

Problem: Which regions of a DNA string code for proteins?

Method: Change point analysis

Change Point Analysis: When does a string of random variables that samples symbols from probability distribution ϕ change to a probability distribution ψ ?

AAAAA|GGGAG

AAAAA

$$P_{\phi}(G) = 0$$

$$P_{\phi}(A) = 1$$

GGGAG

$$P_{\psi}(G) = .8$$

$$P_{\psi}(A) = .2$$

How can we detect the nonstationarity?

Components: A string $S = s_1, \dots, s_h, \dots, s_N$ has disjoint substrings $S^{(1)}, S^{(2)}, \dots, S^{(j)}, \dots, S^{(\ell)}$ where $\sum_{j=1}^{\ell} \text{Len}(S^{(j)}) = N$. Each substring has respective weight $\pi^{(1)}, \pi^{(2)}, \dots, \pi^{(j)}, \dots, \pi^{(\ell)}$. S is constructed from a multivariate random variable X that can take the values x_1, x_2, \dots, x_k with probability

$$P(X = x_i | x_i \in S^{(j)}) = \frac{\# x_i \text{ in segment } S^{(j)}}{\text{Length of } S^{(j)}} = p_i^{(j)}$$

For instance, let $\ell = 3$, $x_1 = A$ and $x_2 = G$. For the segmented string

AAA|AAG|GGAG

$$\begin{array}{lll} p_1^{(1)} = 1 & p_1^{(2)} = 2/3 & p_1^{(3)} = 1/4 \\ p_2^{(1)} = 0 & p_2^{(2)} = 1/3 & p_2^{(3)} = 3/4 \end{array}$$

Let $\mathbf{p}^{(j)} = [p_1^{(j)}, \dots, p_\ell^{(j)}]$ and $p_i = \frac{1}{\ell} \sum_{j=1}^{\ell} p_i^{(j)}$. The Jensen-Shannon divergence is defined as

$$D[\mathbf{p}^{(1)}, \dots, \mathbf{p}^{(\ell)}] = H\left(\sum_{j=1}^{\ell} \pi^{(j)} \mathbf{p}^{(j)}\right) - \sum_{j=1}^{\ell} \pi^{(j)} H(\mathbf{p}^{(j)}) \quad (1)$$

where

$$H(\mathbf{p}^{(j)}) = - \sum_{i=1}^k p_i^{(j)} \log_2 p_i^{(j)} \quad (2)$$

(1) can be rewritten as

$$D = \sum_{j=1}^{\ell} \sum_{i=1}^k p_i^{(j)} \pi^{(j)} \log_2 \frac{p_i^{(j)}}{p_i} \quad (3)$$

In an information theory context, (3) can be considered the mutual information between a symbol and a substring.

Theorem: As $N \rightarrow \infty$, D is related to a χ^2 distribution with $(k - 1)(\ell - 1)$ degrees of freedom.

Sketch of Proof: We use a second order Taylor expansion. The first order term vanishes. A full proof can be found in [5].

$$\begin{aligned}
 D &= \sum_{i,j} p_i^{(j)} \pi^{(j)} \log_2 \frac{p_i^{(j)}}{p_i} \\
 &\stackrel{\text{Taylor}}{\approx} \sum_{i,j} \frac{p_i^{(j)} \pi^{(j)} - p_i \pi^{(j)}}{\ln 2} \sum_{i,j} \frac{\left(p_i^{(j)} \pi^{(j)} - p_i \pi^{(j)}\right)^2}{(2 \ln 2) p_i \pi^{(j)}} \\
 &= \sum_{i,j} \frac{\left(p_i^{(j)} \pi^{(j)} - p_i \pi^{(j)}\right)^2}{(2 \ln 2) p_i \pi^{(j)}}
 \end{aligned}$$

We now take into account a Markov assumption. Using similar notation,

$$P(X = x_{i_2} = s_h | s_{h-1} = x_{i_1}, x_i \in S^{(j)}) = p_{i_2|i_1}^{(j)} \quad (4)$$

