quelques choix de modélisations mathématiques, statistiques et informatiques dans le domaine de la santé et du végétal

Gilles HUNAULT, laboratoire HIFIH

HIFIH : Hémodynamique, Interaction Fibrose et Invasivité tumorale Hépatique

UPRES 3859, IFR 132, Faculté de Médecine, Université d'Angers

Chercheur associé au centre I.N.R.A. d'Angers

Chercheur associé au laboratoire d'informatique L.E.R.I.A (Angers)

16 novembre 2017

Sommaire

Remerciements aphabétiques

Remerciements alphabétiques (1)

C.H.U. / Laboratoire H.I.F.I.H.

Sandrine BERTRAIS, Jérôme BOURSIER, Paul CALÈS.

Centre Paul PAPIN / I.C.O.

Michelle BOIDRON-CELLE, Erick GAMELIN.

I.N.R.A. Angers

Matthieu BARRET, Tristan BOUREAU, Louis GARDAN.

Remerciements alphabétiques (2)

L.E.R.I.A. / Département informatique

Jacques BOYER, Benoit DA MOTA, Frédéric LARDEUX,

David LESAINT, Frédéric SAUBION.

Doctorants devenus docteurs

Fabien CHHEL, Céline ROUSSEAU, Samir REZKI, Sory TRAORÉ.

\+ les technicien(ne)s, ingénieur(e)s et autres personnels B.I.A.T.S.S.

Format général pour chaque rubrique

- décrire le problème en termes métier ;
- situer la problématique mathématique, statistique ou informatique ;
- esquisser les choix effectués et leur *rationnel*.

Site compagnon :

`http://forge.info.univ-angers.fr/~gh/Applis/Choixm/`

Position du problème (1980)

*On veut réaliser une classification **non supervisée** de souches bactériennes phytopathogènes de Pseudomonas en fonction de caractéristiques phénotypiques binaires ou ternaires.*

Taille des données : disons quelques centaines de lignes, quelques dizaines de colonnes.

Les données sont sur cartes perforées à Angers.

Le programme est exécuté à Marseille.

On obtient le listing du dendrogramme en 3 mois après correction des erreurs de saisie.

*Comment faire **mieux** ?*

2. Quelle classification ascendante hiérarchique utiliser ?

Choix mathématiques et informatiques

Effectuer des traitements en local, écrire le programme de classification (mais dans quel langage de programmation ?).

Continuer à réaliser des classifications hiérarchiques non supervisées plutôt que des simples partitionnements et, parmi ces méthodes, retenir des classifications ascendantes plutôt que descendantes afin de rester cohérent avec les précédentes classifications.

Utiliser une distance « standard » pour la matrice d'entrée.

Choisir un critère d'agrégation ***adapté*** *aux données phénotypiques de souches bactériennes.*

Rappel des méthodes CAH

Principe : à partir d'une **matrice de distances** entre éléments [isolés]

- on **choisit** deux éléments qu'on fusionne en un nouvel élément,
- on calcule la **distance** entre les anciens éléments et le nouveau,
- on supprime les deux anciens éléments choisis,
- on recommence jusqu'à avoir regroupé tout le monde.

Liste des choix en CAH dans les années 80

- *choix de la **distance initiale** (ou "coût", "score"...)*

  binaire : jaccard, russel-rao, simpson, sokal...

  comptage : czekanowski, clark, kulczynski...

  info : hamming, wagner, levenstein, édition...

  autre : dice, tanimoto...

- *choix du **critère de sélection*** (ou "indice d'agrégation")

  "linkage" simple ou complet, "pair group"...

  divergence, inertie...

- *choix de la **formule de recalcul des distances***

  min, max, moyenne, pondérée, ultramétrique...

Quelques formules de recalcul des distances

Si $D_{x,y}$ est la distance de x à y, la distance de l'ancien groupe k au nouveau groupe composé de i et de j est

$$D_{ij,k} = a_i D_{i,k} + b_j D_{j,k} + c D_{i,j} + d |D_{i,k} - D_{j,k}|$$

méthode	a_i	b_j	c	d	*interprétation*
single	$1/2$	$1/2$	0	-1/2	minimum
complete	$1/2$	$1/2$	0	$+1/2$	maximum
upgma	$n_i/(n_i+n_j)$	$n_j/(n_i+n_j)$	0	0	moy. pondérée
nj	$1/2$	$1/2$	$-1/2^*$	0	moy. réduite

Choix statistiques et réalisations dans les années 80

Paramètre	***Choix***
Distance initiale	*Jaccard-Sneath*
Critère d'agrégation	*Lien mininimal*
Formule de recalcul	*UPGMA*

Implémentations en FORTRAN puis en PASCAL, en Dbase VII Windows et enfin en PHP.

La distance de Jaccard et Sneath (induite par l'indice de similarité éponyme) est bien adaptée aux problèmes de taxonomie bactérienne sur données binaires.

Le lien minimal est un choix raisonnable car c'est le critère d'agrégation du programme à Marseille.

La formule de recalcul UPGMA était le choix conseillé à l'époque par la communauté taxonomique européenne.

Et si c'était à refaire en 2017 (1) ?

La liste des méthodes s'est allongée (*ward, single, complete, average, mcquitty, median, centroid*) donc encore des choix à faire.

La liste des distances s'est allongée donc encore des choix et des comparaisons à faire.

Par contre il n'y aurait plus de de programmation importante à réaliser : les logiciels scientifiques et statistiques – R, Python, SAS, SPSS, Statistica, Matlab, Scilab – ont tous aujourd'hui des modules, des packages, des procédures de calculs de distances, de classification.

... **encore faut-il savoir utiliser ces logiciels et connaitre les modules.**

Liste des choix de méthodes possibles en 2017

Méthodes non supervisées	*Méthodes supervisées*
classifications hiérarchiques,	réseau de neurones,
k-moyennes, centres mobiles,	algorithmes génétiques,
nuées dynamiques,	algorithmes de colonies, de fourmis,
cartes auto-organisatrices...	algorithmes auto-adaptatifs...

