

Significance Of Inter-Species Matches When Evolutionary Rate Varies

Jia Li

Department of Statistics

Department of Computer Science and Engineering

Collaborator: Webb Miller

Department of Biochemistry and Molecular Biology

Department of Computer Science and Engineering

The Pennsylvania State University

Background

- Align DNA sequences of two species

```
human AAAATTGGTACATAAA
mouse . . .AGGGG--CATAAA
      UUUMNNMMGGMMMMMM
```

- A genomic interval that is highly conserved between the two species can be considered as a candidate for encoding a protein or regulating gene transcription.
- Assess the significance of matched segments by Karlin and Altschul's method:
 - Assign each possible symbol $\gamma \in \{U, M, N, G\}$ a score Z , e.g., $Z_M = 1$, $Z_N = -1$, $Z_U = Z_G = -L$, where $L \gg 1$.
 - The score of a segment of symbols is defined as the sum of the scores of all the positions in the segment.
 - Given a sequence of UMNG modeled by an i.i.d. stochastic process, compute the probability of a sequence generated randomly by the source having its maximal segment score exceeding that of the given segment.

Focused Issues

- Focused issues:
 - Identify different “modes” of background.
 - In a fixed mode of background, model the alignment sequence by a Markov chain.
- Methodology:
 - Develop an extended version of the hidden Markov model (HMM).
 - Apply Karlin and Dembo’s theorem.

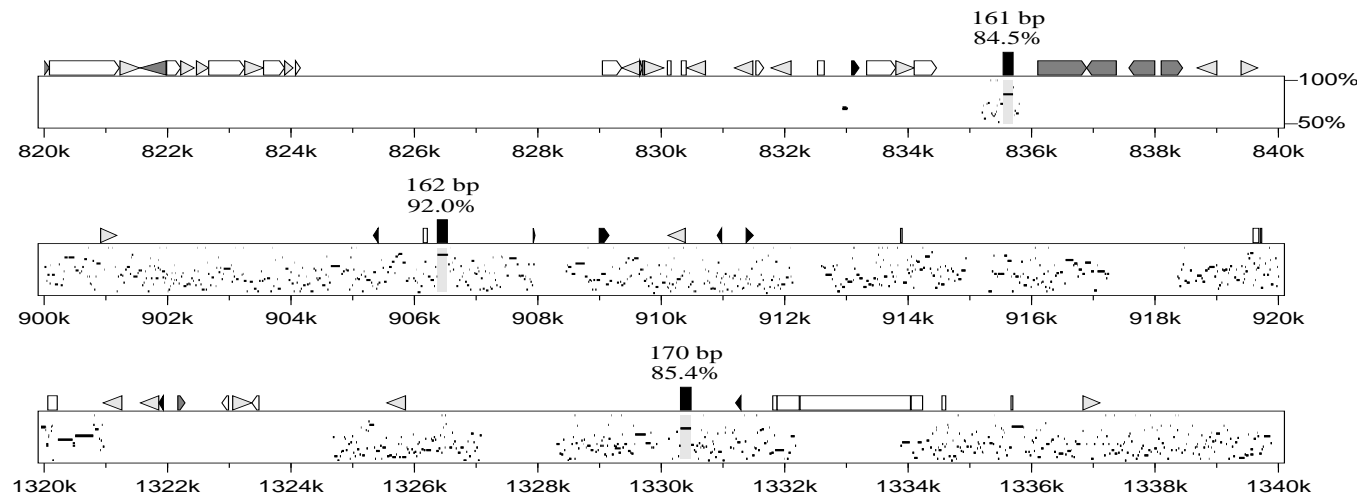
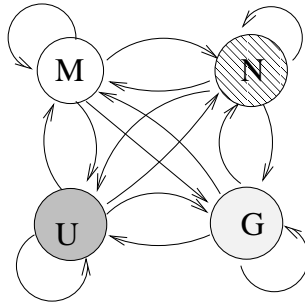


Figure 1: Percent identity plot (PIP) of some human-mouse alignments. Triangles and other icons along the top indicate positions of interspersed repeats and low-complexity regions found in the human sequence by the RepeatMasker program. Each tiny horizontal line in the PIP indicates the human positions and percent nucleotide identity of an interval between consecutive gaps in a local alignment with the mouse genomic sequence.

