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 nonstationary processes generated from a collection of information sources in a cyclic manner, thus exhibiting cyclostationarity. 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. 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 of 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 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 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 (EM) algorithm is developed to learn latent periods and Viterbi algorithm is employed to learn the most likely path through the state space. The algorithm reconstructs the likely insertions and deletions in the ancestral cyclostationary sequence. Results are provided with the simulated data 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. An n symbol long *cyclostationary* symbolic sequence s with period K is generated by K information sources (or random variables) denoted as $X_1, \ldots, X_K$, in a cyclic fashion. Consequently, the likelihood of observing a sequence s can be expressed solely in terms of the emission probabilities of the states. The emission probabilities of X_i are described by a 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 $P^{(n)}$.

The *dominant period* of a K-periodic cyclostationary sequence is defined to be the substring of consensus bases in a period. It is described by the symbolic sequence $\mathrm{s}^* = \mathrm{s}_1^* \ldots \mathrm{s}_K^*$ of length K such that the i^{th} symbol in every period is more likely to be s_i^* than any other symbol from the alphabet. Mathematically, $\mathrm{s}_i^* = \mathcal{A}_{i^*}$ where $i^* = \arg\max_{1 \le j \le |\mathcal{A}|} P_i(j)$.

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 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 afte division of x by y. Then for $1 \le i \le N$, the symbo $\mathbf{s}_i$ is generated by the random variable $X_{\lfloor i \rfloor_k}$. The searc space for k is the set $\mathcal{K} = \{1, \ldots, N_0\}$, for some $N_0 < I$ and for corresponding probabilistic source $\mathbf{Q}^{(k)}$ the searc space is the subset $\mathcal{Q}^{(k)} \subseteq [0,1]^{|\mathcal{A}| \times k}$ of column stochas tic matrices.

2.1. The Maximum Likelihood Estimate

The maximum likelihood estimate of the cyclostationar source is the column-stochastic matrix given by the optimization problem

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

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

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

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 hypothesis $\mathcal{Q}$, 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

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

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

$$K_{\text{MDL}} = \underset{k \in \mathcal{K}}{\text{argmin}}\, 2\lceil \log k \rceil + k|\mathcal{A}| \log \lceil \frac{N}{k} \rceil - \log P(\mathbf{s}|\mathbf{Q}^{(k)}_{\text{ML}}) \qquad (4)$$

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 estimator degrades severely. Consider the sequence

ACT GCT CT ACT ACGAT ACT ACT ACT (5)

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

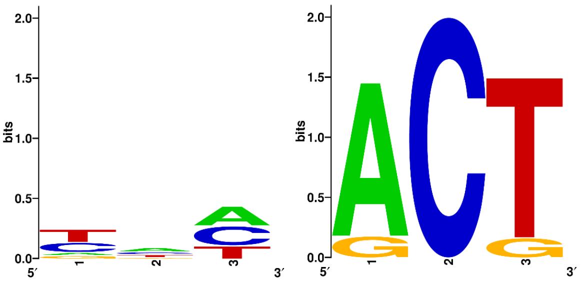

Figure 1. Weblogo depicting mutual information between repeats of 3-periodic DNA sequence (a) ACTGCTCTACTACGATACTACTACT (b) gapped sequence ACTGCT-CTACTACGA-TACTACTACT.

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

The ML probabilistic source with period 3 obtained from the algorithm above is

$$\mathbf{Q}^{(3)}_{\text{ML}} = \begin{bmatrix} 1/9 & 3/8 & 3/8 \\ 1/9 & 2/8 & 0 \\ 3/9 & 2/8 & 3/8 \\ 4/9 & 2/8 & 2/8 \end{bmatrix}. \qquad (6)$$

The correspondence between different periods is depicted by plotting the weblogo [9] which captures the mutual information at each location of the period. Figure 1(a) shows the weblogo for the sequence above. We develop a method which optimally gaps the DNA sequence to mark the possible insertions and deletions. The desired output is the optimally gapped sequence

ACT GCT -CT ACT ACG A-T ACT ACT ACT. (7)

Figure 1(b) shows the Weblogo [9] for the gapped sequence. The dominant period of the gapped sequence is ACT.

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, $\ldots$) 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:

1. $\tau = P(X_{\lfloor i \rfloor_k} \rightarrow X_{\lfloor i+1 \rfloor_k})$,
2. $\epsilon = P(X_{\lfloor i \rfloor_k} \rightarrow I_{\lfloor i \rfloor_k})$,
3. $\delta = P(X_{\lfloor i \rfloor_k} \rightarrow D_{\lfloor i+1 \rfloor_k})$

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 apriori 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 Section 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-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 $\mathbf{s}_i$ being 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

$$
\begin{aligned}
V_j^X(i) &= \log P(X_{\lfloor j \rfloor_k} = x_i) \\
&\quad + \quad \max \left\{ \begin{array}{l} V_{j-1}^X(i-1) + \log t_{X_{j-1}X_j} \\ V_{j-1}^I(i-1) + \log t_{I_{j-1}X_j} \\ V_{j-1}^D(i-1) + \log t_{D_{j-1}X_j} \end{array} \right. \\
V_j^I(i) &= \log e_I(x_i) + \max \left\{ \begin{array}{l} V_j^X(i-1) + \log t_{X_j I_j} \\ V_j^I(i-1) + \log t_{I_j I_j} \\ V_j^D(i-1) + \log t_{D_j I_j} \end{array} \right. \\
V_j^D(i) &= \max \left\{ \begin{array}{l} V_{j-1}^X(i) + \log t_{X_{j-1}D_j} \\ V_{j-1}^I(i) + \log t_{I_{j-1}D_j} \\ V_{j-1}^D(i) + \log t_{D_{j-1}D_j} \end{array} \right. \qquad (8)
\end{aligned}
$$