2. Quelle classification ascendante hiérarchique utiliser ?

51 indices de dissimilarités pour données binaires en 2017

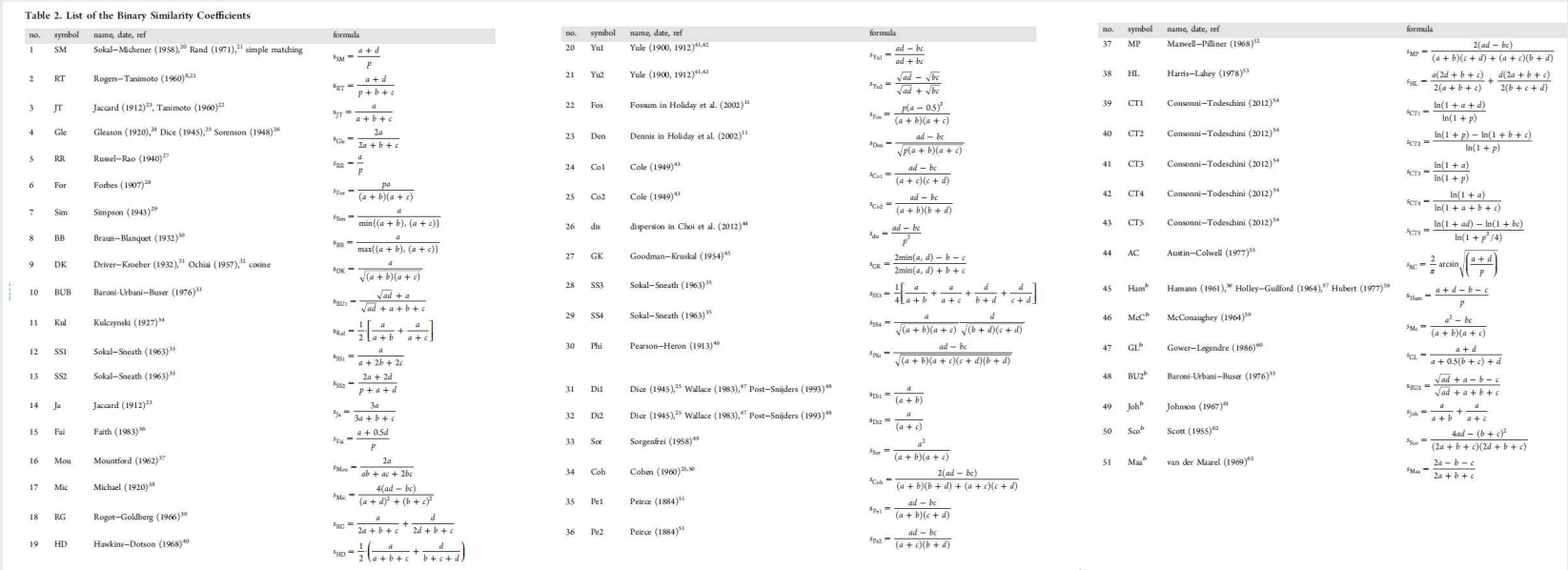


Table 2. List of the Binary Similarity Coefficients

no.	symbol	name, date, ref	formula
1	SM	Sokal–Michener (1958),[20] Rand (1971),[21] simple matching	$s_{SM} = \frac{a+d}{p}$
2	RT	Rogers–Tanimoto (1960)[8,22]	$s_{RT} = \frac{a+d}{p+b+c}$
3	JT	Jaccard (1912)[23], Tanimoto (1960)[22]	$s_{JT} = \frac{a}{a+b+c}$
4	Gle	Gleason (1920),[24] Dice (1945),[25] Sorenson (1948)[26]	$s_{Gle} = \frac{2a}{2a+b+c}$
5	RR	Russel–Rao (1940)[27]	$s_{RR} = \frac{a}{p}$
6	For	Forbes (1907)[28]	$s_{For} = \frac{pa}{(a+b)(a+c)}$
7	Sim	Simpson (1943)[29]	$s_{Sim} = \frac{a}{\min\{(a+b),(a+c)\}}$
8	BB	Braun–Blanquet (1932)[30]	$s_{BB} = \frac{a}{\max\{(a+b),(a+c)\}}$
9	DK	Driver–Kroeber (1932),[31] Ochiai (1957),[32] cosine	$s_{DK} = \frac{a}{\sqrt{(a+b)(a+c)}}$
10	BUB	Baroni-Urbani–Buser (1976)[33]	$s_{BU1} = \frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c}$
11	Kul	Kulczynski (1927)[34]	$s_{Kul} = \frac{1}{2}\left[\frac{a}{a+b}+\frac{a}{a+c}\right]$
12	SS1	Sokal–Sneath (1963)[35]	$s_{SS1} = \frac{a}{a+2b+2c}$
13	SS2	Sokal–Sneath (1963)[35]	$s_{SS2} = \frac{2a+2d}{p+a+d}$
14	Ja	Jaccard (1912)[23]	$s_{Ja} = \frac{3a}{3a+b+c}$
15	Fai	Faith (1983)[36]	$s_{Fai} = \frac{a+0.5d}{p}$
16	Mou	Mountford (1962)[37]	$s_{Mou} = \frac{2a}{ab+ac+2bc}$
17	Mic	Michael (1920)[38]	$s_{Mic} = \frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$
18	RG	Rogot–Goldberg (1966)[39]	$s_{RG} = \frac{a}{2a+b+c}+\frac{d}{2d+b+c}$
19	HD	Hawkins–Dotson (1968)[40]	$s_{HD} = \frac{1}{2}\left(\frac{a}{a+b+c}+\frac{d}{b+c+d}\right)$
20	Yul	Yule (1900, 1912)[41,42]	$s_{Yul} = \frac{ad-bc}{ad+bc}$
21	Yu2	Yule (1900, 1912)[41,42]	$s_{Yu2} = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$
22	Fos	Fossum in Holiday et al. (2002)[11]	$s_{Fos} = \frac{p(a-0.5)^2}{(a+b)(a+c)}$
23	Den	Dennis in Holiday et al. (2002)[11]	$s_{Den} = \frac{ad-bc}{\sqrt{p(a+b)(a+c)}}$
24	Co1	Cole (1949)[43]	$s_{Co1} = \frac{ad-bc}{(a+c)(c+d)}$
25	Co2	Cole (1949)[43]	$s_{Co2} = \frac{ad-bc}{(a+b)(b+d)}$
26	dis	dispersion in Choi et al. (2012)[44]	$s_{dis} = \frac{ad-bc}{p^2}$
27	GK	Goodman–Kruskal (1954)[45]	$s_{GK} = \frac{2\min(a,d)-b-c}{2\min(a,d)+b+c}$
28	SS3	Sokal–Sneath (1963)[35]	$s_{SS3} = \frac{1}{4}\left[\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{b+d}+\frac{d}{c+d}\right]$
29	SS4	Sokal–Sneath (1963)[35]	$s_{SS4} = \frac{a}{\sqrt{(a+b)(a+c)}}\frac{d}{\sqrt{(b+d)(c+d)}}$
30	Phi	Pearson–Heron (1913)[46]	$s_{Phi} = \frac{ad-bc}{\sqrt{(a+b)(a+c)(c+d)(b+d)}}$
31	Di1	Dice (1945),[25] Wallace (1983),[47] Post–Snijders (1993)[48]	$s_{Di1} = \frac{a}{(a+b)}$
32	Di2	Dice (1945),[25] Wallace (1983),[47] Post–Snijders (1993)[48]	$s_{Di2} = \frac{a}{(a+c)}$
33	Sor	Sorgenfrei (1958)[49]	$s_{Sor} = \frac{a^2}{(a+b)(a+c)}$
34	Coh	Cohen (1960)[20,50]	$s_{Coh} = \frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}$
35	Pe1	Peirce (1884)[51]	$s_{Pe1} = \frac{ad-bc}{(a+b)(c+d)}$
36	Pe2	Peirce (1884)[51]	$s_{Pe2} = \frac{ad-bc}{(a+c)(b+d)}$
37	MP	Maxwell–Pilliner (1968)[52]	$s_{MP} = \frac{2(ad-bc)}{(a+b)(c+d)+(a+c)(b+d)}$
38	HL	Harris–Lahey (1978)[53]	$s_{HL} = \frac{a(2d+b+c)}{2(a+b+c)}+\frac{d(2a+b+c)}{2(b+c+d)}$
39	CT1	Consonni–Todeschini (2012)[54]	$s_{CT1} = \frac{\ln(1+a+d)}{\ln(1+p)}$
40	CT2	Consonni–Todeschini (2012)[54]	$s_{CT2} = \frac{\ln(1+p)-\ln(1+b+c)}{\ln(1+p)}$
41	CT3	Consonni–Todeschini (2012)[54]	$s_{CT3} = \frac{\ln(1+a)}{\ln(1+p)}$
42	CT4	Consonni–Todeschini (2012)[54]	$s_{CT4} = \frac{\ln(1+a)}{\ln(1+a+b+c)}$
43	CT5	Consonni–Todeschini (2012)[54]	$s_{CT5} = \frac{\ln(1+ad)-\ln(1+bc)}{\ln(1+p^2/4)}$
44	AC	Austin–Colwell (1977)[55]	$s_{AC} = \frac{2}{\pi}\arcsin\sqrt{\left(\frac{a+d}{p}\right)}$
45	Han[b]	Hamann (1961),[56] Holley–Guilford (1964),[57] Hubert (1977)[58]	$s_{Ham} = \frac{a+d-b-c}{p}$
46	McC[b]	McConaughey (1964)[59]	$s_{Mc} = \frac{a^2-bc}{(a+b)(a+c)}$
47	GL[b]	Gower–Legendre (1986)[60]	$s_{GL} = \frac{a+d}{a+0.5(b+c)+d}$
48	BU2[b]	Baroni-Urbani–Buser (1976)[33]	$s_{BU2} = \frac{\sqrt{ad}+a-b-c}{\sqrt{ad}+a+b+c}$
49	Joh[b]	Johnson (1967)[61]	$s_{Joh} = \frac{a}{a+b}+\frac{a}{a+c}$
50	Sco[b]	Scott (1955)[62]	$s_{Sco} = \frac{4ad-(b+c)^2}{(2a+b+c)(2d+b+c)}$
51	Maa[b]	van der Maarel (1969)[63]	$s_{Maa} = \frac{2a-b-c}{2a+b+c}$

a désigne le nombre de 1 communs aux deux colonnes,

d désigne le nombre de 0 communs aux deux colonnes,

c et d correspondent aux occurences de discordance

(valeur 1 pour une colonne et 0 pour l'autre).

Quelques distances possibles en 2017 via R

Via la fonction **dist()** du package **stats** pour le logiciel **R** :

euclidean maximum manhattan canberra binary minkowski

Via la fonction **vegdist()** du package **vegan** pour le logiciel **R** :

bray	kulczynski	jaccard	gower	altGower	morisita
horn	mountford	raup	binomial	chao	mahalanobis

Logiciel **R**, septembre 2017 :

- environ **14 000** packages,

- soit à peu près **1 million 925 mille** fonctions.

Et si c'était à refaire en 2017 (2) ?

Langage : **R** ou **Python** [2.7.14 *vs* 3.3.7 *vs* 3.6.2] ?

Utilisation via :

- une application ?
- un script ?
- une page Web ?
- un jupyter notebook ?

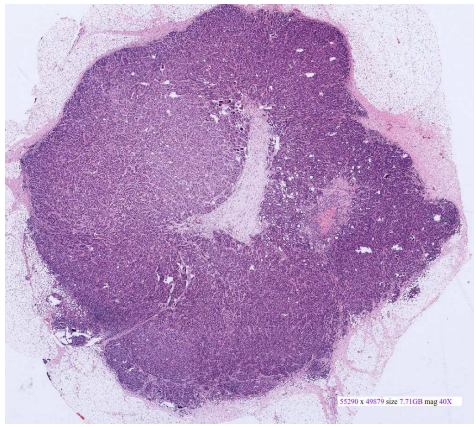
...quid des données ternaires ?

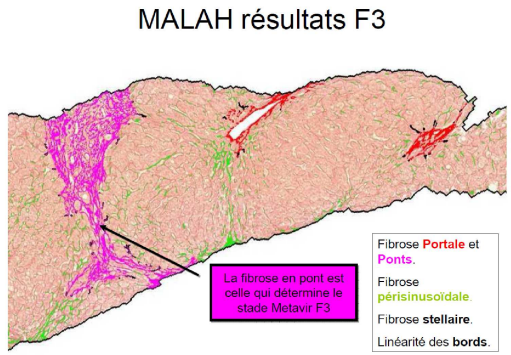
Au passage, thèse GH (1983) en statistiques mathématiques à Paris VI, direction J. P. Benzécri : ***Classification hiérarchique de variables qualitatives.***

Et si c'était à refaire en 2017 (3) ? Script R équivalent

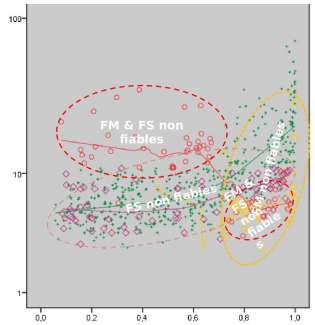
```
library(vegan) # pour vegdist()
library(ape)     # pour as.phylo() et plot.phylo()

# lecture des données

xmp2 <- as.matrix(read.table("xmp2.data",header=FALSE,row.names=1))

# calcul de la matrice des distances

mdc <- vegdist(x=xmp2,method="jaccard")

# classification hiérarchique

cah <- hclust(mdc,method="average") # correspond à UPGMA
desc <- data.frame(cbind(round(cahheight, 3), cahmerge))
names(desc) <- c("Niveau","Aine","Benjamin")

# tracé via plot.hclust et plot.phylo

plot(cah, hang=-1,main="Dendrogramme 1")
plot.phylo(as.phylo(cah), direction="leftwards",
        main="Dendrogramme 2",label.offset=0.01)
```

3. Comment garantir la fiabilité en régression logistique ?





F0	sans fibrose
F1	fibrose portale sans septa
F2	fibrose portale et quelques septa
F3	fibrose septale pré-cirrhose
F4	cirrhose



Position du problème de recherche clinique

*On s'intéresse au **diagnostic** (pas au pronostic) de fibrose hépatique à l'aide de marqueurs biologiques (age, sexe, bilan sanguin...).*

*La **référence** ou « gold standard » – bien que partiellement fausse – se nomme stade METAVIR F et s'exprime en 5 classes progressives de F=0 à F=4.*

On dispose d'images de biopsies hépatiques avec des programmes d'analyse d'images pour définir/vérifier les stades METAVIR.

*On dispose de **deux cibles binaires privilégiées** avec nécessité de traiter, nommées fibrose avancée ($F \geqslant 2$) et cirrhose hépatique (F=4).*

*On a déjà sélectionné des modèles de **régression logistique binaire** à 4, 5, 6 ou 7 variables pour ces cibles suivant l'étiologie (alcool, virus, stéatopathie) via leur performance (AUROC, AIC, YOUDEN...).*

Comment minimiser les faux-négatifs, les faux-positifs et quantifier la fiabilité des résultats pour des scores exprimés sur les 5 classes F_i ?

Choix mathématiques, statistiques et informatiques (1)

Puisque la sélection de variables a déjà été faite, il faut essayer de

- *comprendre du point de vue médical d'où viennent les faux-négatifs, les faux-positifs.*
- *trouver mathématiquement comment prendre en compte ces faux-négatifs et ces faux-positifs.*
- *réussir à transcrire la fiabilité des résultats au praticien et au patient.*

L'idée de base a consisté à

- *encadrer le comportement des régressions logistiques en fonction de seuils d'alertes (débordements) pour les marqueurs.*
- *comparer les scores issus des différentes régressions logistiques.*

Mais hélas...

Difficultés de modélisation

Hélas...