For instance,

AAA|AAG|GGAG

$p_{1 1}^{(1)} = 1$	$p_{1 1}^{(2)} = 2/3$	$p_{1 1}^{(3)} = 0$
$p_{1 2}^{(1)} = 0$	$p_{1 2}^{(2)} = 0$	$p_{1 2}^{(3)} = 1/3$
$p_{2 1}^{(1)} = 0$	$p_{2 1}^{(1)} = 1/3$	$p_{2 1}^{(1)} = 1$
$p_{2 2}^{(1)} = 0$	$p_{2 2}^{(1)} = 0$	$p_{2 2}^{(1)} = 2/3$

A first order Markov assumption is demonstrated above. Note that arbitrary finite Markov order can be assumed.

The *Jensen-Shannon divergence of Markov order m* is defined as

$$D^m[\mathbf{p}^{(1)}, \dots, \mathbf{p}^{(\ell)}] = H^m\left(\sum_{j=1}^{\ell} \pi^{(j)} \mathbf{p}^{(j)}\right) - \sum_{j=1}^{\ell} \pi^{(j)} H^m(\mathbf{p}^{(j)}) \quad (5)$$

where

$$H^m(\mathbf{p}^{(j)}) = - \sum_{i_1=1}^{k'} p_{i_1}^{(j)} \sum_{i_2=1}^k p_{i_2|i_1}^{(j)} \log_2 p_{i_2|i_1}^{(j)} \quad (6)$$

and $k' = k^m$. The original definition of the Jensen-Shannon divergence is the special case $m = 0$. A similar proof shows that D^m is related to a χ^2 distribution.

Theorem: Let

$$D_{\max}^m = \max_h \left\{ D^m[\mathbf{p}^{(1)}, \mathbf{p}^{(2)}] : \begin{array}{l} S^{(1)} = [s_1, \dots, s_h] \\ S^{(2)} = [s_{h+1}, \dots, s_N] \end{array} \right\} \quad (7)$$

As $N \rightarrow \infty$, D_{\max}^m is related to a χ^2 distribution with $k^m(k-1)(\ell-1)$ degrees of freedom.

Ingredients of Proof: (full proof will be in the publication [1])

- A second order Taylor expansion.
- Two hypotheses:
 - H_0 The joint distribution $p_{i_2|i_1}^{(j)}$ is separable into $p_{i_1}^{(j)} p_{i_2|i_1}$.
 - H_A The joint distribution is not separable.
- The $-2 \log \lambda$ theorem [9].

Claim: D_{\max}^m can be used to detect protein coding regions in DNA.

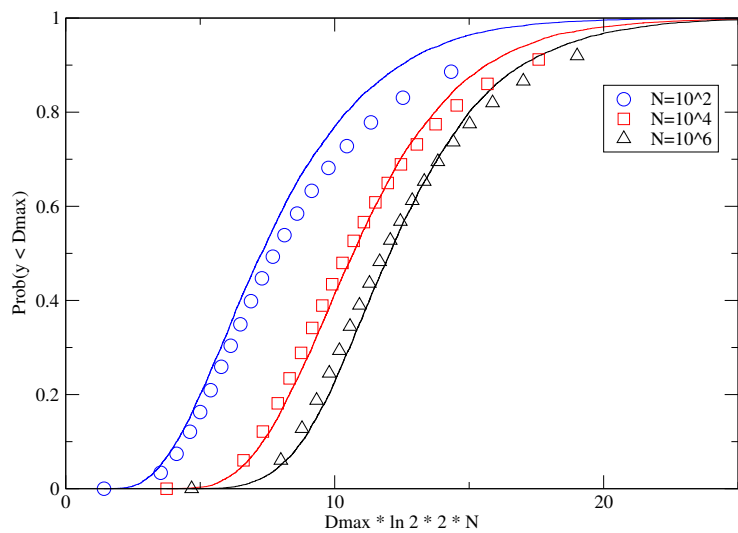
This claim is based on the following:

1. In coding regions, certain codons occur more frequently. In non-coding regions, codon usage is more uniform[4].
2. D_{\max}^m can be used to recursively segment a DNA string according to the inter-heterogeneity of two substrings.