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:

$$
\begin{aligned}
\gamma_i(X_j) &= \arg\max_{\beta} \left[V_{j-1}^{\beta}(i-1) + \log t_{\beta_{j-1}X_j} \right] \\
\gamma_i(I_j) &= \arg\max_{\beta} \left[V_j^{\beta}(i-1) + \log t_{\beta_j I_j} \right] \\
\gamma_i(D_j) &= \arg\max_{\beta} \left[V_{j-1}^{\beta}(i) + \log t_{\beta_{j-1}D_j} \right] \qquad (9)
\end{aligned}
$$

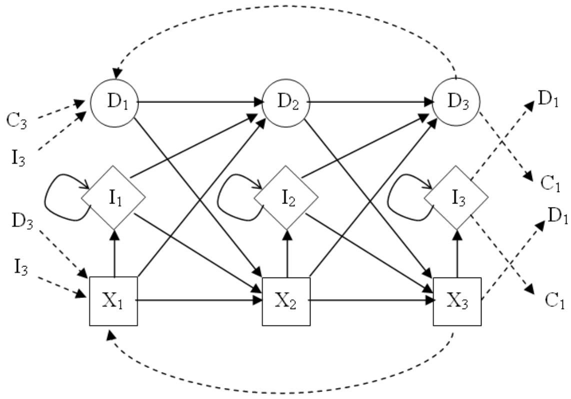

Figure 2. Profile HMM for cyclostationary probabilistic source with period 3.

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:

$$\pi_{i-1}^* = \gamma_I(\pi_i^*), \quad \text{for } i = L, \ldots, 2. \qquad (10)$$

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. Often, the probabilistic source itself needs to be learnt from the sequence and as discussed in Figure 1, with insertions and deletions, $\mathbf{Q}_{\text{ML}}^{(k)}$ is a rather poor estimate. An adaptive approach that iteratively predicts the insertions and deletions may result in better estimates of the cyclostationary source. The goal is to introduce gaps in the given sequence at locations that possibly correspond to deletions in the sequence. Since the actual locations where deletions took place are hidden, Gibbs sampling is proposed to recover those positions.

The algorithm 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 s given $\mathbf{Q}_{\text{ML}}^{(k)}$
    For each position $i$ in s:
        For $j$ from 1 to $k-1$:
            Insert $j$ gaps in 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 s by introducing $j_s$ gaps at location $i_s$
Until convergence or max iterations.
Return cyclo($\mathbf{s}, k$).
```

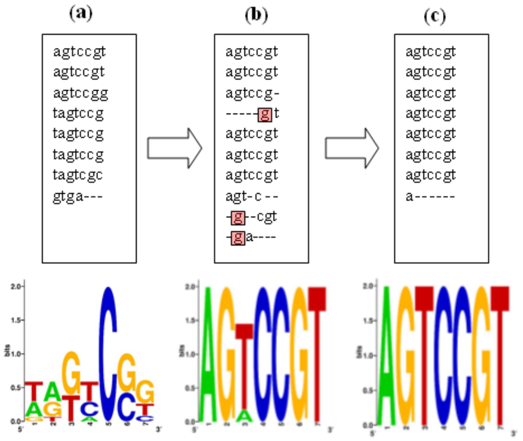

Figure 3. (a) Tandem repeats of AGTCCGT with random insertions and deletions, (b) gapped sequence with '-' denoting likely deletion, (c) sequence reconstructed by estimating the optimal state path using Viterbi algorithm. It is clear from the WebLogos corresponding to sequences in (c) that the tandem repeat pattern was successfully recovered. Going from sequence (b) to (c), the boxed residues in (b) were likely insertions.

The routine perGAP($\mathbf{s}$,k) estimates the cyclostationary source $\mathbf{Q}^{(k)}_{\text{ML,GAP}}$ for the gapped sequence. The likelihood of the gapped sequence is computed in the same manner as for the ungapped sequence but with gaps replaced by symbols that maximize the overall likelihood of the 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 and at each position in the sequence the residue was deleted with probability $\delta_0 = 0.1$ or a new residue was inserted with probability $\epsilon_0 = 0.1$. The emission probability of the insert state was chosen equal for each symbol ($e_I(a) = 0.25$ for $a \in \mathcal{A}$). The maximum likelihood estimate for the probabilistic source was obtained by constructing the gapped sequence using the Gibbs sampling based method and Viterbi's algorithm was employed to ascertain the optimal state path. Figure 3 shows results with the ancestral sequence comprising of tandem repeats AGTCCGT.

The performance of the proposed algorithm at identification of latent periods is studied for severe insertion and deletion rates. Figure 4 plots the average entropy of the estimated cyclostationary source versus the hypothesized periods. 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 average insertion and deletion rate in simulations was chosen to be 1 in every 10 bases.

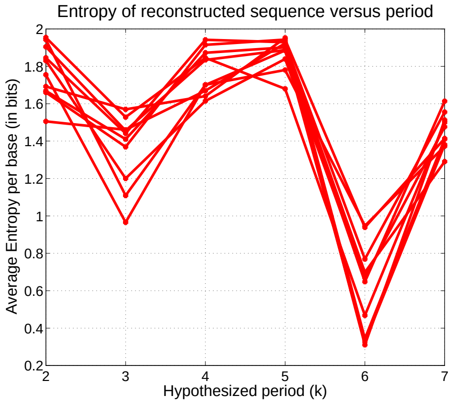

Figure 4. The entropy-per-base is minimized when hypothesized period equals the true period.

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].

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

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

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

[8] Durbin et. al., *Biological Sequence Analysis*, 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] Raman Arora's Homepage, ," Files available at `http://www.cae.wisc.edu/~raman/cgw/`.