Limites fictives de Fibromètres

Variable	Grandat	Tychorée	Balsacine	α-βdine	Globulose	Age	Sexe	V. sup.	Irican
indice	1	2	3	5	6	7	8	9	10
Refus inf	3	0.1	0.1	5	10	1	-1	0.1	0.1
Al. inf maxi	65	-1	-1	75	44	21	-1	-1	-1
Al. inf mini	150	9	-1	86	132	35	-1	-1	9
Moyenne	250	40	5	100	196	45	-1	40	20
Al. sup. mini	332	79	10.2	110	278	58	-1	300	-1
Al. sup. maxi	442	111	16.4	125	345	69	-1	515	-1
Refus sup	2000	20000	80	165	1000	120	2	3000	20000
Unité	G/l	%	UI/l	mmol/l	mg/dl	années		UI/l	UI/l

- il y a beaucoup de cas à modéliser car on peut avoir une, deux...n alertes simultanées ;
- le prescripteur peut disqualifier un marqueur (sauf l'age) pour des raisons de traitements en cours ;
- l'urée comme seule alerte peut être un facteur de confusion ;
- certains marqueurs sont biologiquement dépendants...

Choix mathématiques, statistiques et informatiques (2)

- sur les milliers de cas possibles, des cas standards d'alertes similaires (comme 2R1J) doivent mener aux mêmes calculs ;
- la comparaison des RLB doit pouvoir se faire soit en supprimant la variable soit en la remplaçant par la « normale » clinique ;
- il faut régler des seuils de distance entre RLB pour les comparer ;
- il faut un indice de fiabilité du résultat exprimé en % ;
- il faut injecter à chaque niveau de décision des compétences médicales.

Conclusion : un simple programme de calcul statistiques ne suffit pas, il faut développer un *calculateur/système expert* (en chainage avant), avec environ 250 règles de décision et avec une base de données de cas typiques.

Historique des réalisations

1998 : premiers essais de modèles de régressions logistiques binaires multiples (premier test publié en 1997)

2002 : premières modélisations performantes de régressions logistiques

2004 : dépot du premier brevet associé, création de la startup *BioLiveScale*

2006 : premier Fibromètre payant

2011 : - première vente de licence

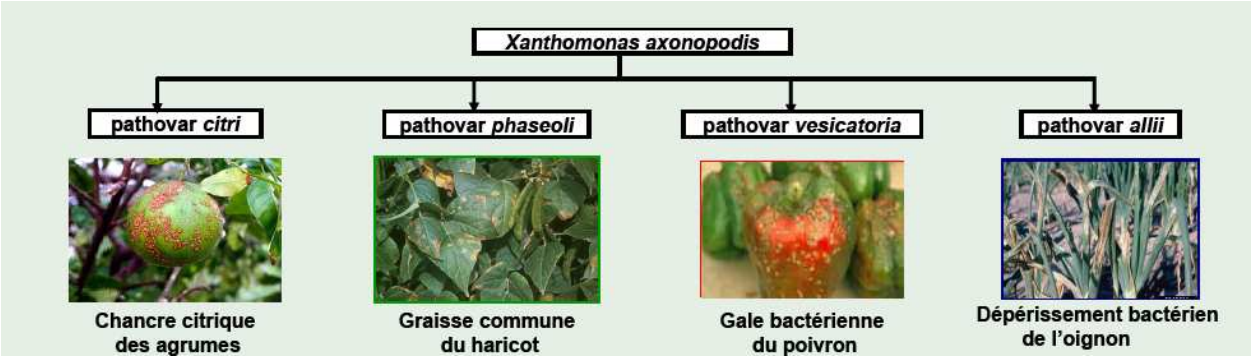
- recommandation nationale (HAS) d'utiliser les tests sanguins non invasifs en première intention dont les Fibromètres

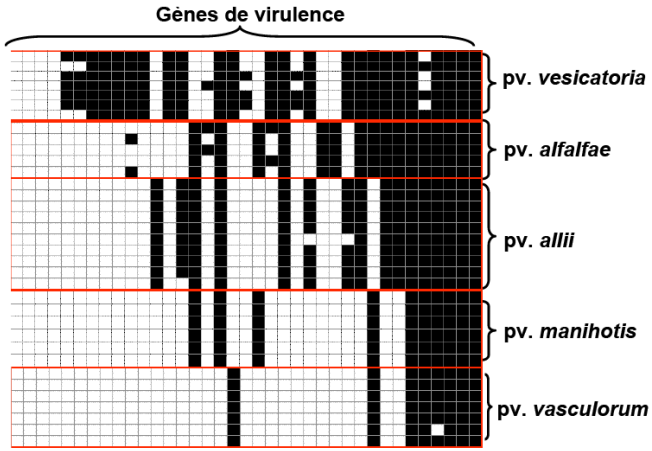
- remboursement par la sécurité sociale d'un Fibromètre par an

Problèmes rencontrés

- réussir à bien définir les échantillons de modélisation et de validation (interne, externe, indépendante) ;
- garantir que les modèles s'appliquent effectivement au « tout-venant » et pas seulement aux patients (malades) du CHU ;
- tester et montrer que des méthodes plus techniques comme les régressions ordinales, polytomiques ne font pas mieux que des RLB ;
- définir un indicateur de fiabilité basé sur une dispersion relative invariante par translation qui sera inclus dans la feuille de résultats ;
- programmer les choix, les faire valider par les cliniciens et trouver comment améliorer les performances jusqu'à un niveau "acceptable" pour une utilisation internationale.

4. Quelle caractérisation spécifique minimale conserver ?





Souche	Groupe	Gènes			
		x_1	x_2	...	x_j
e_1	g_1	1	1	...	0
⋮	⋮	⋮	⋮		⋮
e_m	g_1	1	1	...	1
e_{m+1}	g_2	1	0	...	1
⋮	⋮	⋮	⋮		⋮
e_n	g_k	0	1	...	1

Pour le groupe g_1 : $\phi_{g_1} = DNF_{g_1} \wedge CNF_{g_1}$ avec

- $DNF_{g_1} = (x_1 \wedge x_2 \wedge \cdots \wedge \neg x_j)_1 \vee \cdots \vee (x_1 \wedge x_2 \wedge \cdots \wedge x_j)_m$
- $CNF_{g_1} = (\neg x_1 \vee x_2 \vee \cdots \vee \neg x_j)_{m+1} \wedge \cdots \wedge (x_1 \vee \neg x_2 \vee \cdots \vee \neg x_j)_n$

Dialogue de sourds (1)

[B] *J'ai des groupes de lignes de pathovars de Xanthomonas avec des colonnes de présence/absence de gènes de virulence et je cherche à savoir quelles combinaisons de colonnes caractérisent les groupes à 100 %.*

[M] *C'est quoi Xanthomonas ? et pathovars ?*

[B] *1. des bactéries pathogènes (plus graves que Pseudomonas) ;*
2. une commodité de classement intraspécifique.

[M] *OK, je vous fais une régression logistique multinomiale ou une analyse discriminante...*

[B] *C'est quoi une analyse discriminante ?*

[M] *Une méthode de modélisation/prédiction statistique qui renvoie des probabilités d'appartenance à chaque classe.*

4. Quelle caractérisation spécifique minimale conserver ?

Où est le problème ?

Attention *:*

on ne veut pas un prédicteur de classes pour chaque ligne mais des descripteurs minimaux et spécifiques (exacts à 100 % pour chaque groupe, 0 % pour les autres groupes).

Une solution statistique classique de sélection de variables est-elle adaptée ?

NON. Aucune méthode statistique ne garantit de prédiction exacte à 100 %.

Peut-on essayer d'emboiter des combinaisons de variables avec des coefficients de capacité à diagnostiquer des classes avec des critères d'entropie, d'information au sens de Shannon, de tester des fréquences d'appartenance a priori et a posteriori ?

NON. Là encore, aucune méthode ne garantit de prédiction exacte à 100 %.

4. Quelle caractérisation spécifique minimale conserver ?

Quelques questions à se poser pour résoudre le problème

- est-on sûr d'avoir forcément au moins une solution ?
- que déduire s'il n'y a pas de solution ?
- et au contraire, que faire s'il y a plus d'une solution ?
- comment faire si les statistiques n'ont pas de méthode adaptée pour résoudre ce problème ?
- au fait, s'agit-il d'un problème connu et classique ?

4. Quelle caractérisation spécifique minimale conserver ?

Quelques éléments de réponses après de nombreux essais

- le problème posé n'est pas «simple» car il est équivalent à un problème NP-complet Σ_2^p de classe W2 de minimisation d'expressions logiques DNF (formes normales disjonctives) ;
- cependant en pratique, sur les données fournies par l'INRA on "trouve" souvent des solutions par tatonnement ou par programme avec 5 ou 6 colonnes impliquées ;
- il est très simple de construire un jeu de données sans solution à cause d'au moins une contradiction logique ;
- une approche informatique combinatoire semble pouvoir trouver ces solutions en un temps raisonnable (stage de M2).

Un «solveur» comme solution

Une combinaison de colonnes en 0/1 peut être assimilée à une formule logique avec 0=FAUX et 1=VRAI.

Une solution exhaustive en disjonction de conjonctions existe mais elle est assimilable à du sur-apprentissage : le groupe G_i, c'est l'élément e_{i_1} ou e_{i_2} ou e_{i_3} ... ; l'élément e_{i_j}, c'est sa valeur en colonne 1 et sa valeur en colonne 2 et ...

*Il faut donc inventer un «solveur» pour réduire, simplifier ces formules et trouver des caractérisations à la fois **spécifiques** et **minimales**.*

Mais pour quelle minimalité ?

De quelle minimalité s'agit-il ?

Solution minimale en longueur par groupe (6 colonnes en tout)

Groupe	Formule	Longueur	Taille	Nb. Colonnes
Grp1	C1=1 et C2=1 et C3=0	3	3	3
Grp2	C4=1 et C5=0 et C6=0	3	3	3

Solution minimale en nombre de colonnes (4 colonnes en tout)

Groupe	Formule	Longueur	Taille	Nb. Colonnes
Grp1	C1=1 et C2=1 et C4=0 et C5=1	4	4	4
Grp2	C1=0 et (C2=0 ou C4=1) et C5=0	3	4	4

4. Quelle caractérisation spécifique minimale conserver ?

Stratégie de résolution retenue

- le solveur est capable d'exhiber une solution ou toutes les solutions spécifiques et minimales en colonnes ou en longueur s'il n'y a pas de contradictions dans les données ;

- si les données comportent peu de contradictions :

  $\rightarrow$ création et analyse de sous-ensembles de données sans contradiction

- au cas où il y a dans les données de nombreuses contradictions :

  $\rightarrow$ rejet des données (manque de puissance d'expression)

4. Quelle caractérisation spécifique minimale conserver ?

Réalisations effectuées sur 3 ans

- un solveur **efficace** avec de nombreuses options de calcul pour les données de l'INRA (thèse de F. Chhel) ;
- une interface Web avec plusieurs formats d'entrée/sortie (dont Excel et XML) ;
- plus d'un quinzaine de caractérisations INRA réelles obtenues ;
- 4 articles et 2 posters en biologie et en informatique ;
- dépot d'un brevet et réalisation d'une puce avec la société DIAG-GENE ;

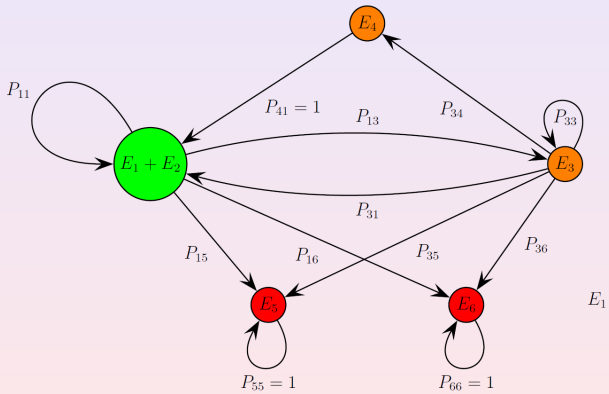
\+ une étude informatique plus poussée en cours sur le problème de caractérisation multiple.

