Statistical modeling of tumorigenesis

Modèles statistiques du développement de tumeurs cancéreuses

Soutenance de thèse de

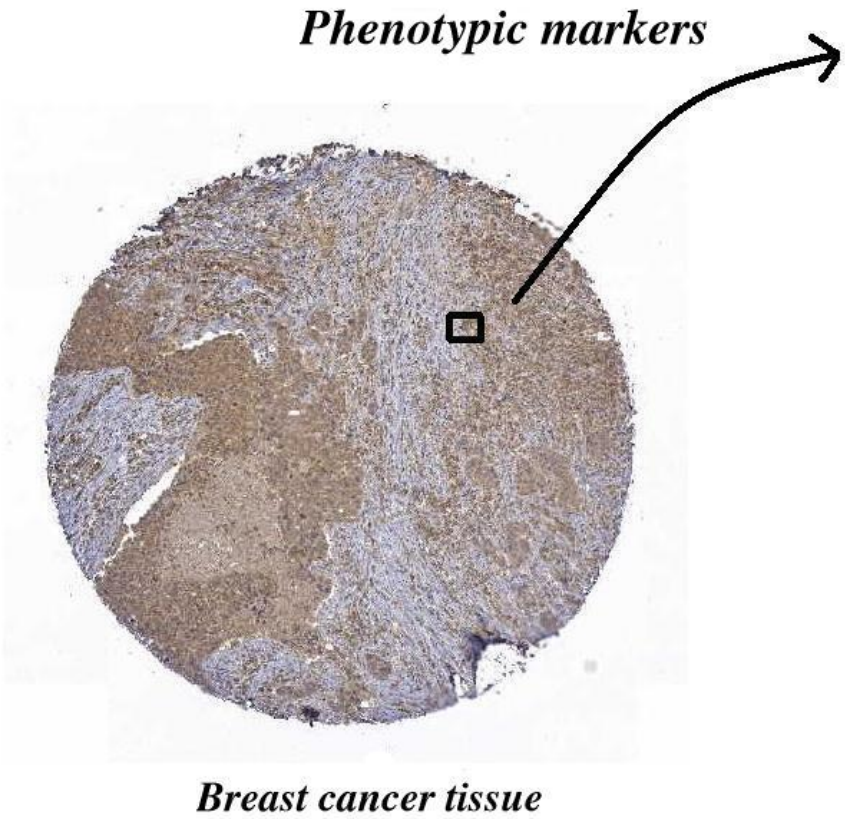
Mathieu Emily

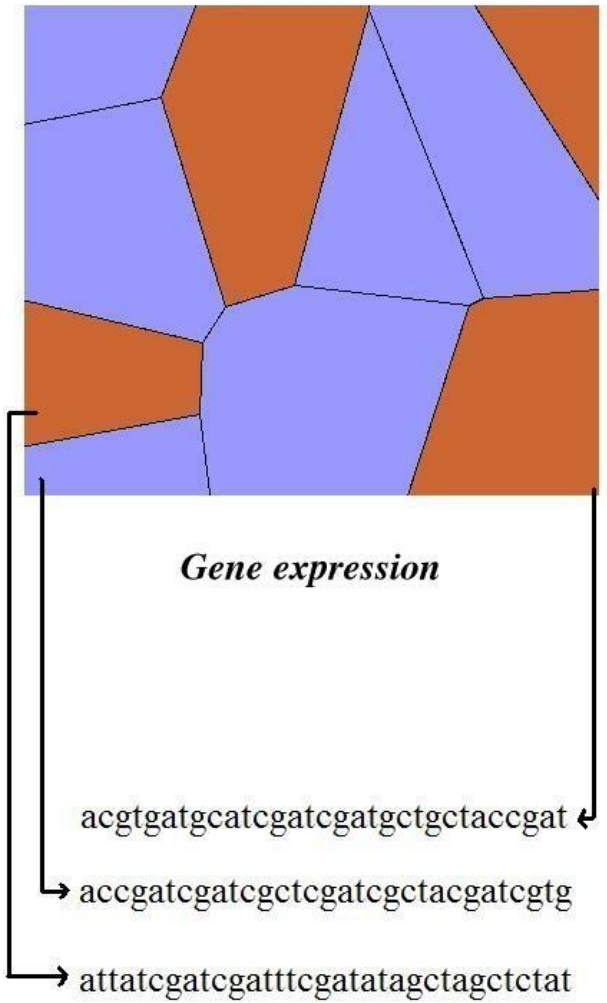
22 Septembre 2006

préparée au
Laboratoire TIMC - Grenoble

- Cancer is a multistage process with at least 3 major steps:
  - Initiation,
  - Promotion,
  - Progression.
- Many mathematical models are dedicated to the study of cancer development (Komarova, 2005):
  - Modeling in the context of epidemiology,
  - Modeling of tumor growth,
  - Modeling of cancer initiation as somatic evolution.

- This work focuses on mathematical models for cancerous tissues at the initiation and the promotion stage.
- It provides statistical tests for early detection of cancer based on:
  - Gene expression measures within a tissue (promotion step).
  - Cell DNA sequences within a tissue (initiation step).


Phenotypic markers
Breast cancer tissue


Gene expression
acgtgatgcatcgatcgatgctgctaccgat
accgatcgatcgctcgatcgctacgatcgtg
attatcgatcgatttcgatatagctagctctat
DNA Sequences
Genotypic markers

- Cell adhesion in cancer at the promotion step. Lower expression of Cellular Adhesion Molecules (CAMs) are correlated with:
  - Breast cancer (Berx and Van Roy, 2001).
  - Lung cancer (Bremnes *et al.*, 2002).

- Cell adhesion in cancer at the promotion step. Lower expression of Cellular Adhesion Molecules (CAMs) are correlated with:
  - Breast cancer (Berx and Van Roy, 2001).
  - Lung cancer (Bremnes *et al.*, 2002).
- Genetic instability at the initiation step.
  - Less accuracy in DNA repair.
  - Genetic instability $\overset{20 \text{ years}}{\longrightarrow}$ tumor manifestation (Bielas and Loeb, 2005).
  - Hereditary Colon Cancer implicates MSH2, MSH6 and MLH1 genes (Fishel *et al.*, 1993).
  - Breast Cancer implicates BRCA1 and BRCA2 genes (Wooster and Weber, 2003).

This thesis contribution

- A model for studying adhesion properties between contiguous cells using gene expression data.
  - Marked Point Processes framework.
  - Estimating an adhesion strength parameter characterizing the tissue.

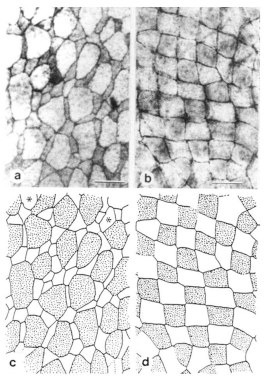
This thesis contribution

- A model for studying adhesion properties between contiguous cells using gene expression data.
  - Marked Point Processes framework.
  - Estimating an adhesion strength parameter characterizing the tissue.
- A model of genetic instability using DNA sequences.
  - Coalescent models of gene genealogies.
  - Testing the occurrence of genetic instability by estimating a raised mutation rate parameter.

