and after it converges switch to MCMC sampling. In the course of sampling we keep track of the optimal and suboptimal consensus sequences, and as a result get not only the best solution but also the suboptimal ones together with their relative probabilities. We call our algorithm ASA (dnA Sequence Alignment). The MCMC approach to multiple alignment in the form of simulated annealing was also considered elsewhere [8,9], where the sampling was performed on alignments rather than consensus sequences as in ASA. Also an MCMC in the form of a Gibbs sampler was applied to block alignment of protein sequences [10].

In Section 3 we give experimental results on synthetic data for different overall error rates $\epsilon \in \{5\%, 7.5\%, 10\%, 15\%, 30\%\}$, where $\epsilon = \lambda + \mu + \gamma$, and different numbers of observed sequences (coverage) N. In all cases the optimal solution [a] found by ASA is never less probable than the true consensus.

[a] This solution may be locally and not globally optimal.

Furthermore, the optimal solution is usually attained in the first few steps of deterministic optimization, and the suboptimal solution is often attained during the subsequent MCMC sampling. As is expected, for low coverage and high error rate the optimal solution sometimes has higher probability than the true consensus. Thus these kinds of experiments empirically establish the accuracy to which one can, in principle, recover the consensus for noisy sequences.

Since in practice the true error parameters are never available, our second set of experiments optimized the posterior probability in (1) with a fixed error parameter $\mathcal{E}^f = (\epsilon^f/3, \epsilon^f/3, \epsilon^f/3)$, $\epsilon^f = 5\%$, whereas the sequence sets were generated with different error rates. Surprisingly, the proportion of correctly inferred consensus sequences was practically the same as in the ideal case, when ASA used the actual error parameters (Table 1.a). In 3.1 we give an example of the ASA alignment and a corresponding SAM HMM alignment[11] for one of the synthetic data sets, which shows some of the effects of the parameter specialization in ASA. It may be argued that other, nonHMM-based consensus estimation methods such as those typically employed in DNA sequencing projects may be more appropriate for the type of data we explore. Therefore we also compare the accuracy of ASA consensus with that of CAP2 assembler[12] and ReAligner[13]. ASA is shown to give superior results.

We also include experimental results of ASA on real genomic data. In Section 4 we explore the differences between ASA, SAM HMM and CLUSTAL alignments of the Alu-Sb subfamily of human repeat sequences. Then we extend our algorithm to allow for cases when sequences have mutual consensus only in certain regions and apply it to the alignment of 252 E.Coli promoters[14].

2 Consensus Optimization in ASA

Here we describe the theoretical ideas behind ASA. Let a nucleotide sequence $\mathbf{C} = c_1, \ldots, c_L$, $c_i \in \{A, C, G, T\}$ be copied independently into N sequences $\mathbf{S}_1, \ldots, \mathbf{S}_N$, $\mathbf{S}_i = s_{i1}, \ldots, s_{iL_i}$,$s_{ik} \in \{A, C, G, T\}$ with errors. When copied, a letter can be deleted with probability μ, if not deleted, substituted with another letter (chosen uniformly from the rest of the letters) with probability γ. Furthermore, during the copy process, before any letter (and after the last one) there can be inserted ψ letters, where ψ is geometrically distributed with parameter λ. These assumptions can be generalized, e.g. to allow residue-dependent substitution rates. The problem is to restore the original sequence $\mathbf{C}$ from its N observed noisy copies.