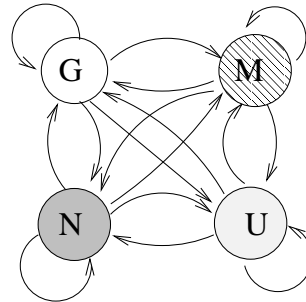
Modeling by HMM

UUUUUUUMNNUUUUUUMMMMMNNMMGMUUUUUUMMGNMUUUUUUUMMUUUUU

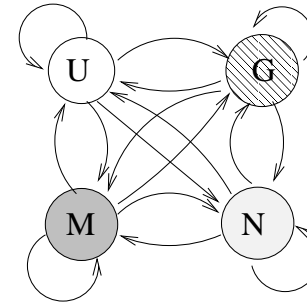
Mode 1



Mode 2

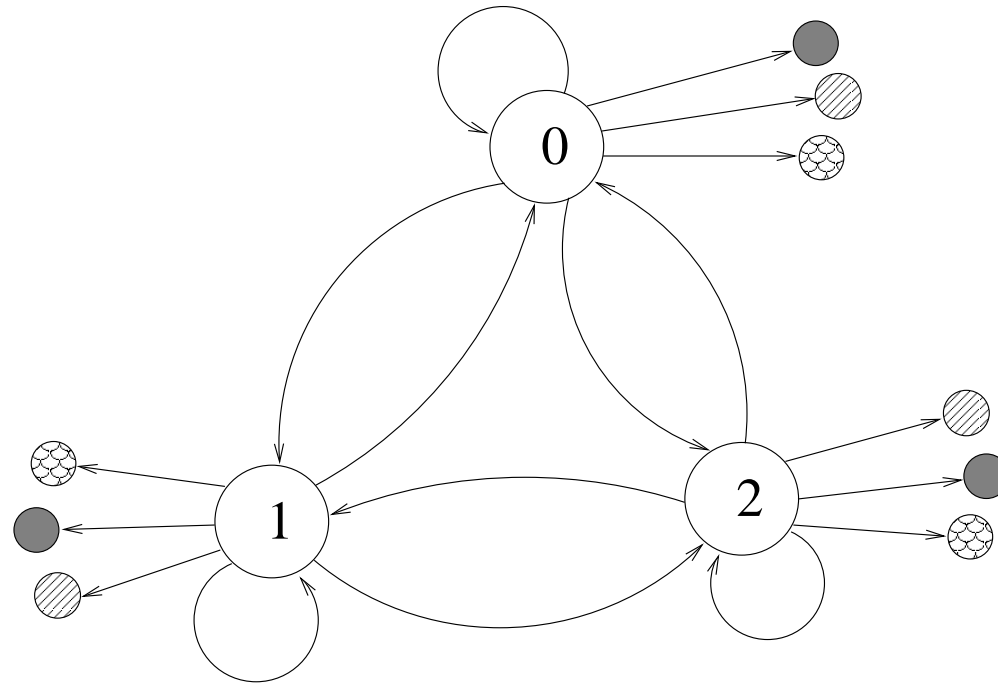


Mode 3



- The modes (states) of background are unknown and have to be extracted based on alignment data.
- The background modes are viewed as the hidden states in an HMM.
- To estimate the HMM, M, N, G are all treated as 1, and U as 0.
- Model the sequence of UMNG by a different Markov chain within a different state of background.
- Statistical properties such as evolutionary divergence rate are reflected by the Markov chains within the various states of background.

Hidden Markov Model



- The underlying state sequence is a Markov chain.
- Every state in the Markov chain emits symbols according to a given *probability mass function* (pmf).

