

Outline

- ① Normalized maximum likelihood model
- ② Models with memory
- ③ DNA duplications in evolution and disease
- ④ Optimizing the representation length of the full locus
- ⑤ Uncovering gene duplications by segmentation
- ⑥ Segmentation of haplotype data
- ⑦ Results

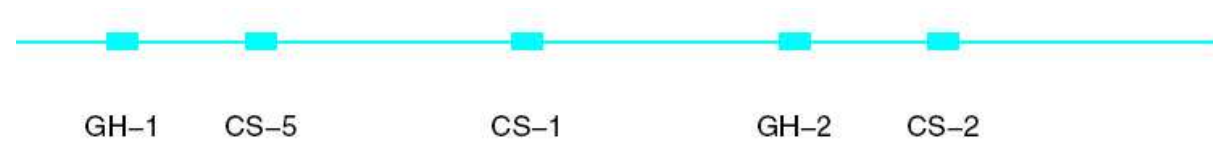
Segmentation by MDL methods

Minimum description length principle:

”Use in inference the model able to represent with the shortest number of bits both the model and its errors”

- segmentation of ECG
- segmentation of audio
- clustering of microarray data
- blocking of haplotype data

Approximate matching



Use a universal coding of the binary mask resulting from matching two candidate sequences.

- Normalized maximum likelihood models
- For memoryless sources (Bernoulli)
- For sources with memory

The memory model

- The matching sequence x^n is a Markov source of order k
- The binary state $[x_{t-k} \dots x_{t-1}]$ identified with the integer j
- Parameters of the model: the conditional probabilities

$$\theta_{ji} = P(x_t | x_{t-1}, \dots, x_{t-k})$$

- The maximum likelihood of the parameters (Bartlett '55)

$$\hat{\theta}_{ji} = \frac{n_{ji}}{n_{j\cdot}}$$

- where n_{ji} are the counts of the pairs (ij) in x^n and $n_{j\cdot} = \sum_i n_{ji}$
- The matrix $\mathbf{n}(x^n)$ of counts n_{ji} is a sufficient statistics for θ

The NML model

- The normalized ML is by definition

$$\hat{P}(x^n | r(x^n) = r) = \frac{\hat{P}(x^n | r(x^n))}{\sum_{y^n} \hat{P}(y^n | r(y^n))}$$

- Can be efficiently computed

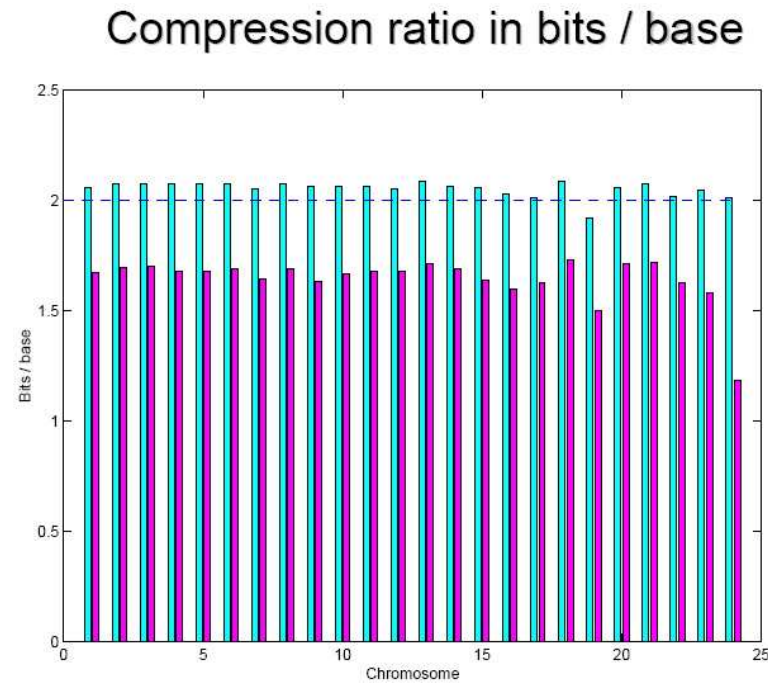
$$\hat{P}(x^n | r(x^n) = r) = \frac{\mathbf{n}(x^n) \mathbf{n}(x^n)}{\sum_{(r,s,\mathbf{n}) \in \Omega_{r,n}} N(r,s,\mathbf{n}) \mathbf{n}^{\mathbf{n}}}$$

where $\mathbf{n}^{\mathbf{n}}$ is a shorthand notation for $\prod_{i \in \mathcal{A}, j \in \mathcal{A}^m} \left(\frac{n_{ji}}{n_j} \right)^{n_{ji}}$