We adopt a Bayesian approach, successfully used in many related kinds of biosequence analysis[15,16]. To reconstruct the consensus we find the maximum *a posteriori* probability (MAP) estimate of $\mathbf{C}$ by solving (1). First, we specify

$P_0(\mathbf{C})$, the *a priori* distribution of the original sequence. We assume that the length of the sequence has uniform distribution on some interval $|\mathbf{C}| \equiv L \sim U([L_{min}, L_{max}])$, and that given the length of the sequence all letter combinations are equally likely. Thus $P_0(\mathbf{C}) \propto 4^{-|\mathbf{C}|}$. More sophisticated priors, e.g. allowing different *a priori* probabilities for the four bases, are also possible. Second, we observe that due to insertions and deletions there are a variety of ways in which a correspondence between letters in $\mathbf{S}_i$ and $\mathbf{C}$ can be assigned. We refer to such a correspondence as an alignment between $\mathbf{S}_i$ and $\mathbf{C}$ and denote it by a. Thus, $P(\mathbf{S}_i|\mathbf{C}, \mathcal{E})$, the probability that $\mathbf{C}$ was copied into $\mathbf{S}_i$ with the error parameter $\mathcal{E} = (\lambda, \mu, \gamma)$, is a sum over all possible alignments a of $P(\mathbf{S}_i, a|\mathbf{C}, \mathcal{E})$. Putting this together we get

$$P(\mathbf{C}|\mathbf{S}_1, \ldots, \mathbf{S}_N, \mathcal{E}) = P_0(\mathbf{C}) P(\mathbf{S}_1, \ldots, \mathbf{S}_N|\mathbf{C}, \mathcal{E}) / P(\mathbf{S}_1, \ldots, \mathbf{S}_N|\mathcal{E}) \propto \qquad (2)$$
$$P_0(\mathbf{C}) \prod_{i=1}^{N} P(\mathbf{S}_i|\mathbf{C}, \mathcal{E}) = P_0(\mathbf{C}) \prod_{i=1}^{N} \sum_a P(\mathbf{S}_i, a|\mathbf{C}, \mathcal{E}).$$

It is easy to see, especially adopting an HMM representation of our model, that the sum over all alignments can be effectively calculated via the forward algorithm[17] and thus for every $\mathbf{C}$ the *a posteriori* probability can be evaluated efficiently up to a constant.

Our approach to the MAP estimation of $\mathbf{C}$ given in (1) starts with a greedy optimization algorithm that has the same flavor as "model surgery"[5]. If one has a tentative consensus $\mathbf{C}$, the sequence itself can give some hints about how to change it so as to increase the *a posteriori* probability (3). Let $e_{ins}(k), e_{del}(k), e_{sub}(k)$ be the expected number of insertions, deletions and substitutions at position $k \in [0, L]$, given the current tentative consensus. These errors can be expressed in terms of sufficient statistics for HMM parameters and can be calculated via the standard forward-backward algorithm[17]. If, say the average number of deletions at position k is the highest relative to other errors $e.(\cdot)$ one can try to propose a new sequence $\mathbf{C}^{new}$ by deleting the k-th letter. If the number of insertions $e_{ins}(k)$ (or substitutions $e_{sub}(k)$) is the highest, one can propose a sequence $\mathbf{C}^{new}$ by inserting before the k-th position (or substituting k-th letter for) the best letter out of $\{A, C, G, T\}$. If the objective function (3) increases on $\mathbf{C}^{new}$ then one *accepts* the new tentative consensus. If not, one tries instead the change with the next highest $e.(\cdot)$.

This greedy algorithm attempts to change only one position of $\mathbf{C}$ at every step, or, as we say, to propose a single $\mathbf{C}$-neighbor. It is clear that due to local changes it might converge to a local maximum of (3) and miss the global one. To avoid convergence to a local maximum, ASA also includes a Metropilis-Hastings algorithm. This stochastic algorithm is capable of moving from one mode to another by accepting, with some small probability, less favorable $\mathbf{C}^{new}$s that bridge the two modes. The Metropolis-Hastings algorithm at step t uses the current tentative consensus $\mathbf{C}^t$ and randomly chooses a sequence $\mathbf{C}^{new}$ according to some proposal distribution $Q(\cdot|\mathbf{C}^t, \mathcal{E})$ defined on the set of $\mathbf{C}^t$-neighbors, i.e. $\mathbf{C}^{new} \sim Q(\cdot|\mathbf{C}^t, \mathcal{E})$. After this it randomly decides whether or not to accept $\mathbf{C}^{new}$ according to the Metropolis-Hastings ratio[18]

