Recherche statistique de biomarqueurs du cancer et de l'allergie à l'arachide

Olivier Collignon

NANCY UNIVERSITE

THESE

Docteur en Mathématiques

Ecole doctorale : Informatique, Automatique, Électronique, Électrotechnique et Mathématiques

Département de Formation Doctorale Mathématiques

Recherche statistique de biomarqueurs du cancer et de l'allergie à l'arachide

Institut Elie Cartan de Nancy et GENCLIS SAS

Olivier COLLIGNON

Soutenue publiquement le 16 octobre 2009 devant le jury composé de :

Président du jury et rapporteur
M. Gilles CELEUX, DR — INRIA Futurs

Rapporteur
M. Christophe BIERNACKI, Pr. — Université des Sciences et Technologies de Lille 1

Examinateurs
M. Bernard E. BIHAIN, DR — GENCLIS SAS
Mme Denise-Anne MONERET-VAUTRIN, Pr. — Centre Hospitalier Universitaire de Nancy
M. Jean-Christophe TURLOT, MCF — Université de Pau et des Pays de l'Adour

Directeurs
M. Jean-Marie MONNEZ, Pr., directeur — Nancy Université
M. Pierre VALLOIS, Pr., co-directeur — Nancy Université

Table des matières

Remerciements **5**

Introduction **7**

1 Introduction à la Biologie Moléculaire **9**
1.1 L'information génétique . . . 9
1.1.1 La cellule . . . 9
1.1.2 L'Acide DésoxyriboNucléique . . . 9
1.2 Des gènes aux protéines . . . 12
1.2.1 Les gènes . . . 12
1.2.2 La synthèse des protéines . . . 12
1.3 La réplication . . . 14
1.3.1 Les mécanismes de la réplication . . . 14
1.3.2 Vérification de la réplication . . . 15
1.4 Les protéines . . . 16
1.4.1 Les acides aminés . . . 16
1.4.2 Les polypeptides . . . 17
1.4.3 Structure des protéines . . . 18
1.4.4 Rôles des protéines . . . 18
1.4.5 Dégradation des protéines . . . 18
1.5 Le système immunitaire . . . 19

I Recherche de biomarqueurs du cancer **21**

2 Introduction à la cancérologie **23**
2.1 Les mutations génétiques . . . 23
2.1.1 Définition . . . 23
2.1.2 Transmission des mutations . . . 23
2.1.3 Les types de mutations . . . 23
2.2 Quelques notions de cancérologie . . . 24
2.2.1 Caractéristiques biologiques du cancer . . . 24
2.2.2 Facteurs de risque . . . 25
2.2.3 L'évolution de la maladie . . . 25
2.2.4 Conséquences du cancer . . . 26
2.3 La médecine du cancer . . . 26
2.3.1 Diagnostic . . . 26
2.3.2 Traitements . . . 26
2.3.3 Prévention . . . 26

3 Contrôle du risque de première espèce dans un ensemble de tests 27
3.1 Introduction . . . 27
3.2 Procédures de tests multiples : éléments de base . . . 28
3.2.1 Règle de décision . . . 28
3.2.2 Faux positifs et faux négatifs . . . 28
3.2.3 Critères de détermination du seuil t . . . 29
3.3 Estimation et contrôle du taux de fausses découvertes . . . 30
3.3.1 Estimation du FDR . . . 30
3.3.2 Contrôle du FDR . . . 31
3.4 Estimation de $\pi_0 = \frac{m_0}{m}$ . . . 31
3.4.1 Location Based Estimator pour une v.a absolument continue . . . 31
3.4.2 Location Based Estimator pour une v.a discrète . . . 32
3.4.3 Autres méthodes . . . 33

4 Comparaison de la probabilité de survenue d'une substitution sur un ARNm sain et sur un ARNm cancéreux 35
4.1 Introduction . . . 35
4.2 Les Expressed Sequences Tags (EST) . . . 35
4.2.1 Synthèse de l'ADNc et séquençage des EST . . . 36
4.2.2 Les erreurs de séquençage . . . 37
4.3 Approche bioinformatique . . . 38
4.4 Analyse statistique . . . 39
4.4.1 Formalisation du problème . . . 39
4.4.2 Impact de l'erreur de séquençage sur le modèle . . . 40
4.4.3 Tests de comparaison de deux probabilités . . . 41
4.4.4 Les p-values . . . 44
4.4.5 Test exact de Fisher . . . 45
4.5 Résultats . . . 49
4.5.1 Tests de comparaison de deux probabilités . . . 50
4.5.2 Etude comparative du test de comparaison de probabilités et du test exact de Fisher . . . 52
4.5.3 Résultats du test de comparaison de probabilités ou du test exact de Fisher 54
4.6 Conclusions . . . 56

