TOWARDS AN ACCURATE EST CONSENSUS [a]

B. LAZAREVA-ULITSKY[1], S. RAHMANN, and D. HAUSSLER[2]
Department of Computer Science,
University of California at Santa Cruz,
Santa Cruz, CA 95064, USA, fax: (831) 459-4829,
[1] *betty@cse.ucsc.edu, tel: (831)459-3576*
[2] *haussler@cse.ucsc.edu, tel: (831)459-2105*

The paper deals with application of a new multiple alignment algorithm to accurate consensus inference in EST assembly. The algorithm, which we call ASA, is based on the maximum *a posteriori* probability (MAP) estimate of a consensus sequence. ASA takes as input a tentative assembly and re-estimates the assembly and its consensus in ambiguous regions in order to find a more accurate consensus. As output, ASA produces an optimal consensus as well as its possible alternatives with their relative probabilities in ambiguous regions. It can also provide a file with detailed reliability values that specify loss in log-probability when any consensus residue is modified using appropriate substitutions, insertions or deletion. When applied to EST assembly, this information on the accuracy of the consensus can be used further to correct possible frame-shifts and reconstruct ORFs. Our algorithm takes care of sequencing artifacts, such as gel compression, and reports single nucleotide polymorphisms (SNPs) it finds in the course of alignment. We compare the accuracy of ASA, CAP2, PHRAP and TIGR on ESTs from UNIGENE clusters with known mRNA. Our experiments show that TIGR outperforms CAP2 and PHRAP, and in turn, ASA is slightly better than TIGR in terms of consensus accuracy, and uses significantly fewer wildcards.

1 Introduction

A probabilistic multiple alignment algorithm (ASA) based on MAP estimation of a consensus sequence was introduced in [1]. It was assumed there that observed sequences were copied from an unknown original sequence with insertion, deletion and substitution errors, and that the probabilities of those errors do not depend on the positions in the original and observed sequences. In the experiments on synthetic data generated according to this model, ASA found the true consensus in almost all cases

[a] *Keywords: EST consensus, consensus reliability, SNP detection*

where it could have been recovered, and significantly outperformed known alignment programs when the error rate was high.

In section 2 we present an extension of ASA for use in consensus inference from EST assemblies. The sequencing errors in ESTs are known to depend highly on the position in observed sequence and on the local context (e.g., see[2,3]). To supply ASA with position-specific errors we trained a neural network to take an EST sequence and context-specific information as input, and estimate the insertion, deletion, and substitution probabilities at every position in the EST sequence. The probabilities computed by the neural net are then used in ASA's model, along with ASA's optimization algorithm, to produce a MAP estimate of the consensus sequence for any particular "noisy" region of an EST assembly.

By locating and reestimating all "noisy" regions, ASA can take a layout of the contig produced by any assembler and find its own MAP consensus sequence along with the optimal alignment that corresponds to it. Since ASA is capable of finding not only optimal, but also suboptimal consensus sequences, it outputs possible alternatives to the consensus in ambiguous regions. It can also output a file with detailed reliability values, that indicate how much the log-probability will be changed if any particular position in the consensus is modified using substitution, insertion, or deletion. Those detailed reliabilities extend the confidence values estimated for every position of a fixed alignment in [10]. ASA's detailed probabilistic model is sensitive enough to use for SNP detection as well. Indeed, it is impossible to completely disentangle the problem of detecting SNPs in EST alignments from the problem of detecting and repairing misalignments and sequencing errors in EST assemblies: one cannot be solved without the other. Using ASA, the SNPs are marked as wild cards in the consensus and are evaluated in the course of alignment.

In section 3 we present experimental results comparing the accuracy of the consensus predicted by ASA, CAP2, PHRAP and TIGR. The comparison was performed on ESTs from UNIGENE clusters that contain a known mRNA for reference. ASA predicted the consensus most accurately, though at the expense of more extensive computations. We believe that this tradeoff of speed versus accuracy will be justified in cases where it is important that protein homologies are not missed because of frame shift errors in consensus estimates of genes from EST assemblies, and where it is important to get more sensitive detection of SNPs from EST assemblies.