5. Ne parlons pas de l'estimateur d'Aalen-Johansen !

Notations

- $X = \{X(t), t \geqslant 0\}$ un processus stochastique à temps continu et à espace d'états fini $E = \{1, \ldots, K\}$;
- $C_j = \{C_j(t), t \geqslant 0\}$ un processus stochastique à temps continu et à valeurs dans $\mathbb{R}_+$, $C_j(t)$ représentant le coût de l'événement de type j à l'instant t ;
- Q_j : $Q_j(t) = \sum_{k \geqslant 1} C_j(T_k) \mathbb{1}_{\{T_k \leqslant t, X(T_k) = j\}}$ un processus stochastique à temps continu et à valeurs dans $\mathbb{R}_+$, $Q_j(t)$ représentant le coût accumulé sur $[0, t]$ généré par les événements de type j, où $\{T_k, k \geqslant 1\}$: ensemble des instants successifs de survenue des événements ;
- $Q(t) = \sum_{j \in E} Q_j(t)$: coût total accumulé sur $[0, t]$;

Soit n observations $i.i.d.$ $(C_{ji}(.), X_i(.), T_{ki})$ de $(C_j(.), X(.), T_k)$



E_1 : sans toxicité,

E_2 : toxicité mineure, E_3 : toxicité majeure, E_4 : interruption du traitement à cause de toxicité, E_5 : décès dû à une toxicité, E_6 : progression ou décès dû à l'évolution de la maladie

Quantité d'intérêt

$$Q_j^*(t) = \int_0^t \mathbb{1}_{\{X(u)=j\}} dQ_j(u)$$

Coût médical moyen accumulé sur $[0, t]$

$$\mu_j(t) = \mathbb{E}[Q_j^*(t)] \qquad \text{et} \qquad \mu(t) = \sum_{j \in E} \mu_j(t)$$

$$\begin{aligned} \mu_j(t) &= \int_0^t \mathbb{P}(X(u) = j) dV_j(u) \quad \text{où } V_j(u) = \mathbb{E}[dQ_j(u) | X(u) = j]. \\ &= \int_0^t \pi_j(u) dV_j(u) \quad \text{où} \quad \pi_j(u) = \mathbb{P}(X(u) = j) \end{aligned}$$

Estimation des probabilités d'états à travers l'estimateur d'A-J

$$\hat{\pi}_j(t) = \sum_{h=0}^{K} \hat{\pi}_h(0) \hat{P}_{hj}(0, t)$$

$(\hat{P}_{hj}(0,t))_{hj \in E}$ forment la matrice $\hat{P}(0,t)$ donnée par :

$$\hat{P}(0,t) = \prod_{T_i \leqslant t} \left(Id_K + \Delta \hat{A}(T_i) \right)$$

Estimation des intensités de transition

$$\hat{A}_{hj} = \begin{cases} \dfrac{d\bar{N}_{hj}(u)}{\bar{Y}_h(u)} & \text{si h}\neq\text{j} \\ \hat{A}_{hh}(t) = -\sum_{j:j \neq h} \hat{A}_{hj}(t) & \text{si h=j} \end{cases}$$

$\bar{N}_{hj}(t) = \sum_{i=1}^{n} \sum_{k \geqslant 1} \mathbb{1}_{\{T_{k,i} \leqslant t, X(T_{k,i}) = j, X(T_{k-1,i}) = h\}}$ et
$\bar{Y}_j(t) = \sum_{i=1}^{n} \sum_{k \geqslant 1} \mathbb{1}_{\{T_{k-1,i} < t \leqslant T_{k,i}, X(T_{k-1,i}) = j\}}$

5. Ne parlons pas de l'estimateur d'Aalen-Johansen !

Le problème posé

- dans un cadre d'évaluation médico-économique, il s'agit d'une analyse de type coût-efficacité ;
- on doit étudier des données incomplètes «censurées» (patients décédés, perdus de vue, sortis de l'étude...) ;
- la censure, ici à droite de type I (fixe ou aléatoire), est *informative*, ce qui rend les modèles de Kaplan-Meier des données de survie seules inadaptés ;
- on doit l'appliquer au bénéfice potentiel d'un dépistage pré-thérapeutique des toxicités induites par le 5-FU (fluorouracile) dans le cas du traitement du cancer colo-rectal ;
- on doit prendre en compte le fait qu'il s'agit d'un modèle multi-états à risques compétitifs et récurrents.

5. Ne parlons pas de l'estimateur d'Aalen-Johansen !

Contexte

On estimait à 37 413 nouveaux cas par an et 16 865 décès par an de cancers colorectaux, en 2005 en France selon l'INCa et l'InVS. Malgré le développement de nouvelles molécules comme l'Irinotécan et l'Oxaliplatine, le 5-fluorouracile (5-FU) reste le traitement de référence en chimiothérapie dans ce type de cancer. Si l'efficacité du 5-FU est définitivement admise tant en terme de réponse qu'en terme d'allongement de la durée de survie des patients (cf. [20, 21, 22]), il n'en demeure pas moins qu'il engendre encore des toxicités très sévères chez certains patients (altération de la qualité de vie, toxicités mortelles). On estime entre $0,3$ % et $1,2$ % la fréquence des décès liés au traitement selon les protocoles habituels à base de 5-FU et entre 25 % et 30 % la fréquence des toxicités graves (grade III-IV OMS), (cf. essai MOSAIC [23], essai NSABP C-07 [24]). En outre, 14 % des patients arrêtent le traitement à cause d'une toxicité [23]. Enfin, les toxicités mobilisent d'importantes ressources : Tsalic *et al.*, 2003 [25] ont observé 13 % d'hospitalisations à cause de toxicités dues au 5-FU dans une étude prospective portant sur 243 patients traités par un protocole standard à base de 5-FU. Le coût par patient à 10 mois est estimé à 2793 € pour une incidence des toxicités majeures à 31 % (Delea et al., 2002 [26]) aux États-Unis d'Amérique.

5. Ne parlons pas de l'estimateur d'Aalen-Johansen !

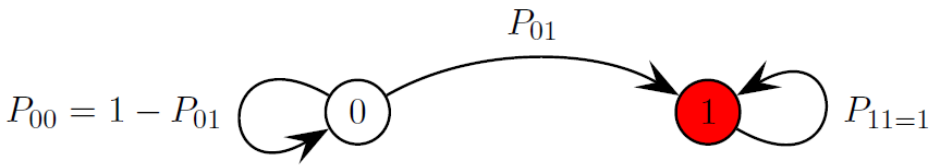
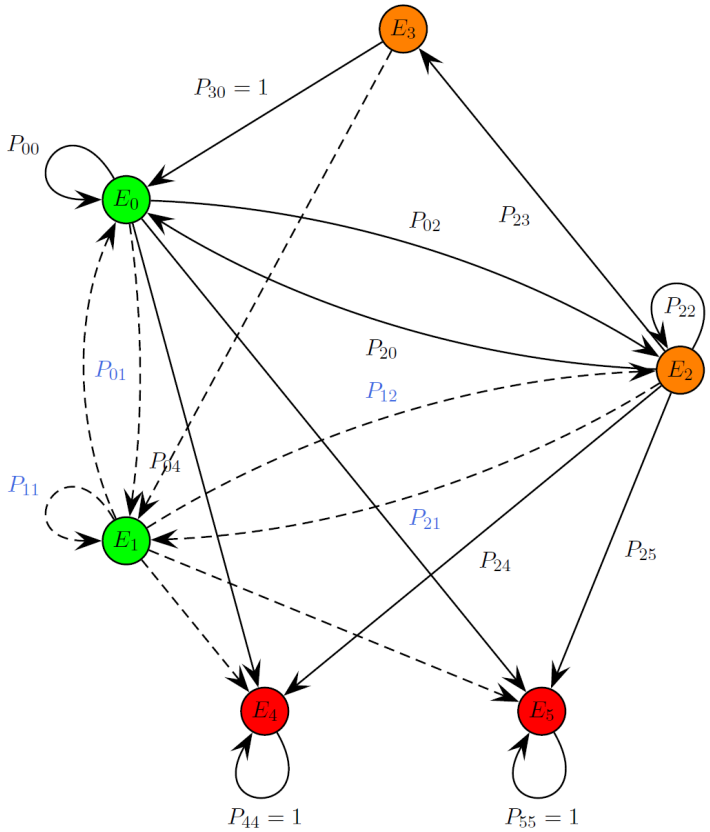


FIGURE 2.6 – Un processus de Markov à deux états (0 : vivant, 1 : décès) : processus de survenue du décès

5. Ne parlons pas de l'estimateur d'Aalen-Johansen !



E_0 : sans toxicité, E_1 : toxicité mineure, E_2 : toxicité majeure, E_3 : interruption du traitement à cause de toxicité, E_4 : décès dû à une toxicité, E_5 : progression ou décès non lié à une toxicité.

5. Ne parlons pas de l'estimateur d'Aalen-Johansen !

Numéro	i	$X(.)$	C_k (€)	CHT
1	1	0	0	A
2	1	2	1032	A
3	1	1	1032	A
4	2	1	0	B
5	2	3	5455	B
6	2	4	5455	B
7	3	0	0	B
8	3	0	0	B
9	3	1	0	B
10	3	0	0	B

TAB. 4.1 – Données extraites pour trois patients parmi 1742

5. Ne parlons pas de l'estimateur d'Aalen-Johansen !

Coût (en €)	État E_0	État E_1	État E_2	État E_3	Total
0	886	0	0	0	886
1032	48	10	4	1	63
1542	26	8	2	0	36
2065	26	4	1	0	31
2562	6	1	3	0	10
4105	6	1	3	0	10
5124	5	1	3	0	9
5455	0	0	1	0	1
6488	0	0	1	0	1
10911	0	0	1	0	1

TAB. 4.2 – Effectifs observés dans les états en fonction du coût total accumulé pour les patients de la cohorte B

Réalisations (1)

Dans le cadre de la thèse de S. Traoré, deux estimateurs non-paramétriques du coût médical cumulé ont été proposés et étudiés théoriquement (biais, consistance, vitesse de convergence et normalité asymptotique). Ces estimateurs ont ensuite été comparés aux méthodes existantes sur des ***données simulées*** *au niveau du biais, des probabilités de couverture de la moyenne théorique (par un intervalle de confiance défini selon une loi normale) avant d'être appliqués à des* ***données réelles****, dont celles du centre Paul PAPIN.*

Proposition

1. $n^{1/2}(\hat{\mu}_j - \mu_j) \xrightarrow{ps} 0$ (consistence et vitesse de convergence)
2. $n^{1/2}(\hat{\mu}_j - \mu_j)$ converge faiblement vers un processus gaussien de moyenne 0 et de fonctions de covariance en (s,t) estimée par $\hat{\xi}_j(s,t) = \frac{1}{n}\hat{\psi}_{ji}(s)\hat{\psi}_{ji}(t)$ où

$$\hat{\psi}_{ji}(t) = \int_0^t \frac{\hat{\pi}_j(u)d\mathcal{M}_i^j(u)}{\bar{Y}_j(u)/n} - \hat{\mu}_j(t)\int_0^t \frac{dM_i^j(u)}{\bar{Y}_j(u)/n} + \int_0^t \frac{\hat{\mu}_j(u)dM_i^j(u)}{\bar{Y}_j(u)/n}$$

$$\mathcal{M}_i^j(t) = Q_{ji}(t) - \int_0^t Y_{ji}(u)d\hat{V}_j(u)$$

$$M_i^j(t) = N_i^j(t) - \int_0^t Y_{ji}(u)d\hat{H}_j(u)$$

Preuve : (extension Gosh et Lin (Biometrics, 2000)).