5 Modélisation de la probabilité de survenue d'une infidélité de transcription sur un ARNm 59
5.1 Problème et données . . . 59
5.2 Hypothèses sur le mécanisme d'infidélité de transcription . . . 61
5.2.1 Hypothèses . . . 61
5.2.2 Impact de l'erreur de séquençage sur le modèle . . . 61
5.3 Tests des hypothèses . . . 64
5.3.1 Test de l'hypothèses (5.5) . . . 64
5.3.2 Test de l'hypothèses (5.6) . . . 66
5.4 Conclusions . . . 69

II Recherche de biomarqueurs de l'allergie à l'arachide 71

6 Introduction aux problématiques de l'allergie à l'arachide 73
6.1 Introduction . . . 73
6.2 Description biologique des variables . . . 74

6.2.1 Les dosages immunologiques . . . 75
6.2.2 Mise en évidence des IgE non circulantes par les prick tests . . . 77
6.2.3 Le test de provocation orale (TPO) . . . 78

7 Méthodes d'analyse discriminante 81
7.1 Formulation . . . 81
7.2 Sélection des variables discriminantes . . . 81
7.2.1 Tests d'homogénéité . . . 82
7.2.2 Sélection pas-à-pas . . . 83
7.3 Méthodes de classement . . . 85
7.3.1 Règle de classement linéaire *(Linear Discriminant Analysis (LDA))* . . . 85
7.3.2 Règle de classement quadratique *(Quadratic Discriminant Analysis (QDA))* 85
7.3.3 Les k plus proches voisins *(k-Nearest Neighbours (k-NN))* . . . 86
7.3.4 La régression logistique . . . 86
7.3.5 Segmentation par arbres de décision . . . 86
7.3.6 Les Support Vector Machine *(SVM)* . . . 88
7.3.7 Les courbes ROC . . . 90
7.4 Mesures de la qualité d'une règle de classement . . . 92
7.4.1 Méthode de l'échantillon-test . . . 92
7.4.2 Validation croisée *(cross-validation)* . . . 93

8 Simplification du diagnostic de l'allergie à l'arachide 95
8.1 Etude descriptive des individus allergiques et des individus atopiques . . . 95
8.1.1 Données cliniques et mesures de la sévérité . . . 95
8.1.2 Analyse en Composantes Principales . . . 96
8.2 Discrimination allergie / atopie à partir des dosages immunologiques . . . 99
8.2.1 Test de comparaison de moyennes . . . 99
8.2.2 Discrimination à l'aide du seuil de détection de la méthode . . . 100
8.2.3 Détermination d'un seuil optimal pour chaque variable . . . 101
8.2.4 Discrimination par l'ensemble des prédicteurs . . . 103
8.3 Conclusion et perspectives . . . 104

9 Prédiction de la sévérité de l'allergie à l'arachide : résumé 107
9.1 Introduction . . . 107
9.2 Approche statistique . . . 108
9.2.1 Les scores du TPO et du premier accident . . . 108
9.2.2 La dose réactogène . . . 109
9.3 Résultats . . . 111
9.4 Conclusions et perspectives . . . 111

10 Discriminant analyses of peanut allergy severity scores 115
10.1 Introduction . . . 115
10.2 Experimental Procedure and Data . . . 116
10.2.1 Immunoassays . . . 117
10.2.2 Skin Prick Tests (SPTs) . . . 117
10.3 Statistical approach . . . 117
10.3.1 Design of the study . . . 117
10.3.2 Multiple Factorial Analysis (MFA) . . . 118
10.3.3 Variable selection . . . 120
10.3.4 An algorithm for simultaneously clustering the response variable and selecting discriminant variables . . . 120

10.3.5 Discriminant analysis . . . 121
10.4 Results . . . 122
10.4.1 Principal Component Analysis . . . 122
10.4.2 First accidental exposure score . . . 122
10.4.3 DBPCFC score . . . 124
10.4.4 Eliciting dose . . . 125
10.5 Conclusions . . . 127

