Computing risks based on family history in genetic diseases

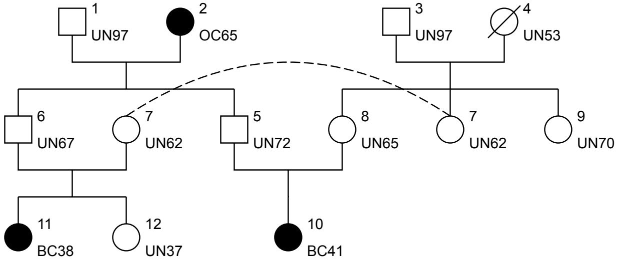
Alexandra Lefebvre, Olivier Bouaziz, Grégory Nuel

LPSM, UPMC, Sorbonne Université, Paris, France

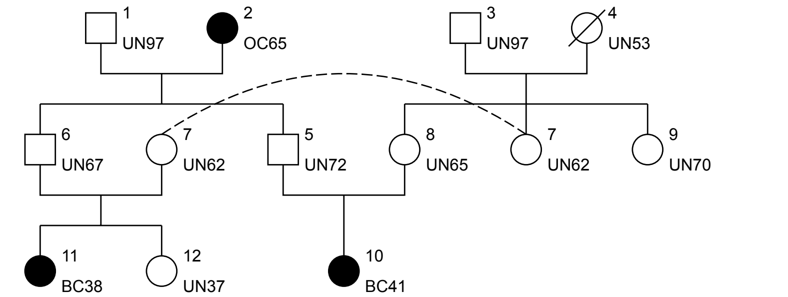
SMPGD 2018
Université de Montpellier

Context of the breast cancer

- 1st cancer in women. 54,000 women in France each year
- Complex disease due to an accumulation of mutations (BRCA 1/2, PALB2, RAD51, etc.)
- **Inherited mutation** in 10 to 15% of the cases



Family History (FH) : Pedigree + Sequencing (seq) + Survival (Y)



Variant data (var)

- Known variant or VUS
- Database
- Functional tests
- Molecular dynamics

Cancer pathology (patho)

- MSI status (Lynch)
- BRAF mutation (Lynch)
- ER, PR, HER2 status (Breast)
- invasive/in situ (Breast)

compute $\Rightarrow$ posterior **variant status** **carrier risk** **tumoral risk**

V : the set of variant status
var : the set of variant data
X : the set of genotypes

$$\left.\begin{array}{l}\text{seq : the set of sequencing data}\\ Y \text{ : the set of phenotypes (survival data)}\end{array}\right\} \rightarrow \mathrm{FH} = \{seq, Y\}$$

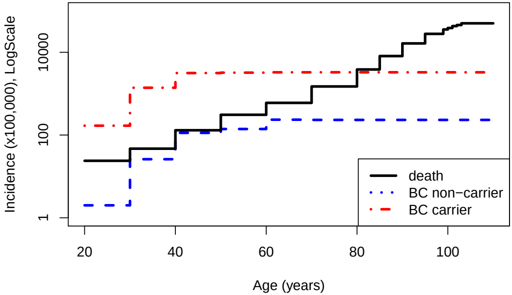
patho : the set of pathology reports

$$\mathbb{P}(\text{var}, V, X, \text{seq}, Y, \text{patho}) = \prod_j \mathbb{P}(\text{var}_j | V_j)\, \mathbb{P}(V_j)$$
$$\prod_i \mathbb{P}\left(X_i | X_{\text{pat}_i}, X_{\text{mat}_i}\right) \mathbb{P}(\text{seq}_i | X_i)\, \mathbb{P}(Y_i | X_i)$$
$$\prod_{i, Y_i \neq \text{UN}} \mathbb{P}(\text{patho}_i | X_i)$$

Our first objective : modelling efficiently the quantity
$$\prod_i \mathbb{P}\left(X_i | X_{\text{pat}_i}, X_{\text{mat}_i}\right) \mathbb{P}(Y_i | X_i)$$
Illustrated with a mendelian model and parameters taken from the breast cancer literature [Claus et al., 1991, Easton et al., 1993].