- $\Omega_{r,n}$ collects all the possible triplets (r, s, \mathbf{n}) ;
- $N(r, s, \mathbf{n})$ is cardinality of the set of strings having the same (r, s, \mathbf{n}) .

Previous applications of the NML model

- The only compressor specialized for DNA running on the whole human genome



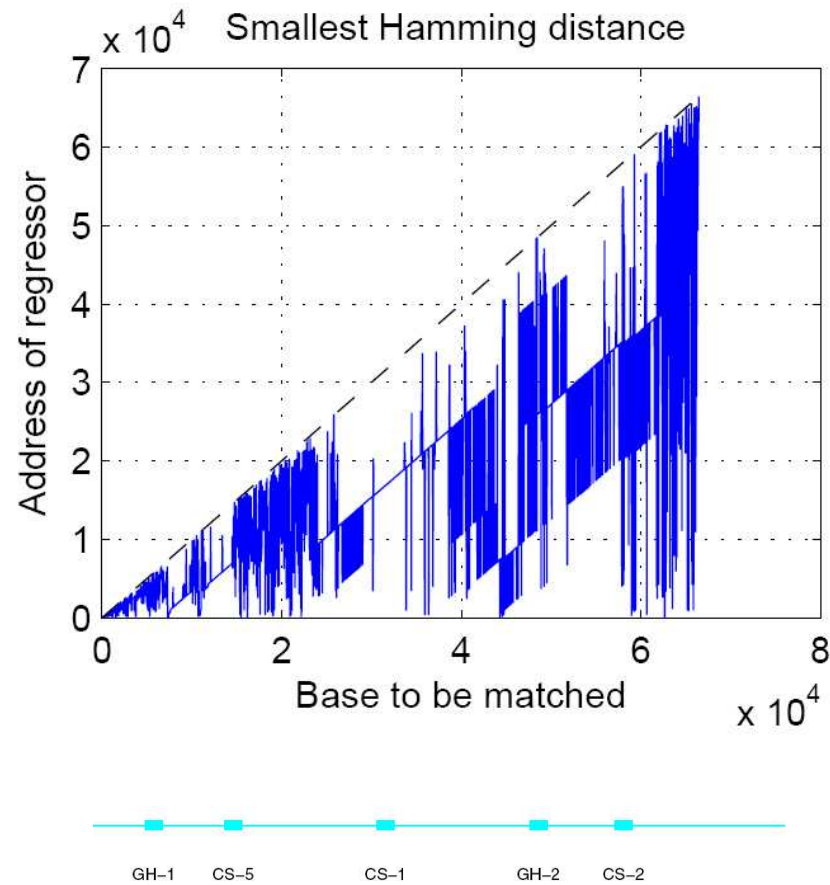
Korodi & Tabus ACM Trans. on Information Systems, 2005

Korodi & Tabus DCC 2007

DNA duplications and their role in evolution and disease

- The genomes evolve mainly by three processes: mutations at a single base pair, rearrangements of chromosomes, and DNA duplication
- When a gene gets duplicated, the new gene may evolve and encode proteins taking over functions other than that of the originally encoded protein
- In humans it is estimated that approximately 5% of the DNA sequence is obtained through duplications
- The size of the duplications ranges from thousands to hundred thousands bases
- The degree of sequence similarity is higher than 90%
- Uncovering the duplications underwent by the genome of a contemporary species may help in understanding the interaction between genes
- Many genetic disorders appear due to similar rearrangements/duplication processes

Traditional approach to gene duplication



Optimizing the overall cost for duplication analysis

- Consider blocks of a fixed size and match them freely in the past seen sequence
- Representation of the block done by three elements:
 - ↳ Encoding the address (pointer) of the matching block in the past (using a universal code for integers)
 - ↳ Encoding the mask of matching between the two blocks using the universal NML codes
 - ↳ Specification of the mismatching bases between the two blocks