11 Un algorithme de classification et de sélection simultanée de variables discriminantes 133
11.1 Introduction . . . 133
11.2 Descriptif de l'algorithme . . . 133
11.2.1 Principe de l'algorithme ascendant . . . 133
11.2.2 Construction des classes . . . 134
11.2.3 Déroulement de la procédure . . . 134
11.3 Validation de l'algorithme sur divers jeux de données . . . 135
11.3.1 Modèle probabiliste . . . 135
11.3.2 Données simulées . . . 135
11.3.3 Les iris de Fisher . . . 139
11.3.4 Les poissons du lac Laengelmavesi . . . 140
11.4 Conclusion et perspectives . . . 142

Conclusion 143

Remerciements

Je tiens tout d'abord à remercier vivement MM. les professeurs Jean-Marie Monnez et Pierre Vallois pour m'avoir encadré tout au long de cette thèse. Ils m'ont grandement aidé à mener à bien les travaux présentés ici en me transmettant leurs connaissances et en prolongeant toujours les questionnements statistiques sous-jacents aux problèmes biologiques et médicaux. J'ai également acquis grâce à eux la rigueur permettant de structurer mes travaux et de concevoir une étude statistique la plus exhaustive possible. La participation au groupe de travail de biostatistiques de l'IECN, dont je salue ici les membres, m'a également été très profitable pour me familiariser avec les problématiques de la statistique et pour élargir ma culture mathématique. En particulier, je remercie Sandie Ferrigno et Pierre Debs d'avoir vérifié que les problématiques biologiques et médicales étaient clairement expliquées et rendues accessibles à un lecteur néophyte.

Je veux ensuite remercier le président de Genclis M. Bernard Bihain pour avoir initié la collaboration avec l'IECN. Il m'a fait confiance en me permettant d'intégrer son laboratoire de recherche et de rejoidre son équipe dynamique de biologistes. Il m'a également appris à persévérer dans mes recherches et à ne pas baisser les bras à chaque obstacle. De plus, je remercie chaleureusement l'ensemble du personnel de Genclis, et en particulier Sandrine Jacquenet pour sa disponibilité et sa gentillesse. En plus de m'avoir expliqué les fondamentaux de l'allergologie, elle a été une oreille attentive à mes questionnements tout au long de cette thèse. Virginie Ogier et Benoit Thouvenot m'ont transmis les rudiments de Biologie nécessaires à l'accomplissement de mes travaux. Je remercie également ce dernier pour la relecture attentive des parties de biologie du manuscrit. Je remercie également Frances Yen-Potin pour les corrections d'anglais. Je remercie enfin tout particulièrement Marie Brulliard et l'équipe de bioinformatique de Genclis pour leurs discussions motivantes, leur aide en programmation, et surtout pour avoir été mes compagnons durant cette thèse.

Merci à l'ensemble des médecins du service d'allergologie de Nancy, pour leur confiance et leur vif intérêt pour les résultats issus du traitement statistique de leur données. Cette amicale collaboration m'a permis d'appréhender les problématiques actuelles de l'allergie et de mieux cerner les problèmes des médecins. Je remercie en particulier Mme le professeur Gisèle Kanny pour ses conseils avisés et la relecture de ce manuscrit.

Je souhaite remercier M. le professeur Christophe Biernacki et M. Gilles Celeux, directeur de recherches, d'avoir accepté d'être les rapporteurs de ma thèse. Je remercie également M. Jean-Christophe Turlot, maître de conférences, et Mme le professeur Denise-Anne Moneret-Vautrin d'avoir accepté d'être mes examinateurs.

Je remercie tous ceux qui de près ou de loin m'ont permis de mener à bien cette thèse de doctorat et qui se sont intéressés à mes travaux.

Un grand merci enfin à ma famille, mon père, ma mère, mon frère et ma soeur pour leur soutien inconditionnel. Je souhaite enfin dédier cette thèse à mes deux grands-pères.

Introduction

Cette thèse de doctorat de mathématiques appliquées est une thèse CIFRE (conventions industrielles de formation par la recherche) et a été encadrée par les professeurs Jean-Marie Monnez et Pierre Vallois de l'Institut Elie Cartan de Nancy. L'entreprise d'accueil, présidée par Monsieur Bernard E. Bihain, directeur de recherches, est le laboratoire GENCLIS, jeune entreprise innovante nancéienne créée au premier janvier 2004. GENCLIS est spécialisée dans la recherche et le développement pré-clinique de nouvelles molécules à visée thérapeutique et diagnostique de deux maladies immunitaires : le cancer et l'allergie. Ainsi l'objectif de cette thèse est de détecter des protéines, qui peuvent être des anticorps, qui permettraient de discriminer un groupe "malade", dont les individus sont des patients atteints du cancer ou d'une allergie à l'arachide, d'un groupe "témoin". Décrivons le plan du manuscrit.