Hidden Markov Model (Continued)

- Maximum likelihood estimation by the EM algorithm
- Recursive algorithm for efficient computation: *forward-backward* algorithm.
- Conceptually, HMM can be regarded as clustering with spacial dependence, or mixture models with spacial dependence.
- Each state corresponds to a “cluster”, or a “mode”, or a representative “trend”.
- Other applications:
 - Speech recognition
 - Image processing
 - Bioinformatics

Hidden Markov Model with Markovian Observations

- In the basic HMM, it is assumed that given all the states, the conditional distribution of observation x_t at position t only depends on the state s_t at the same position, that is,

$$\begin{aligned} & P(x_1, x_2, \dots, x_T \mid s_1, s_2, \dots, s_T) \\ &= P(x_1 \mid s_1)P(x_2 \mid s_2) \cdots P(x_T \mid s_T) . \end{aligned}$$

- In the new model, it is assumed that given all the states, the conditional distribution of observation x_t at position t depends on the observation x_{t-1} and state s_{t-1} at the previous position.

$$\begin{aligned} & P(x_1, x_2, \dots, x_T \mid s_1, s_2, \dots, s_T) \\ &= P(x_1 \mid s_1)P(x_2 \mid x_1, s_1) \cdots P(x_T \mid x_{T-1}, s_{T-1}) . \end{aligned}$$

- The HMM with Markovian observations (HMMMO) takes into consideration the strong inter-position dependence and in the mean time is capable of extracting modes of context.

Background Trends in An Aligned Sequence

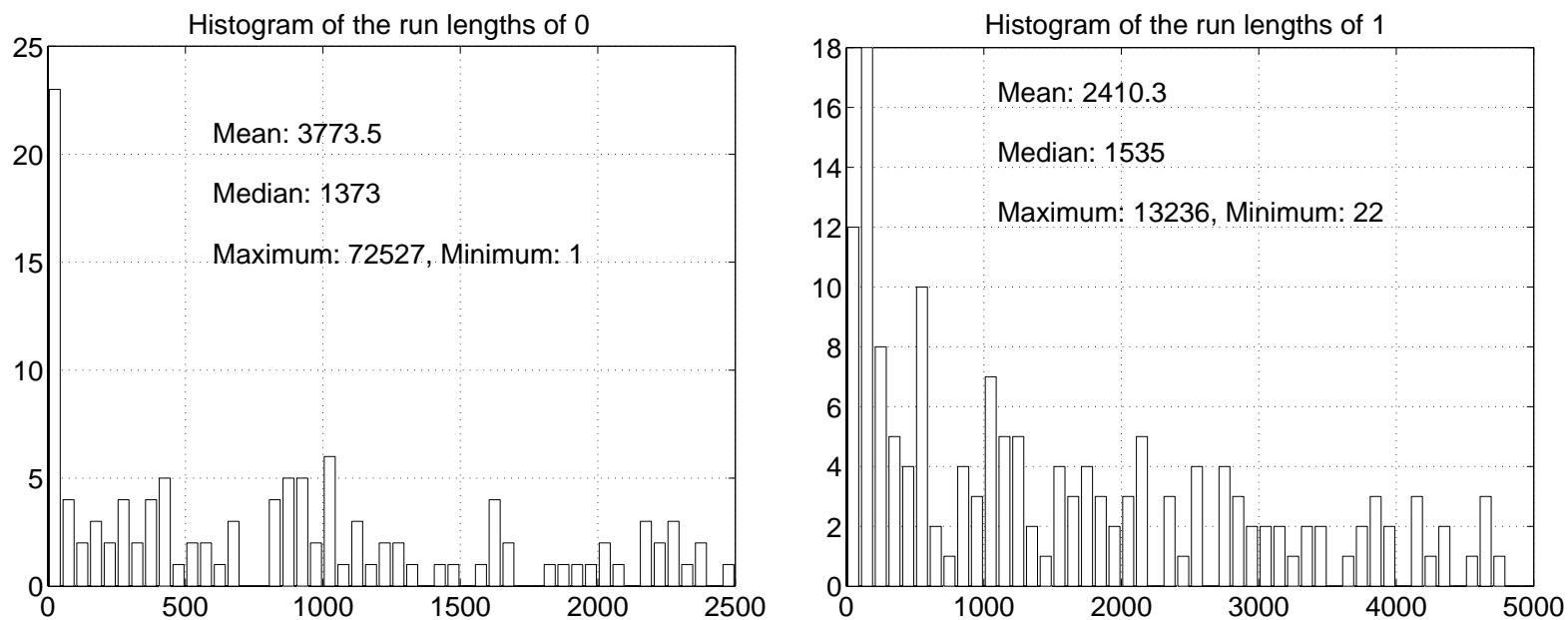


Figure 2: Histograms of the run lengths of 0's and 1's in the alignment sequence. The ranges of the histograms displayed are not complete.

