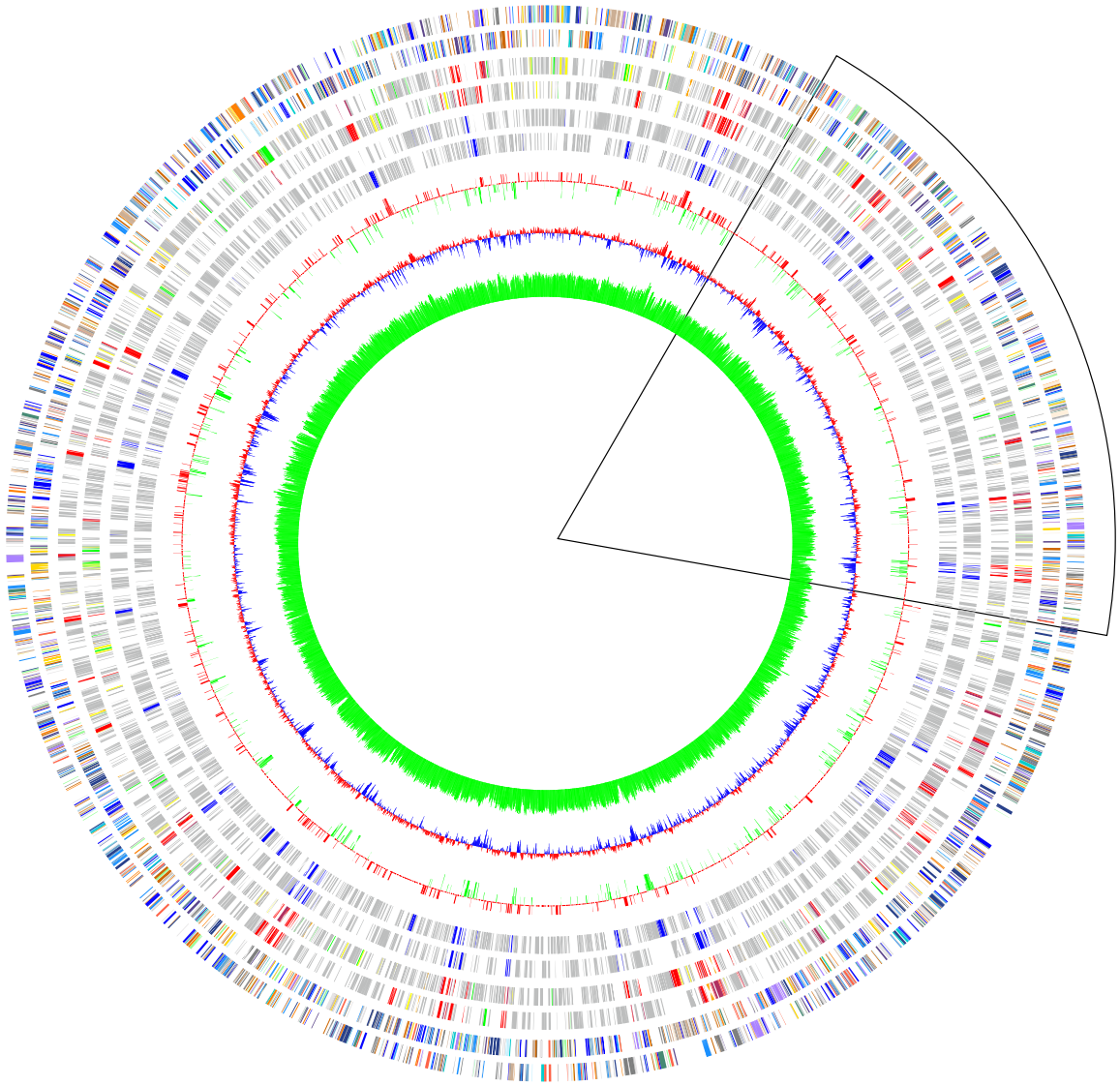
Mean-Field Analysis of Recursive Entropic Segmentation of Biological Sequences

