Test de structure de séquences biologiques basé sur la Chaos Game Representation

Peggy Cénac
INRIA Rocquencourt & Université Paul Sabatier (Toulouse III)

11 octobre 2005

Séminaire Mathématiques pour le Génome

Genopole d'Evry

Plan

Chaos Game Representation

- Definition
- Examples
- Main properties

Construction of the test of structure

- Characterization of structure
- Test
- Numerical experiments

Genomic signature

- Dinucleotide relative abundance profile
- CGR-based relative abundance
- Taxonomy tree

Chaos Game Representation - Definition

Graphical representation of DNA in a bounded set.

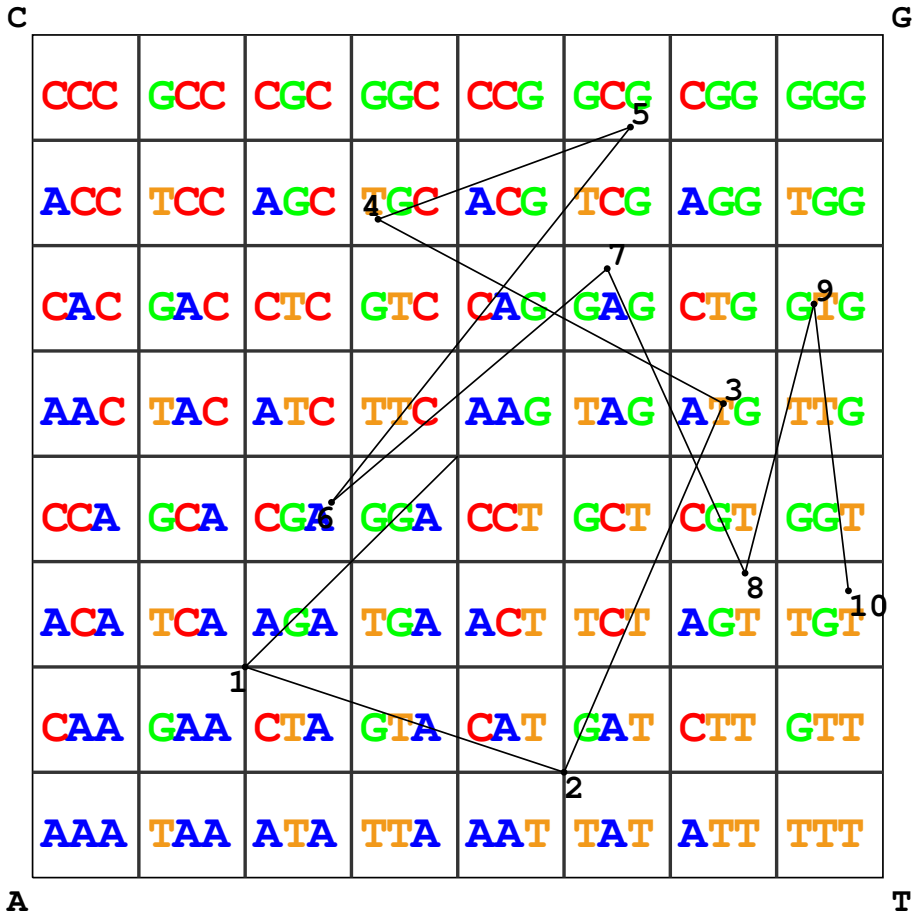
- Storage tool
- Pattern visualization
- Sequences comparison (local/global)

Iterative mapping technique

- DNA sequence $U = (u_i)_{i=1,\ldots,n}$, where $u_i \in \{A, C, G, T\}$.
- The Chaos Game Representation of U, on the unit square S is a sequence $\{X_0, \ldots, X_n\}$ defined by

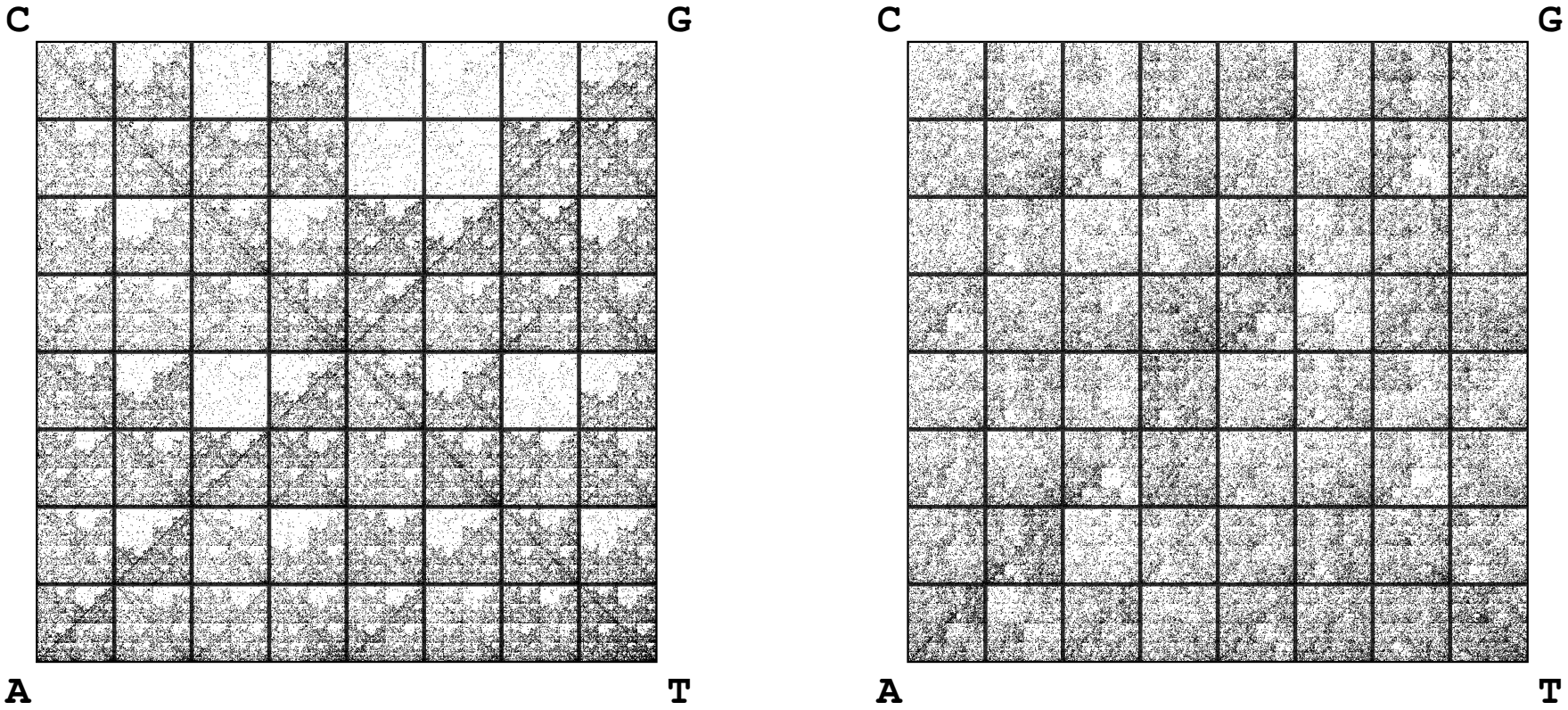
$$\begin{cases} X_0 = (\frac{1}{2}, \frac{1}{2}) \\ X_{i+1} = \frac{1}{2}\big(X_i + \ell_{u_{i+1}}\big), \\ \ell_A = (0,0), \quad \ell_C = (0,1), \quad \ell_G = (1,1), \quad \ell_T = (1,0). \end{cases}$$