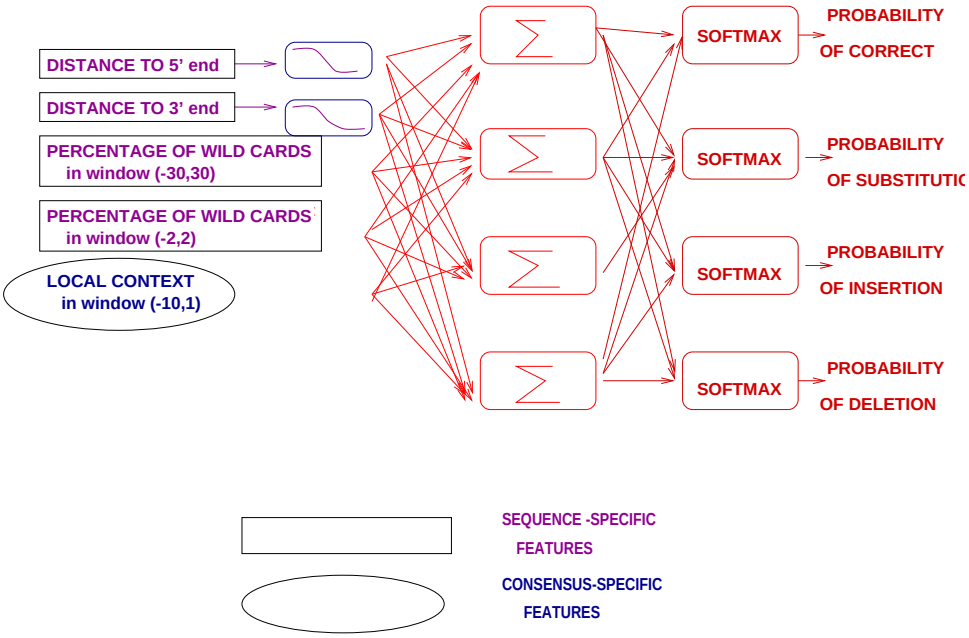

Figure 1: A schematic representation of a NN designed to estimate position-specific probabilities of correct base-calls, substitutions, insertions and deletions. It takes as an input the distance from the EST position being evaluated to the 5' and 3' end of the EST, the proportion of wild cards in two context windows around the read of different sizes (+/- 30 bases and +/- 2 bases), and also information taken from the previous 10 bases, the current and the next bases in the true mRNA from which the read was obtained. Note, that in consensus reconstruction the true mRNA is unavailable to the NN. However, when ASA evaluates posterior probabilities of possible consensus sequences, the NN uses those candidates in place of the true mRNA.

2 ASA Inference of EST Consensus

2.1 Neural Network for Estimation of Position-specific Errors in ESTs

The sequence quality at a particular position in a DNA read is known to depend, among other things, on the location of this position with respect to the 5' and 3' ends, on the quality of base-calls in the vicinity of the position, and on the local context of this position in the true DNA or mRNA. To estimate these position-specific errors we designed and trained a neural network (NN) with one hidden layer (fig.1), using gradient descent with negative log-likelihood as an objective function to be minimized during training.

The training and test examples for the net were obtained from pairwise

comparisons between mRNAs and corresponding ESTs from UNIGENE [11]. Both the training and test data sets contained about 15,000 correct base-call examples and about the same number of false base-call examples. We compared the predictions of our NN with that of a simple homogeneous model that predicts the same probabilities of correct base-calls, substitutions, insertions and deletions at every EST position. Those homogeneous probabilities were estimated from all available examples to be, respectively, { 0.955, 0.021, 0.015, 0.009 }. The advantage of NN model over a homogeneous model can be measured in *bits saved* as a difference in negative log-likelihood scores of the two models. This difference will be negative if NN model outperforms the homogeneous model. The average number of bits saved for one true base-call example, substitution example, insertion and deletion example is found to be, respectively, {-0.0066, -1.586, -2.091, -2.708 } for the training set and { -0.0088, -1.710, -2.357, -2.938 } for the test set, indicating that the neural net model represents a significant improvement over the simple homogeneous error model.