Part A. *Gibbsian spatial point process for tissue organization*

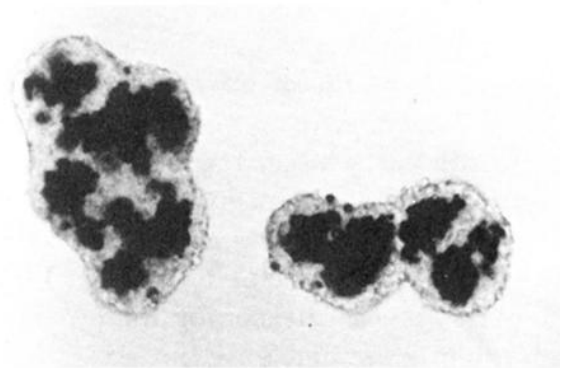
Spatial development of biological tissues

- Cell patterns play a major role in many biological processes:
  - Embryogenesis,
  - Morphogenesis,
  - Tumorigenesis.
- Gene expression data may help to characterize cell patterns within a tissue:



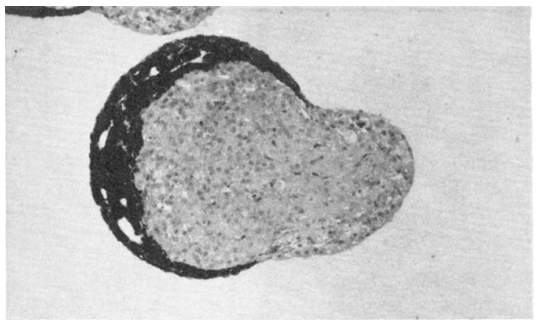
Checkerboard

(Honda *et al.*, 1986)



Cell Sorting

(Armstrong, 1989)

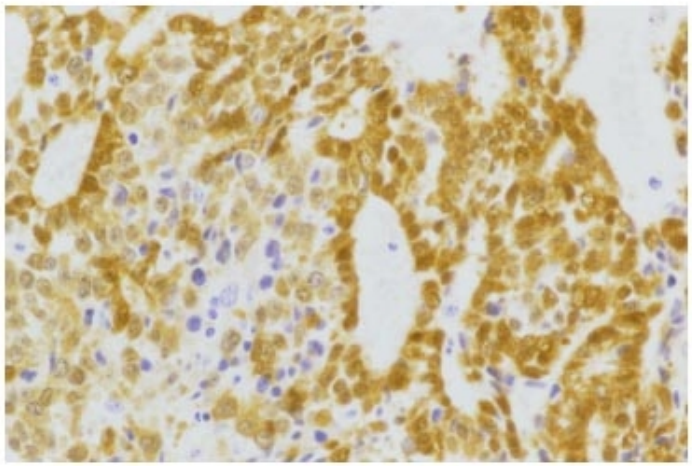


Engulfment

(Armstrong, 1989)

Cell adhesion - DAH

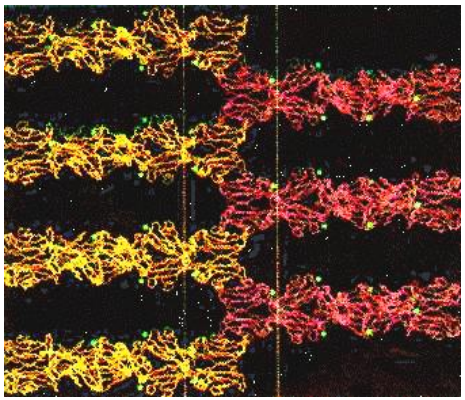
- The Differential Adhesion Hypothesis (DAH) is one of the most robust hypothesis (Steinberg, 1962):
  - Adhesion is function of differential expression of Cellular Adhesion Molecules (CAMs).
  - Cell arrangements minimize the adhesion energy,
- Among the CAMs, the Cadherin-Catenin complex is known to be deeply implicated in tumorigenesis.



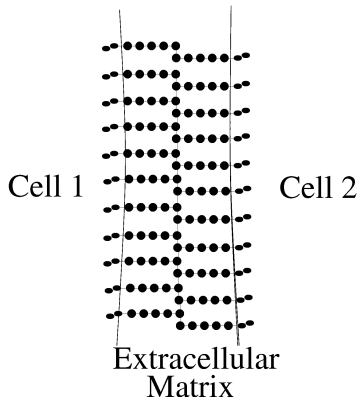
$\beta-$Catenin gene expression in human hepatocellular carcinoma (Lin, 2003).

Cadherin-catenin complex

- A zipper-like structure (Shapiro, 1995):



(a)



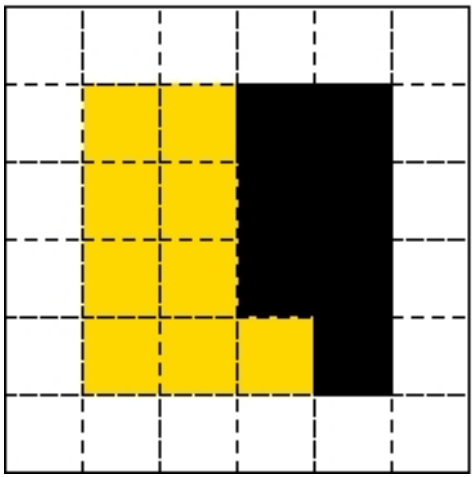
(b)

Crystal structured model (a) and picture (b) of linear zipper adhesion between cadherin-catenin complexes of two cells.

- The adhesion energy is function of the membrane separating contiguous cells.

Mathematical models of the Differential Adhesion Hypothesis (DAH) are classified according their geometry (Brodland, 2004):

- Lattice models (Mochizuki *et al.* 1996, Takano *et al.* 2002).
- Centroid models (Honda *et al.* 1996, Honda *et al.* 2000).
- Vertex models (Nagai *et al.* 1998, Honda *et al.* 2004)
- Sub-cellular lattice model: Graner and Glazier's model (1992).



Example of Graner and Glazier's model configuration with two cells

Graner and Glazier's model

- Each cell, denoted by σ, is a set of pixels and each pixel (i, j) is characterized by a type $\tau_{(\sigma_{ij})}$ (3 different types: ℓ for light cells, d for dark cells and M for extracellular matrix).
- The Energy, H_{GG}, is defined as:

$$H_{GG} = H_{Adh} + Constraint$$

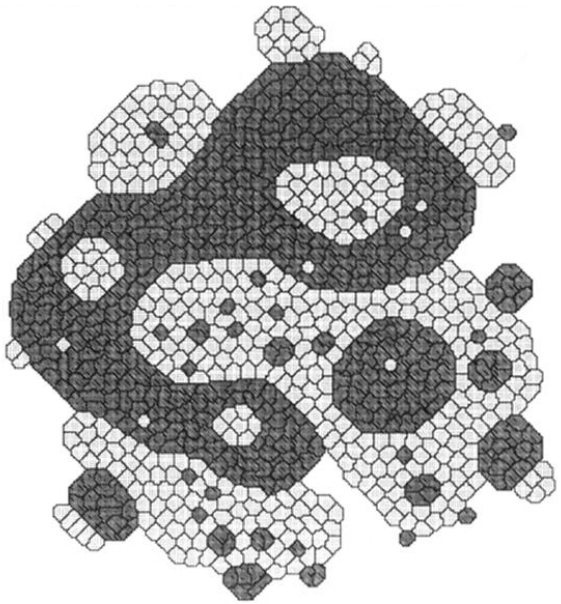
The adhesion term is an extension of the Potts interaction function:

$$H_{Adh} = \sum_{(i,j)\sim(i',j')} J\left(\tau(\sigma_{ij}), \tau(\sigma_{i'j'})\right)\left(1 - \delta_{\sigma_{ij},\sigma_{i'j'}}\right)$$

and $$Constraint = \sum_{\sigma} C(area(\sigma))$$

Graner and Glazier's model

$\ell = light$, $d = dark$ and $M = medium$.



Example of GG's configuration using $J_{\ell,\ell} = 14$, $J_{d,d} = 14$, $J_{\ell,d} = 29$, $J_{\ell,M} = J_{d,M} = 16$ (Glazier and Graner, 1993).