Examples (1)



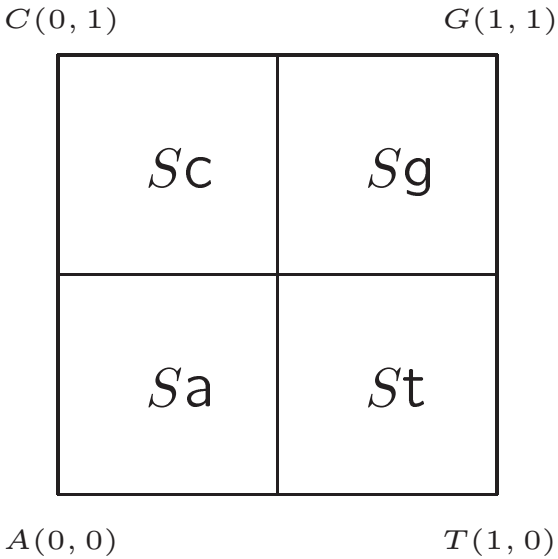
CGR of the word ATGCGAGTGT.

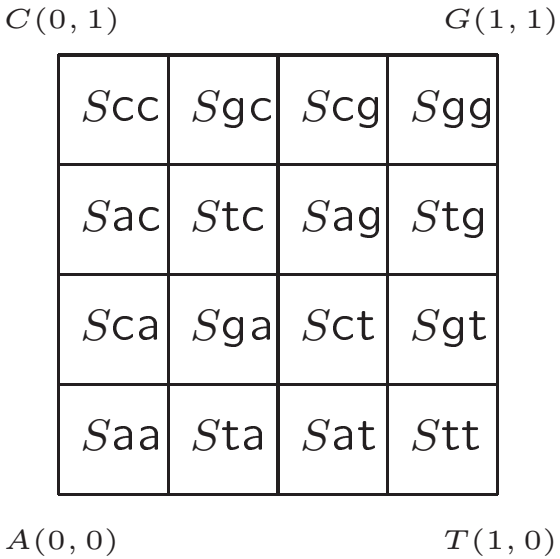
Examples (2)



CGR of 200000 nucleotides of Chromosome 2 of Homo Sapiens (on the left) and of Bacteroides Thetaiotaomicron (on the right).

CGR - Definition (2)





- $Sw \stackrel{\text{def}}{=} \sum_{k=1}^{i} \frac{1}{2^{i-k+1}} \ell_{v_k} + \frac{1}{2^i} S$, where w is the word $v_1 \dots v_i$.
- Counting points in $Sw \Leftrightarrow$ counting occurrences of w.
- Each point contains the whole sequence history.

Stochastic properties of the CGR

- U is supposed to be a stationary ergodic sequence.
- $(X_n)_{n\geq 0}$ is a Markov chain of order 1, and converges almost surely to a random vector X with distribution π.
- When U is i.i.d. and uniformly distributed, π is the Lebesgue measure on S. Whenever U is not uniformly distributed, π is continuous, singular with respect to the Lebesgue measure.
- The law of large number holds, and the empirical measures converge.

Characterization of structure

For any word $w = v_1 \dots v_i$, any set B, let us define

$$Bw \stackrel{\text{def}}{=} T_{v_i} \circ \cdots \circ T_{v_1}(B), \quad \text{where} \quad T_{v_i}(x) \stackrel{\text{def}}{=} \frac{1}{2}(x + \ell_{v_i}).$$

Then the stationary random sequence U is

- an i.i.d. sequence if and only if

$$\pi(Bu) = \pi(B)\pi(Su), \quad \forall u \in \mathcal{A}, \ \forall B \subset S.$$

- a Markov chain of order m if and only if

$$\frac{\pi(Bwu)}{\pi(Bw)} = \frac{\pi(Swu)}{\pi(Sw)}, \quad \forall B \subset S, \quad \forall w \in \mathcal{A}^m, \quad \forall u \in \mathcal{A}.$$

  In particular the ratio $\frac{\pi(Bwu)}{\pi(Bw)}$ does not depend on B.

