RECONSTRUCTING LATENT PERIODS IN GENOME SEQUENCES WITH INSERTIONS AND DELETIONS

Raman Arora[1]*, Colin Dewey*[2] *and William A. Sethares*[1]

University of Wisconsin-Madison
[1] Department of Electrical and Computer Engineering, 1415 Engineering Drive, Madison WI 53706,
[2] Department of Biostatistics and Medical Informatics, 1300 University Ave, Madison, WI 53706,
ramanarora@wisc.edu, cdewey@biostat.wisc.edu, sethares@ece.wisc.edu

ABSTRACT

Tandem and latent repeats in genome sequences provide insight into its various structural and functional roles. Such regions in genome sequences are modeled as cyclostationary processes, generated by a collection of information sources in a cyclic manner. The maximum likelihood (ML) estimates can be easily generated for the cyclostationary profiles and for the statistical period of such subsequences. However, in the presence of insertions and deletions, the ML estimators suffer greatly in their ability to accurately identify the periods. This paper extends the cyclic model to a profile hidden Markov model (PHMM) to account for insertions and deletions. An iterative algorithm is developed to learn parameters of the PHMM and Viterbi algorithm is employed to learn the most likely path through the state space. This reconstructs likely insertions and deletions in the sequence and results in better estimates of the statistical period and cyclostationary profiles than the ML approach. Experimental results are provided with simulated sequences as well as with chromosome 1 sequence from human genome.

1. INTRODUCTION

The sequential structure of a genome has biological implications. Several regularities and base dependencies have been observed in DNA and protein sequences and are associated with various molecular functions. This paper focuses on repetitions and short-range recurring-statistical-dependencies in the symbolic sequences.

Genome sequences are symbolic sequences comprising strings of symbols (representing nucleotides or amino acids) drawn from a finite set (or alphabet), typically with no algebraic structure. These sequences exhibit various kinds of repetitions and regularities, and finding such features is fundamental to understanding the structure of the sequences. Latent periodicities in DNA sequences have been shown to be correlated with several structural and functional roles [1, 2, 3].

Most current approaches to detecting periodicities transform the symbolic sequences into numerical sequences and compute the Fourier transform [4, 5, 6]. Though this is computationally convenient, it imposes a mathematical structure that is not present in the data. In contrast, the formulation in [7] implies no mathematical structure on the alphabet and presents a general *mapping-invariant* approach to the detection of periodicities. Each symbol of the sequence is assumed to be generated by an information source with some underlying probability mass function (pmf) on the alphabet. The sequence is assumed to be generated by drawing symbols from a collection of such sources in a cyclic manner. This is a simple first-order Markov process with a trivial transition matrix. The number of sources is equal to the latent period in the sequence.

This paper extends the cyclic model to a Profile Hidden Markov model [8] to allow for insertions and deletions. An iterative algorithm is proposed for recovering latent periods by reconstructing likely insertions and deletions in the ancestral cyclostationary sequence. Results are provided with simulated as well as real DNA sequences.

2. MODELING PERIODICITIES WITH A MARKOV CHAIN

Let $\mathcal{A} = \{a_1, \ldots, a_L\}$ be a finite alphabet of size L. For DNA sequences, $\mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ where the symbols represent nucleotides Adenine, Cytosine, Guanine and Thymine respectively. A *discrete probabilistic source* is defined to be a sequence of probability mass functions $P^{(1)}, P^{(2)}, \ldots P^{(N)}$ on the alphabet $\mathcal{A}$. A probabilistic source is defined to be *cyclostationary* with period K if $P^{(n)}(a) = P^{(n+K)}(a)$ for all n and $a \in \mathcal{A}$. It can be realized with K information sources (or random variables) denoted as $X_1, \ldots, X_K$ that generate an n symbol long sequence $\mathbf{s}$ in a cyclic fashion. Consequently, the likelihood of observing the sequence $\mathbf{s}$ can be expressed solely in terms of the emission probabilities of the states. The emission probabilities of X_i are given by probability mass function P_i. Collecting the $|\mathcal{A}| \times 1$ dimensional vectors P_i into a matrix $\mathbf{Q}^{(k)} = [P_1, \ldots, P_k]$ gives a compact description of the k-periodic cyclostationary source $P^{(n)}$.