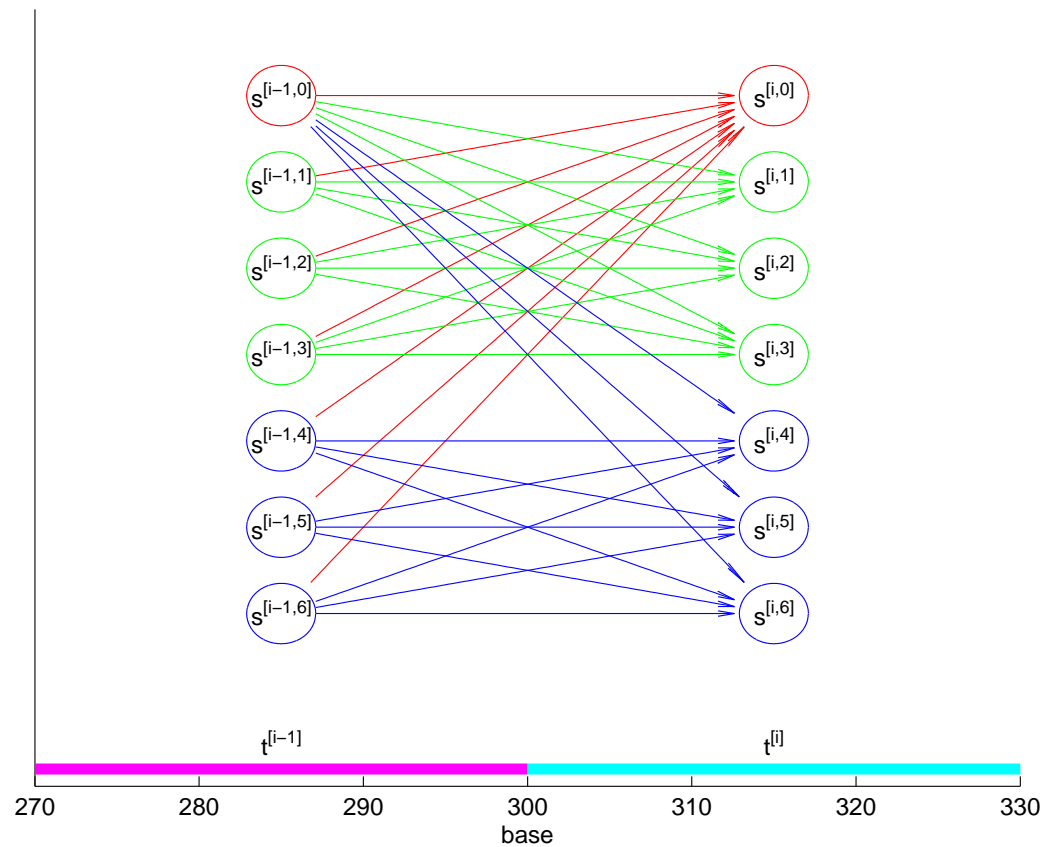
Encoding the pointers and the mask

- The integer address $a_{i,j}$ is encoded:
 - ➡ either directly, using Elias code
 - ➡ or as relative to the best prediction based on previous match $a_{i-1,\ell} + L$
(Golomb-Rice code with parameter 0)
- the matching mask $x^{[i,j]}$ is encoded by the universal model in $-\log_2(P(x^{[i,j]}))$ bits (the initial state r is the all zero string)