Dans le premier chapitre sont présentés les mécanismes biologiques fondamentaux menant du gène à la protéine.

La première partie de la thèse traite de la recherche de biomarqueurs du cancer.

Après une introduction à la cancérologie (Chapitre 2), l'analyse statistique des EST (Expressed Sequence Tag) permet de mettre en évidence dans le Chapitre 4 un mécanisme jusqu'alors insoupçonné en biologie : l'infidélité de la transcription de l'ADN en ARNm. Lors de la transcription, on observe en effet que des nucléotides d'un ARNm peuvent être remplacés par un autre nucléotide. On s'intéresse alors à la comparaison des probabilités de survenue des infidélités de transcription dans des ARNm cancéreux et dans des ARNm sains. Afin de mener une étude à large échelle, les ARNm peuvent être étudiés via les EST, qui sont des fragments d'ADNc séquencés après transcription inverse des ARNm présents dans un tissu. Les EST sont de plus entachées d'erreurs de séquençage. Un modèle est proposé permettant de prendre en compte ces erreurs de séquençage dans la méthode de comparaison des probabilités de survenue d'infidélité de transcription dans les deux types de tissus. La méthode statistique utilisée repose alors sur une procédure de tests multiples (dont les principes sont rappelés Chapitre 3) menée sur les positions de la séquence de référence de 17 gènes d'intérêt, avec laquelle les positions d'ARNm plus fréquemment sujettes à des substitutions dans le cas cancéreux que dans le cas sain sont mises en évidence. Ces erreurs conduiraient ainsi à la production de protéines dites aberrantes, dont la séquence d'acides aminés ne correspond pas à celle définie par l'ADN. Mesurer la quantité de ces protéines dans le sang permettrait par la suite de détecter les patients atteints de formes précoces de cancer. Cette partie a donné lieu à la publication d'un article [1]. Le Chapitre 5 présente ensuite un essai de modélisation de la probabilité de survenue d'une infidélité de transcription de l'ADN en ARNm.

La deuxième partie de la thèse s'attache à l'étude de l'allergie à l'arachide, dont les problématiques sont introduites dans la Chapitre 6.

Afin de diagnostiquer l'allergie à l'arachide, un TPO (Test de Provocation Orale) est réalisé en clinique. Le protocole consiste à faire ingérer des doses croissantes d'arachide au patient, souvent un enfant, jusqu'à l'apparition de symptômes objectifs dont la gravité est inconnue. Ainsi, le TPO peut se révéler dangereux pour le patient. Détecter l'allergie à l'arachide par une

méthode simple et sans danger pour le patient apparaît de ce fait comme un véritable enjeu de santé publique. Après des rappels sur l'analyse discriminante dans le Chapitre 7, une étude clinique menée sur 243 patients recrutés dans deux centres différents est présentée dans le Chapitre 8. On montre en particulier que la mesure des anticorps dirigés contre la protéine *rAra-h2* permet de très bien distinguer les patients allergiques à l'arachide de notre échantillon. Mais la sévérité de l'allergie à l'arachide varie extrêmement d'un individu à l'autre : certains n'ont qu'un léger urticaire, alors que d'autres doivent être accueillis en soins intensifs. Connaître la sévérité de l'allergie d'un patient est donc nécessaire pour l'informer des risques qu'il encourt. La sévérité de l'allergie à l'arachide peut être mesurée par trois indicateurs différents : le score du TPO, le score du premier accident et la dose réactogène. Le Chapitre 9 résume ainsi un article soumis [2] proposant des analyses prédictives de ces mesures de sévérité, qui constitue le Chapitre 10. Dans cette étude, la présence de 36 anticorps est observée sur 93 patients allergiques, dont 6 par dosages immunologiques et 30 par tests cutanés. L'analyse repose sur l'utilisation comme prédicteurs de facteurs issus d'une Analyse Factorielle Multiple, permettant ainsi de donner le même poids aux deux groupes de variables explicatives (dosages immunologiques et tests cutanés) et sur la mise en compétition de plusieurs règles de classement. De plus, la dose réactogène, qui est la dose d'arachide induisant les premiers symptômes, est une variable dont les valeurs sont fixées par paliers définis dans un protocole. Donc pour chaque patient seul un intervalle contenant la dose réactogène est connu, et il n'est pas possible de mener une régression de la dose réactogène pour tenter de prédire sa valeur. Un algorithme de classification des intervalles de doses réactogènes et de sélection simultanée des variables explicatives est proposé, permettant ensuite la mise en compétition de règles de classement. Enfin, le dernier chapitre présente quelques travaux numériques réalisés afin de valider cet algorithme.

