Sequential Modeling for Identifying CpG Island Locations in Human Genome

Nilanjan Dasgupta, Simon Lin, and Lawrence Carin, *Fellow, IEEE*

***Abstract*—We consider several sequential processing algorithms for identifying genes in human DNA, based on detecting CpG ("C proceeds G") islands. The algorithms are designed to capture the underlying statistical structure in a DNA sequence. Sequential processing using a Markov model and a hidden Markov model are shown to identify most CpG islands in annotated (marked) DNA subsequences available from publicly available DNA datasets. We also consider a wavelet-based hidden Markov tree (HMT). In the context of the HMT, we address design of adaptive wavelets matched to CpG islands, this accomplished via lifting and genetic-algorithm optimization.**

***Index Terms*—DNA, genes, HMM, Markov models.**

DNA IS COMPRISED of a sequence of subunits called nucleotides [1]: adenine (*A*), cytosine (*C*), guanine (*G*), and thyamine (*T*). In the human genome, a *C* nucleotide is generally modified chemically by methylation if followed by a *G*. Methyl-*C* mutates into a *T* with a high probability. The methylation process is suppressed in localized segments of the genome, often at the "start" regions of many genes. These regions, characterized by a higher concentration of *C*-*G* dinucleotides than elsewhere, are called CpG islands ("C proceeds G"). Our objective is to produce models for distinguishing variable-length CpG islands from the rest of the DNA sequence. We consider the following algorithms: a Markov model (MM) [2], hidden Markov model (HMM) [3], [4] and a wavelet-based hidden Markov tree (HMT) [5]. These algorithms have been employed in many contexts, including examination of gene sequences. We believe this to be the first examination of such for detecting CpG islands.

In the Markov model, [2], the sequence of nucleotides are modeled as a Markov process. The model is therefore characterized by a 4×4 transition-probability matrix. Separate Markov models are designed for CpG and non-CpG sequences, and the classifier is a likelihood ratio. The HMM [6] employs two hidden states: one characteristic of an underlying CpG region, the other characteristic of non-CpG data. We therefore have a 2×2 state-transition matrix (allowing possible transitions from CpG and non-CpG regions). Each state is characterized by a four-dimensional probability vector, representing the state-dependent probability of observing a given nucleotide.

Manuscript received January 14, 2002; revised July 27, 2002. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Scott T. Acton.

N. Dasgupta and L. Carin are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708-0291 USA (e-mail: lcarin@ee.duke.edu).

S. Lin is with the Duke Bioinformatics Shared Resource, Duke University Medical Center, Durham, NC 27708 USA.

Digital Object Identifier 10.1109/LSP.2002.806062

In the context of the HMT, the four members of the nucleotide alphabet are mapped to discrete numbers (*A*, *T*, *G*, and *C* are mapped to 3/2, 1/2, −1/2, and −3/2, respectively). This mapping is not unique, and one could assign vector values (such as (1,0,0,0) for *A*, (0,1,0,0) for *T*, etc.) to individual nucleotides [7] and implement the HMT model in the vector space for CpG island identification. However, we chose to avoid the HMT implementation in the vector space in order to obtain a single HMT model (unlike multiple HMT models to capture the vector interactions) having a comparable complexity with respect to the MM and HMM. The numerical sequence is subjected to a multilevel wavelet decomposition, to generate a sequence of wavelet trees. The basic assumption in HMT modeling is stationarity of the observed DNA sequence over the support of a wavelet tree. Hence, all the wavelet trees created from annotated CpG islands are modeled by a single HMT. The HMT model parameters are optimized using an expectation–maximization (EM) algorithm [5].

In addition to considering traditional wavelet decompositions, we also address the design of wavelets matched to CpG islands. In this context, Sweldens [8] has developed a general algorithm for designing biorthogonal wavelets, implemented directly in the time domain. This formalism, termed "lifting," yields a simple technique for insertion into a general cost function. In lifting, a discrete sequence is partitioned into its even and odd indexed elements, and prediction (p) and update (u) filters are used to partition the system outputs into a coarse and fine representation of the original signal. We use finite-impulse-response filters to constitute p and u. Here, we design wavelets (p and u filters) specifically for every level of the wavelet decomposition. Hence, for an L-level wavelet decomposition, our objective is the design of L filter pairs $(p_1, u_1), \ldots, (p_L, u_L)$. The coefficients of the p and u filters at all levels are used to constitute a genetic algorithm (GA) chromosome [9]. For each set of filters under consideration by the GA, we perform the associated L-level wavelet decomposition. An HMT is designed based on training data, and the GA cost function quantifies the accuracy of detecting known CpG islands in testing data. The GA attempts to design wavelets to maximize this cost function.