- A mode extracted by an HMM should reflect long range changes and should not be affected by over-localized fluctuations.
- The run lengths of aligned and unaligned intervals are very long.

Estimation of HMMMO

- Parameters to estimate for an HMM with M states:
 - $2M$ probability mass functions: $P(x_t = i \mid x_{t-1} = j, s_{t-1} = m)$, $i, j = 0, 1$, $m = 1, \dots, M$. For notation brevity, we write $p_{j,i}(m) = P(x_t = i \mid x_{t-1} = j, s_{t-1} = m)$.
 - The state transition probability matrix, $\|a_{m,n}\|$, $m, n = 1, \dots, M$.
- Maximum likelihood estimation by the EM algorithm:
 - $L_m(t)$ denote the conditional probability of being in state m at position t given all the observations.
 - $H_{m,n}(t)$ denote the conditional probability of a transition from state m at position t to state n at position $t + 1$ given all the observations.
 - The re-estimation formulae for the transition probabilities $a_{m,n}$, $m, n = 1, \dots, M$:

$$a_{m,n} = \frac{\sum_{t=1}^{T-1} H_{m,n}(t)}{\sum_{t=1}^{T-1} L_m(t)}.$$

- The re-estimation formulae for the probabilities $p_{j,i}(m)$, $i, j = 0, 1$, $m = 1, \dots, M$:

$$p_{j,i}(m) = \frac{\sum_{t=1}^{T-1} L_m(t) I(x_t = j) I(x_{t+1} = i)}{\sum_{t=1}^{T-1} L_m(t) I(x_t = j)}.$$

Modified Forward-Backward Algorithm

- Define the forward probability $\alpha_m(t)$ as the joint probability of observing the first t x_τ 's, $\tau = 1, \dots, t$, and being in state m at position t .

$$\begin{aligned}\alpha_m(1) &= \pi_m p_{x_1}(m) \quad 1 \leq m \leq M \\ \alpha_m(t) &= \sum_{n=1}^M \alpha_n(t-1) p_{x_{t-1}, x_t}(n) a_{n,m} \\ & \quad 1 < t \leq T, 1 \leq m \leq M .\end{aligned}$$

- Define the backward probability $\beta_m(t)$ as the conditional probability of observing x_τ 's after position t , $\tau = t+1, \dots, T$, given the state at position t is m and the observation at t is x_t .

$$\begin{aligned}\beta_m(T) &= 1 \\ \beta_m(t) &= p_{x_t, x_{t+1}}(m) \sum_{n=1}^M a_{m,n} \beta_n(t+1), \quad 1 \leq t < T .\end{aligned}$$

Modified Forward-Backward Algorithm (Continued)

- The probabilities $L_m(t)$ and $H_{m,n}(t)$ are solved by

$$\begin{aligned} L_m(t) &= P(s_t = m \mid \mathbf{x}) = \frac{P(\mathbf{x}, s_t = m)}{P(\mathbf{x})} \\ &= \frac{1}{P(\mathbf{x})} \alpha_m(t) \beta_m(t) \end{aligned}$$

$$\begin{aligned} H_{m,n}(t) &= P(s_t = m, s_{t+1} = n \mid \mathbf{x}) \\ &= \frac{1}{P(\mathbf{x})} \alpha_m(t) a_{m,n} p_{x_t, x_{t+1}}(m) \beta_n(t+1), \end{aligned}$$