2.2 Identification of SNPs and Sequencing Artifacts

ASA was also extended to identify polymorphisms and allow for sequencing artifacts such as gel compression and inversions. The identification of SNPs can be naturally incorporated in our probabilistic approach to consensus inference. At every position that might look like a polymorphic site, ASA computes the posterior probability of a SNP. If this is low, the variation of nucleotides at the site in question is explained by sequencing errors, whose probabilities are intrinsicly estimated by the neural net described in 2.1. If the posterior probability of a SNP is high, then a wild card is used at this position in the consensus sequence.

Our initial experiments were performed on SNPs from HGBASE[5] (Human Genic Bi-Allelic SEquences) at Uppsala University , where the majority of SNPs are mirrored from Whitehead Insitute SNP database [6]. We examined EST assembly regions of length 51 derived from UNIGENE which were known to contain an SNP from HGBASE in the middle and which had coverage higher than 3. Out of this overall set of 110 HGBASE SNPs, only 30 showed polymorphic variation at the corresponding cite of assembly region in the UNIGENE cluster, which can be explained by the unrepresentative sample of libraries that contributed ESTs to these UNIGENE clusters. ASA detected all 30 of these SNPs, and also reported

5 extra SNPs. It is hard to estimate whether these are false positives, since most of the 51-long regions in HGBASE were carefully checked on European population, and UNIGENE comprises ESTs from a variety of populations. Further results on the quality of SNP detection will be described more fully in a forthcoming paper, after additional experiments are performed.

To detect gel compression, ASA specifically interrogates sites where the reads sequenced in opposite directions differ by an indel. Call this a candidate compression site. There are special error probability parameters in ASA to model gel compression, such that at a candidate compression site, bases read in the 'difficult' direction have very high probabilities of deletion. For corresponding bases sequenced in the opposite direction, error probabilities obey regular rules, as calculated by the neural net described above. ASA combines this differential evidence from reads going in opposite directions to evaluate the posterior probability that there is gel compression at each candidate compression site. In addition, that when ASA reports detailed reliability values discussed in section 2.3, *every* position in consensus is checked for possible gel compression, not just candidate compression sites.

In its current implementation ASA uses homogeneous prior probabilities of a SNP and gel compression, but these priors can be adjusted to fine tune the system if needed.

2.3 Example of ASA INPUT/OUTPUT

ASA takes as input a tentative assembly in GDE[4] (Genetic Data Environment) format. The assembly in this format has a simple representation, can be viewed via GDE, and is provided by TIGR Assembler[8].

ASA outputs a file **.cons* with re-estimated optimal consensus and possible alternatives to the consensus in its ambiguous positions. This file is also written in GDE format and can be viewed in GDE. An example of a **.cons* output file is given in Table 1. The consensus corresponds to the region of the EST assembly with coverage 4-2 (see Table 2). The ESTs in the contig are N42072, T49651, T49652, AA169576, T93013. This EST cluster is from the 5-lipoxygenase activating protein (FLAP) gene, and the given consensus segment corresponds to the end of coding region in the mRNA (X52195, M6326).