Let K denote the true period and k be the hypothesized period. The number of complete statistical periods in an N-symbol long k-periodic cyclostationary sequence $\mathbf{s}$ are $M = \lfloor N/k \rfloor$, where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x. Define $\lfloor i \rfloor_k = 1 + ((i-1) \mod k)$, where $(x \mod y)$ denotes the remainder after division of x by y. Then for $1 \leq i \leq N$, the symbol $\mathbf{s}_i$ is generated by the random variable $X_{\lfloor i \rfloor_k}$. The search space for k is the set $\mathcal{K} = \{1, \ldots, N_0\}$, for some $N_0 < N$

and for corresponding probabilistic source $\mathbf{Q}^{(k)}$ the search space is the subset $\mathcal{Q}^{(k)} \subseteq [0,1]^{|\mathcal{A}|\times k}$ of column stochastic matrices.

2.1. The Maximum Likelihood Estimate

The ML estimate of the cyclostationary source is the column-stochastic matrix given by the optimization problem

$$\mathbf{Q}_{\text{ML}}^{(k)} = \arg \max_{\mathbf{Q}\in\mathcal{Q}^{(k)}} \prod_{i=1}^{N} P(X_{\lfloor i \rfloor_k} = \mathbf{s}_i | k, \mathbf{Q}). \quad (1)$$

For fixed k, the $(j, \lfloor i \rfloor_k)^{th}$ element of the matrix $\mathbf{Q}_{\text{ML}}^{(k)}$, for $j = 1, \ldots, |\mathcal{A}|$, is given as [7],

$$\left[\mathbf{Q}_{\text{ML}}^{(k)}\right]_{j,\lfloor i \rfloor_k} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}\{\mathbf{s}_{(m-1)k+\lfloor i \rfloor_k} = \mathcal{A}_j\} \quad (2)$$

where $\mathbf{1}\{\cdot\}$ is the indicator function.

2.2. Regularized maximum likelihood estimator

MDL principle avoids overfitting automatically by trading off complexity with the goodness of fit: Given the data and a collection of hypotheses $\mathcal{Q}$, it picks the model that compresses the data most with respect to the description method. The best estimate of the cyclostationary period of sequence $\mathbf{s}$ is the $k \in \mathcal{K}$ that minimizes the description length

$$\mathbb{L}(\mathbf{s}; k) = \mathbb{L}(\mathcal{Q}^{(k)}) + \mathbb{L}(\mathbf{s}|\mathbf{Q}_{\text{ML}}^{(k)}) \quad (3)$$

where $\mathbb{L}(\mathcal{Q}^{(k)})$ is the description length (in bits) of the hypothesis $\mathcal{Q}^{(k)}$ and $\mathbb{L}(\mathbf{s}|\mathbf{Q}_{\text{ML}}^{(k)})$ is the length (in bits) of the description of the data when encoded by the best ML hypothesis $\mathbf{Q}_{\text{ML}}^{(k)} \in \mathcal{Q}^{(k)}$. The term $\mathbb{L}(\mathcal{Q}^{(k)})$ is the *parametric complexity* of the model and $\mathbb{L}(\mathbf{s}|\mathbf{Q}_{\text{ML}}^{(k)})$ is the *stochastic complexity* of the sequence given the model. The MDL estimator is given as [7],

$$K_{\text{MDL}} = \underset{k\in\mathcal{K}}{\text{argmin}} \left(2\lceil \log k \rceil + k|\mathcal{A}| \log \lceil \frac{N}{k} \rceil - \log P(\mathbf{s}|\mathbf{Q}_{\text{ML}}^{(k)})\right)$$

3. EXTENSION TO A PHMM FOR LATENT PERIODS WITH *INDELS*

The penalized ML estimator given by the MDL principle performs well even with severe mutation rates [7]. But in face of insertions and deletions the performance of the estimator degrades severely. Consider the sequence