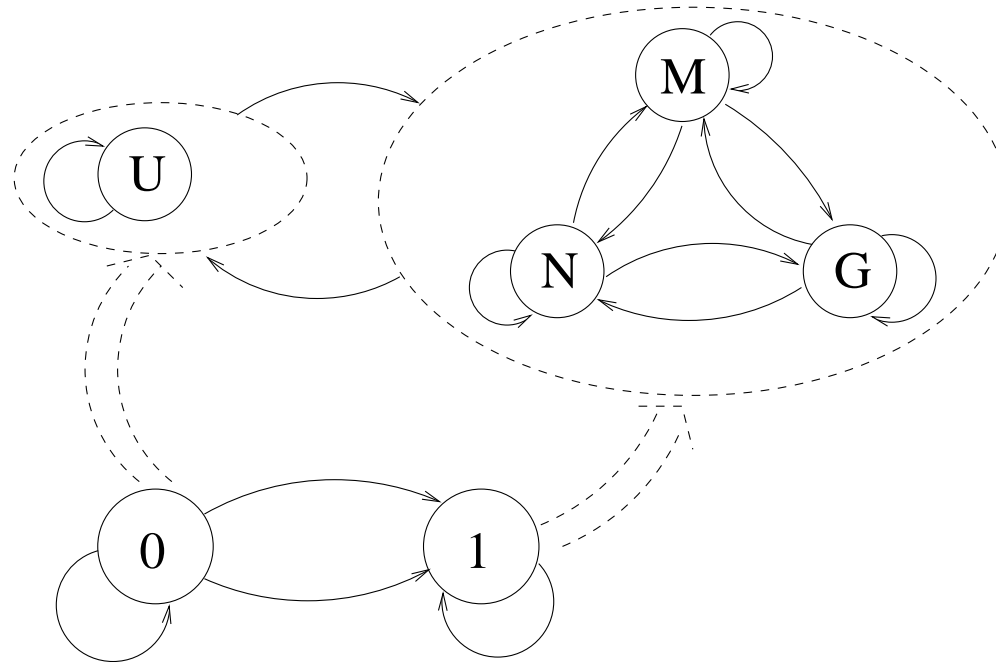
where $P(\mathbf{x})$ is the joint probability of observing all x_t 's, $t = 1, \dots, T$, and for any t ,

$$P(\mathbf{x}) = \sum_{m=1}^M \alpha_m(t) \beta_m(t).$$

Selection of the Number of States

- The number of states in the HMM is chosen by the Bayesian Information Criterion (BIC).
- The optimal model maximizes the penalized log likelihood $\log P(\mathbf{x}) + \frac{k}{2} \log T$, where k is the number of parameters in the HMM and T is the sequence length.
- The total number of parameters in an HMM with M states is $M(M-1) + 2M = M^2 + M$.
 - The number of parameters to specify the transition matrix $\|a_{m,n}\|$ is $M(M-1)$.
 - The number of parameters to describe the Markov chain of 0 and 1 in each state is 2.
- Constraints can also be put on the state transition probabilities $a_{m,n}$ to reduce the complexity of an HMM.
- VCFS: 4 states with $a_{m,n}$ the same for all $n \neq m$.
- The entire human Chromosome 22: 4 states without constraints on $a_{m,n}$.

Modeling the Aligned Sequence



- Within each state of the HMMMO, the two symbols 0 and 1 form a Markov chain.
- Symbols G, M, N are contained in 1.
- Within a run of 1's, G, M, and N form a sub-Markov chain. This Markov chain is estimated by the maximum likelihood criterion.
- Equivalently, U, G, M, and N form a 4-state Markov chain.

Karlin and Dembo's Theorem

- For a sequence modeled by a Markov chain, compute the asymptotic probability of this sequence having its maximal segment score exceeding a threshold.
- Assumptions on the Markov chain:
 - The Markov chain \mathcal{P} is irreducible and aperiodic.
 - The negative drift condition is $E[Z] < 0$.
 - For each symbol ζ_i , there exists a symbol ζ_j such that $p_{\zeta_i, \zeta_j} > 0$, $Z_{\zeta_i, \zeta_j} > 0$; and a symbol ζ_k such that $p_{\zeta_i, \zeta_k} > 0$, $Z_{\zeta_i, \zeta_k} < 0$.