➠ construction of a test family

➠ genomic signature

Testing the structure of a sequence (1)

- H_0 : "$U = u_1 \dots u_n$ is an i.i.d. sequence"
- H_m : "U is a Markov chain of order m"
- H : "U is a stationary ergodic sequence"

Let us denote $\hat{\pi}_n(E) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{E\}}(X_j)$ the empirical measure of π. Let u_α be the $(1-\frac{\alpha}{2})$-quantile of the normal law. Define also

$$\hat{\sigma}_n(B,v) \stackrel{\text{def}}{=} \sqrt{\Big(\hat{\pi}_n(Sv)\big(1-\hat{\pi}_n(Sv)\big)\Big)\Big(\hat{\pi}_n(B)\big(1-\hat{\pi}_n(B)\big)\Big)}.$$

Then the set

$$\left\{ \left| \hat{\pi}_n(Bv) - \hat{\pi}_n(Sv)\hat{\pi}_n(B) \right| > u_\alpha \frac{\hat{\sigma}_n(B,v)}{\sqrt{n}} \right\} \tag{1}$$

is a reject region with asymptotic level α of the null hypothesis H_0 against the hypothesis $H \setminus H_0$.

esting the structure of a sequence (2)

- The choice of the most suitable B and v depends on the distribution (p_u) which is unknown in practice.
- For the test of H_0 against H_m, $m \geq 1$, Reinert et al. proposed to make use of Pearson statistics

$$X^2 \stackrel{\text{def}}{=} \sum_{u,v \in \mathcal{A}} \frac{\big(N(uv) - N(u\cdot)N(\cdot v)/(n-1)\big)^2}{N(u\cdot)N(\cdot v)/(n-1)}, \tag{2}$$

where $N(uv)$ counts the occurences of uv in the sequence, $N(u\cdot)$ (resp. $N(\cdot v)$) is the number of 2-letter words beginning with u (resp. ending with v). This test can be seen as a generalized likelihood ratio test.

esting the structure of a sequence (3)

Let $q_\alpha(d)$ be the $(1-\alpha)$-quantile of the chi-square distribution $\chi^2(d)$.

Then, for any partition $\mathcal{P}$ of S, with $|\mathcal{P}| = K > 1$, where $|\mathcal{P}|$ denotes the size of the partition, the set

$$\left\{ \sum_{\substack{B\in\mathcal{P} \\ v\in\mathcal{A}}} \frac{n\big(\hat{\pi}_n(Bv) - \hat{\pi}_n(B)\hat{\pi}_n(Sv)\big)^2}{\hat{\pi}_n(B)\hat{\pi}_n(Sv)} > q_\alpha\Big((d-1)(K-1)\Big) \right\},$$

is a reject region with asymptotic level α, of a test of H_0 against $H \setminus H_0$.

esting the structure of a sequence (4)

- H_0 : "$U = u_1 \dots u_n$ is an i.i.d. sequence"
- H_m : "U is a Markov chain of order m"
- H : "U is a stationary ergodic sequence"

Moreover, the set

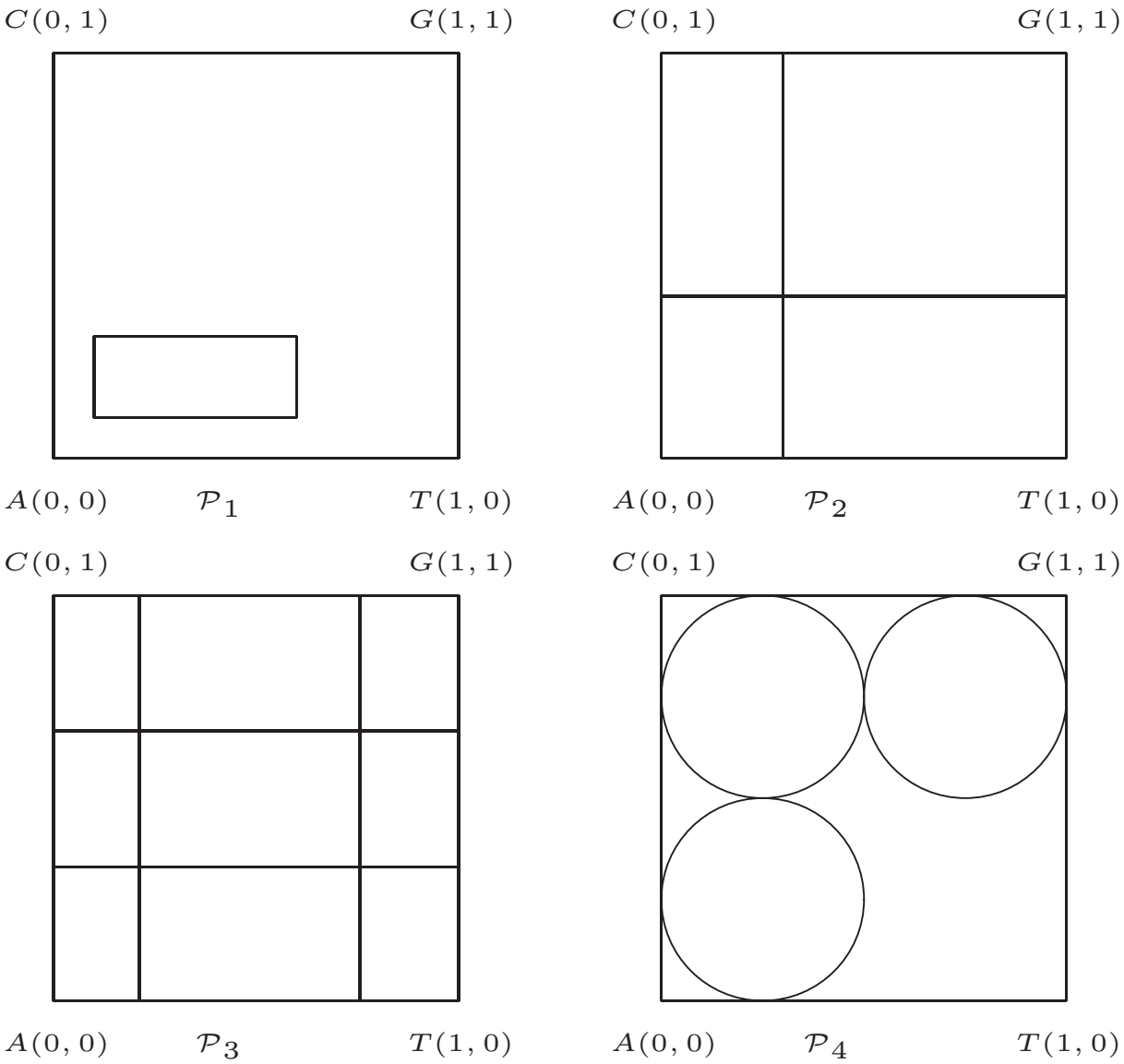
$$\left\{ \sum_{\substack{wu \in \mathcal{A}^m \times \mathcal{A} \\ B \in \mathcal{P}}} \frac{n\big(\hat{\pi}_n(Sw)\hat{\pi}_n(Bwu) - \hat{\pi}_n(Swu)\hat{\pi}_n(Bw)\big)^2}{\hat{\pi}_n(Sw)\hat{\pi}_n(Swu)\hat{\pi}_n(Bw)} > q_\alpha\Big(d^m(d-1)(K-1)\Big) \right\},$$

is a reject region with asymptotic level α, of a test of H_m against $H \setminus H_m$.