Along with the consensus in file **.cons*, ASA also outputs the corre-

```
     ..cpta.........cpta.................cpt.........  progs
     (..@.....).......@...................*..........  indels
351  TGGGCTACATATTTGGGGAAACGCATCATACTCTTCCTGTT-CCTCATGT  Optimal
     ---D-------------d----------------------t--------  Subopt.

     ................................................  progs
     ................................................  indels
400  CCGTTGCTGGCATATTCAACTATTACCTCATCTTCTTTTTCGGAAGTGAC  Optimal
     ------------------------------------------------  Subopt.

     ........cpt......cpta....cpt.....................  progs
     .........*........@.......*.....................  indels
450  TTTGAAAACT-ACATAAAAGACGATCTCC-ACCACCATCTCCCC-TCTAC  Optimal
     ----------t-------d----------c--------------t-----  Subopt.

     ..cpt.....................................  prog
     ...*......................................  indels
497  TTC-TCATTCCCTAA-..........................  Optimal
     ---t-----------a..........................  Subopt.
```

Table 1: Optimal consensus (lines 'Optimal') and its possible alternatives in ambiguous positions (lines 'Subopt'). The consensus corresponds to the end of coding region. On lines 'Subopt' the letters "D" and "d" denote possible deletions, the capital letters denote possible alternatives whose relative probabilities P (with respect to optimal consensus) are greater than .5, while small letters denote the alternatives whose P is less than .5 but greater than .01. The lines 'progs' and 'indels' are typed by hand. The positions where at least one of the programs made an indel are marked by asterisk '*' and '@' on line 'indels'. The list of the programs that made wrong prediction is given above the corresponding indel on line 'progs' ('c','p','t',and 'a' stand, respectively, for CAP2, PHRAP,TIGR, and ASA). The positions marked by '+' are reported incorrectly in the ASA optimal consensus (as well as by PHRAP, CAP2 and TIGR), however the alternatives in ASA's suboptimal consensus give the correct prediction. These were the only ASA errors in the coding region. It should be noted that CAP2 did not include the sequence 'ya78b04.s' in the contig, and thus the last 4 indels in CAP alignment come from the region of coverage 2 and 1 (see Table 2 for comparison).

sponding EST alignment in GDE format. An example of an ASA alignment that corresponds to consensus in Table 1 is shown in Table 2.

ASA can also provide reliability values in file **.rel*. Those values are obtained by evaluating posterior probabilities of all possible consensus sequences that differ from the optimal consensus by any insertion,substitution or deletion. In particular, when we estimate a probability that a base should be inserted in consensus before any position, we take into account the possibility of gel compression at this base. As an example, in Table 3 we give those reliability values for positions marked by '(...)' on the line 'indels' in Table 1.

3 Accuracy comparison of different assemblers

To compare the accuracy of ASA consensus prediction to that of known assemblers, such as CAP2 [7], PHRAP [9] and TIGR [8] we have chosen 4,645 clusters in UNIGENE [11] (Human sequence collection) that along with ESTs contain at least one mRNA. Each of these clusters was split into a file with ESTs and a file with mRNAs. We first ran the TIGR assembler on each EST cluster. From this, for each EST cluster we got several contigs with corresponding consensus sequences. Afterwards, we ran CAP and PHRAP on the ESTs from each contig produced by TIGR and if those assemblers preserved the contig (did not split it into subcontigs), we ran ASA to reestimate the consensus for this contig. Thus, in this experiment the consensus sequences reported by CAP, PHRAP, TIGR and ASA were all predicted from the identical sets of ESTs. To evaluate the accuracy of the consensus prediction, we compared the predicted consensus sequences with one of the corresponding mRNAs, assuming that this mRNA is correct. We performed our experiments on two different validation sets of UNIGENE clusters and got practically identical results. Those results for coding/noncoding regions, different coverage levels, and different error types are summarized in fig.2.

One can see that in terms of frame-shifts, TIGR is substantially more accurate than CAP2 and PHRAP, while ASA is a little more accurate than TIGR. It is also interesting to notice that ASA consistently performs better in noncoding regions. Since TIGR reports a high number of wild cards, its consensus is very ambiguous. It is clear from fig.2 that these ambiguities can be avoided, since ASA produces almost the same number