- If scores Z_{ζ_i, ζ_j} , $i = 1, 2, \dots, r$, are non-lattice, then

$$\lim_{T \rightarrow \infty} P\left\{M_\gamma(T) - \frac{\ln T}{\theta^*} > z\right\} = 1 - \exp(-K^* e^{-\theta^* z}),$$

where K^* and θ^* are constants determined by the transition probability matrix and the scores.

- If scores Z_{ζ_i, ζ_j} are lattice of span δ (δ is the largest number of which all the Z_{ζ_i, ζ_j} 's are multiples), let $K^+ = e^{\theta^* \delta} K^*$.

$$\begin{aligned} 1 - \exp(-K^* e^{-\theta^* z}) &\leq \liminf_{T \rightarrow \infty} P\left\{M_\gamma(T) - \frac{\ln T}{\theta^*} > z\right\} \\ &\leq \limsup_{T \rightarrow \infty} P\left\{M_\gamma(T) - \frac{\ln T}{\theta^*} > z\right\} \leq 1 - \exp(-K^+ e^{-\theta^* z}). \end{aligned}$$

Computing p-values

- Suppose scores Z_{ζ_i, ζ_j} are lattice of span $\delta = 1$.
- When T is sufficiently large, for a high-scoring segment with score $\frac{\ln T}{\theta^*} + z$, the upper bound $1 - \exp(-K^+ e^{-\theta^* z})$ provides a “conservative” p-value for the segment.
- Since $K^+ = e^{\theta^*} K^*$, K^+ is close to K^* when θ^* is close to zero. In this case, the upper bound is close to the real probability $P\{M_\gamma(T) - \frac{\ln T}{\theta^*} > z\}$.

Background Difference in Divergence Rate

- For each “mode” of background evolutionary divergence rate extracted by the HMM, a Markov chain of UMNG is estimated.
- For each state m of the HMM, the p-value as a function of the segment score \bar{z} is

$$p_v(\bar{z}, m) = 1 - \exp\left(-K^+(m)e^{-\theta^*(m)\left(\bar{z} - \frac{\ln \pi_m T}{\theta^*(m)}\right)}\right) .$$

- How to take into account the “background” difference of the two segments in two different regions?
 - For each region, we compute the conditional probability distribution of the state of a randomly selected position from the region given the entire observed sequence.
 - Assume a region ranges from t_1 to t_2 and t is a position randomly selected from the region, the conditional probability $P\{s_t = m \mid y_1, y_2, \dots, y_T\}$ is

$$P\{s_t = m \mid y_1, y_2, \dots, y_T\} = \frac{\sum_{\tau=t_1}^{t_2} L_m(\tau)}{t_2 - t_1 + 1} .$$

- The weighted sum of the p-values $p_v(\bar{z}, m)$ is

$$\bar{p}_v(\bar{z}) = \sum_{m=1}^M P\{s_t = m \mid y_1, y_2, \dots, y_T\} p_v(\bar{z}, m) .$$

Summary on Computing p-values

- Extract modes of background divergence rate using HMMMO.
- Estimate the Markov chain of UMNG within each state of the HMMMO.
- Compute constants K^+ and θ^* for each Markov chain.
- Given a region, compute the conditional probability distribution of the state of a randomly selected position from the region.
- Compute $\bar{p}_v(\bar{z})$ to assess a segment with score \bar{z} in the region.

VCFS Region of Human Chromosome 22

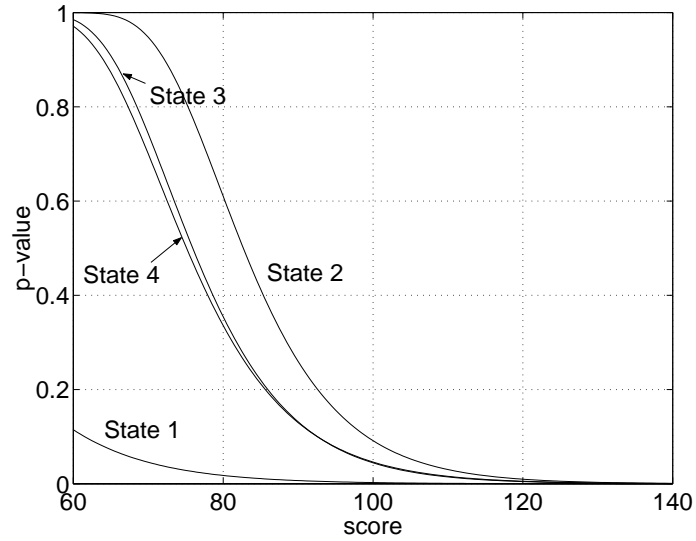


Figure 3: The p-value as a function of the segment score for the Markov chain in each state.