Chapitre 1

Introduction à la Biologie Moléculaire

Ce chapitre a pour but de fournir au lecteur le bagage de biologie moléculaire nécessaire à la compréhension des études menées dans la suite, et d'en présenter les différents acteurs biologiques. On citera trois ouvrages classiques en guise de références globales : [3], [4] et [5].

1.1 L'information génétique

1.1.1 La cellule

La **cellule** (en latin *cellula*, petite chambre) est l'unité structurale constituant toute partie d'un être vivant. Chaque cellule est un être vivant fonctionnant de manière autonome, tout en interagissant avec les autres.

Il existe deux principales structures cellulaires :

- les **procaryotes**, ne possédant pas de **noyau**, qui sont les bactéries,
- les **eucaryotes**, qui possèdent un noyau individualisé. C'est le cas des animaux, des champignons et des plantes.

Dans le corps humain, les cellules sont eucaryotes, et leur nombre est estimé à 50 000 milliards. Les cellules humaines sont subdivisées en 220 types différents, en fonction de leur rôle. Les cellules de même type sont regroupées en tissus, eux-même regroupés en organes.

La cellule contient une multitude de structures spécialisées appelées **organites** ou **organelles**. Parmi elles se trouve le noyau, qui contient la plupart du matériel génétique nécessaire à la cellule. Son diamètre varie de 10 à 20 micromètres.

Ses deux fonctions principales sont :

- le contrôle des réactions chimiques du **cytoplasme**, qui est le "contenu" de la cellule,
- le stockage des informations nécessaires à la division cellulaire.

1.1.2 L'Acide DésoxyriboNucléique

L'Acide DésoxyriboNucléique (ADN), support de l'information génétique, est une macro-molécule présente dans les cellules de tous les êtres vivants. Il contient toute l'information nécessaire au bon fonctionnement de la cellule. Chez les eucaryotes, l'ADN est principalement contenu dans le noyau des cellules.

L'ADN assure trois fonctions :

- stocker l'information génétique,
- transmettre cette information de génération en génération,
- permettre l'évolution biologique de l'espèce.

L'ADN est formé de deux brins enroulés en hélice. Chaque brin est constitué de l'enchaînement de **nucléotides** (Fig. 1.1 a), molécules résultant de l'addition :

- de **désoxyribose**, qui est un pentose, c'est-à-dire un sucre à 5 carbones, numérotés de 1 à 5,
- d'un **groupement phosphate**, fixé sur le carbone 5,
- **d'une base azotée**, qui est une molécule organique formée d'un ou deux cycles où alternent des atomes de carbone et d'azote.

Il existe quatre bases différentes (Fig. 1.1 b) :

- **Adénine (A)**,
- **Thymine (T)**,
- **Guanine (G)**,
- **Cytosine (C)**.

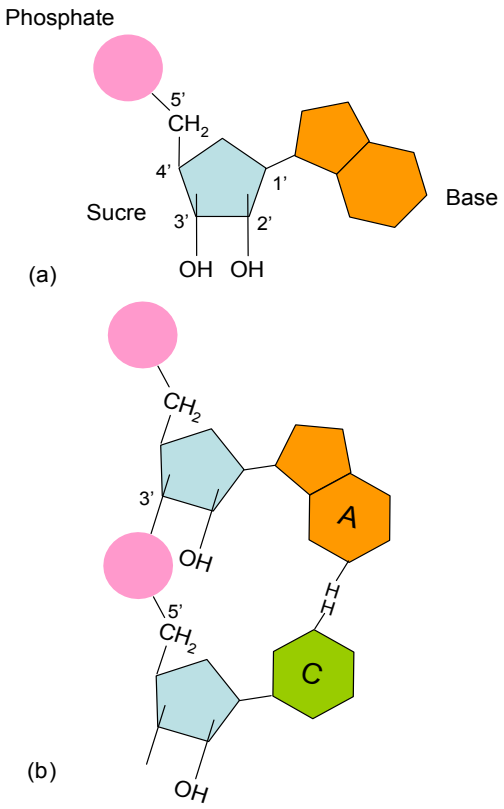


FIG. 1.1 – *Les nucléotides et les bases de l'ADN*

L'adénine et la guanine sont appelées des **purines**, tandis que la cytosine et la thymine sont des **pyrimidines**.

Un brin d'ADN est constitué lorsque des nucléotides se lient les uns aux autres par réaction entre le groupement hydroxyle OH en position 3' du désoxyribose du premier nucléotide, et le groupement phosphate en 5' du nucléotide suivant (Fig. 1.1 b).