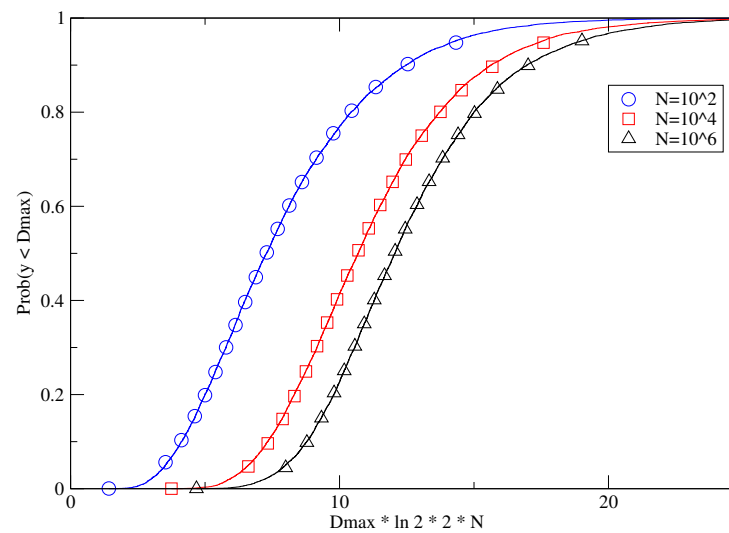
To make this claim a reality, we need power curves for D_{\max}^m .

Theoretical power curves with many assumptions were derived in [7]; however, we turn to simulation and curve fitting to derive more accurate results.

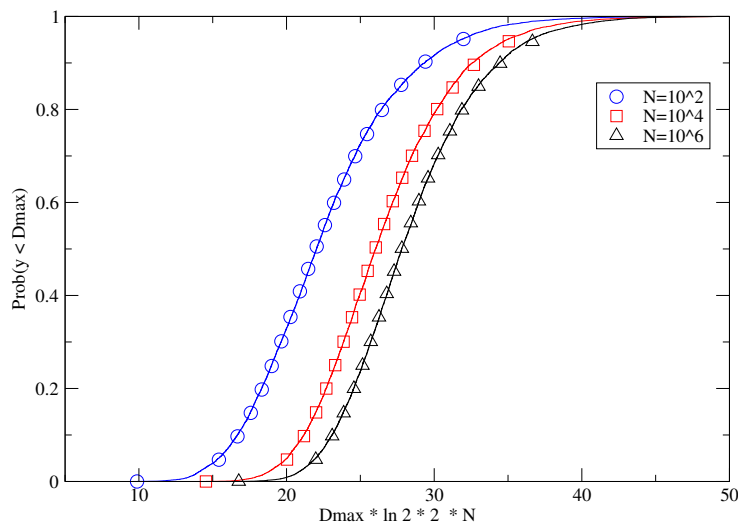
Theoretical Power Curves Without Markov Assumption



