Computational Identification of Noncoding RNA Genes through Phylogenetic Shadowing

Kushal Chakrabarti[1] **and Daniel L. Ong**[2]

Keywords: noncoding RNAs, gene prediction, comparative genomics

1 Introduction

Although fairly accurate databases exist for protein-coding genes, little is known about another important class of genes known as *noncoding RNA genes*. These genes, which have been implicated in a wide variety of critical biochemical pathways including brain development [7] and viral defense [5], are not translated into polypeptides. Instead, their transcribed RNAs fold into stable, base-paired secondary and tertiary structures that confer catalytic ability. For the purposes of this paper, it is especially important to note that these secondary structures cause noncoding RNA genes to contain pseudo-palindromic sequences.

Unfortunately, these pseudo-palindromic and other signals are not statistically sufficient for the computational identification of such genes [9]. Because they are difficult to detect even through biological techniques, it is important that accurate computational approaches be developed [10]. Although many heuristic and specialized methods have been suggested, comparative genomics approaches [8] have shown particular promise. However, even these approaches are primitive. For instance, current comparative genomics approaches are limited to two sequences, despite recent work showing the importance of using several related species [2]. Other problems include various heuristic approximations and poor scaling.

Here, we briefly present a machine learning approach to genome-wide noncoding RNA gene prediction with multiple sequence alignments. The algorithm we describe scales linearly with respect to the length and number of genomes. More importantly, the approach is statistically sound, and allows the direct computation of probabilities through modular protein-coding, noncoding RNA, and intergenic sequence models.

2 Theory

We have developed a graphical model that integrates *probabilistic context-free grammars (PCFGs)* and phylogenetic trees within a *generalized hidden Markov model (GHMM)* framework. The latter of these models, the GHMM, is a straightforward generalization of the HMM in which a state can explicitly choose the length of its emission (according to a specified length prior). Our GHMM can be conceptually considered as three different states, each of which emits sequence alignments based on protein-coding, noncoding RNA, or intergenic sequence models.[3] The protein-coding and intergenic models emit multiple alignment columns according to standard GHMMs [1] and phylogenetic trees.

On the other hand, the noncoding RNA model emits columns according to PCFGs and phylogenetic trees that define a joint probability distribution over pairs of columns. This simultaneous emission of multiple, distant columns allows the PCFG to model the evolutionary dependencies between base-paired positions within noncoding RNA genes [9]. In addition, the inclusion of phylogenetic trees permits us to model the expectation that base-paired

[1]Dept. of Computer Science, Univ. of Calif., Berkeley. E-mail: kushalc@uclink.berkeley.edu
[2]Dept. of Computer Science, Univ. of Calif., Berkeley. E-mail: dlong@ocf.berkeley.edu
[3]In practice, this is not strictly true because of noncoding RNAs encoded within introns.

positions within noncoding RNA genes will undergo *compensatory mutation*. Perhaps more importantly, alignments of several, closely related genomes can be used (*phylogenetic shadowing*) [2]; because noncoding RNA genes tend to mutate faster than protein-coding genes, such sequence sets prevent pathological alignments and increase the accuracy of comparative genomics approaches. Despite these apparent benefits, phylogenetic trees that define joint distributions over pairs of columns have only been recently developed [6].

Unfortunately, because PCFGs are too inefficient for long sequences, they must be approximated. Although Rivas and Eddy [9] use a windowing approach, it is both inefficient and probabilistically unsound. We instead enforce a standard constraint on the GHMM length priors, ie. that it be zero for all lengths beyond some constant C. This constraint allows us to straightforwardly and efficiently compute the necessary PCFG probabilities in time $O(C^3 + LC^2)$, where L is the length of the multiple alignment. For instance, the inside probability of the first C bases can be computed in the standard fashion in time $O(C^3)$, while the necessary probabilities for each remaining base can be incrementally computed in time $O(C^2)$. We briefly note that this approximation is "perfect," ie. the probability computed in this fashion is exactly equivalent to the probability computed naively.

3 Experimental Results

Seven yeast species were downloaded and analyzed. Genome-wide homology maps were first generated for these species [4], which were then used to align the genomes [3]. Aligned genomic DNA for annotated ORFs and noncoding RNAs was then extracted, and the remainder was classified as null sequence. The model was then used to reclassify these sequences; these model reclassifications were then compared against the known (true) classification. Although initial results are very promising, we intend to systematically study the performance of our algorithm against other noncoding RNA genefinders in the future.

References

[1] Alexandersson, M., Pachter, L., and Cawley, S. 2003. SLAM: Cross-species gene finding and alignment with a generalized pair hidden markov model. *Genome Research*, 13:496-502.

[2] Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., and Pachter, L., and Rubin, E.M. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, 299:1391-1394.

[3] Bray, N. and Pachter, L. MAVID: Constrained ancestral alignment of multiple sequences. Genome Research, in press.

[4] Dewey, C. Personal communication.

[5] Hamilton, A., Voinnet, O., Chappell, L., and Baulcombe, D. 2002. Two classes of short interfering RNA in RNA silencing. *The EMBO Journal*, 21(17):46714679.

[6] Jow, H., Hudelot, C., Rattray M., and Higgs, P.G. 2002. Bayesian Phylogenetics Using an RNA Substitution Model Applied to Early Mammalian Evolution. *Molecular Biology and Evolution*, 19(9):1591–1601.

[7] Krichevsky, A.M., King, K.S., Donahue, C.P., Khrapko, K., and Kosik, K.S. 2003. A microRNA array reveals extensive regulation of microRNAs during brain development. *RNA*, 9(10):12741281.s

[8] McCutcheon, J.P. and Eddy, S.R. 2003. Computational Identification of non-coding RNAs in Saccharomyces cerevisiae by comparative genomics. *Nucleic Acids Research*, 31(14):4119-4128.

[9] Rivas, E. and Eddy, S.R. 2000. Secondary structure alone is generally not statistically sufficient for the detection of noncoding RNAs. *Bioinformatics*, 16(7):583-605.

[10] Storz, G. 2002. An expanding universe of noncoding RNAs. *Science*, 296(5571):1260-1263.