The Claus-Easton model [Claus et al., 1991]

- Autosomal, biallelic, dominant, estimated allele frequency f=0.33%
- The hazard functions per genotype (λ_0 and λ_1) estimated from the densities in [Easton et al., 1993]:



$$\frac{\lambda_1(t)}{\lambda_0(t)} \in [20 - 80]$$

Objective : Implement the Claus-Easton model in a Bayesian network (sum/product algorithm) combined with survival data.

Implementation : genotypes

$$\mathbb{P}(X,Y) = \prod_i \underbrace{\mathbb{P}\left(X_i | X_{\text{pat}_i}, X_{\text{mat}_i}\right)}_{\text{genotype}} \underbrace{\mathbb{P}\left(Y_i | X_i\right)}_{\text{phenotypes}}$$

Mode of inheritance : 1 autosomal biallelic gene, f = 0.33%

Founders (**Hardy-Weinberg**) :

$$\begin{cases} \mathbb{P}(X_i = 00) = (1-f)^2 \\ \mathbb{P}(X_i = 10) = \mathbb{P}(X_i = 01) = f(1-f) \\ \mathbb{P}(X_i = 11) = f^2 \end{cases}$$

Offsprings (**Mendel**):

$$\begin{cases} \mathbb{P}(X_i = 00) = (1-\Theta(X_{\text{pat}})) \times (1-\Theta(X_{\text{mat}})) \\ \mathbb{P}(X_i = 10) = \Theta(X_{\text{pat}}) \times (1-\Theta(X_{\text{mat}})) \\ \mathbb{P}(X_i = 01) = (1-\Theta(X_{\text{pat}})) \times \Theta(X_{\text{mat}}) \\ \mathbb{P}(X_i = 11) = \Theta(X_{\text{pat}}) \times \Theta(X_{\text{mat}}) \end{cases}$$

with $\Theta(00) = 0, \quad \Theta(10) = \Theta(01) = 0.5, \quad \Theta(11) = 1$

Implementation : phenotypes

with T_i, the age at disease onset for the individual i.

$$Y_i \text{ (Survival data)} = \begin{cases} \{T_i > \tau_i\} & \text{if } i \text{ is censored (UN) at age } \tau_i \\ \{T_i = \tau_i\} & \text{if } i \text{ is affected (BC or OC) at age } \tau_i \end{cases}$$

Dominant model of disease :

Hazard functions of T : $\begin{cases} \lambda_0(t) & \text{for } X = 00 \text{ (non-carrier)} \\ \lambda_1(t) & \text{for } X \neq 00 \text{ (carrier)} \end{cases}$

Survival functions of T : $\begin{cases} S_0(t) = \exp\left(-\int_0^t \lambda_0(t)\right) dt & \text{for non-carrier} \\ S_1(t) = \exp\left(-\int_0^t \lambda_1(t)\right) dt & \text{for carrier} \end{cases}$

conditional probabilities

- For a **censored** individual at age τ_i

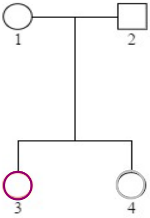
$$\mathbb{P}(Y_i|X_i) = \mathbb{P}(T_i > \tau_i|X_i) = \begin{cases} S_0(\tau_i) & \text{for non-carriers} \\ S_1(\tau_i) & \text{for carriers} \end{cases}$$

- For an **affected** individual at age τ_i

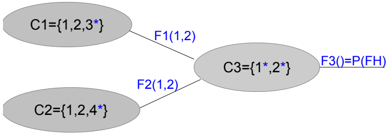
$$\mathbb{P}(Y_i|X_i) = \mathbb{P}(T_i = \tau_i|X_i) = \begin{cases} S_0(\tau_i)\lambda_0(\tau_i) & \text{for non-carriers} \\ S_1(\tau_i)\lambda_1(\tau_i) & \text{for carriers} \end{cases}$$