State	1	2	3	4
unaligned	99.1%	23.8%	72.1%	0.14%
occupied	32.0%	26.8%	35.4%	5.7%

Table 1: Characteristics of the HMM's four states. First row: the percentage of unaligned base pairs in each state. Second row: the stationary frequency of each state.

VCFS Region of Human Chromosome 22 (Continued)

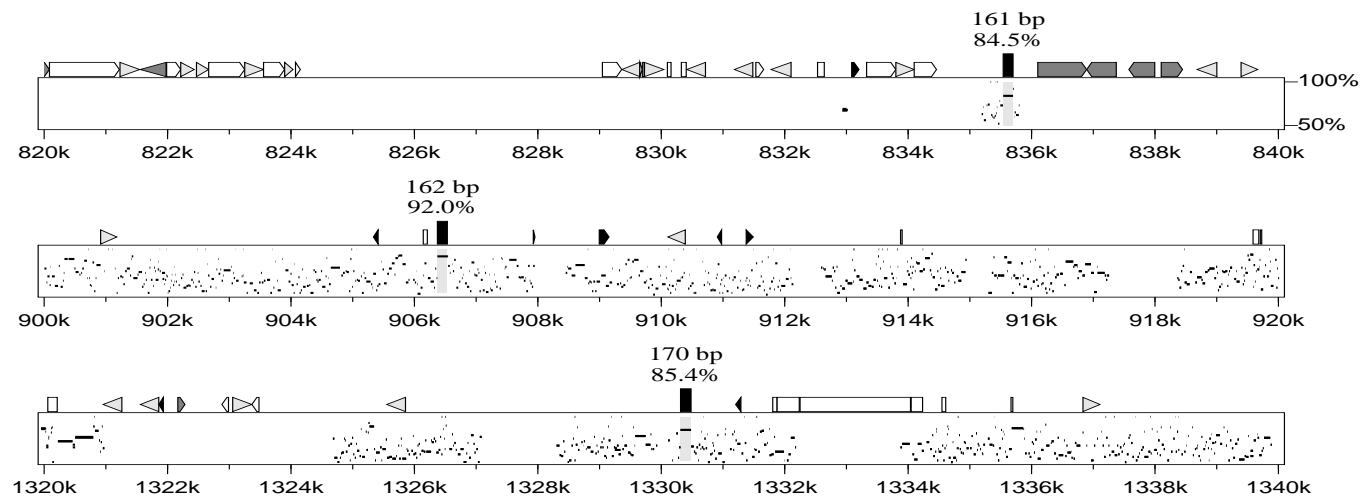


Figure 4: Percent identity plot (PIP) of some human-mouse alignments. Triangles and other icons along the top indicate positions of interspersed repeats and low-complexity regions found in the human sequence by the RepeatMasker program. Each tiny horizontal line in the PIP indicates the human positions and percent nucleotide identity of an interval between consecutive gaps in a local alignment with the mouse genomic sequence.

Segment	1	2	3
Length	161	162	170
Percentage of M	84.5%	92.0%	85.4%
score (M=1, N=-1)	111	137	121
p-value	0.0075	0.0013	0.0081

Table 2: The p-values of the three segments indicated in Figure 4.