Consistence

- Assume that $H \setminus H_0$ (resp. $H \setminus H_m$) holds, and let $B \subset S$, $w \in \mathcal{A}^m$ and $v \in \mathcal{A}$ be such that
  - $$\pi(Bv) \neq \pi(B)\pi(Sv).$$
  - resp.
    $$\pi(Sw)\pi(Bwu) \neq \pi(Bw)\pi(Swu).$$
- If B is one of the set forming the partition, then the test is asymptotically consistent.

Numerical experiments : partitions



The 4 different partitions of the square $[0,1[^2$ arbitrary chosen for the test.

Numerical experiments : results (1)

1000 sequences of length n of Markov chains of order m with random transition matrix, for various values of n have been generated.

						
order	n	Pearson	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$
0	100	4.2%	3.6%	3.4%	2.5%	3.8%
	500	6.1%	4.8%	4.8%	3.8%	5.2%
	1000	5.0%	5.6%	6.2%	5.3%	4.9%
	10000	6.5%	4.8%	5.1%	7.7%	5.8%
1	100	86.4%	12.9%	51.1%	28.9%	69.1%
	500	100%	54.2%	98.7%	94.5%	99.6%
	1000	100%	70.9%	99.9%	99.0%	100%
	10000	100%	97.6%	100%	100%	100%
5	1000	8.6%	6.8%	8.6 %	8.4 %	8.2%
	10000	54.6%	28.7%	55.6%	85.3%	51.6%
	80000	99.4%	84.5%	99.6%	100%	99.6%
2 mixed	500	5.8%	16.5%	49.9%	76.8%	13.7%
	1000	7.0%	26.9%	73.7%	95.1%	20.4%
	10000	7.3%	73.2%	99.8%	100%	74.5%
5 mixed	80000	5.8%	29.7%	76.7%	85.8%	50.5%

Fraction of cases when H_0 is rejected.

Numerical experiments : results (2)

				
order	n	Pearson	$\mathcal{P}_2$	$\mathcal{P}_3$
1	500	5.2%	5.4%	3.6%
	1000	4.1%	5.2%	5.4%
	10000	5.3%	6.4%	6.0%
2	100	60.8%	25.8%	1.3%
	500	100%	98.4%	91.6%
	1000	100%	99.9%	99.6%
	10000	100%	100%	100%
4	500	22.8%	20.2%	30.3%
	1000	51.0%	44.3%	69.7%
	5000	99.9%	99.6%	100%
	10000	100%	100%	100%
3 mixed	500	5.4%	29.6%	48.5%
	1000	8.1%	55.6%	79.3%
	5000	8.3%	96.4%	99.8%
	10000	6.3%	99.0%	100%

Fraction of cases when H_1 is rejected.

Numerical experiments : comments

- The choice of the partition is crucial.
- The reject of long dependence Markov chains increases with the number of sets forming the partition.
- In the special case of markov chains of order-m given by the aggregation of m indepedent markov chains of order 1, the CGR-based test behaves pretty well. This illustrates the strength of the CGR : it does not impose any constraint on the input sequence besides stationarity.

Generalization to several partitions (1)

- In order to minimize the problem related to the choice of a peculiar partition, the following test is a generalization of the previous test, with a collection of partitions. The idea is inspired from the generalization of Bonferroni method described in Baraud et al (2003).
- For any set $\mathcal{P} = \{\mathcal{P}_1, \dots \mathcal{P}_p\}$ of partitions with $K_j \stackrel{\text{def}}{=} |\mathcal{P}_j|$, H_0 is rejected as soon as one of the partition $\mathcal{P}_j$ satisfies

$$\left\{ \sum_{\substack{B \in \mathcal{P}_j \\ v \in \mathcal{A}}} \frac{n\big(\hat{\pi}_n(Bv) - \hat{\pi}_n(B)\hat{\pi}_n(Sv)\big)^2}{\hat{\pi}_n(B)\hat{\pi}_n(Sv)} - q_{\alpha_j}\Big((d-1)(K_j-1)\Big) > 0 \right\}.$$

  One has to chose α_j in order to have a global level α.

Generalization to several partitions (2)

- The set

$$\sup_{1\leq j\leq p}\left\{\sum_{\substack{B\in\mathcal{P}_j\\ wu\in\mathcal{A}^m\times\mathcal{A}}} n\frac{\left(\hat{\pi}_n(Sw)\hat{\pi}_n(Bwu)-\hat{\pi}_n(Swu)\hat{\pi}_n(Bw)\right)^2}{\hat{\pi}_n(Sw)\hat{\pi}_n(Swu)\hat{\pi}_n(Bw)}-q_{\alpha_j}(\delta_j)>0\right\},$$

where $\delta_j \stackrel{\text{def}}{=} d^m(d-1)(K_j-1)$, is a reject region with asymptotic level α, of a test of the null hypothesis H_m against the hypothesis $H\setminus H_m$.