SIEW-ANN CHEONG[1], PAUL STODGHILL[2],
DAVID SCHNEIDER[2], CHRISTOPHER MYERS[1]

[1]Cornell Theory Center, Cornell University
[2]USDA/ARS, Ithaca

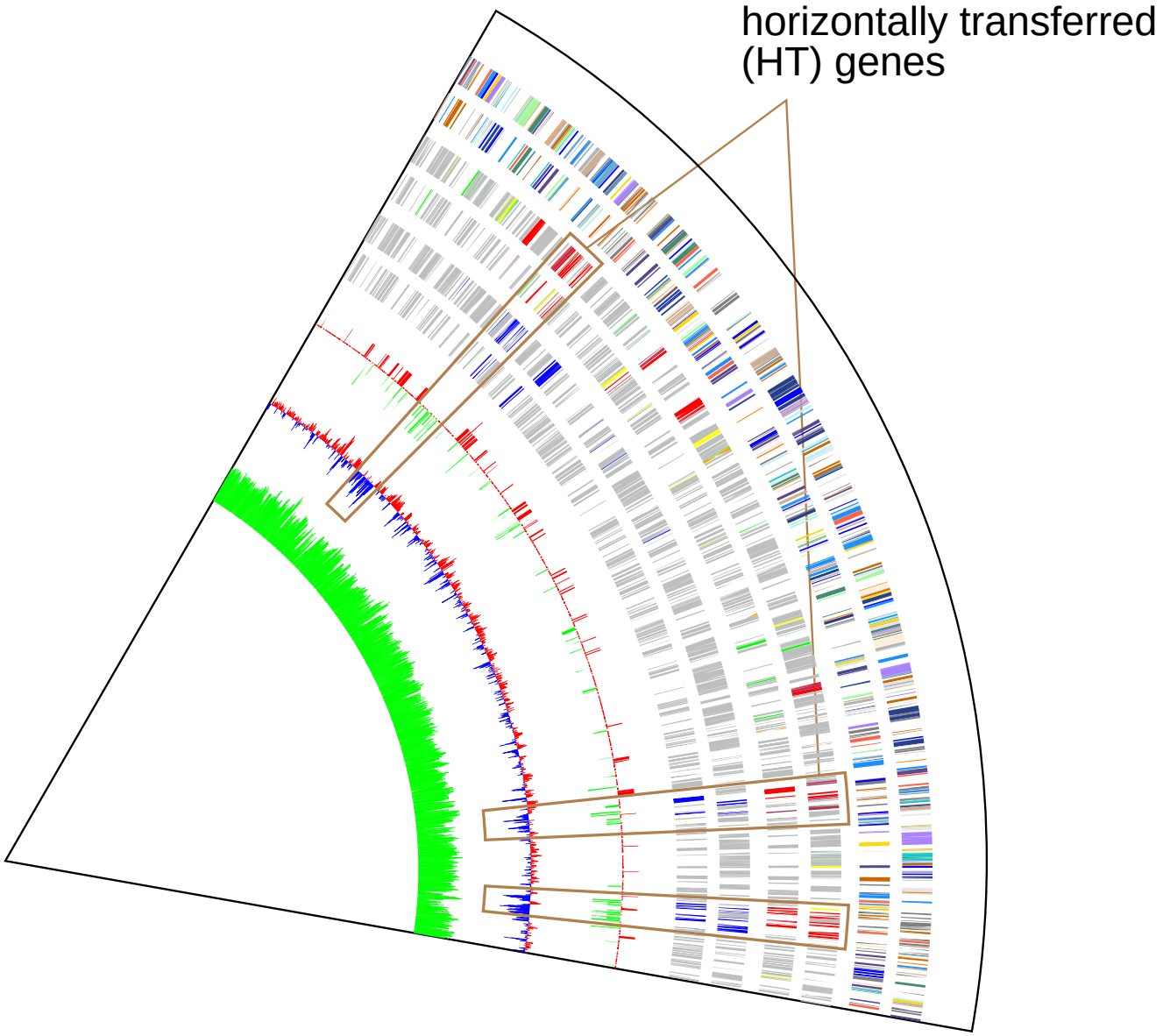
APS March Meeting, March 8, 2007
Denver, Colorado
Research funded by USDA/ARS

Mosaic Nature of Biological Sequences



Reproduced from Ghai, Hain and Chakraborty, *BMC Bioinformatics* **5**, 198 (2004).