$$r = \frac{P(\mathbf{C}^{new}|\mathbf{S}_1, \ldots, \mathbf{S}_N, \mathcal{E})\; Q(\mathbf{C}^t|\mathbf{C}^{new}, \mathcal{E})}{P(\mathbf{C}^t|\mathbf{S}_1, \ldots, \mathbf{S}_N, \mathcal{E})\; Q(\mathbf{C}^{new}|\mathbf{C}^t, \mathcal{E})} \tag{3}$$

by putting $\mathbf{C}^{t+1} = \mathbf{C}^{new}$ with probability $\min(1, r)$ and $\mathbf{C}^{t+1} = \mathbf{C}^t$ with probability $1 - \min(1, r)$.

This procedure constructs an ergodic Markov chain $\{\mathbf{C}^t, t = 0, 1, ..\}$ whose stationary distribution is the distribution (3). This clearly helps in maximizing (3) since the Markov chain is expected to spend the most time in the subset of consensus sequence space where the objective function is high. From a practical standpoint the acceptance ratio (3) should not be too low, and can be adjusted by choosing a proper proposal distribution. In ASA, the proposal distribution $Q(\mathbf{C}^{new}|\mathbf{C}^t, \mathcal{E})$ is a linear combination of the uniform distribution on the set of $\mathbf{C}^t$ and a data-driven distribution, related to the errors $e_{ins}(k), e_{del}(k), e_{sub}(k)$ used in the greedy algorithm.

In practice we combine the deterministic and stochastic approaches. We first run the greedy algorithm starting with one of the sequences in the data set and after it converges switch to the Metropolis-Hastings algorithm. Every iteration of our algorithm is quadratic in L, the length of consensus, and linear in the number of sequences. To speed up these calculations it is possible to consider only near-diagonal alignments.

3 To what extent can the consensus be recovered?

An accurate consensus inference plays a special role in sequence assembly. Certain assemblers, like CAP2[12] and TIGR[19] assembler build a layout of sequences and then refine poorly aligned regions in attempt to find a more reliable consensus, others, like ReAligner[13], refine the overall original layout.

In the problem of consensus restoration, one should distinguish two issues: the capacity of optimization algorithm to find the maximum of the objective function and whether this maximum is attained on the true consensus. For data conforming to the probabilistic models described above, the *a posteriori* probability is the proper objective function for optimization. Therefore, if the true consensus is not the optimal one, it just *can not* be reconstructed from the noisy observations. Similar probabilistic reasoning was used in the assessment of pairwise alignment accuracy[20].

In real-life sequence assemblies the unreliable regions to be realigned are usually fairly short (about $10 - 20$ bps) and very rarely span more than 50 nucleotides. The error rate is estimated to be on average about 5%, but for certain regions it can be substantially higher. In our experiments N copies of a random consensus $\mathbf{C}^*$ of length $L = 20$ were generated according to the model described in Section 2 with error parameter $\mathcal{E}^* = (\lambda^*, \mu^*, \gamma^*)$. Adopting the sequence assembly terminology, we refer to N as *coverage*. We denote by *error rate* the sum of the insertion, deletion and substitution probabilities [b] $\epsilon = \lambda + \mu + \gamma$, and perform the experiments with $\lambda = \mu = \gamma$.

In our first set of experiments the true error parameter $\mathcal{E}^*$ with which sequences were generated was used by the ASA algorithm, and thus the *a posteriori* probability $P(\mathbf{C}|\mathbf{S}_1, \ldots, \mathbf{S}_N, \mathcal{E}^*)$ in (3) was the perfect objective function for consensus optimization (1). The error parameter $\mathcal{E}^*$ was varied from 5% to 30%. For every pair of coverage N and error rate ϵ^*, we used 100 trials to generate diverged sequences and 300 iterations of ASA (Table 1.a). In many trials the optimal solution was attained on the first iteration of the deterministic greedy algorithm (statistics are given in columns 7 and 8), while the suboptimal solution was often attained in the stochastic Metropolis-Hastings phase (data not shown). To get better estimates of the quality of our consensus reconstruction, and to relate our results to simulations in other studies[13], where in the worst case scenario the error rate ϵ^* was 10%, we carried out another set of experiments with $1,000$ trials for every coverage value and the error rate 10%. Since we were primarily interested in the optimal solution rather than suboptimal ones, we reduced the number of iterations to 30 (Table 1.b).