ACT GCT CT ACT ACGAT ACT ACT ACT (4)

which evolved from the tandem repeats of ACT through several insertions, deletions and substitutions. The ML stochastic matrix $\mathbf{Q}_{\text{ML}}^{(k)}$, described by equation (2), is given by the following simple algorithm:

Algorithm: def $\mathbf{Q} = cyclo(\mathbf{s}, k, \mathcal{A})$
for j = 1 to k
  $s_j = \mathbf{s}(j : k : end)$
  for $a \in \mathcal{A}$
    $\mathbf{Q}(a, j) = \#(a \in s_j)/length(s_j, \mathcal{A})$
  end
end
return Q

---

where the function $length(s_j, \mathcal{A})$ returns the number of symbols in s_j from the alphabet $\mathcal{A}$. The correspondence

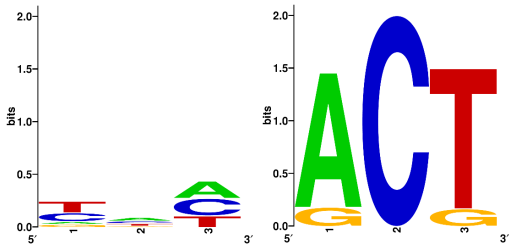

Figure 1. Weblogo depicting mutual information between repeats of 3-periodic DNA sequence (a) given in Equation (4), (b) gapped sequence in Equation (5).

between different periods is depicted by plotting the Weblogo [9] which captures the mutual information at each location of the period. Fig 1(a) shows Weblogo for the sequence in (4). We develop a method that optimally gaps the DNA sequence to mark likely insertions and deletions. The desired output is the optimally gapped sequence

ACT GCT -CT ACT ACG A-T ACT ACT ACT. (5)

Figure 1(b) shows the Weblogo [9] for the gapped sequence. The dominant period of the gapped sequence is ACT. The ML estimate for 3-periodic probabilistic source for the original and the gapped sequence are respectively

$$\mathbf{Q}_{\text{ML}}^{(3)} = \begin{bmatrix} 1/9 & 3/8 & 3/8 \\ 1/9 & 2/8 & 0 \\ 3/9 & 2/8 & 3/8 \\ 4/9 & 2/8 & 2/8 \end{bmatrix}, \mathbf{Q}_{\text{ML,gap}}^{(3)} = \begin{bmatrix} 7/8 & 0 & 0 \\ 1/8 & 0 & 1/9 \\ 0 & 1 & 0 \\ 0 & 0 & 8/9 \end{bmatrix}.$$

3.1. The revised model

In order to account for insertions and deletions when looking for statistical periodicities, a profile hidden Markov model (PHMM) is proposed as shown in Figure 2 for a 3-periodic cyclostationary source. Besides the cyclic transition between the states (X_1, X_2 ,X_3, ...) of the probabilistic source, each state can transition to an *insert* state (which models the symbols that are unlikely to be generated from the sources) or a *delete* state (which accounts for possibly skipped states in a cycle). The insert states have a feedback loop to model variable length block-inserts and a delete state can transition to the next delete state to account for multiple skips. The PHMM is parametrized by transition probabilities: $\tau = P(X_{\lfloor i \rfloor_k} \rightarrow X_{\lfloor i+1 \rfloor_k})$, $\epsilon = P(X_{\lfloor i \rfloor_k} \rightarrow I_{\lfloor i \rfloor_k})$, $\delta = P(X_{\lfloor i \rfloor_k} \rightarrow D_{\lfloor i+1 \rfloor_k})$ and the emission probabilities $e_I(\cdot)$ of the insert state and the probabilistic source $\mathbf{Q}^{(k)}$.

The gapped sequence in Figure 1(b) is reconstructed based on the a priori information that the ancestral sequence had tandem repeats of ACT. In the absence of this prior knowledge, a likely pattern (tandem repeat or a latent period) has to be learnt from the given sequence.

The next subsection briefly describes the Viterbi algorithm to learn the optimal path of states given the knowledge of the probabilistic source, emission probabilities and transition probabilities. A Gibbs sampling based method is outlined in Sec.3.3 to learn the probabilistic source $\mathbf{Q}^{(k)}$.