Deux brins d'ADN peuvent ensuite s'unir par les bases azotées grâce à des liaisons hydrogène : A d'un brin " s'associe " toujours avec T sur l'autre brin, et C toujours avec G. Les couples AT et CG sont appelés paires de bases. Les deux brins d'ADN sont dits **complémentaires** (Fig.1.2).

Chacune des chaînes qui constitue la double hélice présente deux extrémités différenciées (Fig.1.2) :

- l'extrémité 5' : où le groupe phosphate du premier nucléotide est libre *i.e.* non engagé dans une liaison avec un autre nucléotide,
- l'extrémité 3' : où le groupe hydroxyle -OH du dernier nucléotide est libre.

Quand deux chaînes s'apparient pour former une double hélice, l'une s'oriente dans le sens 5'→3', tandis que l'autre s'oriente 3'→5'. Ce phénomène est appelé appariement antiparallèle.

Ce sont les angles particuliers entre les liaisons reliant les nucléotides qui induisent la forme de double hélice de l'ADN (Watson et Crick en 1953) (Fig.1.3).

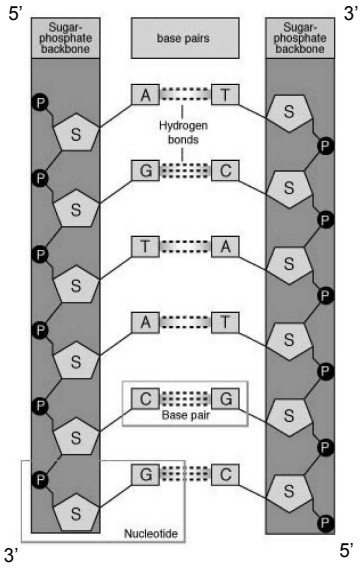


FIG. 1.2 – *Appariement des deux brins d'ADN.* (source : Internet)

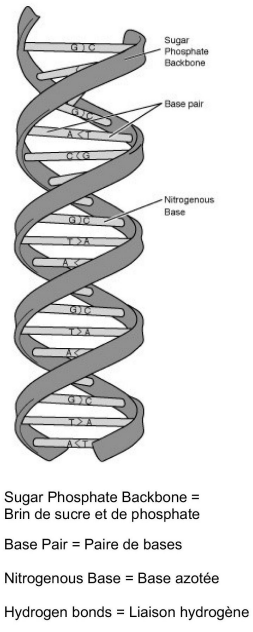


FIG. 1.3 – *Structure en double hélice de l'ADN.* (source : Internet)

Notons enfin que l'ADN est quasiment le même pour tous les individus ; seuls 5% des nucléotides diffèrent d'un individu à l'autre. Ces nucléotides sont appelés des **SNP** (Single Nucleotide Polymorphism) et se trouvent à des positions connues.

1.2 Des gènes aux protéines

1.2.1 Les gènes

Un **gène** est un enchaînement de nucléotides, c'est-à-dire une portion d'ADN, qui commande la synthèse d'une **protéine**. Une protéine est composée d'une chaîne de molécules simples appelées les **acides aminés**.

A titre d'exemple, le génome humain, c'est-à-dire l'ensemble des gènes de l'homme, compte environ 30 000 gènes, tandis que celui d'une drosophile en compte 18 000 gènes.

Toutes les cellules portent tous les gènes, mais leur expression dépend de la nature de la cellule et du stade de développement de l'organisme. L'ensemble des gènes exprimés dans une cellule, et donc des protéines qui seront présentes dans cette cellule, dépendent de facteurs complexes mis en place au cours du développement de l'individu. Certains caractères sont déterminés par un seul gène (comme le groupe sanguin chez l'homme ou comme la couleur des yeux chez la drosophile). Cependant, dans la plupart des cas, un caractère observable dépend de nombreux gènes et éventuellement de l'interaction avec l'environnement.

Sur l'ADN, chaque gène est délimité par des balises qui sont des séquences particulières de plusieurs nucléotides. La séquence qui initie la transcription du gène est appelée **promoteur** ; celle qui met fin à sa transcription est appelée **terminateur** (voir section suivante). Etre deux gènes se trouve de l'ADN ne permettant pas la synthèse de protéine. Dans le génome humain, les gènes codant des protéines n'occupent que 2 à 5% des 3 millions de paires de bases. La fonction des parties non codantes est encore mal connue.

