Mean-Field Analysis of Recursive Entropic Segmentation of Biological Sequences

SIEW-ANN CHEONG\textsuperscript{1}, PAUL STODGHILL\textsuperscript{2}, DAVID SCHNEIDER\textsuperscript{2}, CHRISTOPHER MYERS\textsuperscript{1}

\textsuperscript{1}Cornell Theory Center, Cornell University
\textsuperscript{2}USDA/ARS, Ithaca

APS March Meeting, March 8, 2007
Denver, Colorado
Research funded by USDA/ARS
Mosaic Nature of Biological Sequences

Mosaic Nature of Biological Sequences

The Jensen-Shannon Divergence

- Given length-$N$ sequence $x = x_1x_2 \cdots x_N$, $x_i = A, C, G, T$, assume composed of $M \geq 1$ statistically distinct Bernoulli segments with domain walls at $i_1, \ldots, i_{M-1}$. Determine $M$-segment sequence likelihood

$$P_M(x; i_1, \ldots, i_{M-1}; \hat{p}_1, \ldots, \hat{p}_M) = \prod_{m=1}^{M} \prod_{s=A,C,G,T} (\hat{p}_m^s)^{f_m^s} ; \quad \hat{p}_s^m = \frac{f_m^s}{\sum_{s'} f_{s'}^m}.$$


$$\Delta_M = \log \frac{P_M}{P_1} = - \sum_{s} f_s \log \hat{p}_s + \sum_{m} \sum_{s} f_m^s \log \hat{p}_s^m$$

is symmetric relative entropy providing quantitative measure of ‘goodness-of-fit’ of $M$-segment model over 1-segment model.

- Straightforward to generalize $\Delta_M$ to Markov chains of order $K > 0$. Markov chains model dinucleotide frequencies and codon biases in real genomic sequences better than Bernoulli chains with extended alphabets.
Recursive Jensen-Shannon Segmentation

• STEP 1 (Segmentation):

  – Given sequence \( x = x_1 x_2 \cdots x_N \), compute 2-segment Jensen-Shannon divergence \( \Delta_2(i) \) as function of cursor position \( i \).

  – Find \( i^* \) such that \( \Delta_2(i^*) = \max_i \Delta_2(i) \). The best 2-segment model for \( x \) is \( x = x_L x_R \), where \( x_L = x_1 \cdots x_{i^*} \) and \( x_R = x_{i^*+1} \cdots x_N \).

• STEP 2 (Recursion): Repeat STEP 1 for \( x_L \) and \( x_R \).

• STEP 3 (Termination): 1-segment model selected over 2-segment model if:


  – Model Selection: information criterion (e.g. AIC, BIC) for 2-segment model greater than that for 1-segment model. [Li, Phys. Rev. Lett. 86, 5815 (2001).]
Mean-Field Analysis of Recursive Segmentation

- Distribute sequence statistics uniformly along length. Ignore intra-sequence variations in statistics, i.e. mean-field picture.

Discrete sequence positions, integer counts

Continuous sequence positions, real counts

- Analyze recursive segmentation scheme entirely within mean-field picture:
  - Peaks in mean-field divergence spectrum appear only at domain walls;
  - Domain walls also appear as kinks, or even have vanishing divergence in mean-field divergence spectrum.
  - Recursive segmentation eventually discovers all domain walls.
Context Sensitivity:

- domain wall strengths change as recursion proceeds
- domain walls not discovered according to true order
- incomplete segmentation pick up weak domain walls, but miss stronger ones
- especially severe for repetitive sequences

Need to move or remove existing cuts for better segmentation
Segmentation Optimization

- Two procedures to optimize domain wall $i_m$:

  - **First-order update:** Compute $\Delta^m_2(i)$ for supersegment $(i_{m-1}, i, i_{m+1})$, and choose $i_m = i^*$, such that $\Delta_2(i^*) = \max_{i_{m-1} < i < i_{m+1}} \Delta_2(i)$, to be new position of domain wall.

  - **Second-order update:** Compute $\Delta^{m-1}_2(i)$ for supersegment $(i_{m-2}, i_{m-1}, i)$ and $\Delta^{m+1}_2(i)$ for supersegment $(i, i_{m+1}, i_{m+2})$, and choose $i_m = i^*$, such that $\Delta^{m-1}_2(i^*)\Delta^{m+1}_2(i^*) = \max_{i_{m-1} < i < i_{m+1}} \Delta^{m-1}_2(i)\Delta^{m+1}_2(i)$, to be new position of domain wall.

- Domain walls $\{i_m\}_{m=1}^M$ updated serially, or in parallel.

- Optimized recursive segmentation: Right after STEP 1 (Segmentation), optimize segmentation using first- or second-order update algorithm.
Conclusions and Further Work

• To conclude, we have:
  – Refined recursive segmentation scheme by generalizing Jensen-Shannon divergence to Markov chains of order $K > 0$.
  – Undertaken mean-field analysis of recursive Jensen-Shannon segmentation, and identified possible pitfalls.
  – Developed algorithm for segmentation optimization.

• Further work, completed or in progress:
  – Developed new termination criterion that requires no prior knowledge how many segments to partition sequence into.
  – Derived better understanding of segment Markov-chain order selection problem, within the framework of recursive segmentation.
  – Incomplete segmentation misleading, cluster terminal segments instead to obtain coarser scale description of genome. E.g. to distinguish lineage-specific regions arising from HGT and the genetic backbone.
  – Multiple sequence clustering for comparative, phylogenetic studies.
Jensen-Shannon Divergence of Markov Chains

- For order-$K$ stationary Markov chain, 1-segment sequence likelihood is
  \[ P_1(x; \hat{P}) = \prod_i \hat{p}(x_i|x_{i-1}x_{i-2}\cdots x_{i-K}) = \prod_t \prod_s (\hat{p}_{ts})^{f_{ts}}, \]
  where \( t = t_{-1}t_{-2}\cdots t_{-K} \), and \( s, t_k = A, C, G, T \), and 2-segment sequence likelihood is
  \[ P_2(x; \hat{P}^L, \hat{P}^R) = \prod_t \prod_s (\hat{p}_ts)^{f_{ts}} (\hat{p}_{ts})^{f_{ts}}. \]

- Generalized 2-segment Jensen-Shannon divergence
  \[ \Delta_K = \log \frac{P_2}{P_1} = \sum_t \sum_s [-f_{ts} \log p_{ts} + f_{ts}^L \log p_{ts}^L + f_{ts}^R \log p_{ts}^R]. \]

- Bernoulli sequences are order-($K = 0$) stationary Markov chains.