3.2. Learning the optimal path

Let π denote a path through the state space of the PHMM described in the previous section. Let $V_j^C(i)$ be the log-

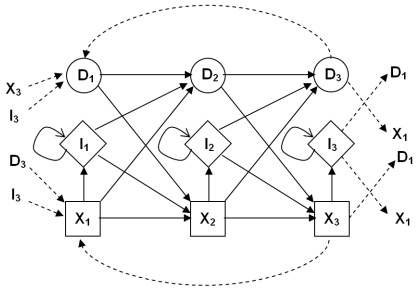

Figure 2. Profile HMM for cyclostationary probabilistic source with period 3.

likelihood of the best path π^* generating the subsequence $\mathbf{s}_{1\ldots i}$ with the symbol $\mathbf{s}_i$ being emitted by the j^{th} information source in the cycle $X_1, \ldots, X_k$. Similarly, $V_j^I(i)$ denotes the log-likelihood of the best path with $\mathbf{s}_i$ emitted by the insert state I_j and $V_j^D(i)$ is the log-likelihood of the best path ending in the delete state D_j. Then

$$\begin{aligned} V_j^X(i) &= \log P(X_{\lfloor j \rfloor_k} = x_i) \\ &\quad + \max \begin{cases} V_{j-1}^X(i-1) + \log t_{X_{j-1}X_j} \\ V_{j-1}^I(i-1) + \log t_{I_{j-1}X_j} \\ V_{j-1}^D(i-1) + \log t_{D_{j-1}X_j} \end{cases} \\ V_j^I(i) &= \log e_I(x_i) + \max \begin{cases} V_j^X(i-1) + \log t_{X_jI_j} \\ V_j^I(i-1) + \log t_{I_jI_j} \\ V_j^D(i-1) + \log t_{D_jI_j} \end{cases} \\ V_j^D(i) &= \max \begin{cases} V_{j-1}^X(i) + \log t_{X_{j-1}D_j} \\ V_{j-1}^I(i) + \log t_{I_{j-1}D_j} \\ V_{j-1}^D(i) + \log t_{D_{j-1}D_j} \end{cases} \end{aligned} \tag{6}$$

where $t_{\alpha\beta}$ denotes the transition probability from state α to state β and can be expressed in term of the parameters τ, ϵ and δ. At each update a pointer is created for each state to the previous state that maximized the likelihood of transitioning to the current state:

$$\begin{aligned} \gamma_i(X_j) &= \arg\max_\beta \left[V_{j-1}^\beta(i-1) + \log t_{\beta_{j-1}X_j} \right] \\ \gamma_i(I_j) &= \arg\max_\beta \left[V_j^\beta(i-1) + \log t_{\beta_j I_j} \right] \\ \gamma_i(D_j) &= \arg\max_\beta \left[V_{j-1}^\beta(i) + \log t_{\beta_{j-1}D_j} \right] \end{aligned} \tag{7}$$

where β can be any of the insert (I), delete (D) or cyclic states (X). The most probable state path π^* ends in the state $\pi_L^* = \arg\max_j \{V_j^X(N), V_j^I(N), V_j^D(N)\}$ and is given by simply tracing back the pointers:

$$\pi_{i-1}^* = \gamma_I(\pi_i^*), \quad \text{for } i = L, \ldots, 2. \tag{8}$$

3.3. Learning the probabilistic source

The knowledge of the underlying probabilistic source is crucial to finding the optimal state path that generated a given sequence. However, the probabilistic source itself needs to be learnt from the sequence. And often times with insertions and deletions, as shown in Figure 1, $\mathbf{Q}_{\text{ML}}^{(k)}$ is a rather poor estimate of the true cyclostationary source. This is due to the mismatch of phase between successive periods. The estimate can be improved using an adaptive approach that iteratively reconstructs the insertions and deletions. The goal is to first introduce gaps in the given sequence at locations that possibly correspond to insertions and deletions in the sequence. An optimal gapping would space out symbols in the sequence such that the total entropy of the cyclostationary source is minimized or equivalently the mutual information between repeats is maximized (also see Figure 1). Given the ML estimate of cyclostationary source from the gapped sequence, the Viterbi algorithm for HMM profile alignment reconstructs both insertions and deletions.