Mosaic Nature of Biological Sequences



Reproduced from Ghai, Hain and Chakraborty, *BMC Bioinformatics* **5**, 198 (2004).

The Jensen-Shannon Divergence

- Given length-N sequence $\mathbf{x} = x_1 x_2 \cdots x_N$, $x_i = A, C, G, T$, 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

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

- Jensen-Shannon divergence [Lin, IEEE Trans. Infor. Theor. **37**, 145 (1991)]

$$\Delta_M = \log \frac{P_M}{P_1} = -\sum_s f_s \log \hat{p}_s + \sum_m \sum_s f_s^m \log \hat{p}_s^m$$

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

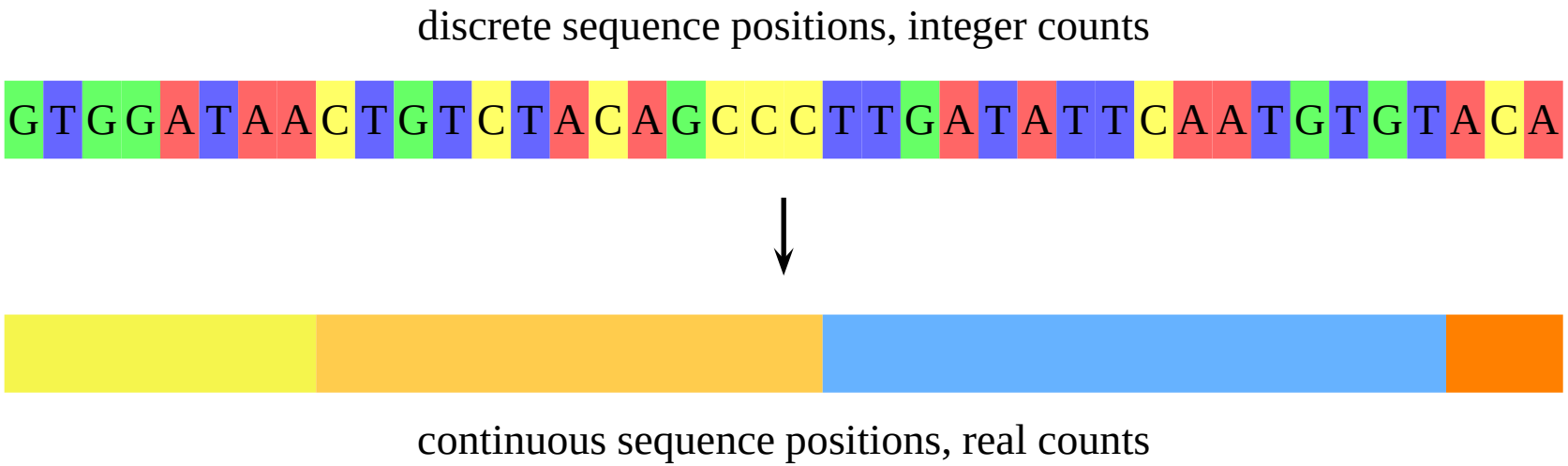
- Straightforward to generalize Δ_M to Markov chains of order $K > 0$. Markov chains model dinucleotide frequencies and codon biases in real genomic sequences better than Bernoulli chains with extended alphabets.