Réalisations (2)

La thèse a permis de tester et de valider deux hypothèses :

- ***Hypothèse 1 (médicale)*** :

Le dépistage pré-thérapeutique réduit l'incidence des toxicités liées au 5-FU, évite les décès dus au 5-FU sans diminuer le délai avant progression tumorale.

- ***Hypothèse 2 (économique)*** :

Le coût supplémentaire que ce dépistage génère est inférieur au coût de prise en charge des toxicités qu'il permet d'éviter.

Applications aux données réelles (1) :

Temps(en 2 semaines) t	$\pi_0(t)$	$\pi_1(t)$	$\pi_2(t)$	$\pi_3(t)$	$\pi_4(t)$	$\pi_5(t)$
1	0,9474	0,0478	0,0046	0	0	0
2	0,9217	0,0689	0,0093	0	0	0
3	-	-	-	-	-	-

TABLE 5.1 – Probabilités d'états ou prévalences des états de santé dans la population A.

Temps(en 2 semaines) t	$\pi_0(t)$	$\pi_1(t)$	$\pi_2(t)$	$\pi_3(t)$	$\pi_4(t)$	$\pi_5(t)$
1	0,5135	0,4209	0,0575	0,0045	0,0011	0,0022
2	0,5045	0,4176	0,0688	0,0033	0,0011	0,0045
3	0,5316	0,4266	0,0316	0,0045	0,0011	0,0045

TABLE 5.2 – Probabilités d'états ou prévalences des états de santé dans la population B.

La cohorte A correspond au dépistage pré-thérapeutique du CPP.

Constituée entre septembre 1993 et août 2007, elle met en jeu 856 patients.

La cohorte B est celle de l'essai C-96 (André et al.). Elle repose sur une la stratégie « standard » qui consiste à administrer les doses traditionnelles de 5-FU selon la posologie du protocole utilisé. 886 patients avaient été inclus entre juillet 1996 et novembre 1999. Tous les patients de cette population étaient en traitement adjuvant. Cette cohorte constitue le groupe de témoins de l'évaluation.

Applications aux données réelles (2) :

$$\hat{P}^{C}(0,1032) = \begin{pmatrix} 0,3886 & 0,3628 & 0,2352 & 0,0087 & 0 & 0,0045 \\ 0,2647 & 0,4763 & 0,2506 & 0,0055 & 0 & 0,0026 \\ 0,1614 & 0,3006 & 0,5379 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Interprétation de cette matrice des probabilités de transitions estimées par l'estimateur d'Aalen-Johansen :

> *sachant qu'un patient quelconque se trouve dans l'état 1 (absence de toxicité) pour un coût médical accumulé de 0 €, la probabilité qu'il se trouve dans l'état 3 (toxicité majeure) après un coût médical accumulé de 1032 € est 0,2352.*

Applications aux données réelles (3) :

Temps t	Sans toxicité $\hat{\mu}_0(t)$	Toxicité mineure $\hat{\mu}_1(t)$	Toxicité majeure $\hat{\mu}_2(t)$	Arrêt à cause de toxicité $\hat{\mu}_3(t)$	toxicité mortelle $\hat{\mu}_4(t)$	Décès ou rechute $\hat{\mu}_5(t)$	Coût Total $\hat{\mu}(t)$
1	0 €	0 €	11,78 €	0 €	0 €	0 €	11,78 €
2	0 €	0 €	37,79 €	0 €	0 €	0 €	37,79 €
3	-	-	-	-	-	-	-

TABLE 5.5 – Coûts moyens accumulés par patient en fonction du temps, population A.

Temps t	Sans toxicité $\mu_0(t)$	Toxicité mineure $\hat{\mu}_1(t)$	Toxicité majeure $\hat{\mu}_2(t)$	Arrêt à cause de toxicité $\hat{\mu}_3(t)$	Toxicité mortelle $\hat{\mu}_4(t)$	Décès ou rechute $\hat{\mu}_5(t)$	Coût Total $\hat{\mu}(t)$
1	0 €	0 €	144,65 €	8,66 €	6,00 €	0 €	159,16€
2	0 €	0 €	256,58 €	8,66 €	6,00 €	0 €	271,00€
3	0 €	0 €	309,73 €	9,82 €	6,00 €	0 €	325,31€

TABLE 5.6 – Coûts moyens accumulés par patient en fonction du temps, population B.

Conclusions

	A	B
Coût du dépistage	152,76 €	0 €
Coût du traitement	X	X
Coût des toxicités	37,79 €	271,00 €
Total	190,54 €	271,00 €
Coût évité si la stratégie A	80,46€	

A : population dépistée ; B : population non dépistée.

TABLE 7.14 – Calcul du coût évité par patient dépisté.

Des résultats à discuter et conforter.

Des calculs en R, mais pas de package.

Pas de relation avec PRISM (probabilistic model checker).

6. Kullback-Leibler ou Bhattacharyya ?

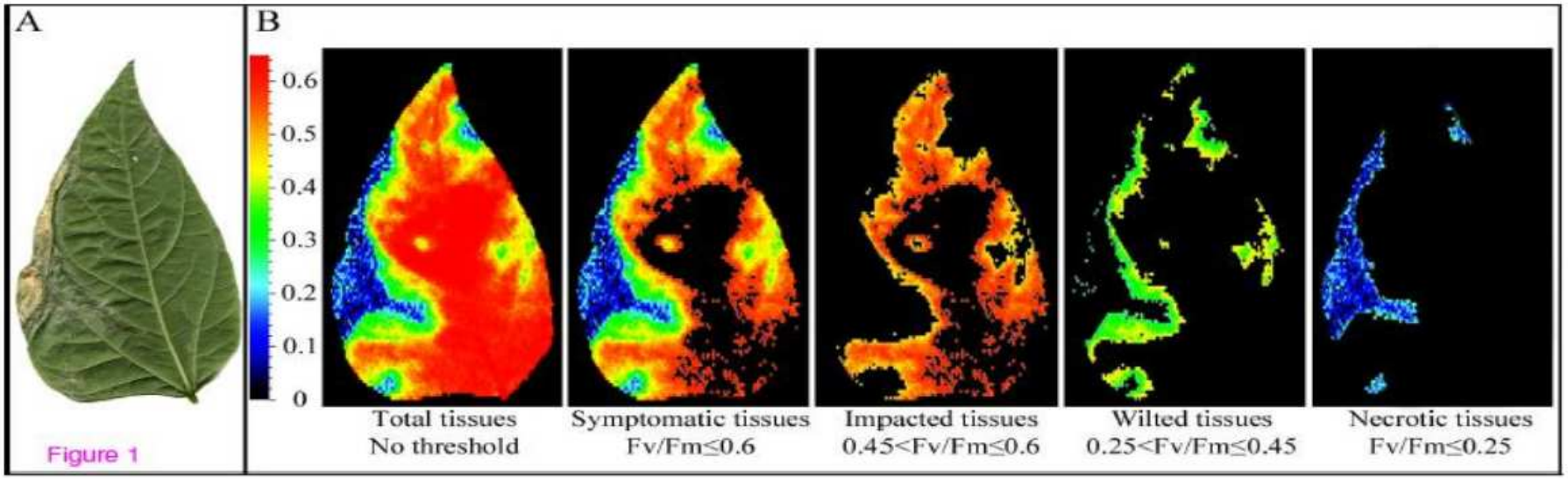


Figure 1

6. Kullback-Leibler ou Bhattacharyya ?

De quoi s'agit-il ?

Dans un contexte d'analyse d'images de fluorescence de chlorophylle produites par l'imageur ***FluoCam7****, on voudrait comparer des images de feuilles saines et de feuilles infectées.*

Ces images sont vues au travers des histogrammes donc on doit comparer des histogrammes.

Quelles distances existent entre histogrammes ?

Y a-t-il des test statistiques pour comparer des histogrammes ?

Au passage : *faut-il normaliser les histogrammes avant comparaison ?*

Quelles comparaisons d'histogrammes ?

On sait classiquement comparer des moyennes, des médianes, des variances, des distributions, mais des histogrammes ? Dans la littérature, on trouve des indices de dissimilarité, des distances comme

- *la distance QF ou "bin-similarity Quadratic-Form distance",*
- *la distance EMD ou "Earth Mover's Distance",*
- *la distance du χ^2 entre histogrammes,*
- *la distance de Hellinger,*
- *la distance de Bhattacharyya,*
- *la distance induite par la divergence de Kullback-Leibler,*
- *la distance de Jenson-Shannon...*

mais attention, il faut parfois entrer la distribution et la fonction utilisée calcule elle-même les histogrammes ; dans certains cas, il faut le même nombre de classes, et/ou avec les mêmes intervalles de boites...