GenBank [11] data contain annotated CpG and non-CpG data, and therefore they are used to test all algorithms. When testing and training are performed on GenBank data, the particular testing and training sequences are partitioned as to be distinct. In addition, we also consider HMT training based on Sanger Institute [10] data, this for CpG islands alone (no non-CpG data). In this case, the HMT training and testing data come from distinct sources. Classification is based on

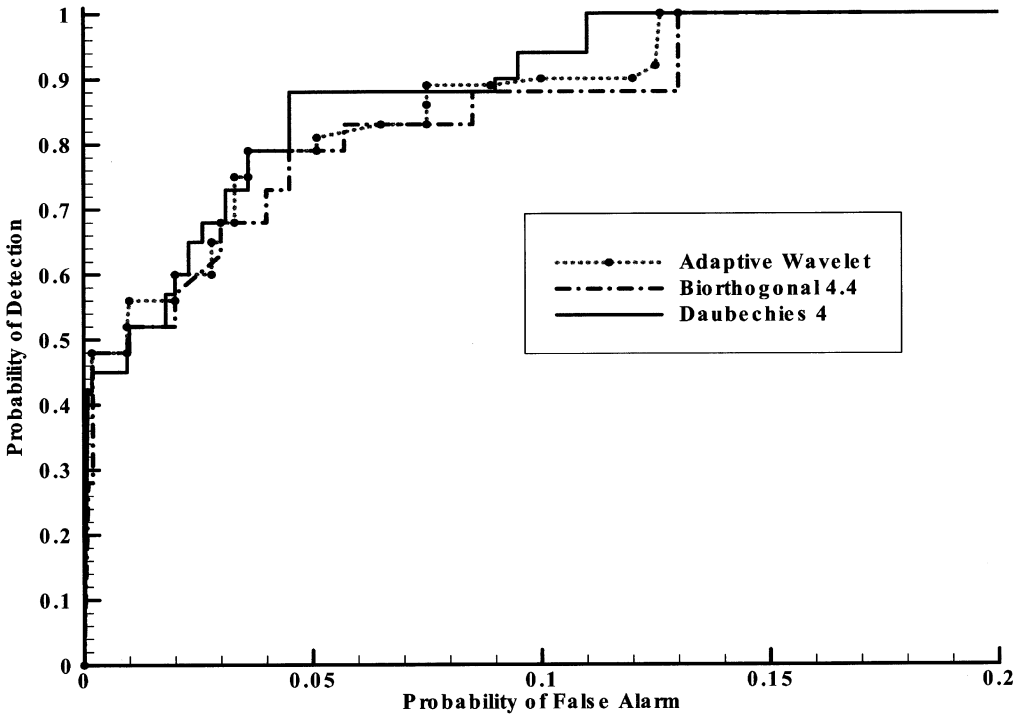

Fig. 1. ROC comparison between HMT model using standard orthonormal (Daubechies 4), biorthogonal (CDF), and GA-optimized sequence-specific adaptive wavelet.

subsequences of length 200, and consecutive windows overlap by 100 units. The MM, HMM, and HMT yield a likelihood (or likelihood ratio in the case of the MM). The associated output is thresholded to make a decision, as to the presence of a CpG island. Using a training set of GenBank data, we determine the threshold required for achieving a misclassification probability of 5% (non-CpG declared CpG), and this threshold is then used on the testing data.

We consider testing on GenBank DNA subsequence AF111 167 (see [11]). The HMT considered has $L = 3$ levels, although results did not vary significantly for more wavelet levels. In Table I, we present expected locations of CpG islands (based on chemical experiments) as well as the predictions of the three algorithms. With regard to the HMT, we achieved identical results independent of whether the training was done with GenBank or Sanger Institute data, this underscoring algorithm robustness (the training CpG data are entirely distinct for this case).

Concerning the results in Table I, we note that the HMT model (Daubechies 4 wavelet [12]) yields a larger set of declared CpG islands compared to the MM and HMM with most declarations in close accord with the empirical expectations. However, each algorithm predicts islands beyond those empirically predicted. These additional predictions are of interest for future chemical studies, of finer sensitivity (i.e., their accuracy cannot be verified at this point). For this case, all three algorithms have missed the island located between elements {152 731–152 951}, while providing comparable ability to detect the remaining expected CpG islands.