Since in practice the true error parameter $\mathcal{E}^*$ is unavailable, in additional experiments we attempted to restore the consensus using the "wrong" objective function, i.e. the *a posteriori* probability (3) conditioned on some fixed error, $\mathcal{E}^f = (\epsilon^f/3, \epsilon^f/3, \epsilon^f/3), \epsilon^f = 5\%$ in our case. Surprisingly, the results of these two types of experiments were almost indistinguishable. The proportion of correctly inferred solutions in the ideal (1.a) and heuristic experiment (1.b) is given in columns 6 and 5. Observe that there is no way to confirm that a solution $\mathbf{C}^0$ is globally optimal. However, in cases when it is less probable than the true consensus $\mathbf{C}^*$, $P(\mathbf{C}^0|\mathbf{S}_1, \ldots, \mathbf{S}_N, \mathcal{E}) < P(\mathbf{C}^*|\mathbf{S}_1, \ldots, \mathbf{S}_N, \mathcal{E})$, the solution is clearly *not* a global optimum but a *local* one. If the proportion of *local* solutions was non-zero we reported it in parenthesis in columns 6 and 5. Otherwise, in the case where we used the true objective function, the quality of reconstruction was marked by asterisk. Since some of solutions

[b] The expected number of errors per position, $\lambda/(1 - \lambda) + \mu + \gamma(1 - \mu)$, is close to ϵ

a: Consensus inference: 100 trials, 300 iterations							
1	2	3	4	5	6	7	8
Cov	ϵ^* (%)	CAP2	ReAligner	ASA	ASA	Iterations	
N				$\epsilon = 5\%$	true ϵ^*	*mean*	*max*
3	5	.81	.81	.81	.81*	1.09	6
	7.5	.71	.71	.79	.79*	1.74	8
	10	.53	.53	.67	.68*	3.34	23
	15	.22	.22	.36	.35*	6.99	69
	30	.01	.01	.02	.03*	22.64	217
5	5	.98	.98	1	1*	0.98	4
	7.5	.94	.94	.99	.99*	1.63	18
	10	.86	.86	.93	.94*	2.67	11
	15	.62	.63	.79(.01)	.79*	5.59	94
	30	.05	.05	.13(.01)	.20*	13.95	107
7	5	1	1	1	1*	0.93	4
	7.5	.99	.99	1	1*	1.41	7
	10	.97	.97	.99	.99*	2.03	8
	15	.85	.86	.94	.92*	5.03	116
	30	.13	.13	.46(.02)	.49*	19.32	290
9	5	1	1	1	1*	0.93	4
	7.5	.99	.99	1	1*	1.36	5
	10	.98	.98	1	1*	1.99	6
	15	.93	.94	.98	.96*	3.58	13
	30	.22	.25	.69	.70*	10.99	183
11	5	1	1	1	1*	0.92	4
	7.5	1	1	1	1*	1.36	5
	10	.98	.98	1	1*	1.97	6
	15	.92	.95	1	1*	3.44	11
	30	.32	.35	.79	.83*	9.45	261
15	30	.39	.45	.90	.92*	5.73	15
b: Consensus inference: 1000 trials, 30 iterations							
3	10	.573	.574	.671(.002)	.671(.002)	3.037	23
5	10	.870	.870	.933	.934*	2.423	24
7	10	.963	.966	.990	.989*	2.079	16
9	10	.981	.983	.999	.999*	1.959	8
11	10	.985	.989	.999	.999*	1.912	7