Graner and Glazier's model

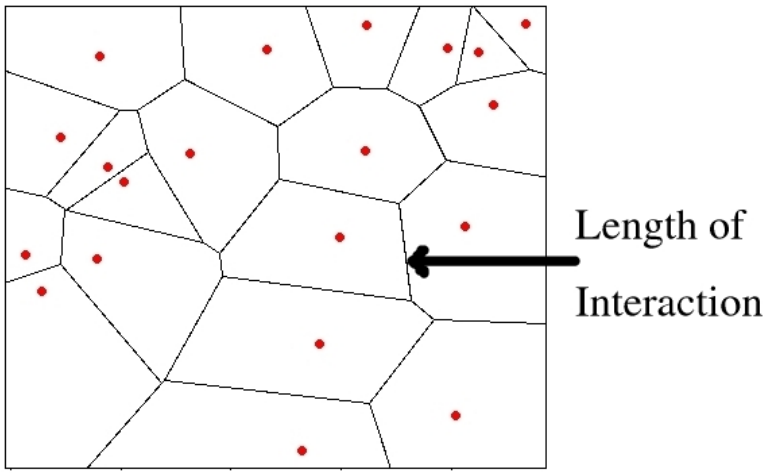
- GG's model has been extended to cancerous processes:
  - Avascular tumor growth (Scott *et al.*, 1999).
  - Tumor invasion (Turner and Sherratt, 2002).
- Despite the large success of this model, there exist some limitations:
  - Loss of cell connexity.
  - Algorithm sensitive to the lattice discretization.
  - No convergence for the algorithm.
  - Lack of mathematical framework for estimating parameters.

Objectives of our model

- Continuous geometry for cells.
- Simulation algorithm with good convergence properties.
- Statistical framework for estimating the strength of adhesion: marked point processes theory.

Geometrical modeling

- According to Honda's studies (Honda 1978, 1983), cells can be modeled by a Dirichlet tiling based on cell nuclei.



Example of a tissue modeled by a Dirichlet tiling

Energy functional

$$H_{CC}(\underline{\varphi}) = H_{Adh} + Constraint$$

with:

$$H_{Adh} = \sum_{i \sim j} \text{length}(i,j) J(\tau_i, \tau_j)$$

and:

$$Constraint = \sum_i C(\text{area}(x_i))$$

and where:

- $\underline{\varphi} = \{\underline{x_1}, \ldots, \underline{x_n}\}$ and $\underline{x_i} = (x_i, \tau_i)$, x_i is the center of the cell i and τ_i the type of cell i ($\underline{x_i}$ is marked point).

Adhesion strength parameter

- With respect to the Poisson process, the density of a configuration $\underline{\varphi}$ can be written as:

$$f(\underline{\varphi}) \propto \exp(-\theta H_{CC}(\underline{\varphi}))$$

  where θ quantifies the strength of adhesion within a tissue.
- Estimating the strength of adhesion is of particular interest.

Mathematical study

Theorem

Let $H_{CC}(\underline{\varphi})$ be the energy function of the following form:

$$H_{CC}(\underline{\varphi}) = \sum_{i \sim j} g(\text{length}(i,j)) J(\tau_i, \tau_j) + \sum_i C(\text{area}(x_i))$$

Assume that g, J and C are bounded on $\mathbb{R}$. Then, there exists a Gibssian marked marked point process that satisfies the local specifications derived from H_{CC}.

Mathematical study - sketch of the proof

Let $E(\underline{x}, \underline{\varphi}) = H_{CC}(\underline{\varphi} \cup \underline{x}) - H_{CC}(\underline{\varphi})$ denotes the energy needed to insert a new point $\underline{x}$ in a configuration $\underline{\varphi}$.

Proposition - Sufficient conditions for existence (Bertin *et al.*, 1999)

- **Local Stability**. For all $\underline{x}$ and $\underline{\varphi}$, it exists $K > 0$ such as:

$$E(\underline{x}, \underline{\varphi}) > -K$$

- **Quasilocality**. For all $\underline{x}$, $\underline{\varphi}$ and Δ bounded set:

$$|E(\underline{x}, \underline{\varphi}) - E(\underline{x}, \underline{\varphi_\Delta})| < \varepsilon(d(x, \Delta^c))$$

where $\varepsilon(x) \longrightarrow 0$ when $x \longrightarrow \infty$.
Then, there exists a Gibssian marked marked point process that satisfies the local specifications derived from H_{CC}.

Algo: Insertion-Deletion Metropolis-Hastings

Algorithm

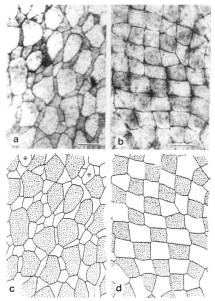
- If $Random < 1/2$: Insertion
  - Random choice of x_{n+1} and τ_{n+1}.
- else: Deletion
  - Uniform choice of a point within the configuration.
- Acceptance probability: $p = \min[1, \exp(-\theta(\Delta H))]$

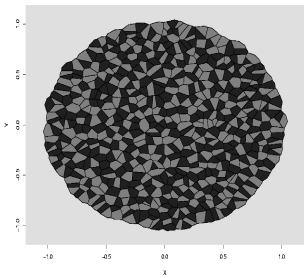
Theorem

Under the same conditions (g, J and C bounded), the Markov chain generated by the Metropolis-Hastings algorithm is ergodic (Harris-Recurrent and aperiodic).

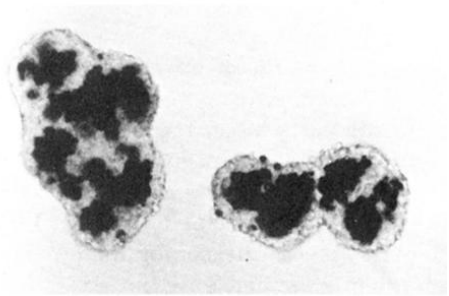
Proof: Using local stability and results from Geyer and Møller (1994).

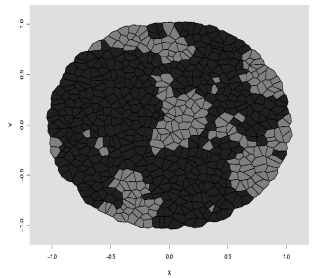
Examples of simulations



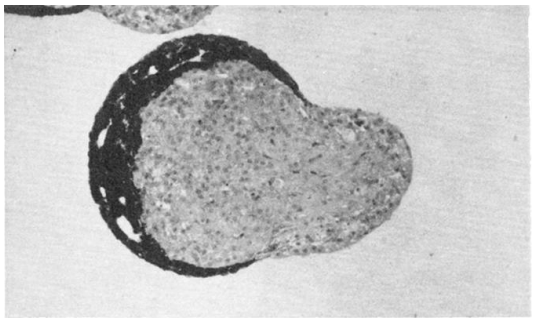


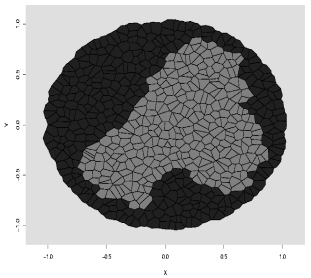
(a)
Checkerboard
(Honda *et al.*, 1996)





(b)
Clustering
(Armstrong, 1989)

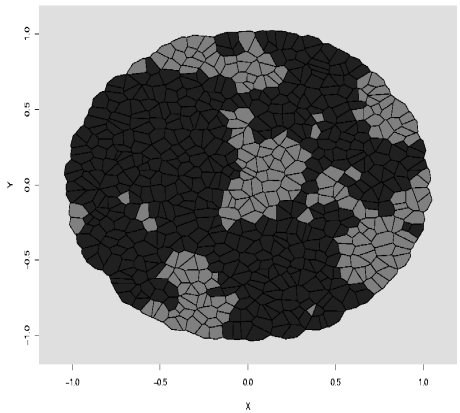


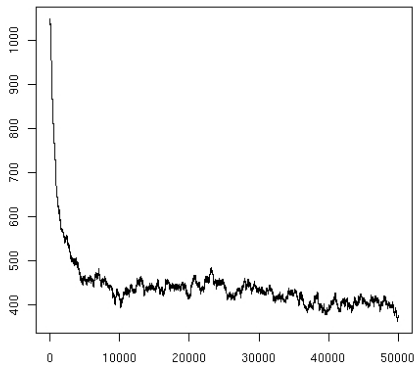


(c)
Engulfment
(Armstrong, 1989)

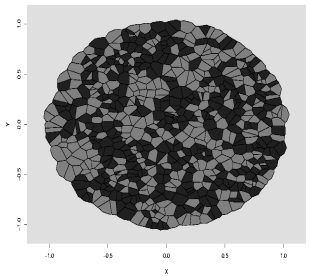
The algorithm performances

- Fast thanks to local the properties of insertion and deletion in the Dirichlet tessellation.
- Convergence: 50000 iterates for around 1000 cells starting from a random configuration (180 sec).