In Fig. 1, we present a comparison of HMT classification results based on a Daubechies 4 [12] orthonormal wavelet, on a similar biorthogonal wavelet (Cohen–Daubechies–Faveau (CDF) [13]) of the same support as the Daubechies 4, and based on a GA-designed biorthogonal wavelet of the same support as the CDF. The training of the HMT was done on the Sanger Institute data, with testing on GenBank data. The results are shown as the probability of detecting CpG islands Pd as a function of the probability of falsely predicting the presence of CpG islands Pf. This represents the well-known receiver-operating characteristic (ROC). As indicated above, "truth" as to the presence of CpG islands in the GenBank data is based on chemical measurements. However, it is possible that there are more CpG islands present than currently known. Therefore, the "truth" used to define Fig. 1 is not absolutely certain. Nevertheless, this does present a reasonable comparison of the HMT based on three distinct wavelets. These results indicate that the HMT performance based on standard wavelets is comparable to that based on an optimized wavelet. This suggests 1) that the CpG islands may be too heterogeneous to

TABLE I
PREDICTION OF CpG ISLANDS IN GENE SUBSEQUENCE "AF111 167" FROM GENBANK. SHOWN ARE EXPERIMENTALLY KNOWN CpG ISLANDS AS WELL AS THE PREDICTIONS OF THE THREE ALGORITHMS

True Position	MM Prediction	HMM Prediction	HMT Prediction
3999 - 4356	4252 - 4777	4052 - 4351	4002 - 4301
21918 - 22806	22327 - 23227	22352 - 22729	22102 - 23001
23126 - 25035	23602 - 25552	23702 - 24001	23502 - 24602
38568 - 38844	26527 - 26752	25052 - 25351	24902 - 25401
39157 - 39414	39127 - 40102	26501 - 26701	26302 - 26801
152731 - 152951	79102 - 79702	39002 - 39301	38902 - 39801
172378 - 173780	172552 - 174427	78902 - 79235	78504 - 79601
	184177 - 184402	172652 - 172951	121602 - 122101
	198727 - 198802	184151 - 184351	172302 - 174401
			183902 - 184401
			198601 - 198801

be particularly well characterized by a single wavelet, and 2) it demonstrates the ability of the HMT to adjust its parameters to suite the wavelet under consideration.

In summary, we have considered three sequential-processing algorithms for analyzing sequential DNA data, with the goal of detecting CpG islands. The MM, HMM, and HMT models yield similar results, with the number of potential CpG islands detected increasing with respective model complexity (for a fixed threshold). Chemical experiments are required to further examine potentially new CpG island discoveries. The HMT results were found to be weakly dependent on the particular wavelet chosen, even when the wavelet was optimized via lifting-based GA design.

REFERENCES

[1] T. A. Brown, *Genomes*. New York: Wiley, 1999.

[2] D. R. Cox and H. D. Miller, *The Theory of Stochastic Processes*. London, U.K.: Chapman & Hall, 1965.

[3] A. C. Camproux *et al.*, "Hidden Markov model approach for identifying the molecular framework of the protein backbone," *Protein Eng.*, vol. 12, no. 12, pp. 1063–1073, Dec. 1999.

[4] L. R. Rabiner and B. H. Juang, "An introduction to hidden Markov models," *IEEE Acoust., Speech, Signal Processing Mag.*, vol. 3, pp. 4–16, Jan. 1986.

[5] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 46, pp. 886–902, Apr. 1998.

[6] R. Hughley and A. Krogh, "Hidden Markov models for sequence analysis: Extension and analysis of the basic method," *Comput. Appl. Biosci.*, vol. 12, pp. 95–107.

[7] D. Anastassiou, "Frequency-domain analysis of biomolecular sequences," *Bioinformatics*, vol. 16, no. 12, pp. 1073–1081, 2000.

[8] W. Sweldens, "The lifting scheme: A custom-design construction of biorthogonal wavelets," *J. Appl. Comput. Harmon. Anal.*, vol. 3, pp. 186–200, 1996.

[9] E. Jones, P. Runkle, N. Dasgupta, L. Couchman, and L. Carin, "Genetic algorithm wavelet design for signal classification," *IEEE Pattern Anal. Mach. Intell.*, vol. 23, pp. 890–895, Aug. 2001.

[10] Sanger Institute. (2002). [Online]. Available: ftp://ftp.sanger.ac.uk/pub/human/sequences/CpG_Island_tag_sequences/CpG_island_library_reads

[11] NIH. (2002). GenBank. [Online]. Available: http://www.ncbi.nlm.nih.gov

[12] I. Daubechies, "Ten lectures on wavelets," *CBMS-NSF Lecture Notes*, vol. 61, 1992.

[13] C. S. Burrus, R. A. Gopinath, and H. Guo, *Introduction to Wavelets and Wavelet Transforms*. Englewood Cliffs, NJ: Prentice-Hall, 1998.

[14] J. P. Fitch and B. Sokhansanj, "Genomic engineering: Moving beyond DNA sequence to function," *Proc. IEEE*, vol. 88, Dec. 2000.

[15] R. L. Claypoole, R. G. Baraniuk, and R. D. Nowak, "Adaptive wavelet transforms via lifting," in *Proc. ICASSP*, 1998.

[16] N. Dasgupta, P. Runkle, L. Couchman, and L. Carin, "Dual hidden Markov model for characterizing wavelet coefficients from multi-aspect scattering data," *Signal Process.*, May 2001.