		...........@.............@.....	
343		........-TGGGCTA-C-ATATTTGGGGAAACGCATCATA-CTC-TT-C	CONSENS.
347	+	........CTGGCCTA-CNATATTTGGGGAAACGCATCATA-CTCCTT-C	yw94h04.r
203	+	........-TGGGCTAAC-ATATTTGGGGAAACGCATCATAACTC-TTTC	ye26b09.r
196	+	........-TGGGCTA-C-ATATTTGGGGAAACGCATCATA-CTC-TT-C	zo89g07.r
190	+	........-TGG-CTA-C-ATATTTGGG-AAACGCATCATA-CTC-TT-C	ya78b04.r
		..*.....	
387		CTG-TT-CC-TCATG-TCCGTTG-CTGG-CATATT-CAAC-TATTA-CCT	CONSENS.
394	+	CTGGTTTCC-TCATG-TCCGTTG-NTGG-CATATT-CAACCTANTA-CCT	yw94h04.r
250	+	CTG-TTTCC-TCATG-TCCGTTGGNTGGGCATATTTCAAC-TATTAACCT	ye26b09.r
240	+	CTGGTT-CCNTCATGGTCCGTTG-CTGG-NANATT-CAAC-TATTA-CCT	zo89g07.r
232	+	CTG-TT-CC-TCATG-TCCGTTG-CTGG-CATATT-CAAC-TATTA-CCT	ya78b04.r
		.....*.....@...	
428		CAT-CTTCTTTTT-CGG-AAGTG-ACTTT-GAAAA-CT-ACATAAAAG-A	CONSENS.
438	+	CATt.....	yw94h04.r
296	+	CAT-CTTCTTTTTTCGGGAAGTGGANTTTTGAAAAANT-ACATAAAAGGA	ye26b09.r
284	+	NAT.....	zo89g07.r
273	+	CAT-CTTCTTTTT-CGG-AAGTG-ACTTT-GAAAA-CTTACATAAAAG-A	ya78b04.r
1	-	-------CTTTTT-CGG-AAGTG-ACTTT-GAAAA-CT-ACAT-AAAG-A	ya78b04.s
		.....*.....*.....	
470		CGATCT-CCACCACCATCTCCCCT-CTACTTCT-CATTCC CTAA-.....	CONSENS.
344	+	CGATCTTCCACCACCATCT.....	ye26b09.r
316	+	CGATCT-CCACCACCATCTCCCCTTCTACTTCTTCATTCC CTAAA.....	ya78b04.r
36	-	CGATCTCCCACCACCATCTCCCCT-CTACTTCT-CANTCC CTAA-.....	ya78b04.s

Table 2: The segment of ASA alignment that corresponds to consensus region in Table 1. The signs '+'and '-' denote the sequencing direction. The asterisk '*' and '@' mark corresponding positions in Table 1.

pos	bp	del	insertion				substitution			
			A	C	G	T	A	C	G	T
350	T	-21.72	-8.84	-6.36	-8.90	-8.91	-26.16	-20.81	-24.15	0.00
351	G	-0.61	-8.41	-8.41	-8.41	-8.41	-8.79	-8.81	0.00	-8.66
352	G	-0.61	-8.36	-8.36	-8.36	-8.36	-8.30	-8.31	0.00	-8.32
353	G	-0.61	-8.39	-7.73	-8.39	-8.39	-12.69	-3.79	0.00	-12.00
354	C	-17.09	-10.28	-8.21	-10.28	-10.28	-23.24	0.00	-25.91	-24.67
355	T	-32.50	-8.52	-8.51	-8.52	-8.51	-23.43	-21.67	-24.10	0.00
356	A	-26.53	-7.36	-8.21	-8.20	-7.98	0.00	-24.76	-25.95	-29.15
357	C	-18.36	-7.20	-6.96	-6.76	-6.87	-19.78	0.00	-22.74	-21.86
358	A	-20.00	-5.17	-5.14	-4.99	-4.95	0.00	-21.94	-23.11	-22.54
359	T	-31.02	-5.17	-8.40	-8.40	-8.40	-26.12	-26.72	-27.86	-0.00