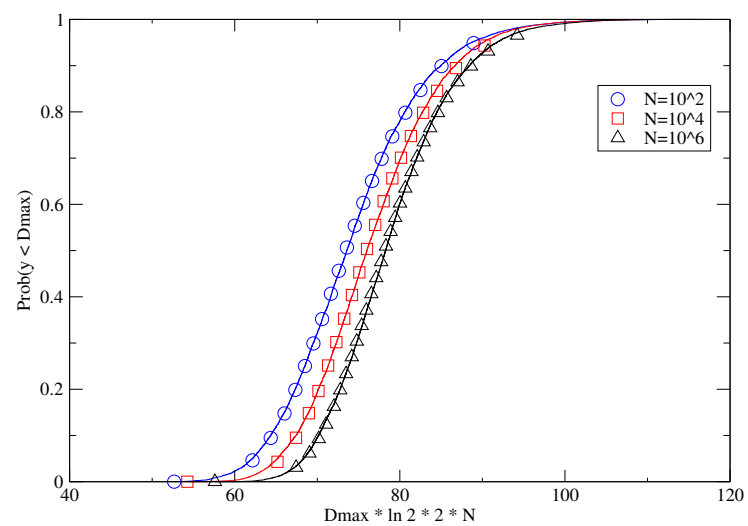
Power Curves Without Markov Assumption



Power Curves With First Order Assumption

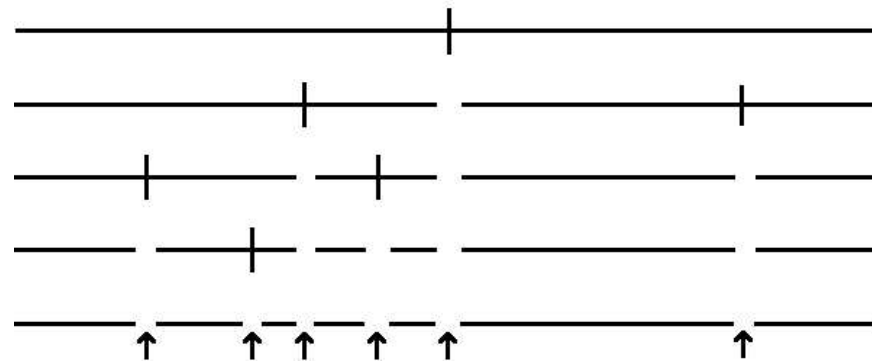


Power Curves With Second Order Assumption



The segmentation algorithm is as follows:

1. Obtain D_{\max}^m for the $N - 1$ possible $\ell = 2$ splits on the string
2. Based on a specified confidence level and N , determine what would be a sufficiently large D_{\max}^m
3. If D_{\max}^m is sufficient, recurse on both substrings; if D_{\max}^m is not sufficient, halt.



This segmentation would indicate 6 coding borders with four levels of recursion.

Results: Our extension of the Jensen-Shannon divergence finds approximately 40% more true coding borders than the previous $m = 0$ method on several test DNA regions. False positives are approximately equal to the $m = 0$ method.

Conclusions: D_{\max}^m is related to a χ^2 distribution. Based on results, higher order modeling better reflects the composition of DNA.

Future Work: Several other problems, e.g. detecting complex repeats in telomeres, have been partially resolved by the $m = 0$ method. Higher order methods are expected to perform better.

References

- [1] A. Arvey, A. Raval, and R. Azad. Higher order Jensen-Shannon divergence applied to symbolic sequences. *In Preparation*, 2006.
- [2] R. K. Azad, P. Bernaola-Galván, R. Ramaswamy, and J. Subba Rao. Segmentation of genomic DNA through entropic divergence: Power laws and scaling. *Physical Review E*, 65:051909–1–6, May 2002.
- [3] P. Bernaola-Galván, I. Grosse, P. Carpena, J. L. Oliver, R. Román, and H. Eugene Stanley. Findings borders between coding and noncoding DNA regions by an entropic segmentation method. *Physical Review Letters*, 85:1342–1345, August 2000.
- [4] R. Grantham, C. Gautier, M. Gouy, M. Jacobzone, and R. Mercier. Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res.*, 9:r43–r74, January 1981.
- [5] I. Grosse, P. Bernaola-Galván, P. Carpena, Ramón Román-Roldán, J. Oliver, and H. E. Stanley. Analysis of symbolic sequences using the Jensen-Shannon divergence. *Physical Review E*, 65:1–15, March 2002.
- [6] L. Horváth. The limit distributions of likelihood ratio and cumulative sum tests for a change in a binomial probability. *Multivariate Analysis*, 31:148–159, October 1989.
- [7] L. Horváth and M. Csorgo. *Limit Theorems in Change Point Analysis*. Probability and Statistics. Wiley, 2002.
- [8] W. Li, P. Bernaola-Galván, F. Haghghi, and I. Grosse. Applications of recursive segmentation to the analysis of DNA sequences. *Computers and Chemistry*, 26:491–510, 2002.
- [9] S.S. Wilks. *Mathematical Statistics*. Wiley, New York, 1962.