Dynamic programming problem

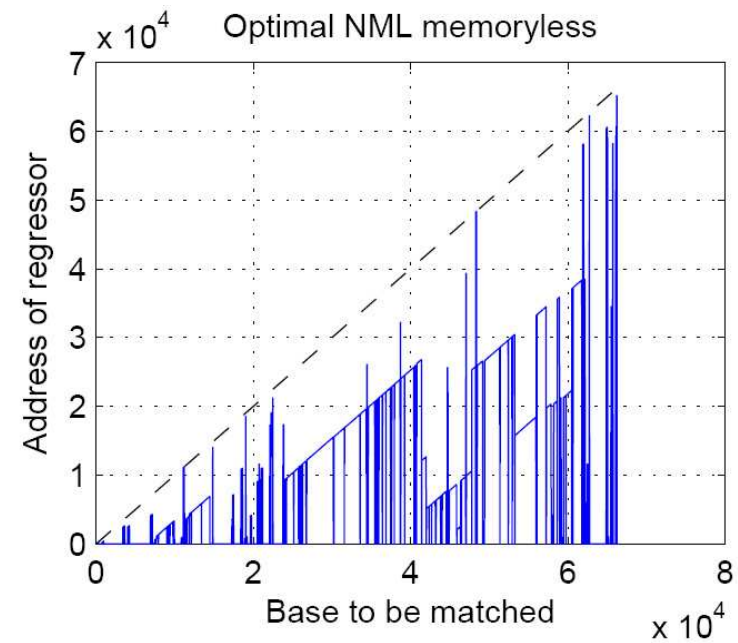
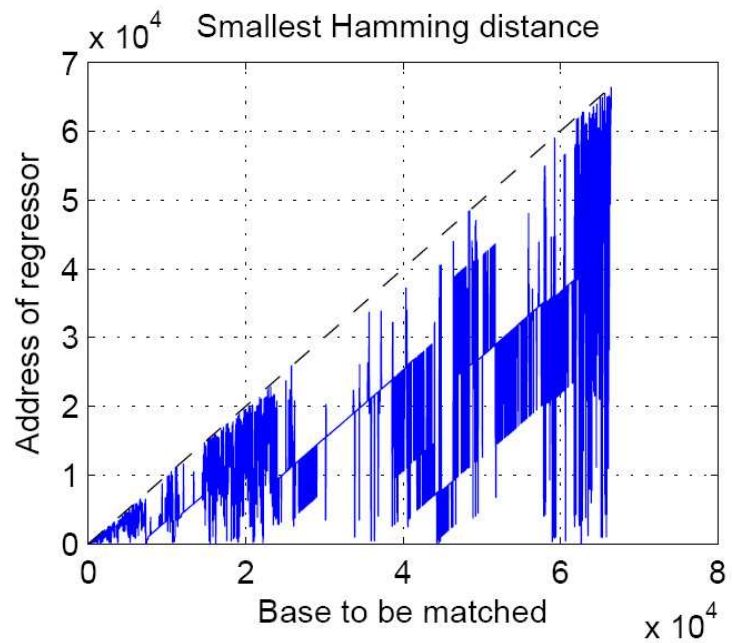
- At each segment $t^{[i]}$ we have $2K_{max} + 1$ states $s^{[i,j]}$ in the dynamic programming trellis.
 - ↳ the state, $s^{[i,0]}$, corresponds to encoding $t^{[i]}$ in clear using $2L + 1$ bits
 - ↳ K_{max} states, $s^{[i,j]}$, $j = 1, \dots, K_{max}$, correspond to encoding $t^{[i]}$ relative to the candidate direct match, $c^{[i,j]}$
 - ↳ the last K_{max} states, $s^{[i,j]}$, $j = K_{max} + 1, \dots, 2K_{max}$, correspond to encoding $t^{[i]}$ relative to the candidate palindrome match, $d^{[i,j]}$

Universal models with memory



Two sections of the trellis for dynamical programming, when $L = 30$, $K_{max} = 3$, $i = 9$. The various ways to encode the pointers will lead to various costs for the transitions between the stages of the trellis.