Since the actual locations where insertions or deletions took place are hidden, these have to be recovered from the search space of all possible insertions and deletions. For a sequence of length n and an insertion and deletion rate of p symbols per base, the search space is of the order $\mathcal{O}\left(\binom{n}{np}^2\right)$. Clearly, the search space becomes un-manageable for long sequences. To address this computational problem, we propose a Gibbs sampling based approach that is much more efficient and has a search space that grows linearly in n. At each location in the sequence, we compute the relative decrease in entropy of the ML estimate of the cyclostationary source after introduction of $k' \leq k$ gaps at that location. Recall that k is the hypothesized period and any occurrence of k consecutive gaps can be pruned. The relative gains are normalized to give a $(N+1)\times(k-1)$ dimensional probability mass function over the cartesian product of sequence indices and the counts of gaps. Sampling from this distribution gives the cartesian pair (i, j) and the sequence is updated by introducing j gaps at location i. Note that Gibbs sampling reconstructs the gaps at one base location at a time. This is equivalent to taking small steps in the search space, where the size and the direction of the step are sampled from a distribution. Also, it should be emphasized that sampling the maximum of the distribution may lead to the search algorithm getting stuck at a local maxima.

The algorithm for Gibbs sampling is described below, with input being the symbolic sequence $\mathbf{s} = \mathbf{s}_1, \ldots, \mathbf{s}_N$ and period k. The output is the probabilistic source for optimally gapped symbolic sequence.

Algorithm: def perGAP($\mathbf{s}, k$)
Do
  $\mathbf{Q}_{\text{ML}}^{(k)}$ = cyclo($\mathbf{s}, k$)
  Compute likelihood L_1 of observing $\mathbf{s}$ given $\mathbf{Q}_{\text{ML}}^{(k)}$
  For each position i in $\mathbf{s}$:
    For j from 1 to $k-1$:
      Insert j gaps in $\mathbf{s}$ at position i to generate $\mathbf{s}_2$
      $\mathbf{Q}_{\text{ML}}^{(k)}$ = cyclo($\mathbf{s}_2, k$)
      Find likelihood L_2 of observing $\mathbf{s}_2$ given $\mathbf{Q}_{\text{ML}}^{(k)}$
      Calculate the likelihood $a_{ji} = L_2/L_1$
  Normalize the weights a to get a distribution.
  Sample (j_s, i_s) from the distribution
  Update $\mathbf{s}$ by introducing j_s gaps at location i_s
Until convergence or max iterations.
Return cyclo($\mathbf{s}, k$).

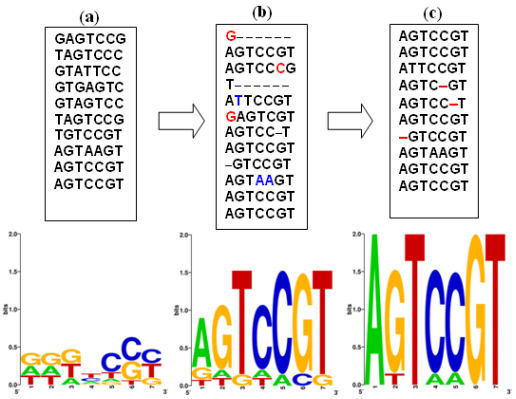

Figure 3. (a) Tandem repeats of AGTCCGT with random insertions, deletions and substitutions (b) intermediate gapped sequence with possible insertions marked red and substitutions marked blue, (c) final sequence reconstructed by estimating the optimal state path using Viterbi algorithm. It is clear from the corresponding WebLogos that repeat pattern is recovered successfully.

The routine perGAP(s,k) estimates the cyclostationary source $\mathbf{Q}^{(k)}_{\text{ML,GAP}}$ for the gapped sequence. This also minimizes the sum-total entropy of the probabilistic source. The gapped sequence in Figure 1 was obtained using the routine perGAP.

