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.
Introduction - Objectives

• 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).
Introduction - Biological levels

Breast cancer tissue

Phenotypic markers

Gene expression

DNA Sequences
Genotypic markers

acgtgatgatgatgatgatgacgtgcga
acctgatgatgatgatgatgacgtgcga
attatcgatcgatcgtacgtacgtacgt
• **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 $\xrightarrow{20\text{ years}}$ tumor manifestation (Bielas and Loeb, 2005).
  - Hereditary Colon Cancer implicates MSH2, MSH6 and MLH1 genes (Fishel et al., 1993).
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**
A. Spatial model - Biological context

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.

\[ \beta - Catenin \text{ gene expression in human hepatocellular carcinoma (Lin, 2003)}. \]
A. Spatial model - Biological context

Cadherin-catenin complex

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

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.
A. Spatial model - Mathematical Models

Mathematical models of the Differential Adhesion Hypothesis (DAH) are classified according their geometry (Brodland, 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 $\sigma$, is a set of pixels and each pixel $(i, j)$ is characterized by a type $\tau_{ij}$ (3 different types: $\ell$ for light cells, $d$ for dark cells and $M$ for extracellular matrix).

- The Energy, $H_{GG}$, is defined as:

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

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 $\text{Constraint} = \sum_{\sigma} C(\text{area}(\sigma))$
Graner and Glazier’s model

\[ \ell = \text{light}, \quad d = \text{dark} \quad \text{and} \quad M = \text{medium}. \]

Example of GG’s configuration using \( J_{\ell,\ell} = 14, \quad J_{d,d} = 14, \quad J_{\ell,d} = 29, \quad J_{\ell,M} = J_{d,M} = 16 \) (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
A. Spatial model - Modeling

Energy functional

\[ H_{CC}(\varphi) = H_{Adh} + \text{Constraint} \]

with:

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

and:

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

and where:

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

• With respect to the Poisson process, the density of a configuration $\varphi$ can be written as:

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

where $\theta$ quantifies the strength of adhesion within a tissue.

• Estimating the strength of adhesion is of particular interest.
A. Spatial model - Modeling

Mathematical study

Theorem

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

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

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

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

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

  $$E(x, \varphi) > -K$$

- **Quasilocality.** For all $x$, $\varphi$ and $\Delta$ bounded set:

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

where $\varepsilon(x) \to 0$ when $x \to \infty$.

Then, there exists a Gibbsian marked marked point process that satisfies the local specifications derived from $H_{CC}$. 
A. Spatial model - Simulation

**Algo: Insertion-Deletion Metropolis-Hastings**

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

**Theorem**
Under the same conditions \((g, J \text{ and } C \text{ bounded})\), 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)
A. Spatial model - Simulation

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).
A. Spatial model - Estimation

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

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

The characteristic patterns emerge for large $\theta$.  

\[ \theta = 1 \quad \theta = 5 \quad \theta = 10 \]
Estimation: Conditional Pseudo-Likelihood

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

$$PL^\Lambda_C(\theta) = \prod_i \text{Prob}(\tau_i | \varphi, \tau \setminus \{\tau_i\}, \theta)$$

**Definition**

An estimator for the adhesion strength parameter ($\theta$) is given by:

$$\hat{\theta}_C = \arg\max_\theta PL^\Lambda_C(\theta)$$
A. Spatial model – Estimation

Estimation: Pseudo-Likelihood

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

\[
PL^\Lambda(\theta) = \exp \left( -\int_\Lambda \int_M \exp(-H_{CC}(x|\varphi)) d\tau_x dx \right) \prod_{x \in \varphi^\Lambda} \exp \left( -H_{CC}(x|\varphi \setminus x) \right)
\]

Definition

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

\[
\hat{\theta} = \arg\max_\theta PL^\Lambda(\theta)
\]
A. Spatial model - Performances for estimators $\hat{\theta}_C$ and $\hat{\theta}$

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>Mean</th>
<th>Variance</th>
<th>$\theta$</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.98</td>
<td>0.70</td>
<td>1.03</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5.01</td>
<td>0.57</td>
<td>4.94</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>10.47</td>
<td>1.20</td>
<td>9.80</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>14.58</td>
<td>2.22</td>
<td>15.03</td>
<td>1.20</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>Mean</th>
<th>Variance</th>
<th>$\theta$</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.01</td>
<td>0.91</td>
<td>0.97</td>
<td>1.3</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5.17</td>
<td>1.12</td>
<td>4.93</td>
<td>1.05</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>10.24</td>
<td>2.24</td>
<td>10.30</td>
<td>1.38</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>15.43</td>
<td>3.87</td>
<td>15.58</td>
<td>2.55</td>
<td></td>
</tr>
</tbody>
</table>

Mean and Variance from 100 replicates for $\hat{\theta}$
A. Spatial model - Performances for estimators $\hat{\theta}_C$ and $\hat{\theta}$