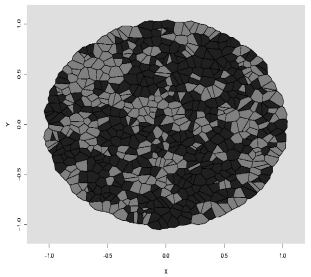




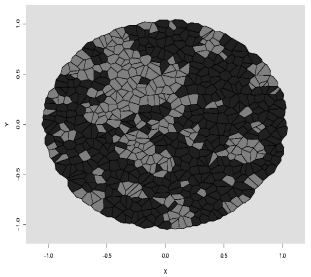
Clustering - $J(\tau_1, \tau_1) = 0$, $J(\tau_2, \tau_2) = 0$ and $J(\tau_1, \tau_2) = 1$



$\theta = 1$

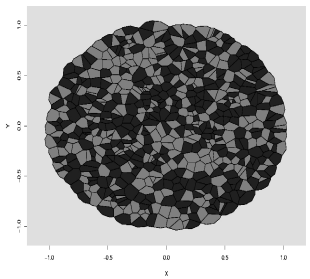


$\theta = 5$

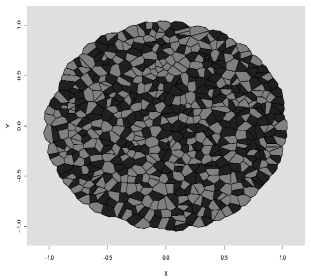


$\theta = 10$

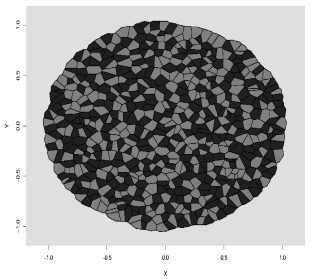
Checkerboard - $J(\tau_1, \tau_1) = 1$, $J(\tau_2, \tau_2) = 1$ and $J(\tau_1, \tau_2) = 0$



$\theta = 1$



$\theta = 5$



$\theta = 10$

The characteristic patterns emerge for large θ.

Estimation: Conditional Pseudo-Likelihood

Let Λ be a bounded set in $\mathbb{R}$. Conditional to the point locations, we have:

$$\mathrm{PL}_{\mathrm{C}}^{\Lambda}(\theta) = \prod_{i} \mathrm{Prob}(\tau_i | \varphi, \tau \backslash_{\{\tau_i\}}, \theta)$$

Definition

An estimator for the adhesion strength parameter (θ) is given by:

$$\widehat{\theta_{\mathrm{C}}} = \mathrm{argmax}_{\theta} \mathrm{PL}_{\mathrm{C}}^{\Lambda}(\theta)$$

Estimation: Pseudo-Likelihood

According to Jensen and Møller (1991), Pseudo-likelihood estimation for Gibbsian point processes is defined by:

$$\begin{aligned} \mathrm{PL}^{\Lambda}(\theta) &= \exp\left(-\int_{\Lambda}\int_{M}\exp(-H_{CC}(\underline{x}|\underline{\varphi}))d\tau_x dx\right) \\ &\quad \prod_{\underline{x}\in\underline{\varphi}_{\Lambda}}\exp\left(-H_{CC}(\underline{x}|\underline{\varphi}\backslash\underline{x})\right) \end{aligned}$$

Definition

An estimator for the adhesion strength parameter (θ) is given by:

$$\widehat{\theta} = \mathrm{argmax}_{\theta}\mathrm{PL}^{\Lambda}(\theta)$$

	Checkerboard		Clustering	
	Mean	Variance	Mean	Variance
$\theta = 1$	0.98	0.70	1.03	0.4
$\theta = 5$	5.01	0.57	4.94	0.94
$\theta = 10$	10.47	1.20	9.80	1.00
$\theta = 15$	14.58	2.22	15.03	1.20

Mean and Variance from 100 replicates for $\widehat{\theta_{\mathrm{C}}}$

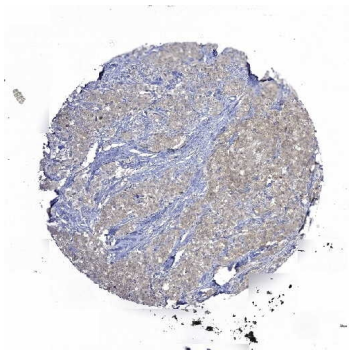
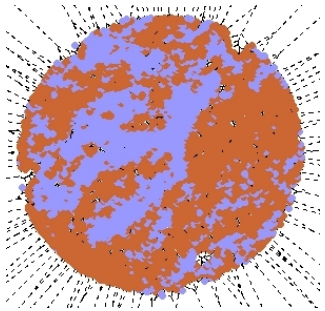
	Checkerboard		Clustering	
	Mean	Variance	Mean	Variance
$\theta = 1$	1.01	0.91	0.97	1.3
$\theta = 5$	5.17	1.12	4.93	1.05
$\theta = 10$	10.24	2.24	10.30	1.38
$\theta = 15$	15.43	3.87	15.58	2.55

Mean and Variance from 100 replicates for $\widehat{\theta}$

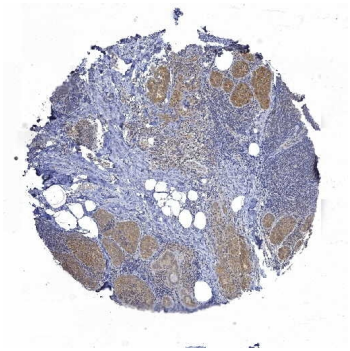
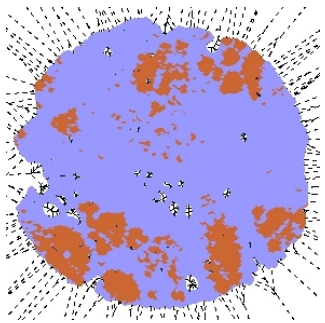
Comments

- Conditional and unconditional estimators seem to be weakly biased.
- Variances increase with θ.
- The conditional estimator is computationally faster than the unconditional estimator.
- Theoretically, $\widehat{\theta}$ should be better than $\widehat{\theta_C}$.
- In practice, we observe the reverse (integral approximations may be a problem).