Recursive Jensen-Shannon Segmentation

- STEP 1 (Segmentation):
  - Given sequence $\mathbf{x} = x_1 x_2 \cdots x_N$, 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 $\mathbf{x}$ is $\mathbf{x} = \mathbf{x}_L \mathbf{x}_R$, where $\mathbf{x}_L = x_1 \cdots x_{i^*}$ and $\mathbf{x}_R = x_{i^*+1} \cdots x_N$.
- STEP 2 (Recursion): Repeat STEP 1 for $\mathbf{x}_L$ and $\mathbf{x}_R$.
- STEP 3 (Termination): 1-segment model selected over 2-segment model if:
  - Hypothesis Testing: probability of obtaining divergence beyond than observed Δ_2 greater than prescribed tolerance ϵ; [Bernaola-Galván *et al*, Phys. Rev. E **53**, 5181 (1996); Román-Roldán *et al*, Phys. Rev. Lett. **80**, 1344 (1998).]
  - 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).]

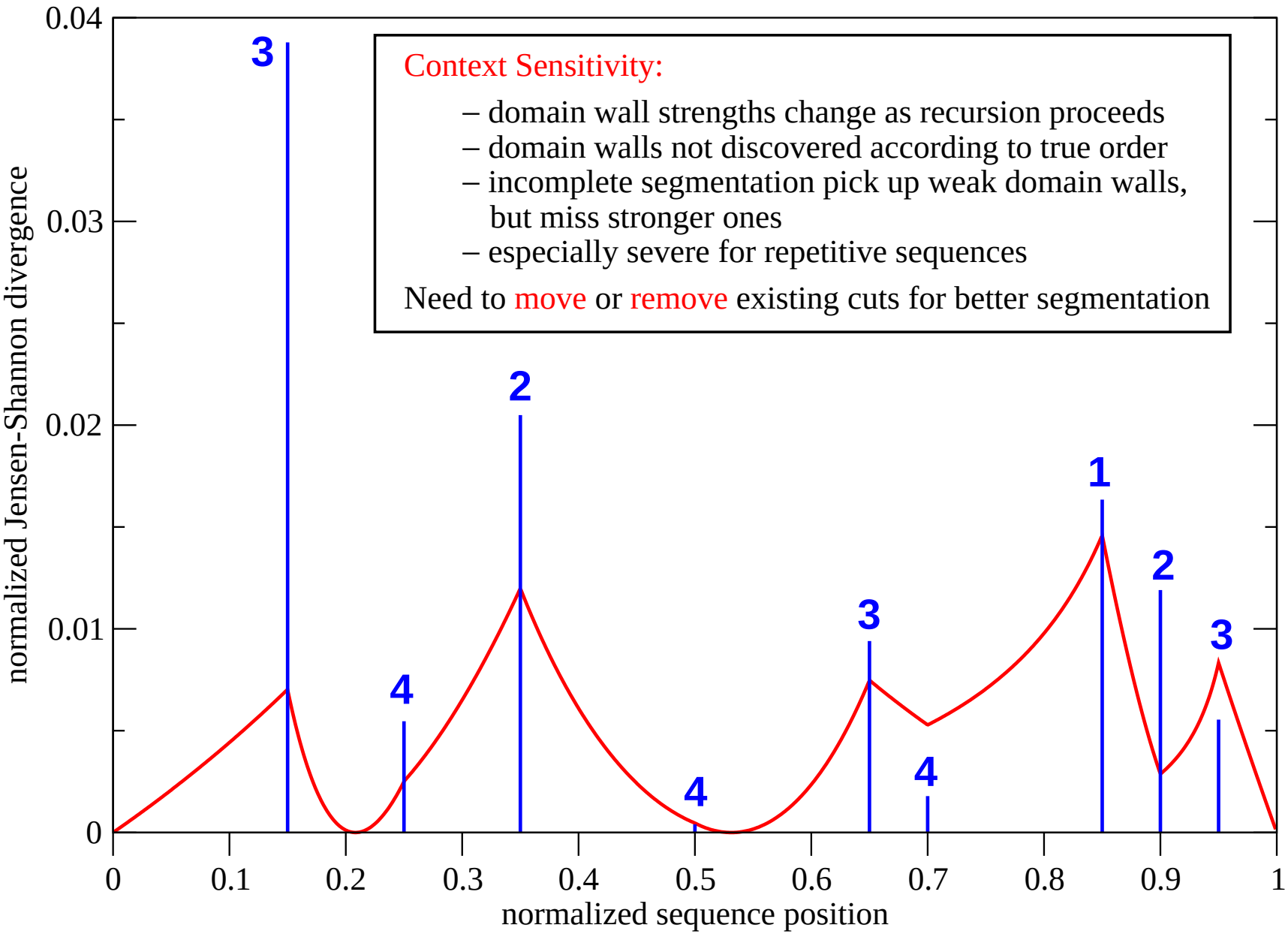
Mean-Field Analysis of Recursive Segmentation