Table 1: *The length of true consensus $L = 20$. 1: Coverage. 2: Error parameter ϵ, $\lambda = \mu = \gamma = \epsilon/3$. 3-6: The proportion of trials with consensus correctly inferred by CAP2 (3), ReAligner (4), ASA with fixed error parameter (5), ASA with true error parameter (6). The asterisk marks the best achievable. 7-8: The mean and max of the number of iterations until the optimal solution is found.*

that converged to the true consensus may be not globally optimal, the quality marked by asterisk is the upper bound for the *ideal* MAP-based performance on the simulated sequence sets. The true consensus just cannot be restored in certain cases because there is too much noise in the observed sequences. We also examined the situation when the insertion, deletion and substitution errors were not equal to each other, but the results did not differ qualitatively from those in column 5 (data not shown).

We also compared the performance of ASA with that of the CAP2[12] assembler and with ReAligner [13], which we used to refine the output of CAP2 (columns 3 and 4 of Table 1). To make CAP2 produce an assembly we added identical flanking sequences to the noisy sequences generated in our experiments. Though the quality of ASA (column 5) is almost ideal, our algorithm needs a larger number of iterations (columns 7 and 8) than ReAligner. Hence ASA may be more appropriate when the main issue in consensus reconstruction is the accuracy.

It is natural to expect that as consensus length increases, the number of steps required by ASA to find the optimal solution should grow, and the efficiency (column 5 of Table 1) should diminish. The results of simulations for different consensus lengths L and coverage N are shown in fig. 1.

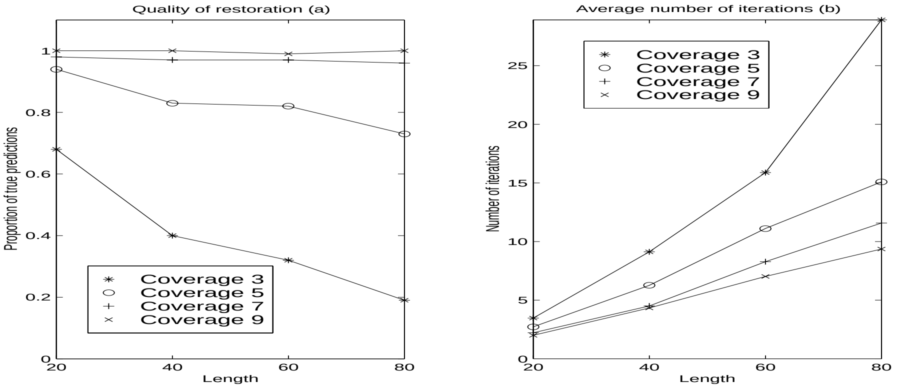

Figure 1: Actual error rate is $\epsilon^* = 10\%$, error rate used in the optimization $\epsilon^f = 5\%$ (a) Proportion of correctly inferred consensus sequences vs. the length of true consensus. (b) Average number of iterations until optimal solution is found.

3.1 HMM alignment vs. ASA

As has been mentioned already, our model is a special case of a profile HMM. Though HMM models are very successful in protein alignments, especially when certain blocks of sequences are aligned almost without indels, it has been noted that their application for DNA alignment can pose certain difficulties. To illustrate this we compare an HMM alignment produced by SAM[5] with default parameters and the corresponding ASA alignment (Table 2). As seen, the SAM alignment method tends to merge alignment columns, allowing less peaked and more varied distributions in Match states. Of course, this is only one setting of the SAM parameters, and further tests would need to be done to find the optimal SAM parameters for this kind of sequence data.