- Data: breast cancer - Two diseased tissues.
- Clustering pattern: $J_{1,1} = 0, J_{2,2} = 0$ and $J_{1,2} = 1$

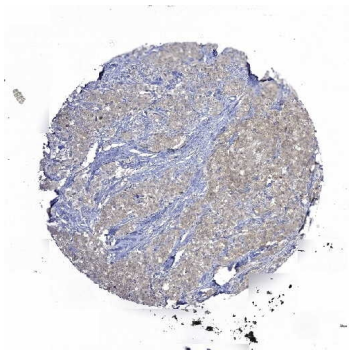
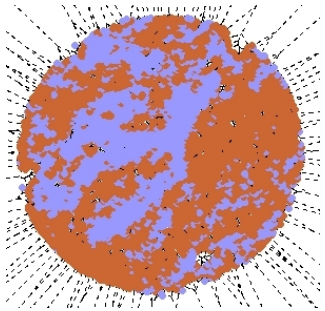
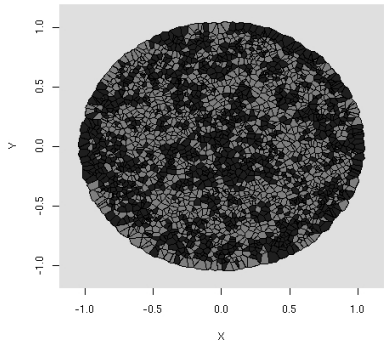



$\widehat{\theta_C} = 14.9$ and $\widehat{\theta} = 15.4$

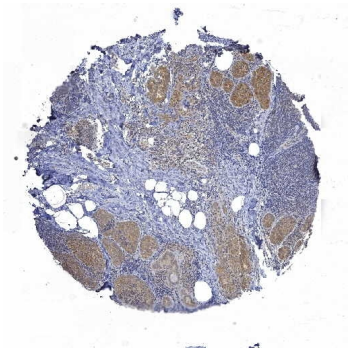
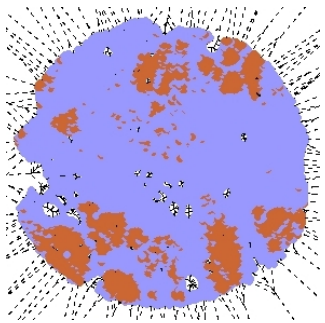
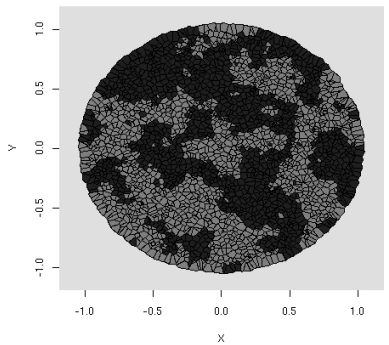



$\widehat{\theta_C} = 31.9$ and $\widehat{\theta} = 33.7$

- Data: breast cancer - Two diseased tissues.
- Clustering pattern: $J_{1,1} = 0, J_{2,2} = 0$ and $J_{1,2} = 1$

$\widehat{\theta_C} = 14.9$ and $\widehat{\theta} = 15.4$

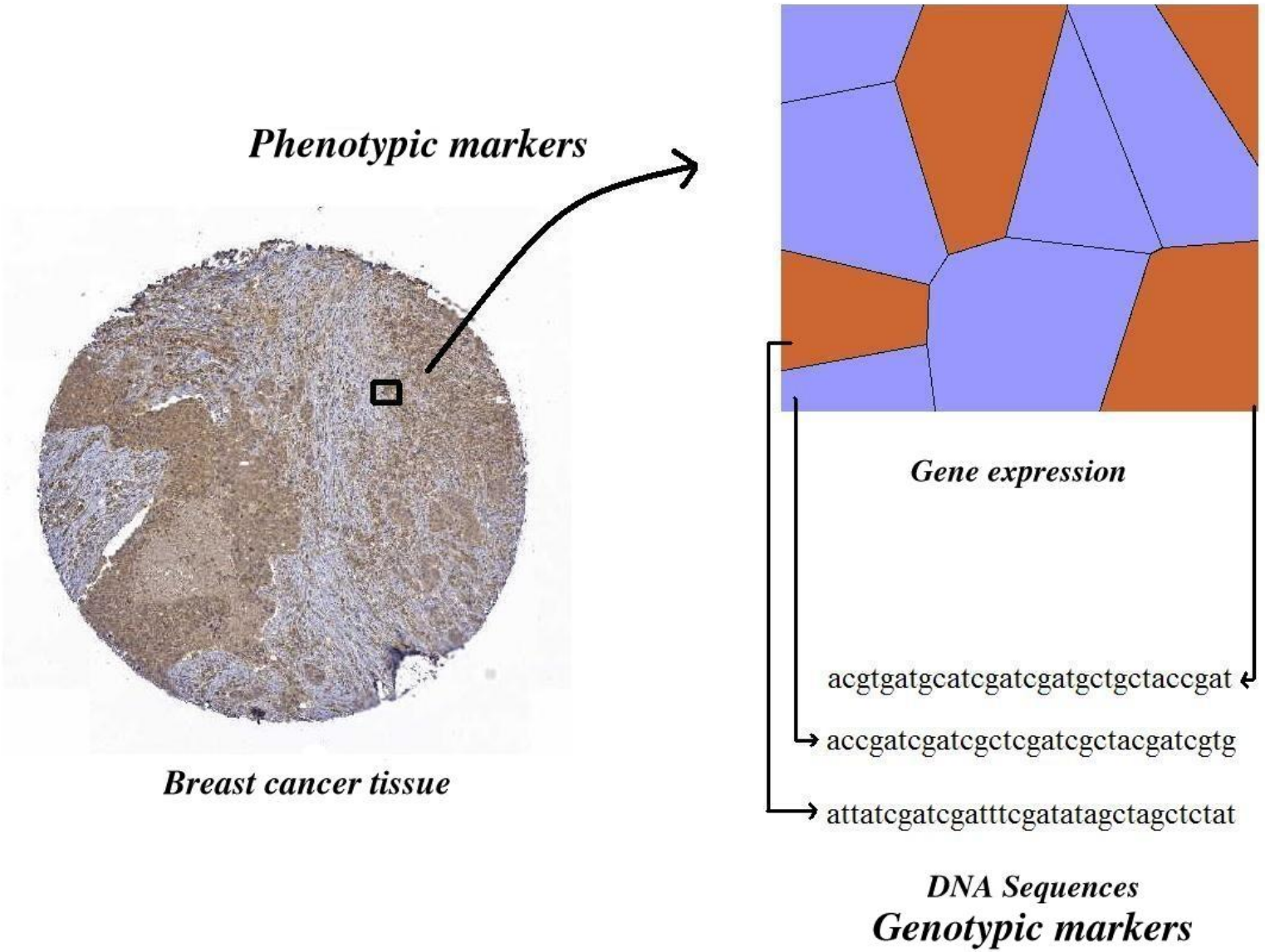




$\widehat{\theta_C} = 31.9$ and $\widehat{\theta} = 33.7$

Comments

- Capacity to discriminate between various cell patterns.
- Simulations with estimated parameters provide patterns consistent with real data.

Part B. *Conditional coalescent model for genetic instability*


Phenotypic markers
Breast cancer tissue
Gene expression
acgtgatgcatcgatcgatgctgctaccgat
accgatcgatcgctcgatcgctacgatcgtg
attatcgatcgatttcgatatagctagctctat
DNA Sequences
Genotypic markers

Genetic instability in tumors

- Theory introduced by Loeb *et al.* in 1974.
- Tumors are characterized by a large number of mutations.
- A loss of genome stability functions occurs early in tumor development.
- Genetic instability as the initiating event is still a matter of debate (Loeb *et al.*, 2003). Alternative theories are:
  - Aneuploidy (Duesberg *et al.*, 1998).
  - Clonal selection (Tomlinson and Bodmer, 1999).

Loss of MMR (Mismatch Repair)

- More than 130 genes are involved in DNA repair (Anderson *et al.*, 2001).
- Alteration of genes involved in:
  - fidelity of DNA replication.
  - efficacity of DNA repair.
- Consequence: increase from 10 to 10000 fold in the mutation rate (Bhattacharyya *et al.* 1994, Tomlinson *et al.*, 1996).
  - Overall mutation rate in somatic human cells: 1.4×10^{-10} nucleotides per cell per division (Loeb, 1991).
  - Genetic instability $10^{-10} \longrightarrow 10^{-6}$ shift.