Impl. in a Bayesian network [1] (sum-product algorithm)

$$\mathbb{P}(X, \mathrm{FH}) = \prod_i \underbrace{\sum_{Y_i \in \mathrm{FH}} \mathbb{P}(X_i | X_{\mathrm{pa}_i}) \mathbb{P}(Y_i | X_i)}_{\mathrm{K}_i(X_i, X_{\mathrm{pa}_i})} \rightarrow \mathbb{P}(\mathrm{FH}) = \sum_X \prod_i \mathrm{K}_i(X_i, X_{\mathrm{pa}_i})$$



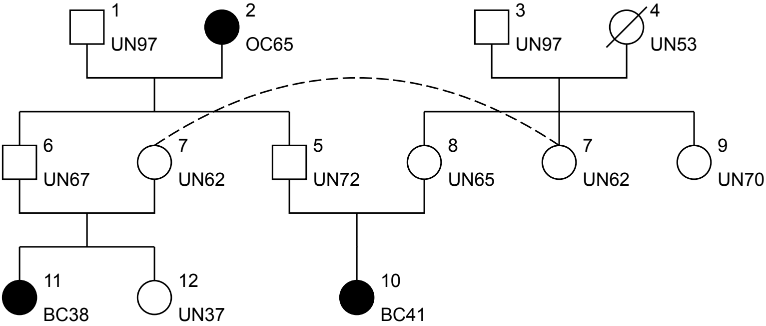
With $X \in \{00, 10, 01, 11\}^n \rightarrow 4^n$configurations