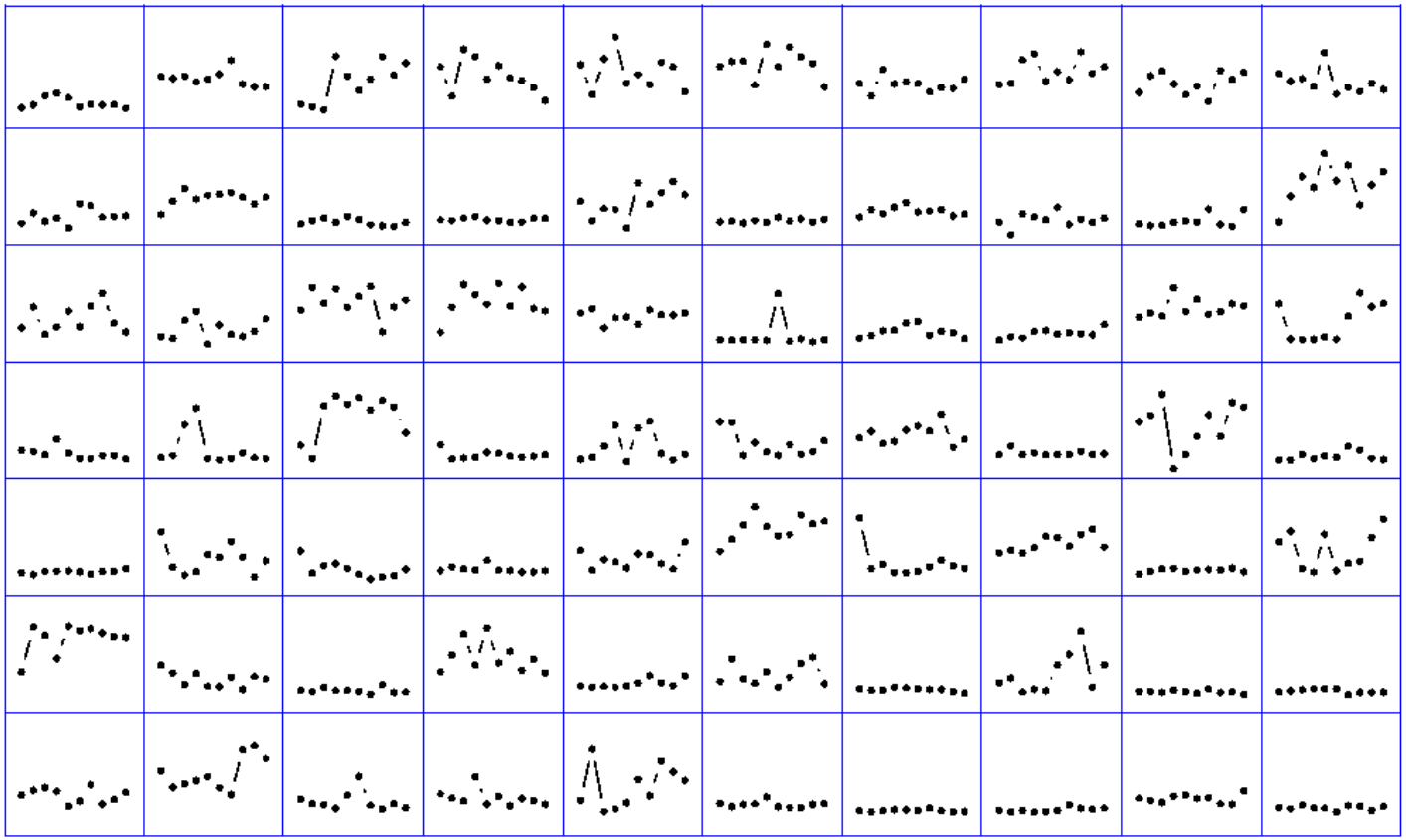
6. Kullback-Leibler ou Bhattacharyya ?

Un choix pragmatique

Pour l'instant (novembre 2017) aucun choix n'a été défini, des analyses sont en cours pour tester les différentes distances et ce qu'elles montrent...

si quelqu'un dans la salle peut nous aider,
il/elle sera le/la bienvenu/e...

7. Faut-il choisir un modèle de prédiction à trois classes ?



7. Faut-il choisir un modèle de prédiction à trois classes ?

Position du problème

On s'intéresse à nouveau au diagnostic de fibrose hépatique.

*Les **données** fournies proviennent de deux centres, Angers (n=522) et Bordeaux (n=532), avec des caractéristiques différentes.*

Pour chaque patient, on dispose de 10 mesures d'élastométrie et d'un diagnostic d'expert en trois classes de niveau de fibrose : FL=1, FL=2 et FL=3 (possiblement FL=1 vs le reste et FL=3 vs le reste).

***Question 1** : Comment modéliser et reproduire l'avis de l'expert ?*

***Question 2** : Combien de mesures faut-il utiliser ?*

Remarque : prendre une mesure dure environ 3 minutes.

7. Faut-il choisir un modèle de prédiction à trois classes ?

Quelques pistes de réflexion

*Il s'agit de **classification supervisée**.*

Il faut certainement synthétiser les données, mais quel(s) indicateurs ou combinaison d'indicateurs sont pertinents ?

moyenne, médiane, minimum, maximum, autre quantile, autre indicateur de tendance centrale ?

écart-type, iqr, mad, autre indicateur de dispersion ?

cdv, iqrr, madr, autre indicateur de dispersion relative ?

7. Faut-il choisir un modèle de prédiction à trois classes ?

Après une revue de plusieurs centaines de modèles

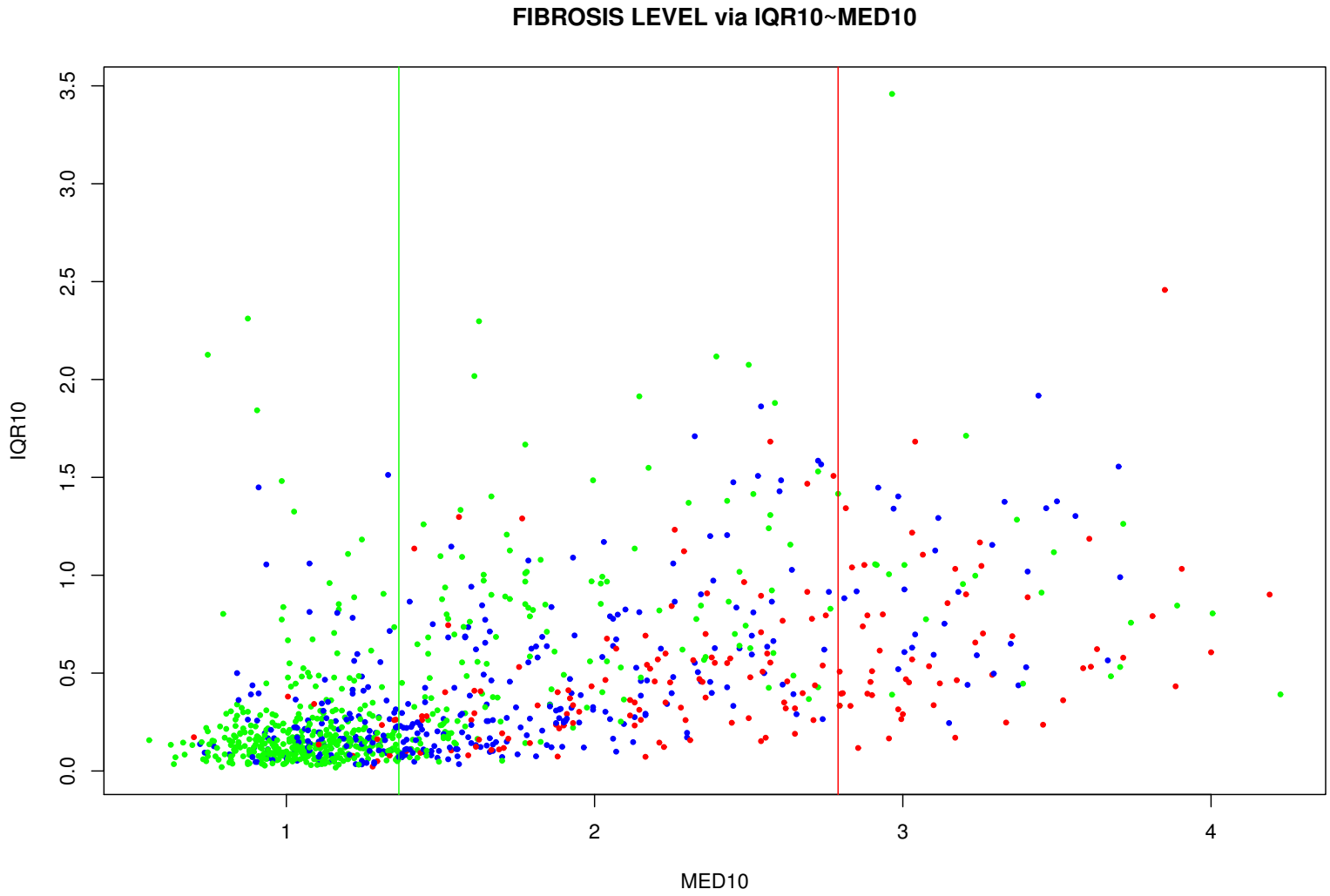
Les performances sont toutes très proches (et pas terribles ! pct BC global 64 %, 44 % pour FL=3). Une stratégie qui modélise les cibles binaires semble donner de meilleurs résultats (AUROC unitaires 0.74 et 0.8, mais AUROC modèle saturé 0.90 avec 98 variables, meilleur forward avec 68 variables, meilleur forward avec 40 variables).

La médiane sur 10 valeurs et l'IQR sur 10 valeurs semblent les meilleurs paramètres capables de prédire les classes.

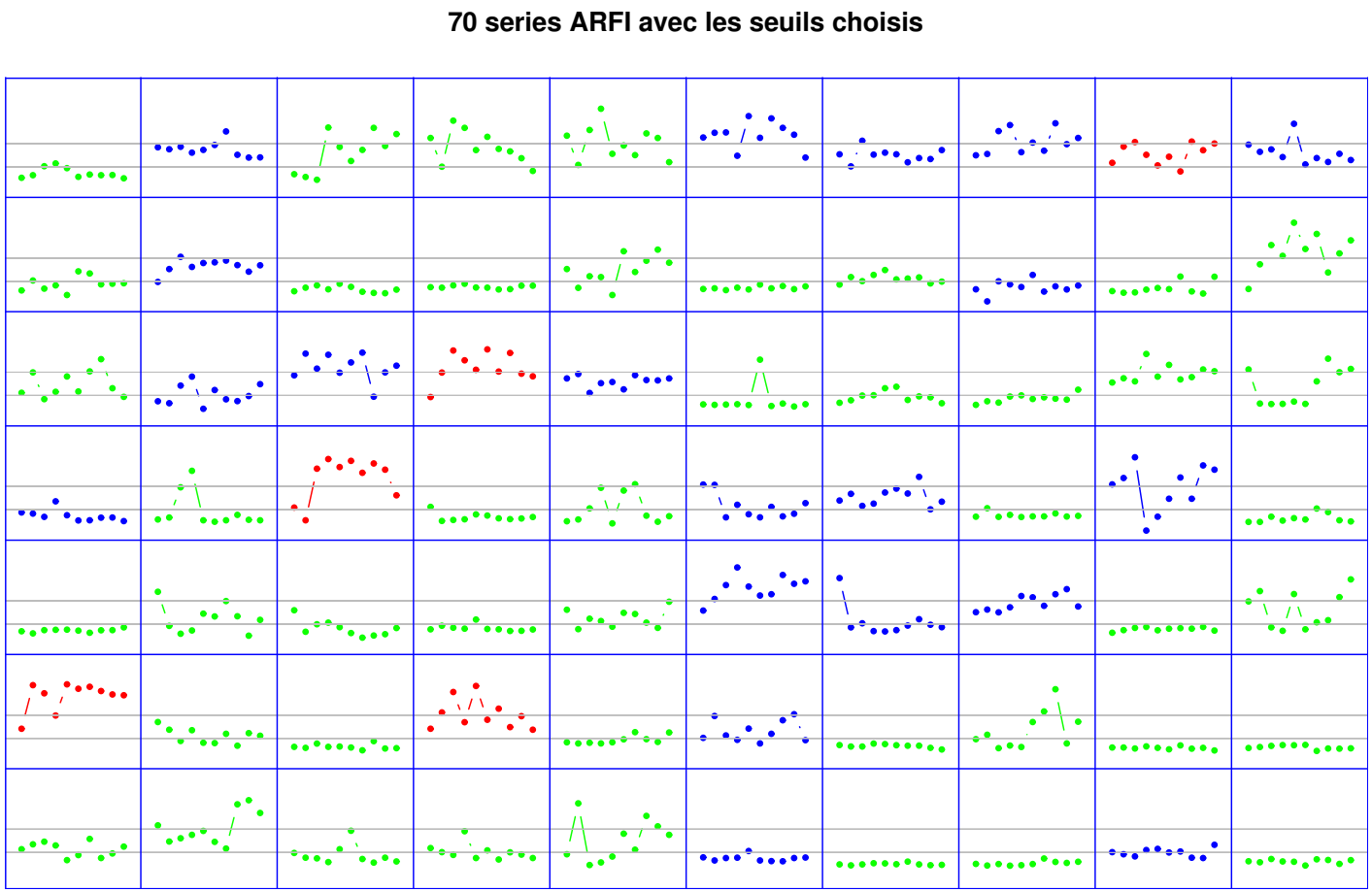
Deux valeurs de la médiane pour cette médiane peuvent servir de seuils pour une classification simplifiée. L'ajout de variables biologiques améliore un peu les performances.