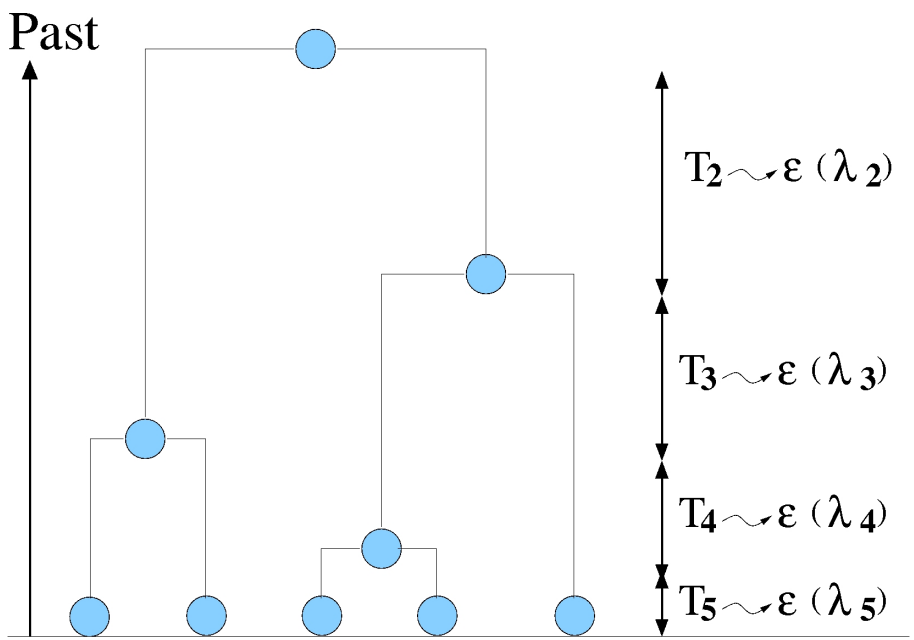
Modeling hypothesis - Loss of MMR

- The sample of genes has two mutation rates. Some cells have a normal mutation rate and the others have a raised mutation rate.
- The number of affected cells is unknown.
- Cell genealogy can be modeled by a coalescent process arising as the limit of a Moran process (Moran 1962, Kingman 1982).
- Neutrality: mutation process is independent on the genealogical process.

- Our goal: testing the occurrence of the loss of MMR.

Neutral coalescent (Kingman 1982, Hein *et al.* 2005)

- Let T_i for $i = 2, \ldots, n$ denote the inter-coalescing times and assume that T_i's are independent and of exponential distribution of parameter $\lambda_i = \frac{i(i-1)}{2}$.



Example of a coalescent tree with $n = 5$

Mutations model

- Infinitely-many sites model (Watterson, 1975).
- Mutations occur according to independent Poisson processes of rate $\theta/2$ along the branches of the tree.
  - $\theta = 4N\mu$ where μ is the mutation rate per base per mitotic division and N is the total number of cells.
- Classical unbiased estimators for θ: Watterson's estimator and Tajima's estimator.

Watterson's estimator

- Let S be the number of segregation sites.
- S is equal to the total number of mutations under the *infinitely many sites* model.

Sequence ♯1	acagttacat
Sequence ♯2	agagctacat
Sequence ♯3	agagttgcgt
	-●--●-●-●-

Example with three DNA sequences where $S = 4$

- Watterson's estimator for θ is defined as:

$$\widehat{\theta_W} = \frac{2S}{\mathrm{E}[L]} = \frac{S}{\sum_{i=1}^{n-1} 1/i},$$

where $L = \sum_{i=2}^{n} iT_i$ is the total length of the tree.

Tajima's estimator

- Let $\Pi(i,j)$ be the number of pairwise differences between sequence i and sequence j.
- Tajima's estimator for θ is defined as:

$$\widehat{\theta_T} = \frac{2}{n(n-1)} \sum_{i<j} \Pi(i,j)$$

Seq1 vs Seq2	Seq1 vs Seq3	Seq2 vs Seq3
acagttacat	acagttacat	agagctacat
agagctacat	agagttgcgt	agagttgcgt

Example with three DNA sequences where $\widehat{\theta_T} = 2.67$ $(\widehat{\theta_W} = 2.67)$

Back to genetic instability - Modeling constraints

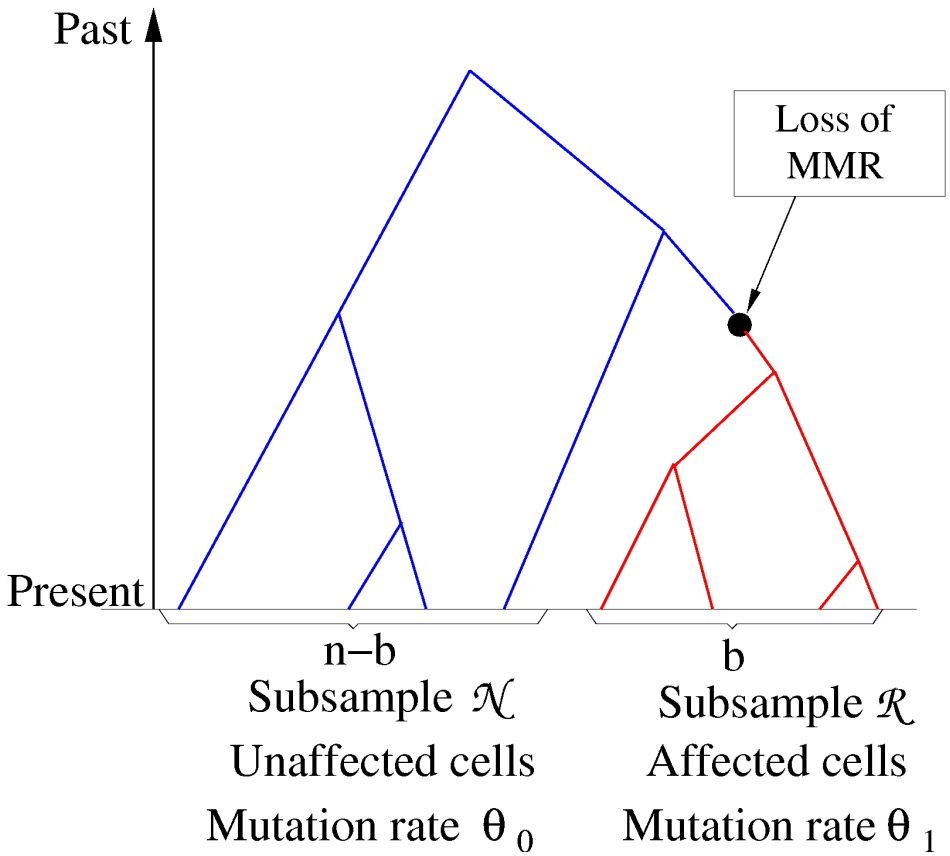
- The event "Loss of MMR", denoted by Δ, occurs once and only once in the genealogy of the sample.

  $\Rightarrow$ Constraints on mutation rates along the Coalescent tree.

- Our sample is divided into 2 subsamples:
  - $\mathcal{N}$ in which the mutation rate θ_0 is "normal",
  - $\mathcal{R}$ in which the mutation rate θ_1 is "raised" $(\theta_1 > \theta_0)$.

  $\Rightarrow$ Topological constraints on the Coalescent tree.

- Our goal: correcting Watterson's and Tajima's estimators for the raised mutation rate knowing the normal mutation rate.



- Mutations follow Poisson processes of rates:
  - $\theta_0/2$ along the blue branches.
  - $\theta_1/2$ along the red branches.