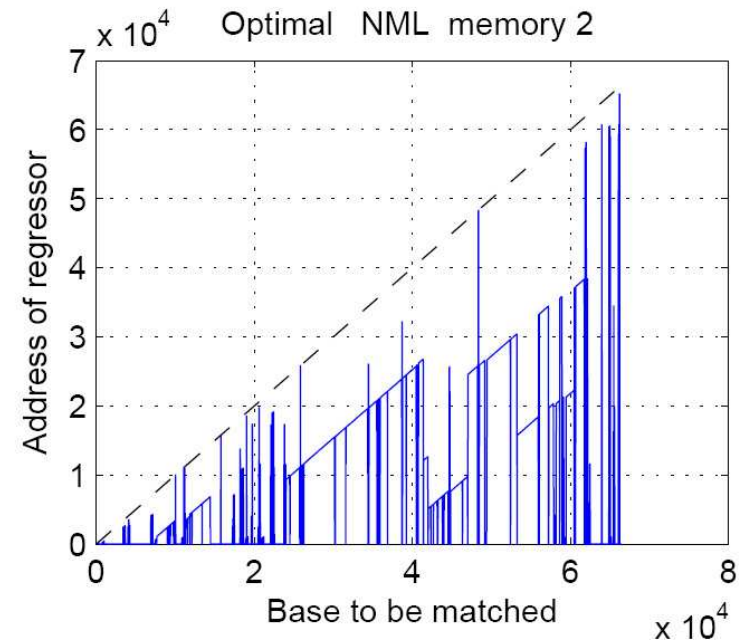
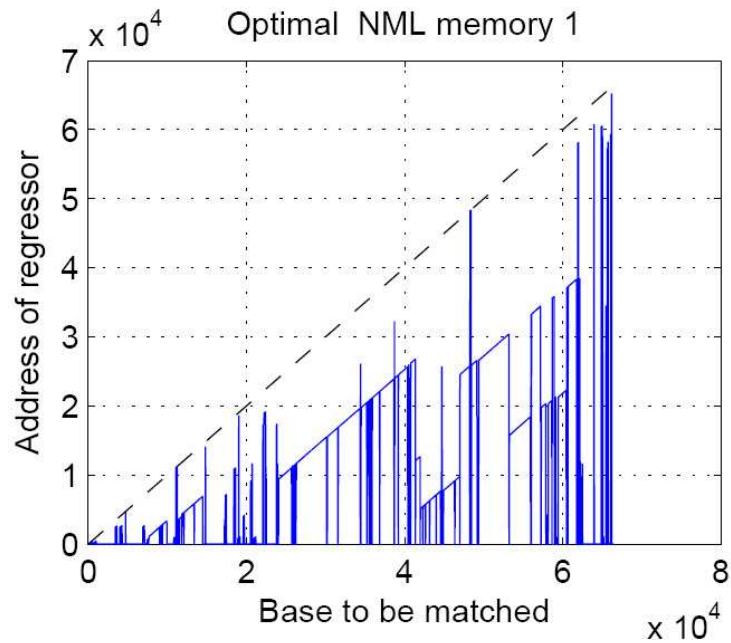
Universal models with memory



Addresses of best matching regressors found by optimizing:

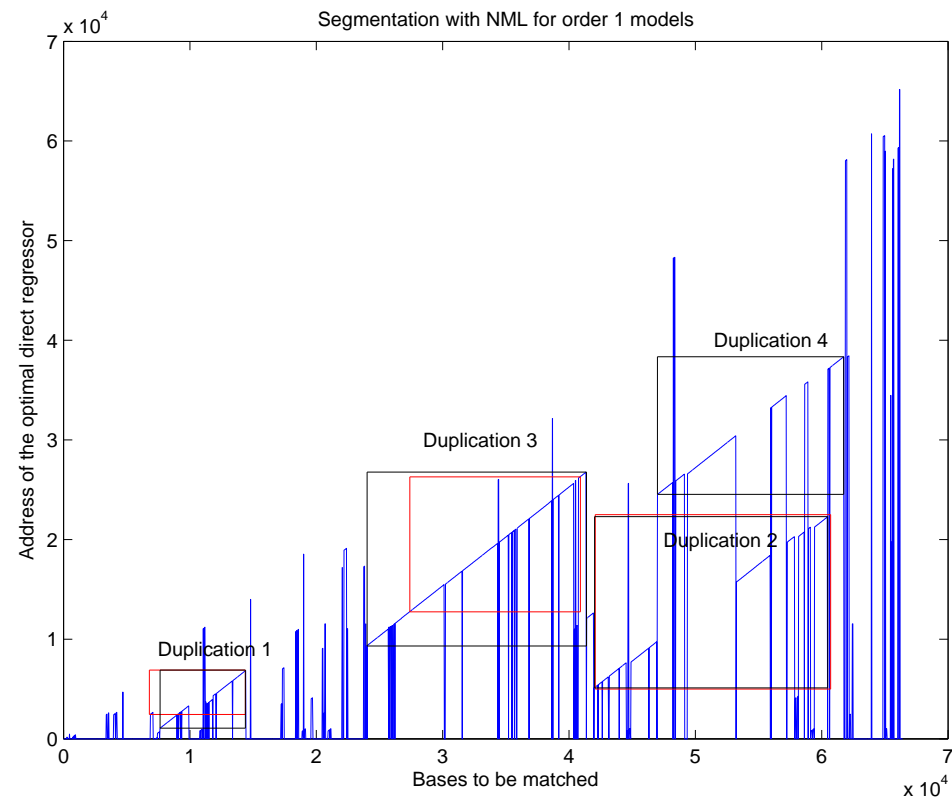
- the Hamming distance
- the representation cost with NML model of order 0.