Table 3: The detailed reliability values at every consensus position (pos) with the base (bp) give the loss in log-probability of consensus if it is modified at this according to corresponding change (deletion, insertion or substitution). For example, the log-probability of the consensus will be decreased by only 0.61 if any of the 'G's at positions 351-353 are deleted. Thus this change is quite possible. On the other hand, if 'C' is inserted before position 353 , the log-probability of consensus will drop down by 7.73, and thus an extra 'C' at this position is much more unlikely. However, note that that the cost of insertions is still lower than the cost of most substitutions and deletions. This can be explained by non-neglectable posterior probability of gel compression at these sites. Indeed, because all sequences are reads in the same direction (see Table 2), and any evidence from the opposite direction is missing, the posterior probability of compression will be close to its prior probability.

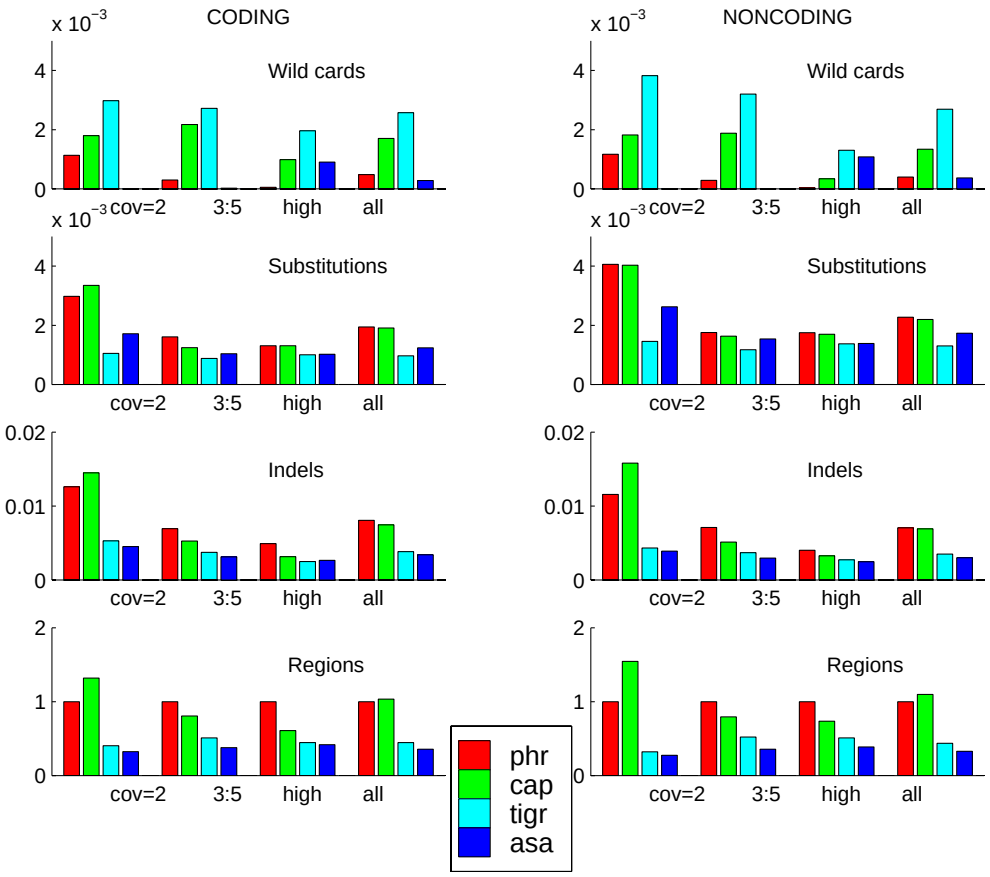

Figure 2: Accuracy comparison of CAP2, PHRAP, TIGR, and ASA for coding/noncoding regions. The performance of every program is considered separately for regions with different coverage level. The first cluster in every picture corresponds to coverage 2, the second one to coverage 3 to 5, the third one to coverage higher than 5, and the final one to overall assembly. The first row of histograms give average frequency of wild cards reported by every program. Only in ASA wild cards denote polymorphic sites, in the rest of the programs wild cards denote uncertainties. The next two rows of histograms give average frequencies of substitutions and indels produced by the four programs. In our calculations of substitutions we assumed that a wild card matches the nucleotides it denotes (e.g., 'N' matches everything). Since TIGR reported a very high number of wild cards, it was thus able to score a relatively low number of substitution errors. To look at the frame-shift problem, in the last row of histograms we give the proportion of noisy regions realigned by ASA whose length was reported incorrectly by the four programs.