Human Chromosome 22

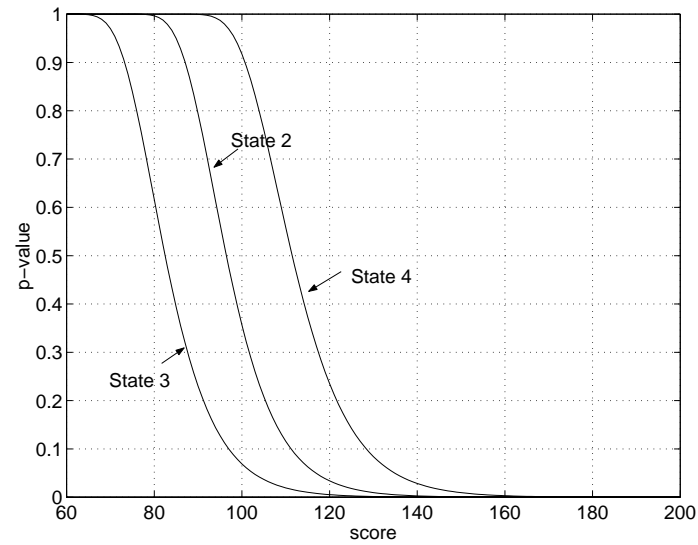


Figure 5: The p-value as a function of the segment score for the Markov chains in State 2, 3, and 4 of the chromosome 22 sequence.

State	1	2	3	4
unaligned	99.94%	63.85%	86.84%	28.33%
occupied	28.89%	31.22%	19.73%	20.16%

Table 3: Characteristics of the four states in the HMM trained on the chromosome 22 sequence. First row: the percentage of unaligned base pairs in each state. Second row: the stationary frequency of each state.

Human Chromosome 22 (Continued)

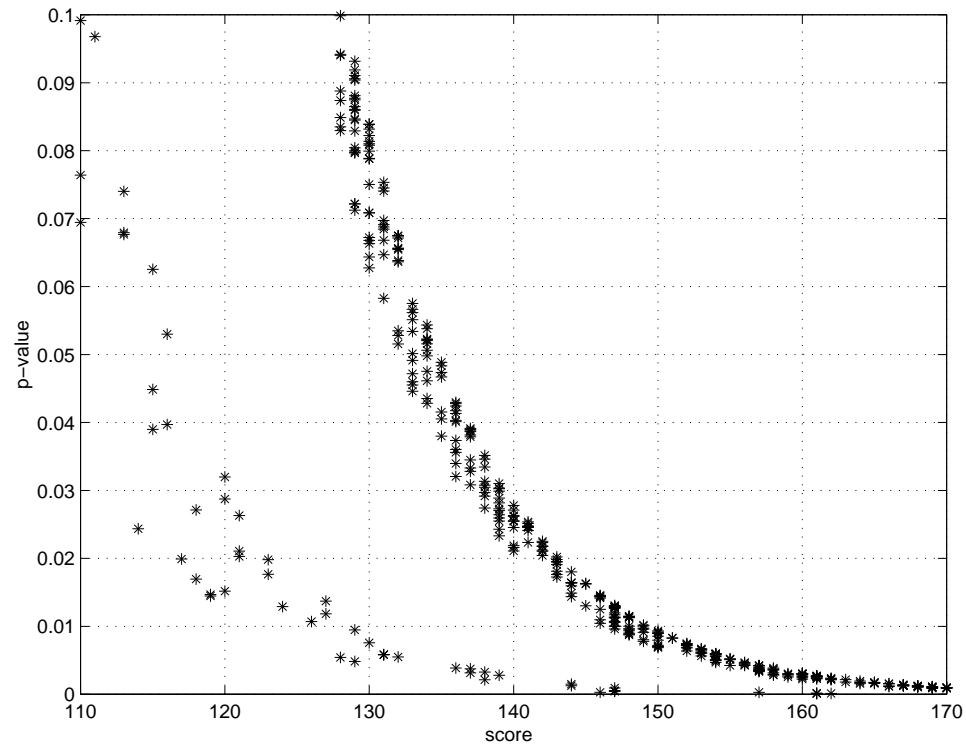


Figure 6: The p-values of a set of high scoring segments.

Discussion

- Other applications of HMMMO in genomics for capturing trends in biological sequences
 - Comparison with *isochores*, i.e., genomic regions of more-or-less constant percentage of C and G nucleotides.
- Using this tool, we can further study the correlation between divergence rate and other genomics properties, e.g., GC level, recombination rate, gene density.