Universal models with memory



Addresses of best matching regressors found by optimizing the representation cost with NML model of orders 1 and 2.

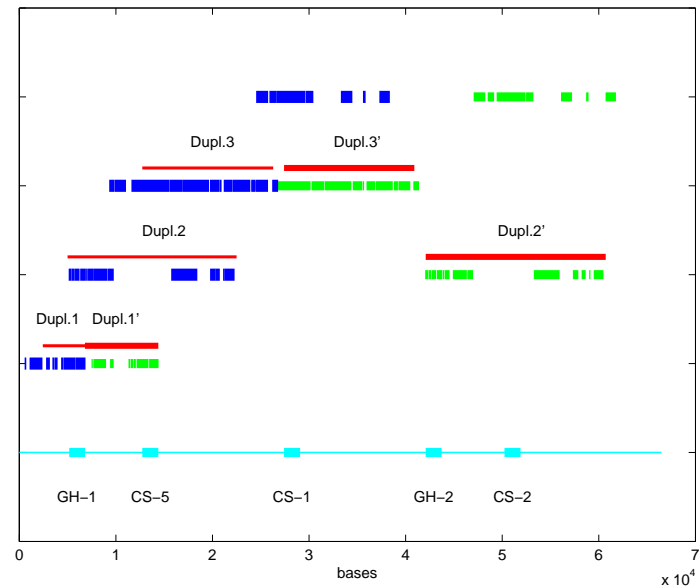
Universal models with memory



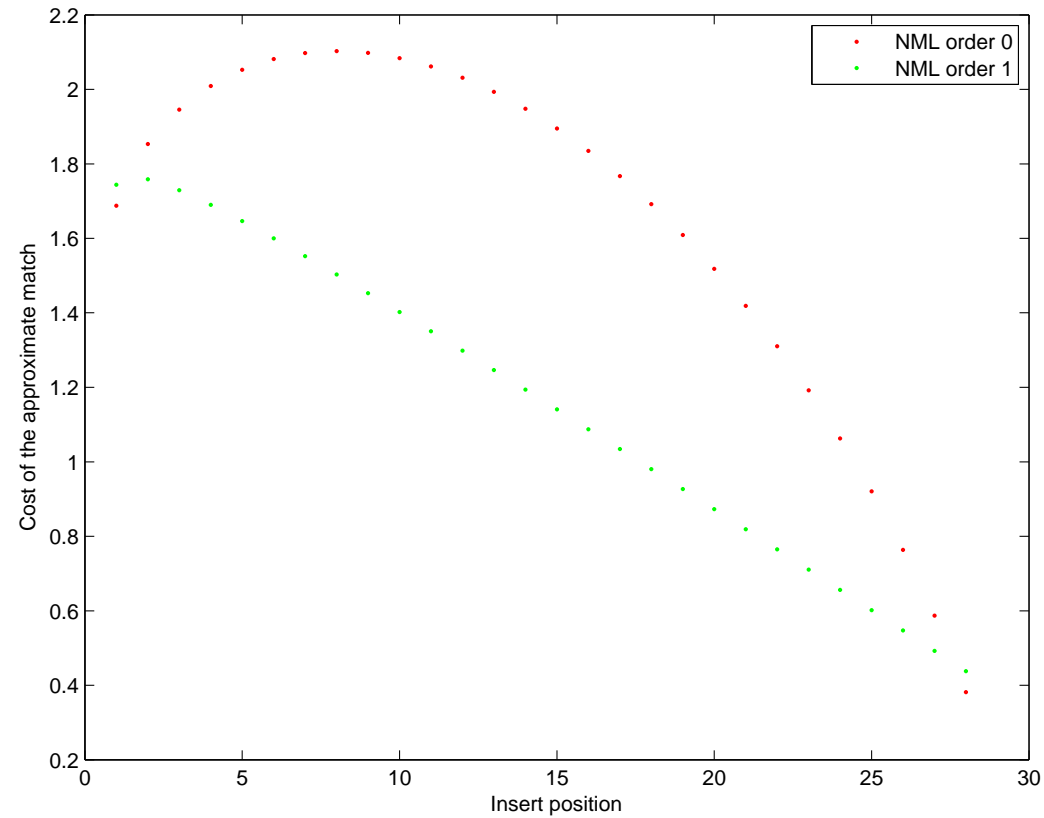
Using the segmentation for finding possible sites of duplication:

- Black rectangles: putative duplication.
- Red rectangles: duplication described previously in the literature.

Universal models with memory



The possible four duplication boxes displayed along the HUMGHCSA sequence. The bottom row: GH-1 (5163-6798), CS-5 (12742-14392), CS-1 (27409-29068), GH-2 (42080-43726), CS-2 (50233-51892).



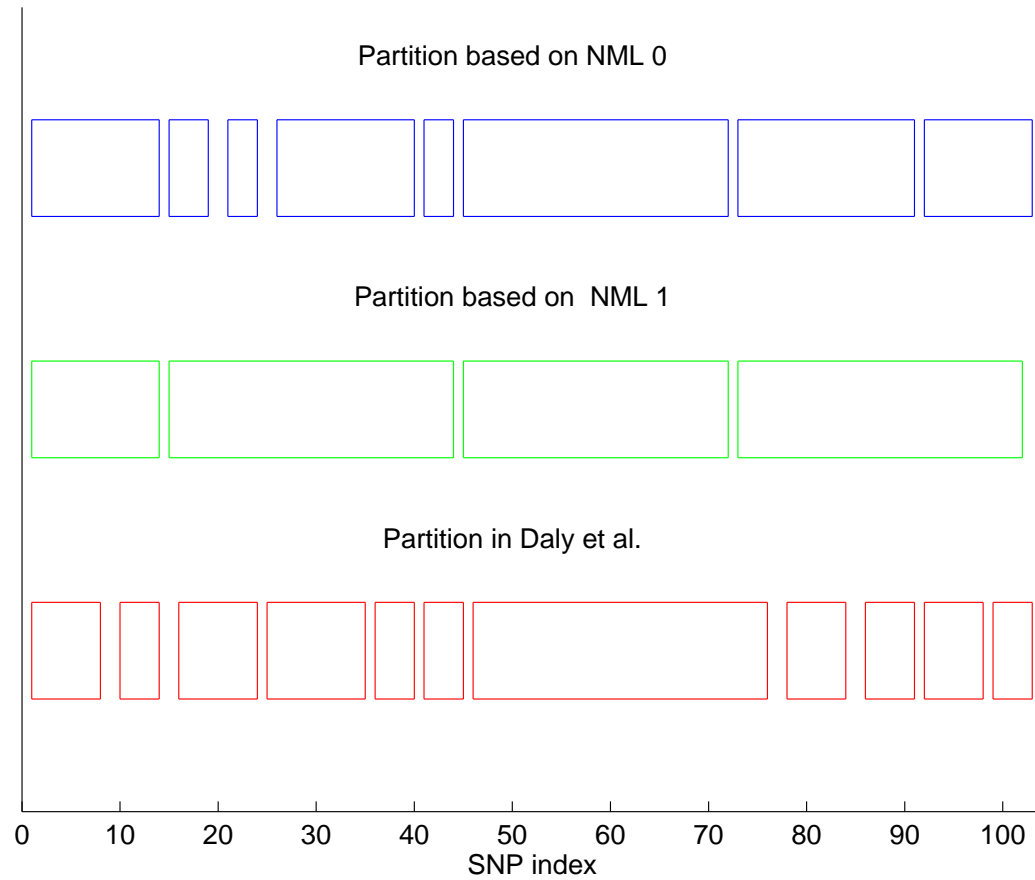
Codelength, $(-\log_2(P(x^L)) + (L - i + 1) * \log_2(3))/L$, in bits per symbol, for representing an approximate match of length $L = 28$, a perfect match with an extra symbol.