of substitutions as TIGR does, using almost no wild cards. In this sense, ASA is significantly better than TIGR.

4 Discussion

In our research we restricted ourselves to consensus inference without trace data and estimated the quality of individual base-calls in ESTs using a neural net. This allows the method to be applied to UNIGENE and other EST clusters/assemblies made from dbEST, which contain neither trace data nor quality assessments. However, information from trace data, e.g., PHRAP reliability values, when available, can be very helpful in improving the accuracy of the consensus sequence. In future work we plan to incorporate this information in ASA's consensus estimation. One of the possible ways to do this is to convert reliability values reported by PHRAP into probabilities of substitutions, insertions, and deletions, used in our algorithm.

The comparison of different assemblers in terms of consensus accuracy on EST data without quality assessments derived from traces, showed that TIGR substantially outperforms CAP2 and PHRAP, and ASA is slightly more accurate than TIGR. While TIGR assembler outputs a high number of wild cards to denote uncertain bases, ASA uses wild cards only to denote SNPs and has almost the same rate of substitution errors as TIGR. Though ASA is more computationally intensive, it should be used when the accuracy of consensus prediction is the main concern.

The probabilistic approach implemented in ASA provides important information on the confidence of base-calls in an EST assembly consensus. This information can be used to reconstruct an ORF and search for homologs. For example, one could use this information in a dynamic programming algorithm that exploits the detailed reliability values reported by ASA, along with coding/noncoding statistics or protein homology, to parse an EST consensus into coding/noncoding regions with possible frame shift corrections.

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 and NSF grant IRI/9123692. We are

also thankful to P. Agarwal for helpful guidance, to G. Churchill for fruitful discussions, to B. Pearce (UCSC) who did the preprocessing of UNIGINE clusters, and to M. Cline (UCSC) for her help with NN code.

1. B. Lazareva, and D. Haussler (1999). A Probabilistic Approach to Consensus Multiple Alignment. *PSB 99*, 150-161. http://www.cse.ucsc.edu/research/compbio/papers/PSB99.ps
2. B. Koop et.al (1993). Sequence Length and Error Analysis of Sequenase and Automated *Taq* Cycle Sequencing Methods. *BioTechniques*, 14(3):442-447.
3. C. Lawrence, and V. Solovyev (1994). Assignment of Position-specific Error Probability to Primary DNA Sequence Data. *Nucleic Acids Research*, 22(7):1272-1280.
4. S. Smith et.al (1994). The Genetic Data Environment: An Expandable GUI for Multiple Sequence Alignment. *CABIOS*, 10:671-675.
5. http://hgbase.interactiva.de/
6. http://www.genome.wi.mit.edu/SNP/human/
7. X. Huang (1996). An Improved Sequence assembly Program. *Genomics*, 33:21-31.
8. G. Granger et.al (1995). TIGR Assembler: A new Tool for Assembling Large Shotgun Sequencing Projects. *Genome Science & Technology*, 1(1):9-19.
9. P. Green et.al. http://www.genome.washington.edu/UWGC/alysistools/phrap.htm
10. G. Churchill, and M. Waterman (1991). The Accuracy of DNA Sequences: Estimating Sequence Quality. *Genomocs*, 14:89-98.
11. G. Schuler et.al. (1996). A Gene Map of the Human Genome. *Science* 274:540-546. ftp://ftp.ncbi.nlm.nih.gov/repository/UniGene/Hs.seq.all