SAM HMM alignment	**ASA alignment**
.............*.......................*.	
ACTCG-----Aa/tTC-CGAATAGATAGAAGTCTGGT	A-CTCG---AATTC-CGA-ATAGATA-GAA-GTCTGCT
AC-CG.....A A TT.CCAATAGATAGAAATCTGGTAA	A-C-CG---AATTC-C-A-ATAGATA-GAA-ATCTGGTAA
ACTCG.....G C TCaCGAATATATAGAAGTCTGCT	A-CTCG---GCT-CACGA-ATATATA-GAA-GTCTGCT
ACTC-.....A T TC.CGAAATGATAGAAGT-TGGT	A-CTC----A-TTC-CGAAAT-GATA-GAA-GT-TGGT
AGTTGggttaA T TC.CGAATACATATGAATGCTGC	AGTTGGGTTAATTC-CGA-ATACATATGAATG-CTGC-
ACTCG.....A A TC.CGAATAGATAGAAGTCTGCT	A-CTCG---AA-TC-CGA-ATAGATA-GAA-GTCTGCT

Original:	ACTCG-AATTCCGAATAGATAGAAGTCTGCT
ASA optimal:	ACTCG-AATTCCGAATAGATAGAAGTCTGCT
ASA suboptimal:	ACTCGGAATTCCGAATAGATAGAAGTCTGCT

Table 2: Comparison of SAM and ASA alignments. The sequences are generated from the original sequence with error rate $\epsilon = 30\%$. The consensus for HMM alignment was deduced by hand. The stars mark the erroneous positions.

4 Alignment of human Alus and E.Coli promoters

We applied ASA to the alignment of 10 sequences that constitute Alu-Sb subfamily of human repeat sequences (ftp://ncbi.nlm.nih.gov/pub/jmc/alu/). The ASA consensus coincides basically with the consensus implied by SAM HMM SAM[5] and ClustalW[21] alignments. However, the alignments themselves differ dramatically. We illustrate the difference with alignments of one small region in Table 3. Again, as in Section 3.1 we see that the ASA alignment tends to produce conserved columns and to find plausible indels. In this respect ASA is close to an assembler and nevertheless allows one to align relatively diverged sequences such as Alu repeats. The CAP2 assembler split the Alu-Sb subfamily into several contigs, and so could not be used on this data. The same is true for the TIGR assembler.

As a generative model, an HMM is much more flexible in representing diverse kinds of sequence distributions than the model used in ASA, with its limited parameterization. However, the parameter estimation methods used for HMMs may be more prone to local convergence than the optimization methods used in ASA. One possibility we plan to explore is, first, to produce an ASA alignment with good agreement in the columns and then use the corresponding model as an initial model for full HMM parameter estimation.

ASA can also be applied to consensus inference for a set of DNA functional sites. Since in alignment of functional sites, certain sub-regions of sequences can be unrelated, and the corresponding regions in different sequences might have varying length, we generalize our model by allowing an extra letter N in

	ASA alignment	SAM HMM alignment
Consensus	-AGCTTGCAGTG-AGCC-G-AG-AT--C-GCGCCA--CTGC-A---C----T-	AGCTTGCAGTG-AGCCG--AGAT-----CGCGCCACT-------GCAc/tT
HSU14568	-AGCTTGCAGTG-AGCC-G-AG-AT--C-CCGCCA--CTGC-A---C----T-	AGCTTGCAGTG.AGCCG..AGAT.....CGCGCCACT.......GCA C T
HSU14570	-AGCTTGCAGTG-AGCC-G-AG-AT--T-GCGCCA--CTGC-AGTCCGCAGT-	AGCTTGCAGTG.AGCCG..AGAT.....TGCGCCACTgcagtccGCA G T
HSU14569	-AGCTTGCAGTG-AGCC-G-AG-AT--C-GCGCCA--CTGC-A---C----T-	AGCTTGCAGTG.AGCCG..AGAT.....CCCGCCACT.......GCA C T
gb—M63005	CA-CT-GCACTCCAGCCTG-GG--TGACAGAGCGAGGCTCCGT---C------	CACTGCACTCC.AGCCT..GGGTgacagAGCGAGGCT.......CCG T C
gb—M20902	-AGCTTGCAGTG-AGCC-G-AG-AT--C-GCGCCA--CTGC-A---C----T-	AGCTTGCAGTG.AGCCG..AGAT.....CGCGCCACT.......GCA C T
gb—M64231	CA-CT-GCACTCCAGCCTG-GGCAG--CAGAGC-A--A-G--A---C----TG	CACTGCACTCC.AGCCT..GGGC.....AGCAGAGCAa......GAC T G
gb—M26939	-CGCCTGCAGTCTAGCCTGGAG-AG--A-GGGCGA--C-CC-----CG---TA	CGCCTGCAGTCtAGCCTggAGAG.....AGGGCGACC.......CCG T A
gb—M20556	-AGTTTGCAGCG-AGCC-G-AG-AT--T-GCGCCACACTGC-A----------	AGTTTGCAGCG.AGCCG..AGAT.....TGCGCCACA.......CTG C A
gb—S54330	-AGCTTGCAGTG-AGCC-A-AC-AT--C-GCGCCA--CTGC-A--------T-	AGCTTGCAGTG.AGCCA..ACAT.....CGCGCCACT.......GCA T -