Generalization to several partitions (3)

1000 sequences of length n of Markov chains of order m with random transition matrix, for various values of n have been generated.

order	n	$\{\mathcal{P}, \mathcal{P}_2\}$	$\{\mathcal{P}, \mathcal{P}_4\}$	$\{\mathcal{P}, \mathcal{P}_2, \mathcal{P}_4\}$
M0	100	3.9 (4.8/5.3)	4.8 (4.0/5.3)	3.7 (4.0/5.3)
	500	4.7 (4.8/5.0)	5.5 (4.8/5.0)	4.9 (4.8/5.0)
	1000	4.0 (4.9/5.0)	5.8 (5.0/5.8)	4.9 (4.9/5.8)
	10000	4.3 (4.7/5.0)	4.7 (4.7/5.0)	4.4 (4.7/5.0)
M1	100	83.0 (54.8/86.4)	82.3 (39.2/86.4)	81.3 (39.2/86.4)
	500	99.7 (97.7/99.8)	99.8 (96.3/99.8)	99.8 (96.3/99.8)
	1000	100 (99.9/100)	100 (99.6/100)	100 (99.6/100)
	5000	100 (100/100)	100 (100/100)	100 (100/100)
M5	1000	8.3 (8.5/8.9)	10.6 (8.5/10.7)	7.4 (8.5/10.7)
	10000	61.7 (54.5/55.0)	84.1 (55.0/83.5)	79.6 (54.5/83.5)
	80000	100 (99.5/99.5)	100 (99.5/100)	100 (99.5/100)
2 mixed	500	41.0 (7.6/48.6)	72.6 (7.6/78.2)	70.3 (7.6/78.2)
	1000	66.4 (5.9/73.2)	92.1 (5.9/94.3)	91.1 (5.9/94.3)
	10000	99.7 (6.8/99.9)	100 (6.8/100)	100 (6.8/100)

Fraction of cases when H_0 is rejected (in %).

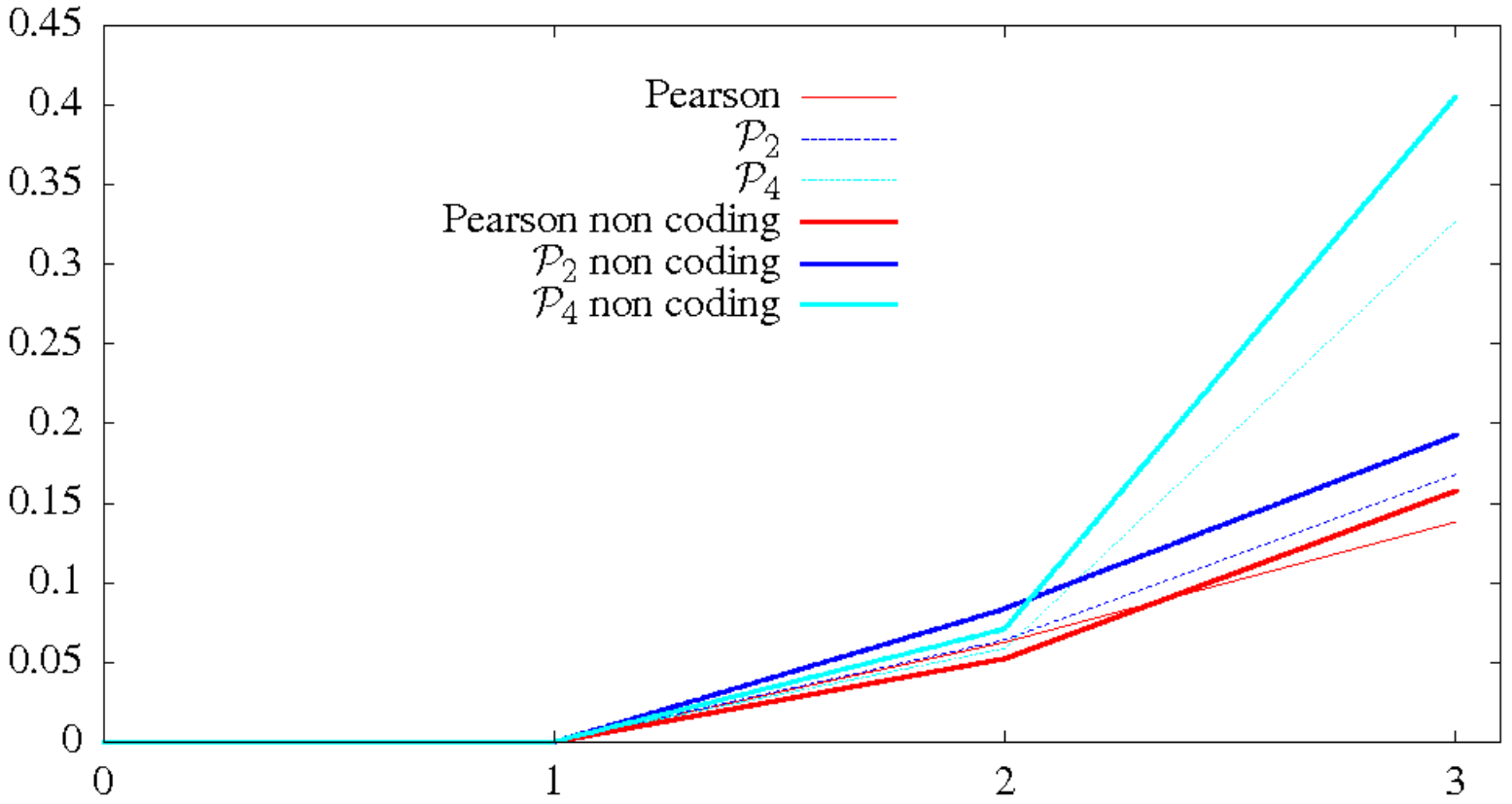
Application to genomes (1)

- First, we compare the structure of non-coding sequences with complete sequences of Homo Sapiens and Mus Musculus, and compute the probability of acceptance of H_m as a function of m.

- In a second set of runs involving Homo Sapiens sequences and three partitions, the probability of acceptance is computed as a function of the length and of the order.