Frequency spectrum

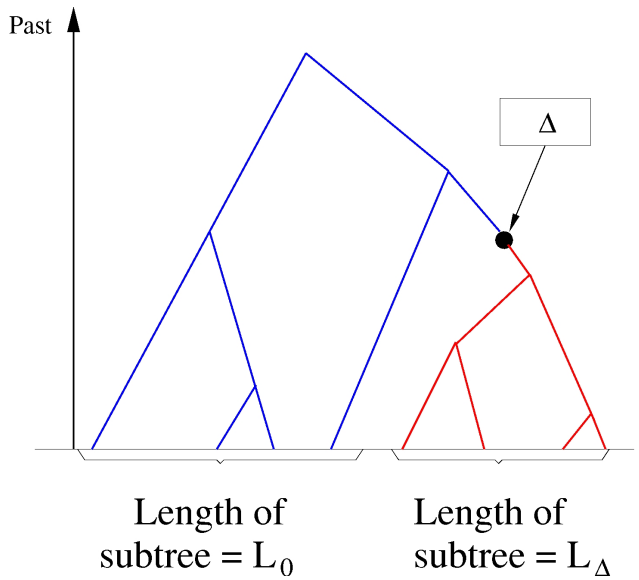
- The genealogy of the sample is a *conditional coalescent tree* (Griffiths and Tavaré 1998, Wiuf and Donnelly 1999).
- The number B of descendants of Δ has the following distribution:

$$P(B = b) = \frac{1}{bH_{n-1}} \quad b = 1, \ldots, n-1.$$

where H_n is the n^{th} harmonic number.

Correction of Watterson's estimator

- S_n, the number of segregating sites, is a random variable equal to the total number of mutations.
- Two contributions for S_n, S_{0_n} and S_{1_n} where:
  - $\mathbf{E}[S_{0_n}] = \mathbf{E}[L_0]\theta_0/2$
  - $\mathbf{E}[S_{1_n}] = \mathbf{E}[L_\Delta]\theta_1/2$



An unbiased estimator of θ_1 is:

$$\widehat{\theta_{1,W}} = \frac{S_n - \mathbf{E}[L_0]\theta_0/2}{\mathbf{E}[L_\Delta]/2}$$

Correction of Watterson's estimator - $\mathbf{E}[L_\Delta] = \mathbf{E}[L_1] + \mathbf{E}[\eta_n]$

Proposition

Let L_1 be the total length of the red sub-genealogy (Griffiths and Tavaré, 2003):

$$\mathbf{E}[L_1|B=b] = \sum_{j=2}^{n-b+1} p_j^\Delta \sum_{k=j+1}^{n} \frac{2}{k(k-1)} c_{jk},$$

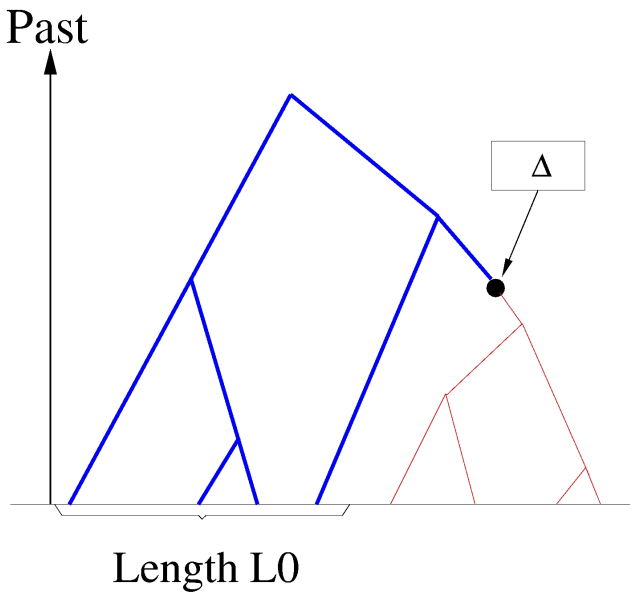
Proposition

Let η_n be the time that separates the MRCA of red sub-sample to Δ (Wiuf and Donnelly, 1999):

$$\mathbf{E}[\eta_n|B=b] = 2 \sum_{k=2}^{n-b+1} \frac{p_k^\Delta}{k}.$$

Correction of Watterson's estimator - L_0

- $\mathbf{E}[L_0]$ and $\mathbf{E}[L_0|B]$ are unknown in the litterature.
- $L_0 = L - L_\Delta$ where:
  - L is the total length of the tree.
  - L_Δ is the length of the red subtree.



Correction of Watterson's estimator - L

Proposition

Assume that the mutation Δ has $B = b$ descendants.
In a conditional coalescent tree we have:

$$\frac{1}{2}\mathbf{E}[L|B=b] = H_{n-1} + \frac{1}{H_{n-1}} \sum_{k=2}^{n-b+1} \frac{p_k^{\Delta}}{b(k-1)}$$

Sketch of the proof: $L = \sum_{i=2}^{n} iT_i$ where T_i are the inter-coalescing times.

Sketch of the proof

Theorem - Inter-coalescing times in a conditional coalescent tree

Assume that the mutation Δ has $B = b$ descendants. The joint probability distribution of $(T_2, \ldots, T_n)$ has a density equal to:

$$f(t_2, \ldots, t_n) = \sum_{k=2}^{n-b+1} p_k^{\Delta} \lambda_k t_k \prod_{\ell=2}^{n} f_\ell(t_\ell)$$

where $f_\ell(t_\ell)$ is the probability density function of the exponential distribution of rate λ_ℓ and:

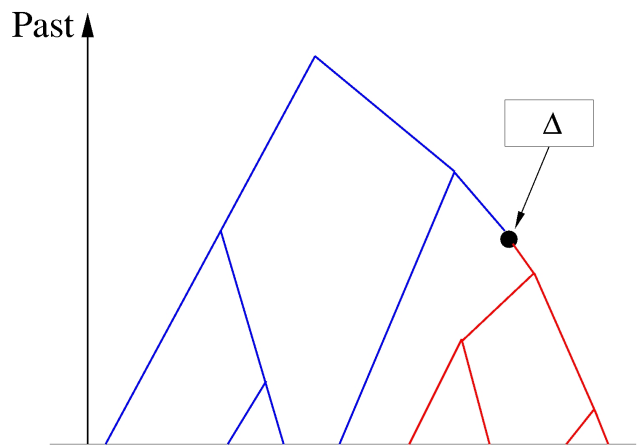
$$p_k^{\Delta} = \binom{n-k}{b-1} \binom{n-1}{b}^{-1} \quad k = 2, \ldots, n-b+1$$

Correction of Tajima's estimator

- Mean number of pairwise differences between genes: Π.
- An unbiased estimator of θ_1 is:

$$\widehat{\theta_{1,T}} = \frac{\Pi - C_n\theta_0}{D_n}$$

- C_n and D_n were founded by considering 3 average coalescing times between two sequences:
  - within $\mathcal{R}$ (in the red subtree),
  - within $\mathcal{N}$ (in the blue subtree),
  - one in each subsample.



Correction coefficients

n	5	10	15	20	25	30	35	40	45
A_n	2.171	2.693	3.024	3.265	3.455	3.612	3.747	3.864	3.967
B_n	0.595	0.68	0.713	0.732	0.746	0.756	0.764	0.771	0.776

Tables for $A_n = \mathbf{E}[L_0]/2$ *and* $B_n = \mathbf{E}[L_\Delta]/2$

n	5	10	15	20	25	30	35	40	45
C_n	0.996	1.019	1.021	1.02	1.02	1.019	1.019	1.018	1.018
D_n	0.253	0.218	0.199	0.187	0.178	0.171	0.166	0.161	0.156

Tables for C_n *and* D_n

Algorithm for simulating a conditional coalescent tree

Algorithm

- Draw B according to the *frequency spectrum*.
- Draw J_Δ, the number of ancestors at the time Δ occurs (Cf. Stephens, 2000).
- Draw the total number of ancestors at the time the subsample $\mathcal{R}$ first has r ancestors ($1 < r < b - 1$) (Tavaré, 2004).
- Sample T_ℓ from the exponential ditribution Gamma$(1, \lambda_\ell)$, for $\ell \neq J_\Delta$ and T_{J_Δ} from the Gamma distribution Gamma$(2, \lambda_{J_\Delta})$.