	CLUSTAL alignment
Consensus	AGCTTGCAGTGAGCCGAGATCGCGCCAC--TGCA-c/tC
HSU14570	AGCTTGCAGTGAGCCGAGATTGCGCCACTGCAGT- C C
gb—M20556	AGTTTGCAGCGAGCCGAGATTGCGCCACACTGCA- C C
HSU14568	AGCTTGCAGTGAGCCGAGATCGCGCCACTGCACT- C C
HSU14569	AGCTTGCAGTGAGCCGAGATCCCGCCACTGCACT- C C
gb—M26939	GACTTGCCGTGAGCCAG-ATTGCGCC----TGCAG T C
gb—S54330	AGCTTGCAGTGAGCCAACATCGCGCCAC--TGCA- T C
gb—M20902	AGCTTGCAGTGAGCCGAGATCGCGCCAC--TGCAC T C
gb—M63005	----TGCAGTGAGCCGAGATCATGCCAC--TGCAC T C
gb—M64231	----TGCAGTGAGCTGAGATCGTGCCAC--TGCAC T C

Table 3: Similar regions in different alignments of Alu-Sb subfamily sequences. Boundary of the sequences may differ in different alignments. The consensus for HMM and CLUSTAL alignments was deduced by hand.

the consensus alphabet, as well as additional parameters λ_{ext}, μ_{ext}, probabilities of insertion and deletion extension.

We applied this extended ASA to the alignment of 252 E.Coli promoters[14], with $\lambda = \mu = \lambda_{ext} = \mu_{ext} = .1$ and different probabilities of substitution (fig. 2). When the requirement of agreement within assembly columns becomes less stringent the consensus reveals more significant consensus letters. Even in case (a), in addition to the familiar -10 and -35 boxes, the consensus contained T +3 bps downstream from -35 box with 60% conservation. When sequences were aligned to both boxes without indels a weak conservation of T was reported[14] at positions +3 and +4. Our alignment shows that these two weak conserved positions can be explained by different spacing between a conserved T and the -35 box. One can find other consensus phenomena of this kind in the data, but in this case it is not clear which have biological significance, and which may be attributable to artifacts resulting from the very specific form of the parametric model we are using.

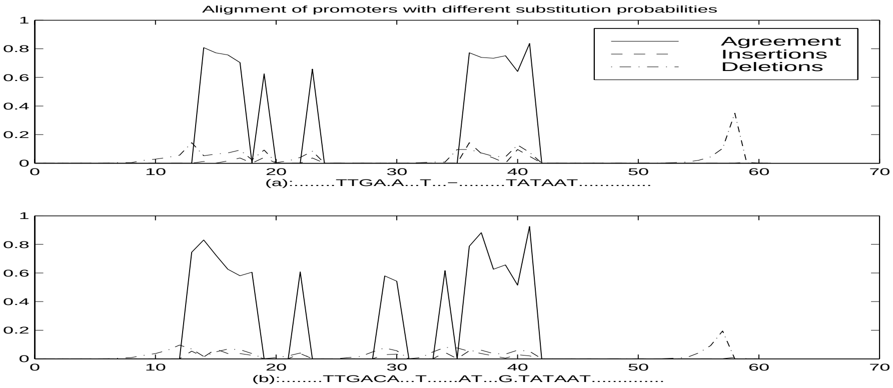

Figure 2: Alignment of E.Coli promoters with substitution probability .2(a) and .3(b) and the rest of the error parameters equal to .1. Every "N" in consensus is substituted by a dot. The figure shows the average number of insertions and deletions, as well as conservation, in the significant consensus positions.