De plus, chez les eucaryotes, un gène est constitué d'une alternance de séquences codantes, appelées **exons**, et de séquences non codantes, les **introns**.

1.2.2 La synthèse des protéines

Trois mécanismes mènent du gène à la synthèse des protéines et peuvent être schématisés de la manière suivante :

$$ADN_{(gènes)} \overset{transcription}{\longrightarrow} ARNpm \overset{épissage}{\longrightarrow} ARNm \overset{traduction}{\longrightarrow} protéine \tag{1.1}$$

1. **La transcription :** Une enzyme appelée **ARN polymérase** produit à partir d'un brin d'ADN un **ARN (Acide RiboNucléique)**, dit **pré-messager**, qui contient à la fois les exons et les introns du gène.

   Quatre caractéristiques fondamentales de l'ARN le distingue de l'ADN :
   - le sucre désoxyribose est remplacé par un ribose (présence d'un groupement -OH sur le carbone n°2) ;
   - la Thymine est remplacée par un **Uracile (U)** (ces deux bases contiennent une information équivalente) ;
   - l'ARN est simple brin, tandis que l'ADN est double brin avec une structure en double hélice ;
   - l'ARN est court (en général de 50 à 5000 nucléotides et non pas des millions comme dans l'ADN).

De plus, la base A de l'ADN est transcrite en U sur le brin d'ARN, T en A, C en G et G en C. L'ARN est ainsi une copie du brin 5'→3' de l'ADN synthétisé par complémentarité à partir du brin 3'→5' ; il est donc orienté dans le sens 5'→3'. La transcription obéit aux mêmes règles d'appariement des bases définies par Watson et Crick.

Remarque : La transcription se déroule à l'intérieur même du noyau d'une cellule eucaryote.

2. **L'épissage :** Les introns, séquences non codantes des gènes, sont éliminés, tandis que les exons, séquences codantes, sont mis bout à bout pour former le brin d'ARN **messager** (ARNm) (Fig.1.5).

   Remarque : Parmi l'ensemble des exons disponibles, certains peuvent ne pas être conservés après épissage, ce choix pouvant différer d'un type de cellule à l'autre. Un gène unique permet donc de synthétiser plusieurs protéines selon les exons conservés dans l'ARNm.

3. **La traduction :** Le brin d'ARNm est traduit en une chaîne d'acides aminés par une grosse molécule appelée **ribosome**. Les acides aminés, au nombre de 20, sont tous caractérisés sur l'ARNm par un triplet de nucléotides appelé **codon**. La séquence d'ARNm à traduire commence toujours par le codon "START" (AUG, la méthionine), et se termine par un codon "STOP" (il en existe trois : TAG, TAA et TGA). Entre ces deux bornes, le brin d'ARNm est ainsi traduit codon par codon (*i.e.* par pas de trois nucléotides). La séquence des acides aminés concaténés forme la protéine synthétisée (Fig.1.6).

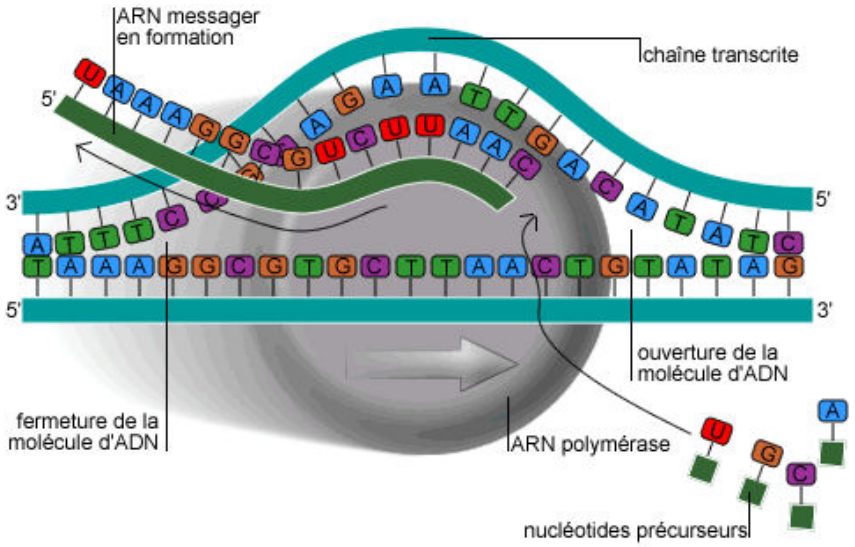


FIG. 1.4 – *Transcription de l'ADN en ARNm* (source : Wikipedia)