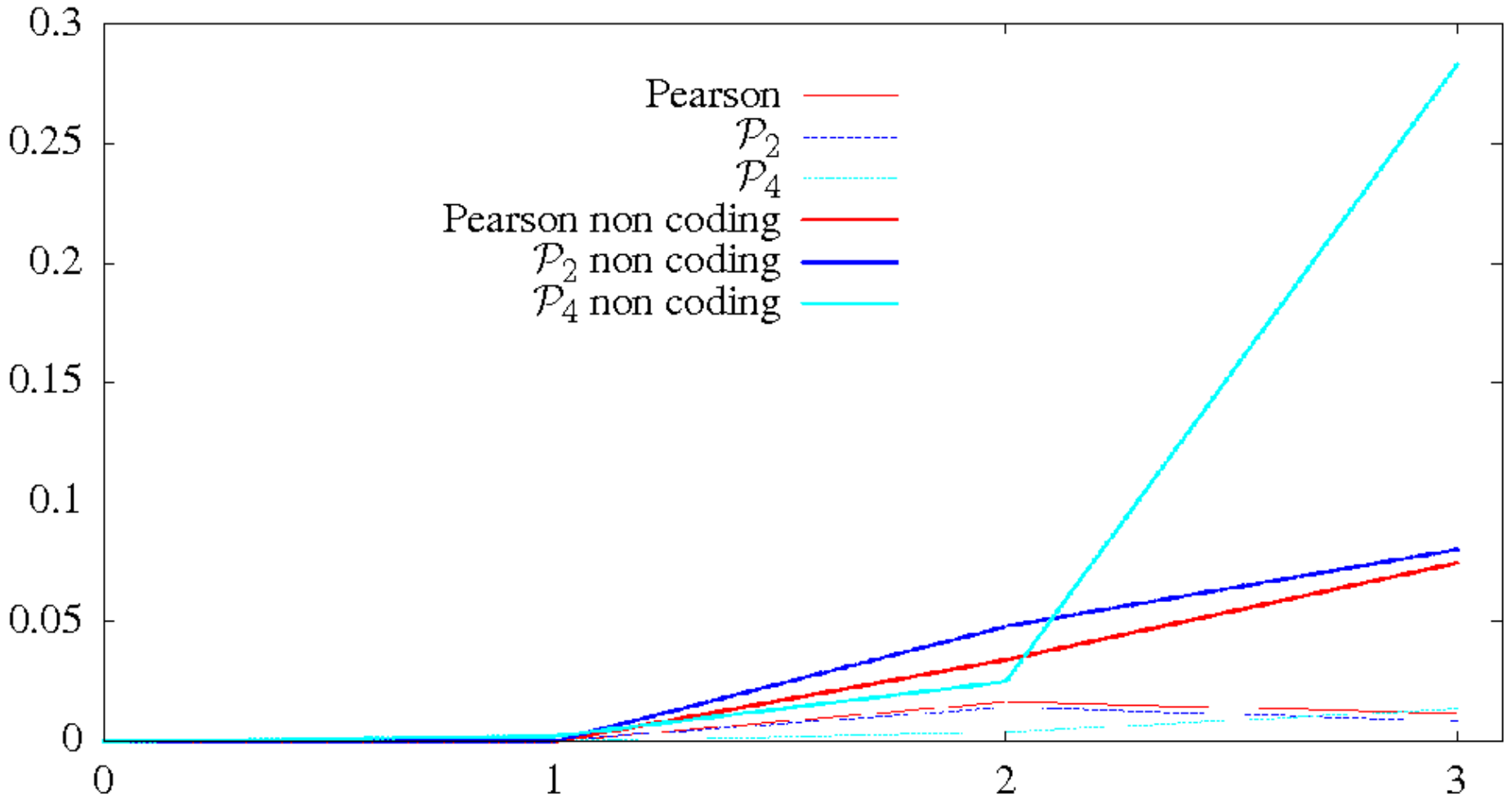
- In a last set of experiments, the probability of acceptance is computed for several sequences of length 10000 taken from various genomes as a function of m.

Application to genomes (2)



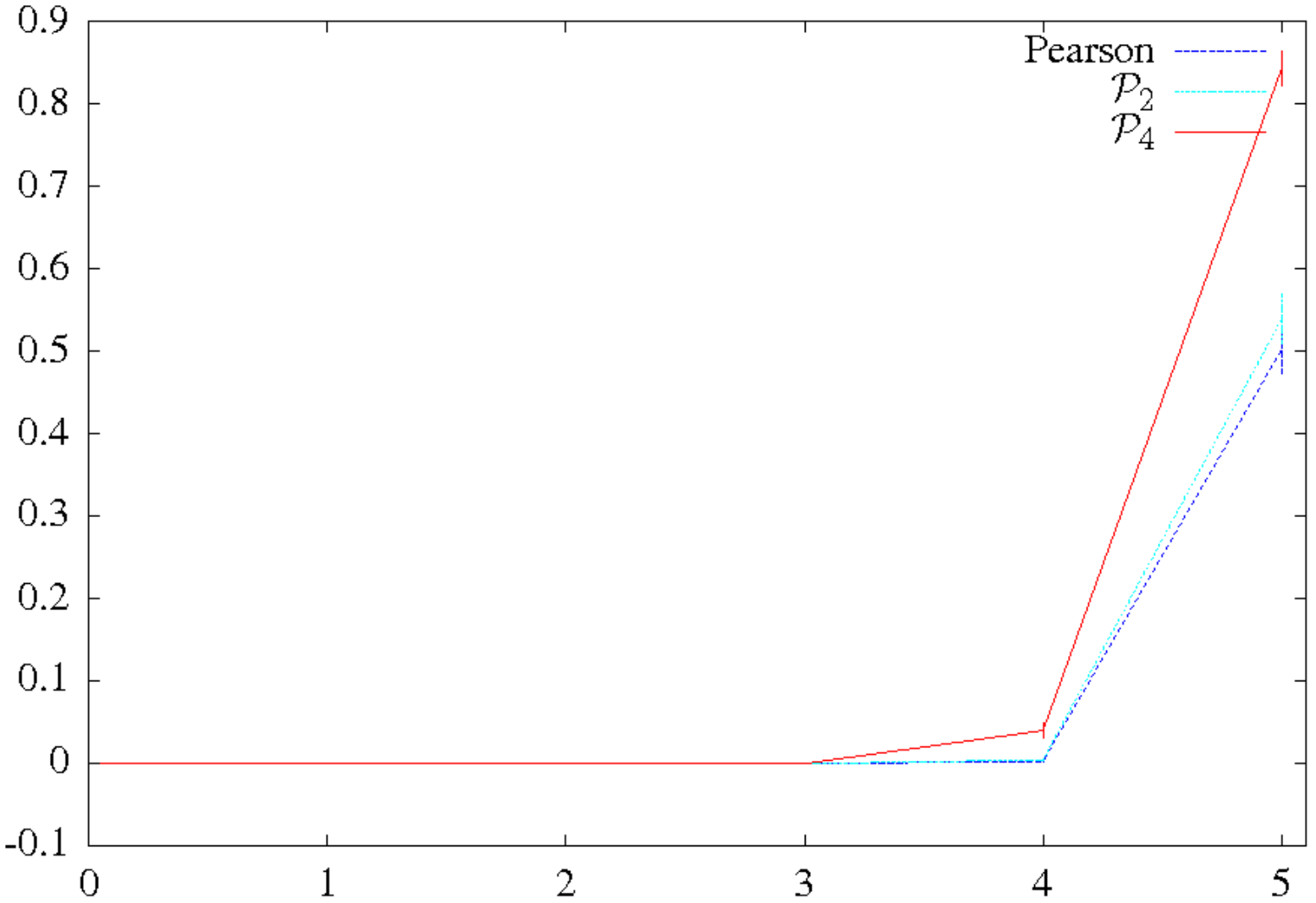
Probability of acceptance of H_m as a function of m for Homo Sapiens. This is the mean of 100 sequences of length 10000.