5 Conclusion

We presented a probabilistic approach to multiple sequence alignment, embodied in the ASA algorithm. An ASA alignment differs from those produced by SAM and CLUSTAL, in that it has higher agreement in columns and looks more like an assembly.

In our experiments on synthetic data, ASA inferred the correct consensus with almost maximal possible accuracy and showed the robustnes to parameter choice. The latter indicates that in HMM training for this type of DNA data, the values of the continuous parameters that characterize the agreement with the consensus in Match states and transitions into Insertion and Deletion states do not play any significant role, and the effort should instead be concentrated on accurate architecture inference. On synthetic data ASA by far outperformed CAP2 and ReAligner at the expense of more extensive computations. Currently, ASA is being implemented in an EST assembly refinement that realigns noisy assembly regions. Further work needs to be done to determine the applicability of this algorithm to genomic data.

Acknowledgements

B. Lavareva-Ulitsky gratefully acknowledges the support of a postdoctoral fellowship from SmithKline Beecham, and D. Haussler the support of DOE grant DE-FG03-95ER6211.

1. D. Sankoff and R.Cedergren in *Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison*, eds. D. Sankoff and J. Kruskal (Addison-Wesley, 1983).
2. M. Waterman, *Introduction to Computational Biology* (Chapman&Hall, 1995).
3. P. Bucher and K. Hoffman in *ISMB-4*, eds. D. States et al.(AAAI Press, 1996).
4. R. Durbin, S. Eddy, A. Krogh and G. Mitchison, *Biological Sequence Analysis* (Cambridge University Press, 1998).
5. A. Krogh, M. Brown, I. Mian, L. Sjolander, and D. Haussler, *J.Mol.Biol* **235**, 1501 (1994).
6. S. Eddy, *Current Opinion in Structural Biology* **6**, 361 (1996).
7. T. Yada, Y. Totoki, M. Ishikawa, and K. Asai, in *Genome Informatics*, eds. T. Akutsu et al. (Universal Academy Press, 1996).
8. A. Lukashin, J. Engelbrecht and S. Brunak, *Nucleic Acids Res.* **20**, 10, 2511 (1992).
9. J. Kim , S. Pramanik, and M.J. Chung, *CABIOS* **10**, 4, 419 (1994).
10. J. Liu, A. Newald, and C. Lawrence, *JASA* **90**, 432,1156 (.)
11. R. Hughey and A. Krogh, *CABIOS* **12**, 95 (1996).
12. X. Huang, *Genomics* **33**, 21 (1996).
13. E. Anson and E. Myers, *J. of Computational Biology* **4**, 3, 369 (1997).
14. C. Harley and R. Reynolds, *Nucl. Acids Res.* **15**, 5, 2343 (1980).
15. J. Zhu, J. Liu, and C. Lawrence, in *ISMB-5*, eds. T. Gaastered et al. (AAAI Press, 1997).
16. Milosavljevic A., *Categorization of Macromolecular Sequences by Minimal Length Encoding* (PhD thesis, UCSC, 1990).
17. L. Rabiner, *Proceedings of the IEEE* **77**, 257 (1989).
18. W. Hastings, *Biometrika* **57**, 1, 97 (1970).
19. G. Granger, O. White, M. Adams, and A. Kerlavage, *Genome Science & Technology* **1**, 1, 9 (1995).
20. I. Holms and R. Durbin, in *RECOMB 98*, eds. S. Istrail et al. (ACM Press, 1998).
21. J. Thompson, D. Higgins, T. Gibson, Nucl. Acids Res. **22**, 4673 (1994).