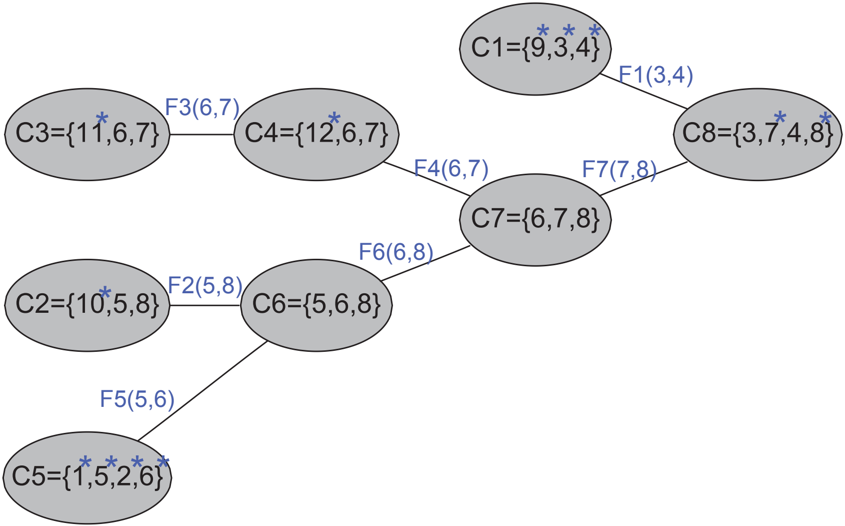


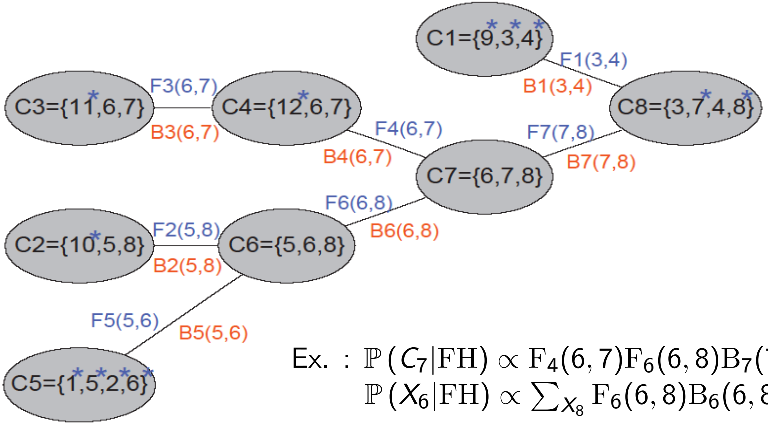
$$\mathbb{P}(\mathrm{FH}) = \sum_{X_1} \mathrm{K}_1(X_1) \sum_{X_2} \mathrm{K}_2(X_2) \overbrace{\sum_{X_3} \mathrm{K}_3(X_3, X_1, X_2)}^{\mathrm{F}_1(X_1, X_2)} \overbrace{\sum_{X_4} \mathrm{K}_4(X_4, X_1, X_2)}^{\mathrm{F}_2(X_1, X_2)}$$

$$\mathbb{P}(\mathrm{FH}) = \mathrm{F}_3(\emptyset) = \sum_{X_1, X_2} \mathrm{K}_1(X_1) \mathrm{K}_2(X_2) \mathrm{F}_1(X_1, X_2) \mathrm{F}_2(X_1, X_2)$$

[1] [Koller and Friedman, 2009]


1
UN97
2
OC65
3
UN97
4
UN53
6
UN67
7
UN62
5
UN72
8
UN65
7
UN62
9
UN70
11
BC38
12
UN37
10
BC41


C1={9,3,4}
F1(3,4)
C3={11,6,7}
F3(6,7)
C4={12,6,7}
C8={3,7,4,8}
F4(6,7)
F7(7,8)
C7={6,7,8}
F6(6,8)
C2={10,5,8}
F2(5,8)
C6={5,6,8}
F5(5,6)
C5={1,5,2,6}

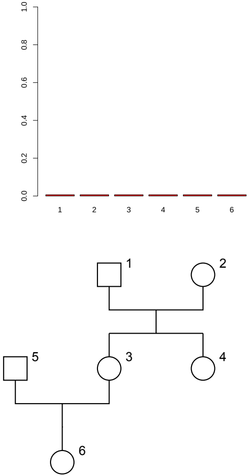


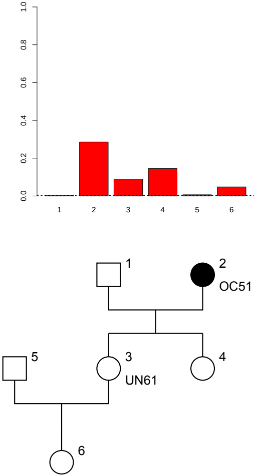
Ex. : $\mathbb{P}(C_7|\mathrm{FH}) \propto \mathrm{F}_4(6,7)\mathrm{F}_6(6,8)\mathrm{B}_7(7,8)$

$\mathbb{P}(X_6|\mathrm{FH}) \propto \sum_{X_8} \mathrm{F}_6(6,8)\mathrm{B}_6(6,8)$

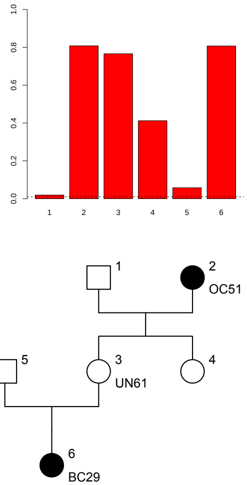
F & B computed once for any later marginal or joint distribution needed
Complexity $\mathcal{O}(4^n) \to \mathcal{O}(n \times 4^k)$, k:tree-width

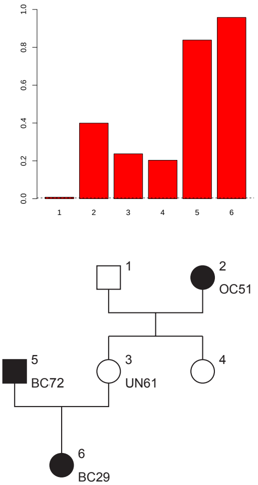
Results : Carrier risk, posterior marginal carrier distribution