Application to genomes (3)



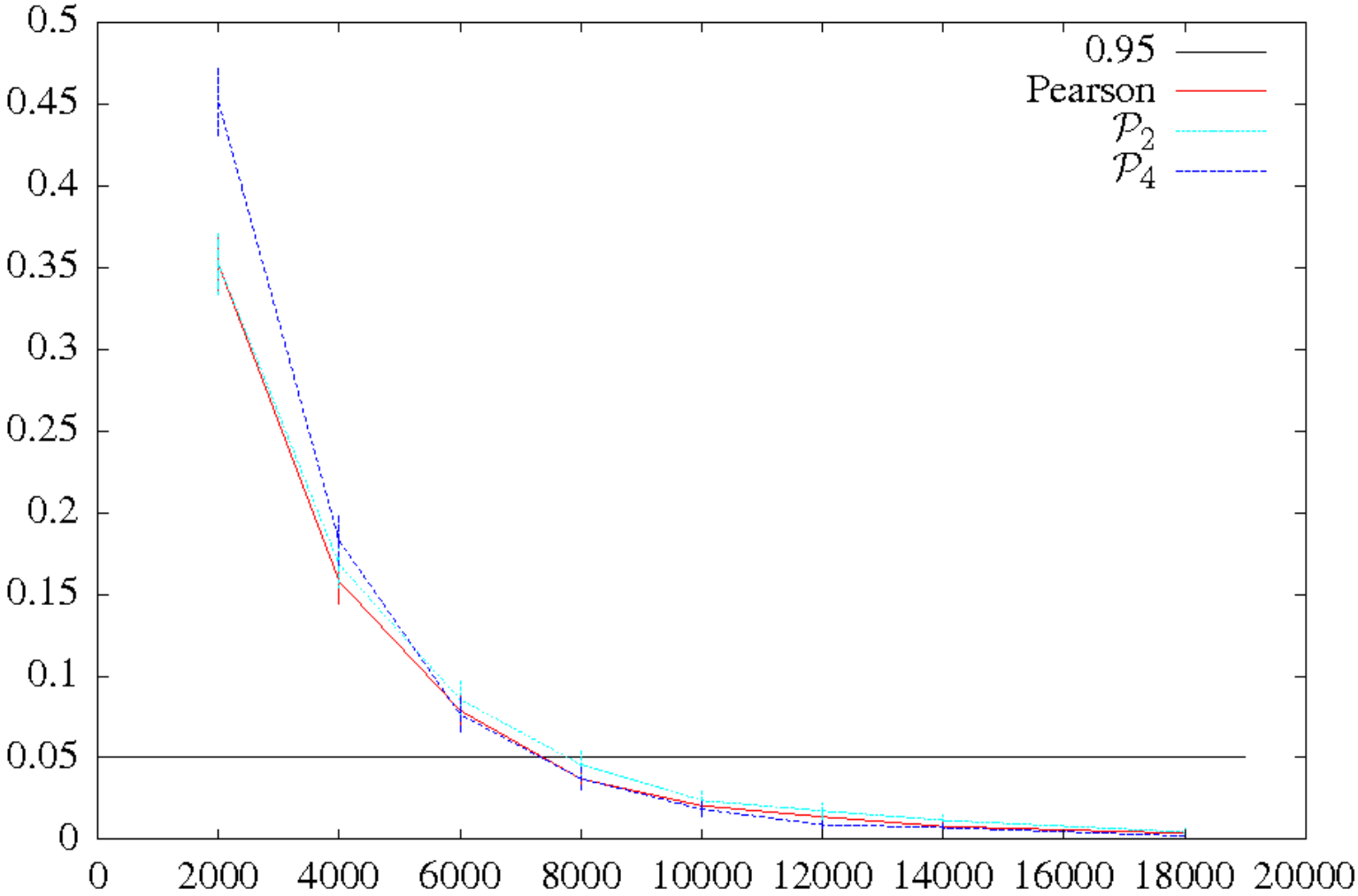
Probability of acceptance of H_m as a function of m for Mus musculus. This is the mean of 100 sequences of length 10000.