Comments

- Conditional and unconditional estimators seem to be weakly biased.
- Variances increase with $\theta$.
- The conditional estimator is computationally faster than the unconditional estimator.
- Theoretically, $\hat{\theta}$ should be better than $\hat{\theta}_C$.
- In practice, we observe the reverse (integral approximations may be a problem).
A. Spatial model - An application to data

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

\[ \hat{\theta}_C = 14.9 \text{ and } \hat{\theta} = 15.4 \]

\[ \hat{\theta}_C = 31.9 \text{ and } \hat{\theta} = 33.7 \]
A. Spatial model - An application to data

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

$\hat{\theta}_C = 14.9$ and $\hat{\theta} = 15.4$

$\hat{\theta}_C = 31.9$ and $\hat{\theta} = 33.7$
A. Spatial model  -  An application to data

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*
B. Coalescent model - Biological levels

*Phenotypic markers*

*Breast cancer tissue*

*Gene expression*

**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).
B. Coalescent model - Biological context

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.
  - efficacy 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: $1.4 \times 10^{-10}$ nucleotides per cell per division (Loeb, 1991).
  - Genetic instability $10^{-10} \rightarrow 10^{-6}$ shift.
B. Coalescent model - Biological context

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 $\lambda_i = \frac{i(i-1)}{2}$.
B. Coalescent model - Mathematical background

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 $\mu$ is the mutation rate per base per mitotic division and $N$ is the total number of cells.
- Classical unbiased estimators for $\theta$: 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 $\theta$ is defined as:

\[
\hat{\theta}_W = \frac{2S}{E[L]} = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}},
\]

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 $\theta$ is defined as:

$$\hat{\theta}_T = \frac{2}{n(n-1)} \sum_{i<j} \Pi(i, j)$$

Example with three DNA sequences where $\hat{\theta}_T = 2.67$ ($\hat{\theta}_W = 2.67$)

<table>
<thead>
<tr>
<th>Seq1 vs Seq2</th>
<th>Seq1 vs Seq3</th>
<th>Seq2 vs Seq3</th>
</tr>
</thead>
<tbody>
<tr>
<td>acagttacat</td>
<td>acagttacat</td>
<td>agagctacat</td>
</tr>
<tr>
<td>agagctacat</td>
<td>agagtgcgt</td>
<td>agagttgcgt</td>
</tr>
</tbody>
</table>
Back to genetic instability - Modeling constraints

- The event “Loss of MMR”, denoted by $\Delta$, 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 $\theta_0$ is “normal”,
  - $\mathcal{R}$ in which the mutation rate $\theta_1$ 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.
B. Coalescent model  -  Conditional coalescent modeling

- Mutations follow Poisson processes of rates:
  - $\theta_0/2$ along the blue branches.
  - $\theta_1/2$ along the red branches.
B. Coalescent model - Conditional coalescent modeling

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 $\Delta$ has the following distribution:

$$P(B = b) = \frac{1}{bH_{n-1}}$$

$b = 1, \ldots, n - 1$.

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 $S_n$, $S_{0n}$ and $S_{1n}$ where:
  - $E[S_{0n}] = E[L_0] \theta_0 / 2$
  - $E[S_{1n}] = E[L_\Delta] \theta_1 / 2$

An unbiased estimator of $\theta_1$ is:

$$\hat{\theta}_{1,W} = \frac{S_n - E[L_0] \theta_0 / 2}{E[L_\Delta] / 2}$$
B. Coalescent model - Results

Correction of Watterson’s estimator - \( \mathbb{E}[L_\Delta] = \mathbb{E}[L_1] + \mathbb{E}[\eta_n] \)

**Proposition**

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

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

**Proposition**

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

\[
\mathbb{E}[\eta_n | B = b] = 2 \sum_{k=2}^{n-b+1} \frac{p_k^\Delta}{k}.
\]
B. Coalescent model - Results

Correction of Watterson’s estimator - $L_0$

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

![Diagram showing past and length L0 with red subtree and total length L]
Correction of Watterson’s estimator - $L$

**Proposition**

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

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

**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 $\Delta$ 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 $\lambda_\ell$ and:

$$p_k^\Delta = \left( \begin{array}{c} n - k \\ b - 1 \end{array} \right) \left( \begin{array}{c} n - 1 \\ b \end{array} \right)^{-1} \quad k = 2, \ldots, n - b + 1$$
Correction of Tajima’s estimator