Statistical errors of $\widehat{\theta_{1,W}}$ and $\widehat{\theta_{1,T}}$ for $\theta_0 = 1$

($N = 2.5 \times 10^9$ and $\mu = 10^{-10}$)

	$\theta_1 = 10$		$\theta_1 = 100$		$\theta_1 = 1000$	
n	**E**	SD	**E**	SD	**E**	SD
10	9.9	12.0	97.4	112.4	947.5	1109.7
30	10.2	12.8	102.9	126.1	1060.3	1286.1
50	10.4	13.5	102.0	131.7	1045.7	1235.9

Expectation and Standard Deviation for $\widehat{\theta_{1,W}}$ using 1000 replicates.

	$\theta_1 = 10$		$\theta_1 = 100$		$\theta_1 = 1000$	
n	**E**	SD	**E**	SD	**E**	SD
10	9.9	13.7	107.3	133.9	1006.2	1243.5
30	9.5	15.5	100.9	147.9	1040.0	1589.5
50	10.3	17.6	106.5	164.6	1039.7	1598.1

Expectation and Standard Deviation for $\widehat{\theta_{1,T}}$ using 1000 replicates.

Statistical errors of $\widehat{\theta_{1,W}}$ and $\widehat{\theta_{1,T}}$ for $\theta_0 = 1$

- Watterson and Tajima's corrected estimators are unbiased.
- They behave like the classical Watterson and Tajima's estimator (high variance).
- The corrected estimators may not be consistent.
- Watterson's corrected estimator seems to has less variance than Tajima's corrected estimator.

Testing the absence of the "Loss of Mismatch Repair"

- H_0: Absence of Δ.
- H_1: Occurrence of Δ and $\theta_1 > \theta_0$.

Assume the knowledge of the sample genealogy and that the data set consists of all intercoalescing times (T_k). The likelihood ratio can be described as:

$$r = \frac{L(\mathrm{H}_1)}{L(\mathrm{H}_0)} = \sum_{k=2}^{n-b+1} \lambda_k p_k^{\Delta} t_k$$

Powers for type I error: $\alpha = 0.05$:

- $1 - \beta = 0.2$ when $b \approx n$ and dropped to 0.1 when $b/n \approx 0.5$, where b is the number of affected cells.

Testing the absence of Δ (LMMR) - $\theta_0 = 1$

- H_0: Absence of Δ.
- H_1: Occurrence of Δ and $\theta_1 > \theta_0$.

n	$\theta_1 = 10$	$\theta_1 = 100$	$\theta_1 = 1000$
20	0.44	0.74	0.90
40	0.42	0.73	0.88

Power of tests for $\widehat{\theta}$ estimator

n	$\theta_1 = 10$	$\theta_1 = 100$	$\theta_1 = 1000$
20	0.44	0.69	0.84
40	0.34	0.64	0.79

Power of tests for Π estimator

Testing the occurrence of Δ (LMMR) - $\theta_0 = 1$

- H_0: Occurrence of Δ and $\theta_1 > \theta_0$.
- H_1: Absence of Δ.

n	$\theta_1 = 10$	$\theta_1 = 100$	$\theta_1 = 1000$
20	0.06	0.18	0.70
40	0.11	0.24	0.59

Power of tests for $\widehat{\theta_{1,W}}$

n	$\theta_1 = 10$	$\theta_1 = 100$	$\theta_1 = 1000$
20	0.12	0.29	0.54
40	0.12	0.19	0.35

Power of tests for $\widehat{\theta_{1,T}}$

Comments

- Watterson's test statistic is more powerful than Tajima's test.
- Power is law when the ratio between the normal and the raised mutation rate is less than 1000 ($\theta_0 < \theta_1$).
  - In agreement with biological experiments: detecting occurrence of the Loss of Mismatch Repair is hard when $\theta_1/\theta_0 < 1.000$ (Boland *et al.*, 1998).
- Conditional on the occurrence of the loss of MMR, powers are decreasing as the sample size increases.
  - Monitoring several loci to increase power of tests.

Publications

- M. Emily, D. Morel, R. Marcelpoil and O.Francois. Spatial correlation of gene expression measures in Tissue Microarray core analysis, *Journal of theoretical Medicine*, Vol. 6, No. 1, Mars 2005, pages 33-39.
  NB: the journal *Journal of Theoretical Medicine* has been moved to *Computational and Mathematical Methods in Medicine*.
- M. Emily and O.Francois. A continuous stochastic model for cell sorting, arXiv q-bio.TO/0605035.
- M. Emily and O.Francois. Conditional coalescent trees with two mutation rates and their application to genomic instability, *Genetics*, Vol. 172, Mars 2006, pages 1809-1820.

Part C. *Conclusion*

Summary

- Two stochastic models were proposed:
  - A Gibbsian spatial model based on gene expression data within tissues.
  - Conditional coalescent model using DNA sequences data.
- Results: new statistical procedures:
  - To estimate the differential adhesion between cells in normal and tumoral tissues.
  - To test the occurrence of the Loss of MMR and to detect genetic instability.

Future works

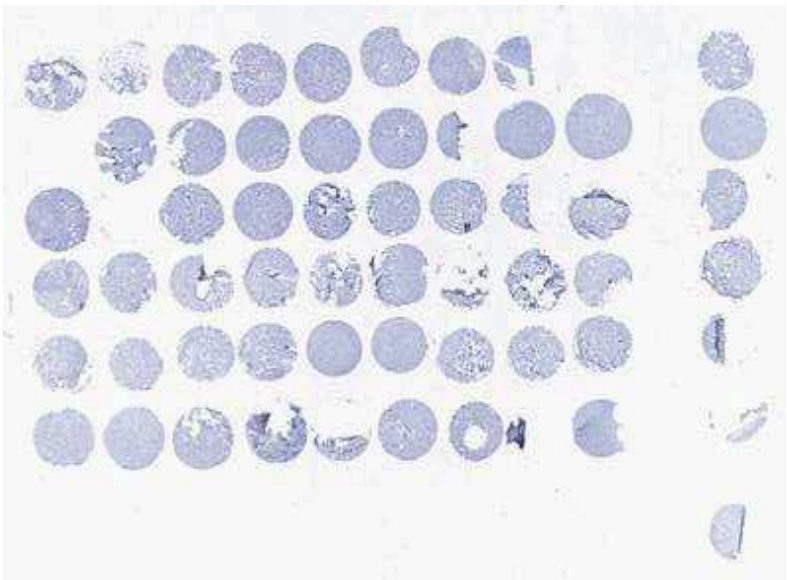
- Spatial point process
  - Mathematical properties of estimators (Billiot *et al.*, 2006).
  - Study the phase transition of our model (Haggström, 2000).
  - Include cell division dynamics (Thom's criterion, 1972).
  - Adapt our model to other issues
    (interaction between trees - Gourlet-Fleury *et al.*, 2004).
- Coalescent model
  - Increase power of tests using a multilocus approach
    (Kühner *et al.*, 1995).
  - Include clonal selection
    (Ancestral Selection Graph - Neuhauser and Krone, 1997).

Impact on early diagnosis of cancer

- In the near future, Polymerase Chain Reaction (PCR) will be standard routine during medical diagnosis.
- High-throughput data such as Fluorescence In Situ Hybridation will make tissue DNA contents easier to analyze.
- Goal: Reduce the time of detection by several years in hereditary cancers (HNPCC, hereditary breast cancer).

Tissue Microarrays

- High-throughput data of gene expression markers is an important emerging technology (Kononen, 1998).
- Perspective: Model-based statistical procedures.



Tissue Microarrays