Application to genomes (4)



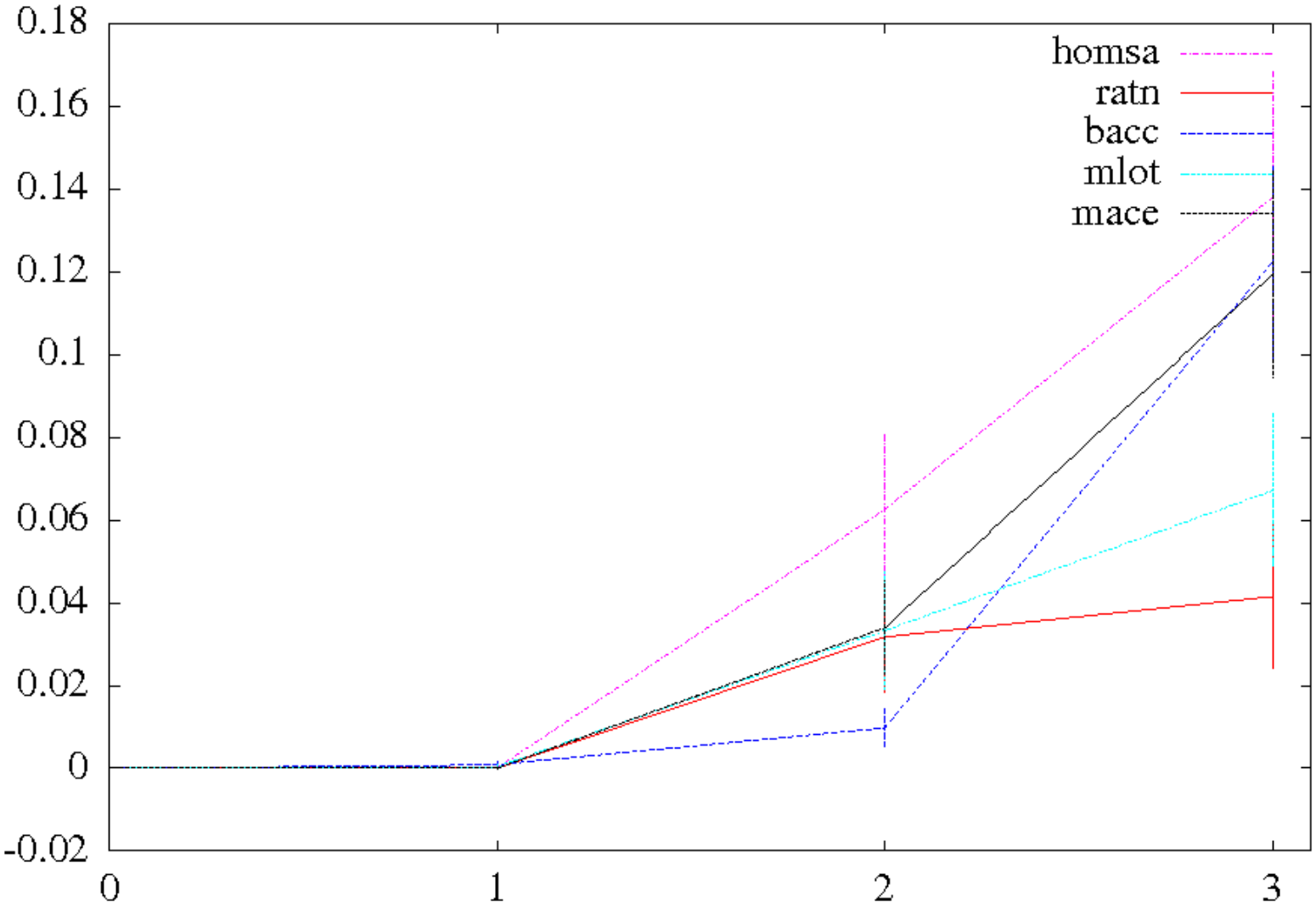
Probability of acceptance of H_m functions of m for Homo Sapiens sequences and different partitions. The mean probability is computed for 1000 sequences of length 50000 and various orders.

Application to genomes (5)



Probability of acceptance of the length n for Homo Sapiens sequences and different partitions. The mean probability is computed for the order $m = 2$ and various lengths.

Application to genomes (6)



Probability of acceptance of H_m as a function of m for various species, using the test built from $\mathcal{P}_2$ on 200 sequences of length 10000. We report the mean probability for each order.

Genomic signature

- Deschavanne et al. use the CGR with a view to characterizing and classifing species.

- Karlin and Burge and Karlin and Mràzek use profile of dinucleotide relative abundance values as a genomic signature, and build taxonomy trees based on these profiles.

- Dinucleotide relative abundance for nucleotide uv can be written as

$$\rho_{uv} \stackrel{\text{def}}{=} \frac{\pi(Suv)}{\pi(Su)\pi(Sv)}.$$

- It is therefore tempting to define a more general CGR-based relative abundance as

$$\rho(B, v) \stackrel{\text{def}}{=} \frac{\pi(Bv)}{\pi(B)\pi(Sv)},$$

Genomic signature (2)

- Karlin and Mràzek, Campbell et al. use the profile

$$\hat{\rho}_{uv} \stackrel{\text{def}}{=} \frac{\hat{\pi}_n(Suv)}{\hat{\pi}_n(Su)\hat{\pi}_n(Sv)}, \qquad \forall u, v \text{ nucleotides}$$

  where the empirical measures are computed from the sequence concatenated with its inverted complement.

- local stability

- advantages
  - ➢ alignments of long sequences are generally not feasible,
  - ➢ a tree based on distance matrix of profiles is independent of the genome segment of 50kb used in its construction,
  - ➢ the signature pervades both coding and non coding DNA.

Genomic signature (3)

- Empirical CGR-based relative abundance

$$\hat{\rho}(B, v) \stackrel{\text{def}}{=} \frac{\hat{\pi}_n(Bv)}{\hat{\pi}_n(B)\hat{\pi}_n(Sv)}.$$

- From N (resp. N') sequences of a species Σ (resp. Σ'), $\hat{\rho}_i(B, v)$ (resp. $\hat{\rho}'_i(B, v)$) is associated to sequence i. For a given partition $\{B_1, \ldots, B_K\}$, the CGR-based relative abundance difference is defined as

$$\delta(\Sigma, \Sigma') = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{N'} \sum_{j=1}^{N'} \frac{1}{K} \sum_{k=1}^{K} \frac{1}{4} \sum_{v \in \mathcal{A}} \left| \hat{\rho}_i(B_k, v) - \hat{\rho}'_j(B_k, v) \right|.$$

- Karlin and Mràzek build taxonomy trees from the matrices of measures of differences. The trees are generated with Neighbor Joining method (Saitou and Nei, Perrière and Gouy).

Perspectives

- Optimal choice of the partition
- Test for hidden markov models
- Model selection