Results : Carrier risk, posterior marginal carrier distribution





Implementation : Disease risk prediction

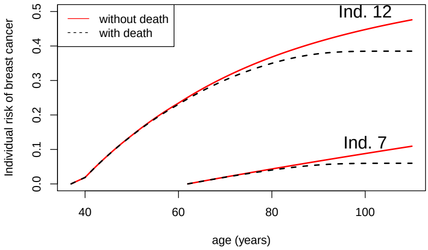
$\pi(\tau) = \mathbb{P}(\text{carrier}|\mathrm{FH})$

- Breast cancer risk with no competing risk of death.
  $\mathbb{P}(T \leq t|\mathrm{FH}) = 1 - S(t|\mathrm{FH})$ with
  $S(t|\mathrm{FH}) = \sum_{X_i} \mathbb{P}(T > t, X_i|\mathrm{FH}) = \pi(\tau)\frac{S_1(t)}{S_1(\tau)} + (1 - \pi(\tau))\frac{S_0(t)}{S_0(\tau)}$

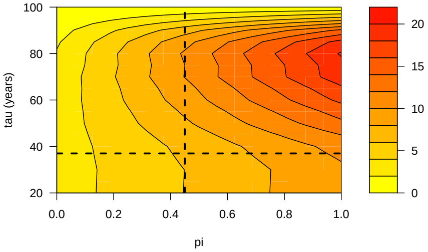
- With competing risk of death.
  $T^* = \min(T_{\text{disease}}, T_{\text{death}})$
  $\lambda_{\text{both}}(t|\mathrm{FH}) = \boldsymbol{\lambda}_{\textbf{disease}}(\boldsymbol{t}|\mathbf{FH}) + \lambda_{\text{death}}(t)$
  $$\begin{aligned}\mathbb{P}(T \leq t|\mathrm{FH}) &= \int_\tau^t S_{\text{both}}(u)\boldsymbol{\lambda}_{\textbf{disease}}(\boldsymbol{u})\boldsymbol{du} \\ &= \int_\tau^t \exp\left(-\int_\tau^u \lambda_{\text{both}}(v)dv\right)\boldsymbol{\lambda}_{\textbf{disease}}(\boldsymbol{u})\boldsymbol{du}\end{aligned}$$

  $\boldsymbol{\lambda}_{\textbf{disease}}(\boldsymbol{t}|\mathbf{FH}) = \pi(t|\mathrm{FH})\lambda_1(t) + (1 - \pi(t|\mathrm{FH}))\lambda_0(t)$
  $\pi(t|\mathrm{FH}) = \frac{\pi(\tau)S_1(t)}{S(t|\mathrm{FH})S_1(\tau)}$

Results : Disease risk



Individual disease risk
with vs without
competing risk of death



Difference in % of disease risk at 100
with vs without
competing risk of death

Perspectives

- Adaptation to Microsatellite Instability (MSI cancer / Lynch syndrome) (A. Duval, Saint-Antoine Hospital)
- Taking into account the sequencing data, the pathology reports, variant data especially to help classify the Variants of Uncertain Significance (VUS).
- Parameters estimations
- Complex distributions (number of carriers in the family) with probability generating functions (polynomials) $\rightarrow$ familial risk
- Multi-state and frailty survival models

Claus, E. B., Risch, N., and Thompson, W. D. (1991).
Genetic analysis of breast cancer in the cancer and steroid hormone study.
American journal of human genetics, 48(2):232.

Easton, D., Bishop, D., Ford, D., and Crockford, G. (1993).
Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. the breast cancer linkage consortium.
American journal of human genetics, 52(4):678.

Koller, D. and Friedman, N. (2009).
Probabilistic graphical models: principles and techniques.
MIT press.

Thank you for your attention