Mais...

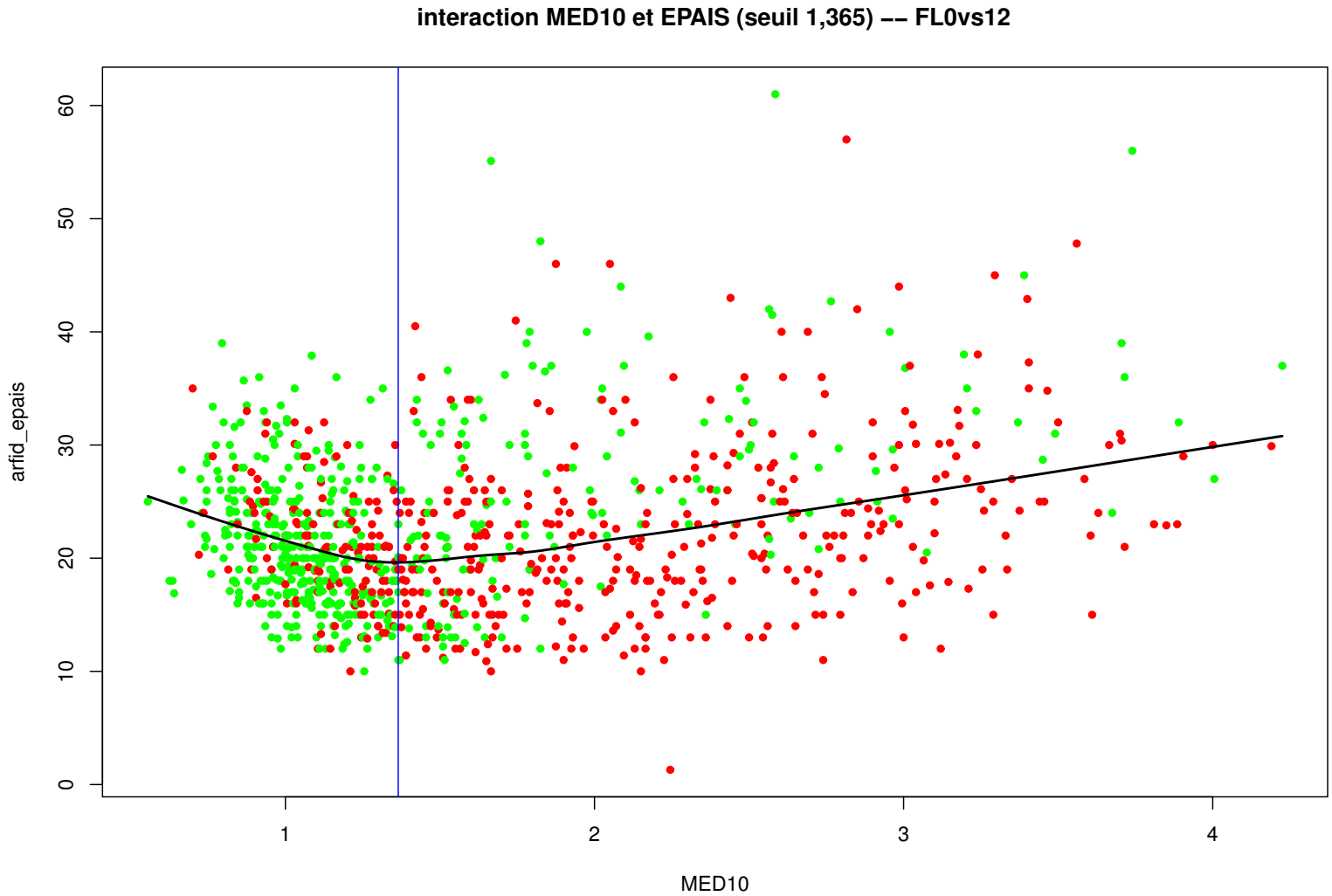
7. Faut-il choisir un modèle de prédiction à trois classes ?



7. Faut-il choisir un modèle de prédiction à trois classes ?



7. Faut-il choisir un modèle de prédiction à trois classes ?



7. Faut-il choisir un modèle de prédiction à trois classes ?

Conclusions

Aucune, étude en cours !

7. Faut-il choisir un modèle de prédiction à trois classes ?

Pourtant...

Une première proposition de deux résultats de de régression logistique binaire (avec quantification de la fiabilité des résultats) ne convient pas aux cliniciens ayant commandé l'étude.

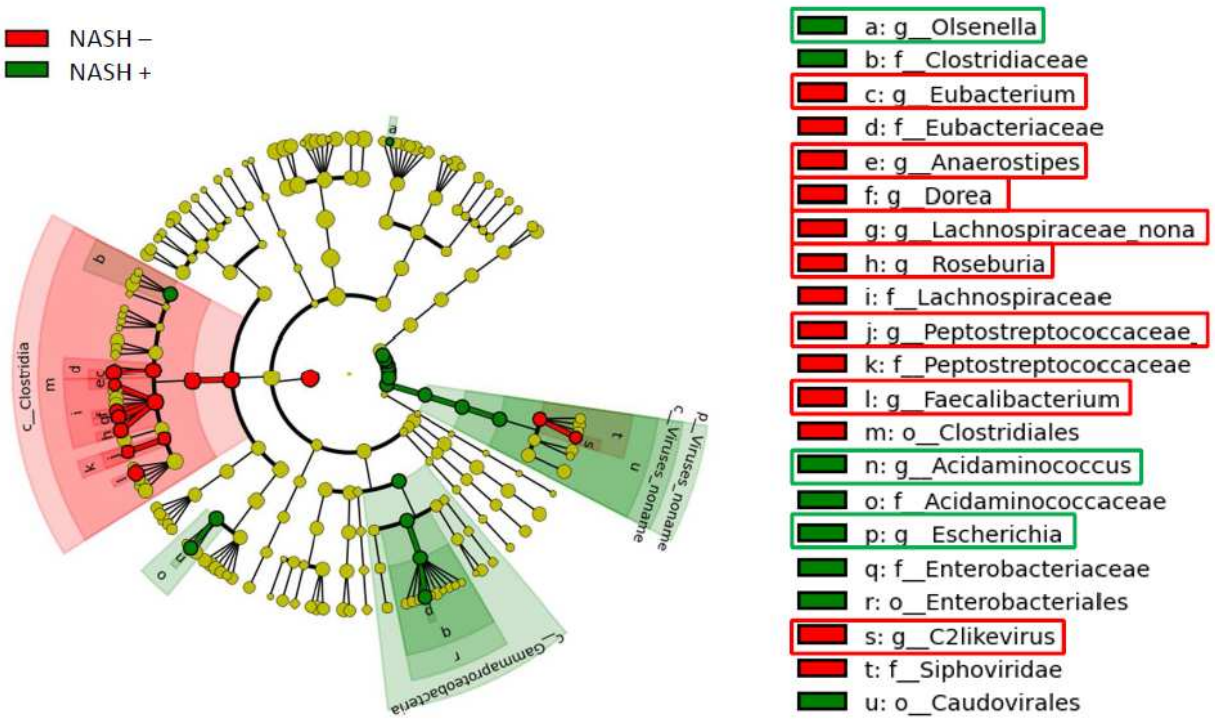
Une seconde série de modélisations plus avancées met en avant une analyse discriminante quadratique «rebinarisée» sur la somme des deux probabilités de classe extrêmes.

La taille des échantillons est sans doute trop faible, mais il n'est pas possible d'attendre 5 ans de plus pour avoir plus de patients (FL=3, 170 patients soit 16 %).

Donc, à suivre...

Identification de taxons associés à la NASH

- LEfSe (Linear discriminant analysis effect size, Segata et al. 2011)



Enfin une question simple ! (1)

En métagénomique, en écologie des microbiotes bactériens/fongiques où on traite des abondances [relatives] d'espèces, IL FAUT (souvent) utiliser :

- *des distances non euclidiennes ;*
- *des positionnements non métriques basés sur l'ordination (les rangs).*

à cause de la «sémantique du double zéro» lorsque les ***co-absences*** *sont non informatives.*

Enfin une question simple ! (2)

Donc exit la norme L^2 et les ACP, et bonjour

- *à l'indice de dissimilarité de Bray-Curtis (qui n'est pas une distance car la propriété d'inégalité triangulaire n'est pas vérifiée) pour les comptages absolus,*
- *à l'indice relatif de Sørensen pour les abondances relatives ;*
- *à la distance UniFrac si on dispose d'informations phylogénétiques ;*
- *au NMDS (Non-metric Multidimensional Scaling)...*

c'est donc simple ?

Quoique...

Quitte à travailler sur des données pour espérer pourquoi ne pas recourir à la topologie et à la TDA (Topological Data Analysis) ?

- *cette méthode n'a pas le défaut de devoir choisir arbitrairement le nombre d'axes principaux comme en ACP,*
- *elle n'oblige pas à projeter dans un espace à n dimensions déterminé arbitrairement par la valeur de stress comme en NMDS.*

Mais... cette théorie repose sur l'homologie persistante (de la topologie algébrique ?) et principalement sur le théorème suivant

Theorem 4.5 *If f is a Morse function on a manifold M and M^a is compact for each a, then M is homotopy-equivalent to a CW complex, with each critical point of index λ corresponding to a different λ-cell.*

donc utilisons les CW-complexes simpliciaux ! ? !