- Mean number of pairwise differences between genes: $\Pi$.
- An unbiased estimator of $\theta_1$ is:

$$\hat{\theta}_{1,T} = \frac{\Pi - C_n \theta_0}{D_n}$$

- $C_n$ 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

<table>
<thead>
<tr>
<th>n</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>35</th>
<th>40</th>
<th>45</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B_n$</td>
<td>0.595</td>
<td>0.68</td>
<td>0.713</td>
<td>0.732</td>
<td>0.746</td>
<td>0.756</td>
<td>0.764</td>
<td>0.771</td>
<td>0.776</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>n</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>35</th>
<th>40</th>
<th>45</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_n$</td>
<td>0.996</td>
<td>1.019</td>
<td>1.021</td>
<td>1.02</td>
<td>1.02</td>
<td>1.019</td>
<td>1.019</td>
<td>1.018</td>
<td>1.018</td>
</tr>
<tr>
<td>$D_n$</td>
<td>0.253</td>
<td>0.218</td>
<td>0.199</td>
<td>0.187</td>
<td>0.178</td>
<td>0.171</td>
<td>0.166</td>
<td>0.161</td>
<td>0.156</td>
</tr>
</tbody>
</table>

*Tables for $C_n$ and $D_n*
B. Coalescent model  -  Results

Algorithm for simulating a conditional coalescent tree

Algorithm

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

Statistical errors of $\hat{\theta}_{1,W}$ and $\hat{\theta}_{1,T}$ for $\theta_0 = 1$
($N = 2.5 \times 10^9$ and $\mu = 10^{-10}$)

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\theta_1 = 10$</th>
<th>$\theta_1 = 100$</th>
<th>$\theta_1 = 1000$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$E$</td>
<td>SD</td>
<td>$E$</td>
</tr>
<tr>
<td>10</td>
<td>9.9</td>
<td>12.0</td>
<td>97.4</td>
</tr>
<tr>
<td>30</td>
<td>10.2</td>
<td>12.8</td>
<td>102.9</td>
</tr>
<tr>
<td>50</td>
<td>10.4</td>
<td>13.5</td>
<td>102.0</td>
</tr>
</tbody>
</table>

Expectation and Standard Deviation for $\hat{\theta}_{1,W}$ using 1000 replicates.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\theta_1 = 10$</th>
<th>$\theta_1 = 100$</th>
<th>$\theta_1 = 1000$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$E$</td>
<td>SD</td>
<td>$E$</td>
</tr>
<tr>
<td>10</td>
<td>9.9</td>
<td>13.7</td>
<td>107.3</td>
</tr>
<tr>
<td>30</td>
<td>9.5</td>
<td>15.5</td>
<td>100.9</td>
</tr>
<tr>
<td>50</td>
<td>10.3</td>
<td>17.6</td>
<td>106.5</td>
</tr>
</tbody>
</table>

Expectation and Standard Deviation for $\hat{\theta}_{1,T}$ using 1000 replicates.
B. Coalescent model - Results

Statistical errors of $\hat{\theta}_{1,W}$ and $\hat{\theta}_{1,T}$ 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 have less variance than Tajima’s corrected estimator.
B. Coalescent model - Results

Testing the absence of the “Loss of Mismatch Repair”

- \( H_0 \): Absence of \( \Delta \).
- \( H_1 \): Occurrence of \( \Delta \) 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:

\[
r = \frac{L(H_1)}{L(H_0)} = \sum_{k=2}^{n-b+1} \lambda_k p_k^\Delta t_k
\]

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 $\Delta$ (LMMR) - $\theta_0 = 1$

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

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\theta_1 = 10$</th>
<th>$\theta_1 = 100$</th>
<th>$\theta_1 = 1000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.44</td>
<td>0.74</td>
<td>0.90</td>
</tr>
<tr>
<td>40</td>
<td>0.42</td>
<td>0.73</td>
<td>0.88</td>
</tr>
</tbody>
</table>

*Power of tests for $\hat{\theta}$ estimator*

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\theta_1 = 10$</th>
<th>$\theta_1 = 100$</th>
<th>$\theta_1 = 1000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.44</td>
<td>0.69</td>
<td>0.84</td>
</tr>
<tr>
<td>40</td>
<td>0.34</td>
<td>0.64</td>
<td>0.79</td>
</tr>
</tbody>
</table>

*Power of tests for $\Pi$ estimator*
Testing the occurrence of $\Delta$ (LMMR) - $\theta_0 = 1$

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

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\theta_1 = 10$</th>
<th>$\theta_1 = 100$</th>
<th>$\theta_1 = 1000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.06</td>
<td>0.18</td>
<td>0.70</td>
</tr>
<tr>
<td>40</td>
<td>0.11</td>
<td>0.24</td>
<td>0.59</td>
</tr>
</tbody>
</table>

Power of tests for $\hat{\theta}_{1,W}$

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\theta_1 = 10$</th>
<th>$\theta_1 = 100$</th>
<th>$\theta_1 = 1000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.12</td>
<td>0.29</td>
<td>0.54</td>
</tr>
<tr>
<td>40</td>
<td>0.12</td>
<td>0.19</td>
<td>0.35</td>
</tr>
</tbody>
</table>

Power of tests for $\hat{\theta}_{1,T}$
### 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.
B. Coalescent model - Publications

Publications


• M. Emily and O. Francois. A continuous stochastic model for cell sorting, arXiv q-bio.TO/0605035.

Part C. **Conclusion**
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).
C. Conclusion - Future

Tissue Microarrays

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