Segmentation of haplotypes

- The haplotype data is listing only selected bases from the human genome, each called a single polymorphism nucleotide (SNP) site
- At each SNP one can find one of two possible bases
- a number of neighboring SNPs (say m sites) seen as a block will be found to take a small number of all 2^m possible combinations
- Finding the block boundaries is interesting for association with diseases or for study of population migration.

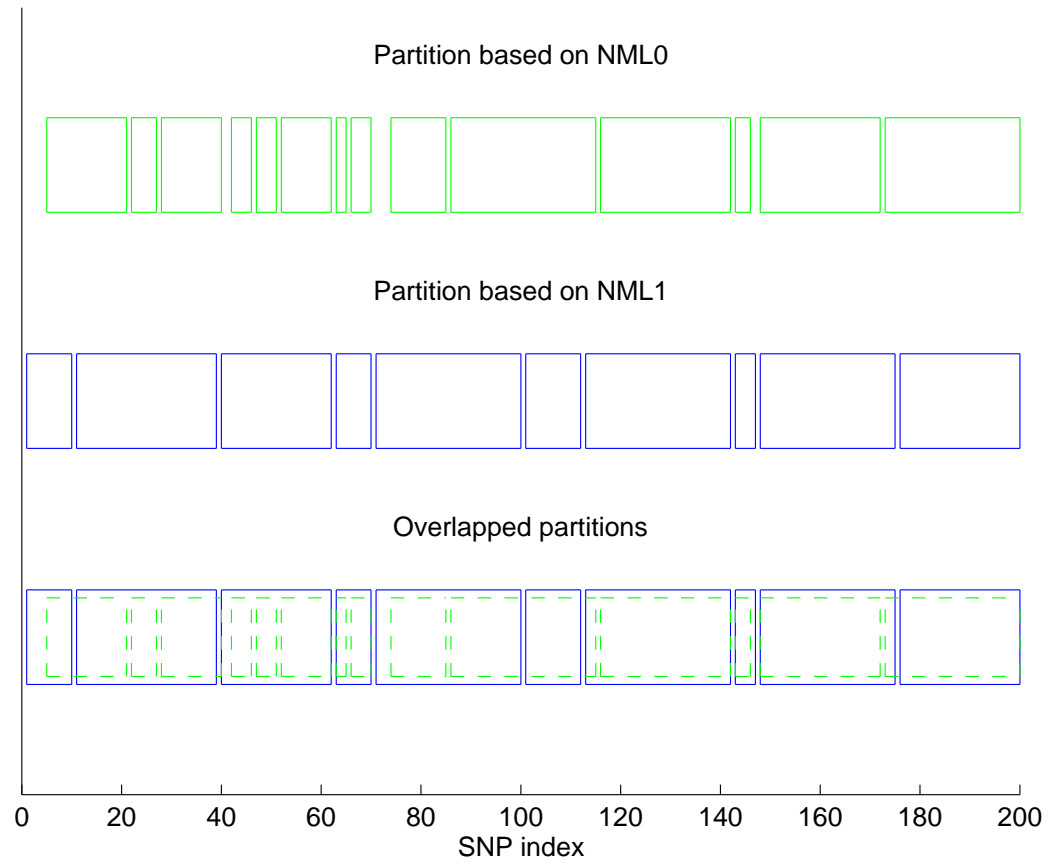
Segmentation of haplotypes

- Jointly clustering and encoding algorithm:
 - ↳ selects the boundaries of the haplotypes
 - ↳ so that the segmentation provides the best description length of the overall data
 - ↳ The matching of the same candidate block from various individuals to the chosen center of the cluster is performed using the NML codes
 - ↳ We do not need to set any prior bounds on the number of common patterns for each block or on the number of blocks



Haplotype segmentation using NML for orders 0 and 1 for the data set from Daly 2000.

Universal models with memory



Haplotype segmentation using NML for orders 0 and 1. A 200 long subsequence from the Hapmap project.

Conclusions

- Universal models provide efficient representation tools for genomic sequences
- More refined model order selection procedures may better account for non-stationarity along sequences.
- The techniques are easy to extend to more adaptive tools