Un exemple de TDA

Using Topological Data Analysis to find discrimination between microbial states in human microbiome data

Mehrdad Yazdani*[1,2], Larry Smarr[1,3] and Rob Knight[4]

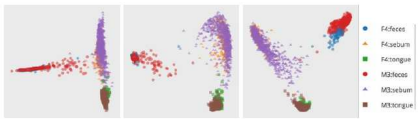
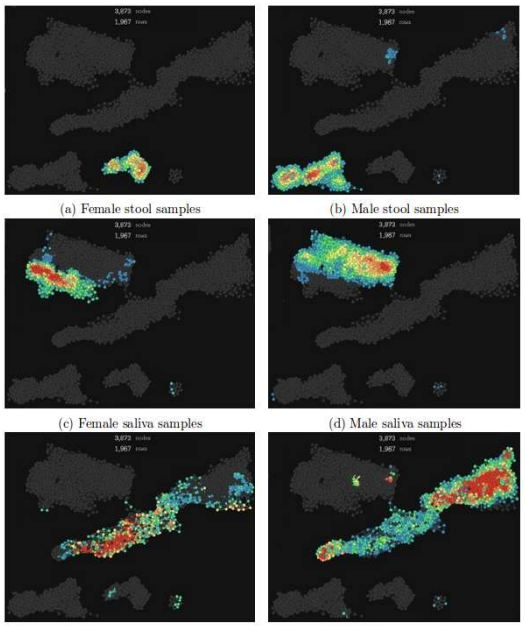


Figure 1: Left panel shows the PCA using of the family relative abundances of our data set. Middle panel shows the MDS of using the Bray-Curtis distance and the right panel is MDS using the UniFrac distance.From this analysis it appears that the samples from the male and female are not significantly different within body sites



(a) Female stool samples (b) Male stool samples

(c) Female saliva samples (d) Male saliva samples

(e) Female skin samples (f) Male skin samples

Figure 2: Unsupervised Topological Data Analysis (TDA, computed using Ayasdi [1]): the color of nodes indicates the proportion of data samples corresponding to the specific subject and body site (red means higher). In (a) and (b) there are two connected components corresponding to female and male stool samples. In (c) and (d) there is little overlap in the samples corresponding to the female and male saliva samples. In (e) and (f), the significant proportions of samples between female and male samples are in shifted in the graph.

4

9. Bowtie2 : *fast* ou *sensitive* ?

Bowtie 2 version 2.2.6 by Ben Langmea

(langmea@cs.jhu.edu, www.cs.jhu.edu/~langmea)

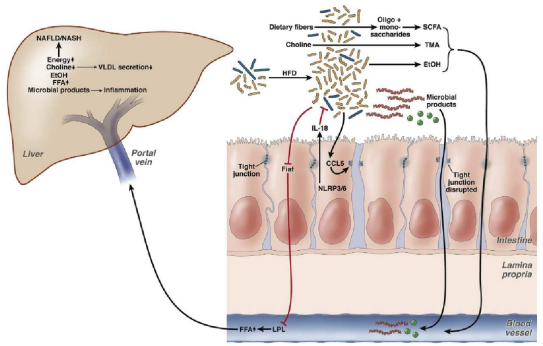
```
Usage:
  bowtie2 [options]* -x <bt2-idx> {-1 <m1> -2 <m2> | -U <r>} [-S <sam>]

  <bt2-idx>  Index filename prefix (minus trailing .X.bt2).
             NOTE: Bowtie 1 and Bowtie 2 indexes are not compatible.
  <m1>       Files with #1 mates, paired with files in <m2>.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <m2>       Files with #2 mates, paired with files in <m1>.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <r>        Files with unpaired reads.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <sam>      File for SAM output (default: stdout)

  <m1>, <m2>, <r> can be comma-separated lists (no whitespace) and can be
  specified many times.  E.g. '-U file1.fq,file2.fq -U file3.fq'.

Options (defaults in parentheses):

 Input:
  -q                 query input files are FASTQ .fq/.fastq (default)
  --qseq             query input files are in Illumina's qseq format
  -f                 query input files are (multi-)FASTA .fa/.mfa
  -r                 query input files are raw one-sequence-per-line
  -c                 <m1>, <m2>, <r> are sequences themselves, not files
  -s/--skip <int>    skip the first <int> reads/pairs in the input (none)
  -u/--upto <int>    stop after first <int> reads/pairs (no limit)
  -5/--trim5 <int>   trim <int> bases from 5'/left end of reads (0)
  -3/--trim3 <int>   trim <int> bases from 3'/right end of reads (0)
  --phred33          qualities are Phred+33 (default)
  --phred64          qualities are Phred+64
  --int-quals        qualities encoded as space-delimited integers

 Presets:                 Same as:
  For --end-to-end:
   --very-fast           -D 5 -R 1 -N 0 -L 22 -i S,0,2.50
   --fast                -D 10 -R 2 -N 0 -L 22 -i S,0,2.50
   --sensitive           -D 15 -R 2 -N 0 -L 22 -i S,1,1.15 (default)
   --very-sensitive      -D 20 -R 3 -N 0 -L 20 -i S,1,0.50

  For --local:
   --very-fast-local     -D 5 -R 1 -N 0 -L 25 -i S,1,2.00
   --fast-local          -D 10 -R 2 -N 0 -L 22 -i S,1,1.75
   --sensitive-local     -D 15 -R 2 -N 0 -L 20 -i S,1,0.75 (default)
   --very-sensitive-local -D 20 -R 3 -N 0 -L 20 -i S,1,0.50

 Alignment:
  -N <int>           max # mismatches in seed alignment; can be 0 or 1 (0)
  -L <int>           length of seed substrings; must be >3, <32 (22)
  -i <func>          interval between seed substrings w/r/t read len (S,1,1.15)
  --n-ceil <func>    func for max # non-A/C/G/Ts permitted in aln (L,0,0.15)
  --dpad <int>       include <int> extra ref chars on sides of DP table (15)
  --gbar <int>       disallow gaps within <int> nucs of read extremes (4)
  --ignore-quals     treat all quality values as 30 on Phred scale (off)
  --nofw             do not align forward (original) version of read (off)
  --norc             do not align reverse-complement version of read (off)
  --no-1mm-upfront   do not allow 1 mismatch alignments before attempting to
```

```
Scoring:
  --ma <int>         match bonus (0 for --end-to-end, 2 for --local)
  --mp <int>         max penalty for mismatch; lower qual = lower penalty (6)
  --np <int>         penalty for non-A/C/G/Ts in read/ref (1)
  --rdg <int>,<int>  read gap open, extend penalties (5,3)
  --rfg <int>,<int>  reference gap open, extend penalties (5,3)
  --score-min <func> min acceptable alignment score w/r/t read length
                     (G,20,8 for local, L,-0.6,-0.6 for end-to-end)

Reporting:
  (default)          look for multiple alignments, report best, with MAPQ
   OR
  -k <int>           report up to <int> alns per read; MAPQ not meaningful
   OR
  -a/--all           report all alignments; very slow, MAPQ not meaningful

Effort:
  -D <int>           give up extending after <int> failed extends in a row (15)
  -R <int>           for reads w/ repetitive seeds, try <int> sets of seeds (2)

Paired-end:
  -I/--minins <int>  minimum fragment length (0)
  -X/--maxins <int>  maximum fragment length (500)
  --fr/--rf/--ff     -1, -2 mates align fw/rev, rev/fw, fw/fw (--fr)
  --no-mixed         suppress unpaired alignments for paired reads
  --no-discordant    suppress discordant alignments for paired reads
  --no-dovetail      not concordant when mates extend past each other
  --no-contain       not concordant when one mate alignment contains other
  --no-overlap       not concordant when mates overlap at all

Output:
  -t/--time          print wall-clock time taken by search phases
  --un <path>           write unpaired reads that didn't align to <path>
  --al <path>           write unpaired reads that aligned at least once to <path>
  --un-conc <path>      write pairs that didn't align concordantly to <path>
  --al-conc <path>      write pairs that aligned concordantly at least once to <path>
  (Note: for --un, --al, --un-conc, or --al-conc, add '-gz' to the option name, e.g.
  --un-gz <path>, to gzip compress output, or add '-bz2' to bzip2 compress output.)
  --quiet            print nothing to stderr except serious errors
  --met-file <path>  send metrics to file at <path> (off)
  --met-stderr       send metrics to stderr (off)
  --met <int>        report internal counters & metrics every <int> secs (1)
  --no-unal          supppress SAM records for unaligned reads
  --no-head          supppress header lines, i.e. lines starting with @
  --no-sq            supppress @SQ header lines
  --rg-id <text>     set read group id, reflected in @RG line and RG:Z: opt field
  --rg <text>        add <text> ("lab:value") to @RG line of SAM header.
                     Note: @RG line only printed when --rg-id is set.
  --omit-sec-seq     put '*' in SEQ and QUAL fields for secondary alignments.

Performance:
  -p/--threads <int> number of alignment threads to launch (1)
  --reorder          force SAM output order to match order of input reads
  --mm               use memory-mapped I/O for index; many 'bowtie's can share

Other:
  --qc-filter        filter out reads that are bad according to QSEQ filter
  --seed <int>       seed for random number generator (0)
  --non-deterministic seed rand. gen. arbitrarily instead of using read attributes
  --version          print version information and quit

  -h/--help          print this usage message
```

Description des données

Dans le cadre d'une coopération CHU/INRA on veut analyser la composition taxonomique et fonctionnelle du microbiote intestinal (selles) de 96 patients atteints de NASH (Non Alcoholic Steatotic Hepatitis).

Suite au séquençage de l'ADN (HiSeq3000, 2×150 pb) on dispose en moyenne de 15 millions de *reads* pairés par échantillon, soit presque 1 To de données.



Schnabl, Gastroenterology 2014

Position du problème

Après assemblage des reads en contigs, on veut regrouper les contigs en bins pour reconstruire les génomes bactériens.

Quel niveau de profondeur et quel taux de couverture faut-il utiliser ?

Questions annexes :

- quel(s) logiciel(s) utiliser ? sur quelle machine ?

  Sanger (16s/10c/2T/70T) ; Gargantua (4s/16c/1.5T/5T+150T).
- si on décide d'exécuter bowtie2,

  faut-il choisir l'option *fast* ou *sensitive* ?
- combien de temps cela va-t-il durer ?

Dialogue de sourds (2)

[M] *Je n'y comprends rien, ce n'est pas des mathématiques.*

[B] *Ce sont des alignements donc des calculs,*

donc c'est des mathématiques !

[M] *Je me suis renseigné, c'est du FM-index donc de la transformée*

de Burrows-Wheeler donc ***sensitive*** *est la meilleure option.*

[B] *? ? ? ! ? ? ? Vous êtes-sûr ?*

[M] *Non.*

10. Quelques remarques pour conclure (1)

Souvent « *l'usage fait loi* » et « *l'expérience est reine* », ce qui ne facilite pas les choix.

Il faut être très modeste car les théories mathématiques sont nombreuses et parfois ardues à mettre en oeuvre.

Reconnaître son incapacité à résoudre un problème (plutôt que «fuir/louvoyer») permet d'avancer car cela oblige à aller chercher des compétences complémentaires.

Commencer par analyser les données présentes se révèle, au fil des années, très pertinent.

La route est souvent longue de la position du problème à une première modélisation et encore plus longue jusqu'à une solution exploitable.

Seule la **collaboration interdisciplinaire** permet d'y aboutir.

10. Quelques remarques pour conclure (2)

Pour démontrer un théorème de mathématiques, on a besoin

- *d'un papier*
- *d'un crayon*
- *un/des mathématicien(ne)s*

Pour réaliser une étude en recherche clinique, il faut

- *des patients*
- *des secrétaires*
- *des infirmières*
- *des TEC et autres ARC*
- *des médecins, des cliniciens, des experts*
- *un/des biostatisticien(ne)s, un/des mathématicien(ne)s*

Merci de votre attention !