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



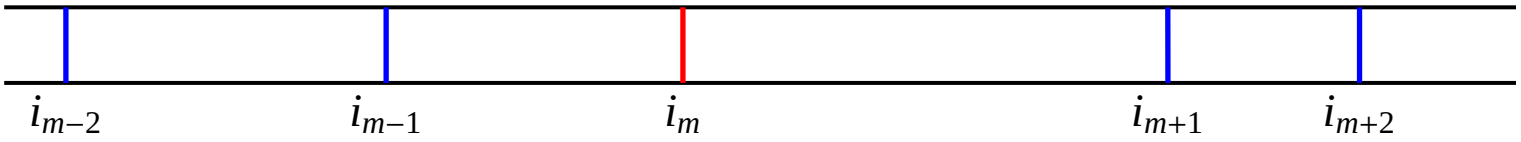
- Analyze recursive segmentation scheme entirely within mean-field picture:
  - Peaks in mean-field divergence spectrum appear only at domain walls;
  - Domain walls also appear as kinks, or even have vanishing divergence in mean-field divergence spectrum.
  - Recursive segmentation eventually discovers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

Pitfalls of Recursive Segmentation



Segmentation Optimization

- Two procedures to optimize domain wall i_m:



  - First-order update: Compute $\Delta_2^m(i)$ 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 $\Delta_2^{m-1}(i)$ for supersegment (i_{m-2}, i_{m-1}, i) and $\Delta_2^{m+1}(i)$ for supersegment (i, i_{m+1}, i_{m+2}), and choose $i_m = i^*$, such that $\Delta_2^{m-1}(i^*)\Delta_2^{m+1}(i^*) = \max_{i_{m-1}<i<i_{m+1}} \Delta_2^{m-1}(i)\Delta_2^{m+1}(i)$, 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

$$P_1(\mathbf{x};\hat{\mathbb{P}}) = \prod_i \hat{p}(x_i|x_{i-1}x_{i-2}\cdots x_{i-K}) = \prod_{\mathbf{t}}\prod_s (\hat{p}_{\mathbf{t}s})^{f_{\mathbf{t}s}}\,,$$

  where $\mathbf{t} = t_{-1}t_{-2}\cdots t_{-K}$, and $s, t_k = A, C, G, T$, and 2-segment sequence likelihood is

$$P_2(\mathbf{x};\hat{\mathbb{P}}^L,\hat{\mathbb{P}}^R) = \prod_{\mathbf{t}}\prod_s \left(\hat{p}^L_{\mathbf{t}s}\right)^{f^L_{\mathbf{t}s}} \left(\hat{p}^R_{\mathbf{t}s}\right)^{f^R_{\mathbf{t}s}}.$$

- Generalized 2-segment Jensen-Shannon divergence

$$\Delta_K = \log\frac{P_2}{P_1} = \sum_{\mathbf{t}}\sum_s \left[-f_{\mathbf{t}s}\log p_{\mathbf{t}s} + f^L_{\mathbf{t}s}\log p^L_{\mathbf{t}s} + f^R_{\mathbf{t}s}\log p^R_{\mathbf{t}s}\right].$$

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