4. EXPERIMENTAL RESULTS

4.1. Simulated data

This section discusses experimental results with simulated DNA sequences. Symbolic sequences with tandem repeats were simulated with independent insertions, deletions and substitutions at each base location. Each evolutionary event involves one character at a time. Figure 3 shows results with the ancestral sequence comprising tandem repeats AGTCCGT. The insertion, deletion and substitution rates are 0.05, resulting in a total erosion of 15%. The emission probability of the insert state was chosen equal for each symbol ($e_I(a) = 0.25$ for $a \in \mathcal{A}$).

The performance of the proposed algorithm at identification of latent periods is studied for severe insertion and deletion rates. Figure 4 plots the description length (see Section 2.2), averaged over 10 simulations (for each parameter setting), of the estimated cyclostationary source against the hypothesized periods for various insertion and deletion rates. The original cyclostationary source in the simulations corresponds to the tandem repeats of ATGACT. The period of the cyclostationary source that best fits the sequence, after reconstruction of likely insertions and deletions, matches the true period. The sequences were 100 symbols long and average substitution rate was 10%.

4.2. Genomic sequences

The proposed method was applied on chromosome 1 of human genome in a sliding window size of 300 base pairs with an overlap of 150 base pairs. Various new periods were discovered and are tabulated in the files uploaded at [10]. Latent and tandem repeats were also observed in protein sequences. Some of these sequences are uploaded in FASTA format at [10].

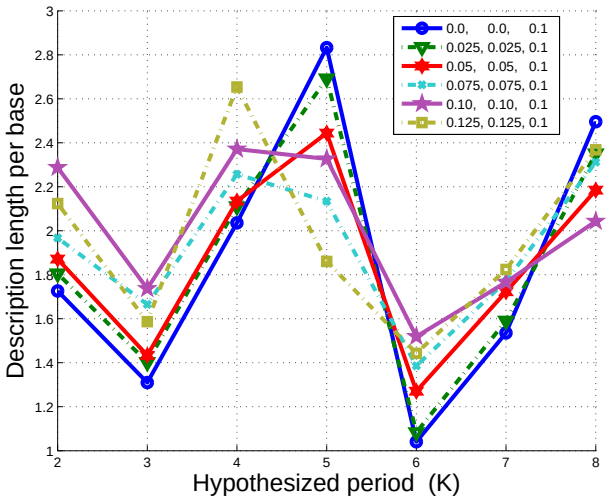

Figure 4. The description length per base is minimized when hypothesized period equals the true period. The curves correspond to various insertion, deletion and substitution rates.

5. CONCLUSIONS

This paper extends the cyclic Markov model proposed in [7] for finding latent periods to a profile hidden Markov model that is robust to insertions and deletions (indels) in the genome sequences. Extensive simulations show that the method was effective even with combined insertion and deletion rate of one in every four base locations.

References

[1] E. V. Korotkov and N. Kudryaschov, "Latent periodicity of many genes," *Genome Informatics*, 2001.

[2] HDCRG, "A novel gene containing a trinucleotide repeat that is expanded and unstable on huntington's disease chromosomes," *Cell*, vol. 72, 1993.

[3] C. M. Hearne, S. Ghosh, and J. A. Todd, "Microsatellites for linkage analysis of genetic traits," *Trends in Genetics*, vol. 8, p. 288, 1992.

[4] S. Tiwari, S. Ramachandran, A. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences," *Comp. App. in Biosciences*, vol. 13, pp. 263–270, 1997.

[5] W. Wang and D. H. Johnson, "Computing linear transforms of symbolic signals," *IEEE Trans. Sig. Proc.*, vol. 50, no. 3, pp. 628–634, March 2002.

[6] M. Akhtar, J. Epps, and E. Ambikairajah, "On DNA numerical representations for period-3 based exon prediction," in *GENSIPS*, 2007.

[7] R. Arora, W. A. Sethares, and J. Bucklew, "Latent periodicities in genome sequences," *IEEE Journal on Sel. Topics in Sig. Proc*, June 2008.

[8] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, 1st ed. Cambridge University Press, 1998.

[9] G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner, "Weblogo: A sequence logo generator," *Genome Res.*, vol. 14, pp. 1188–1190, 2004.

[10] R. Arora. [Online]. Available: http://www.cae.wisc.edu/~sethares/gensips09/gensips09.html