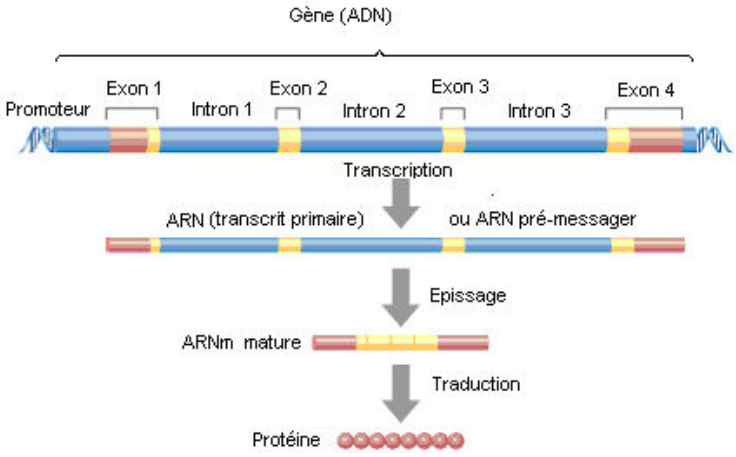


FIG. 1.5 – *Epissage de l'ARN pré-messager* (source : Internet)

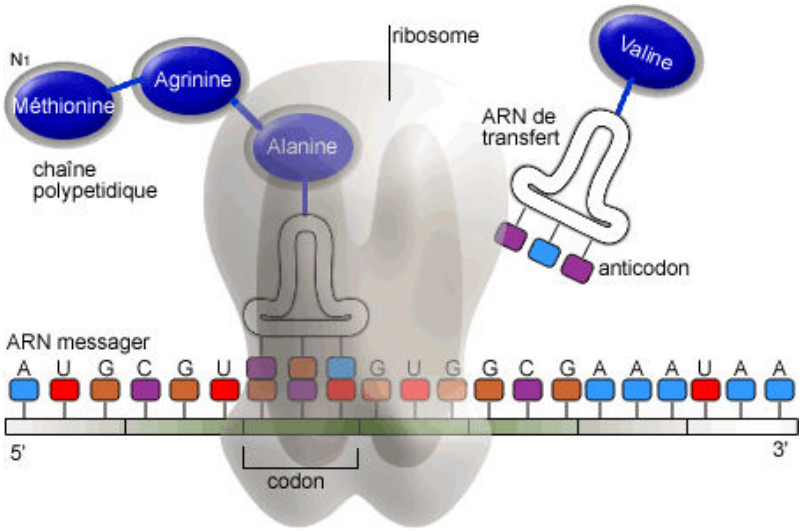


FIG. 1.6 – *Traduction de l'ARNm en protéine* (source : Wikipedia)

1.3 La réplication

1.3.1 Les mécanismes de la réplication

La synthèse des protéines étant régie par les séquences de nucléotides de l'ADN, il est fondamental pour l'organisme d'obtenir des copies précises de ces nucléotides pour les transmettre à d'autres cellules.

Pour que l'ADN double brin, dit **parental**, puisse être répliqué lors de la division cellulaire, il faut d'abord que les deux brins soient déroulés par des enzymes spécialisées. Ce déroulement commence à partir de positions repérées sur l'ADN par des séquences particulières, souvent riches en A et en T, appelées **origines de réplication**. L'endroit à partir duquel les deux brins sont écartés, appelé **fourche de réplication** ou **d'élongation**, s'éloigne de l'origine à mesure que la réplication progresse.

Les deux brins sont copiés de deux manières différentes. En effet, les deux brins parentaux d'ADN sont antiparallèles et l'enzyme responsable de la synthèse des nouveaux brins, appelée **ADN polymérase**, ne peut ajouter des nucléotides que dans le sens $5' \rightarrow 3'$. De ce fait, le premier brin, orienté $3' \rightarrow 5'$, est synthétisé de manière continue par complémentarité à partir

d'une **amorce**, c'est-à-dire une courte séquence d'ARN préexistante. Le nouveau brin est appelé **brin précoce**.

L'autre brin, orienté $3' \rightarrow 5'$ et appelé **brin tardif**, est quant à lui synthétisé de manière discontinue à partir de plusieurs amorces d'ARN, créées à mesure de l'avancement de la fourche de réplication. L'élongation d'une amorce est appelée **fragment d'Okazaki**. Deux fragments sont réunis lorsque le nouveau fragment approche de l'amorce précédente. La répétition de ce processus conduit ainsi à la synthèse du brin tardif (Fig. 1.7).

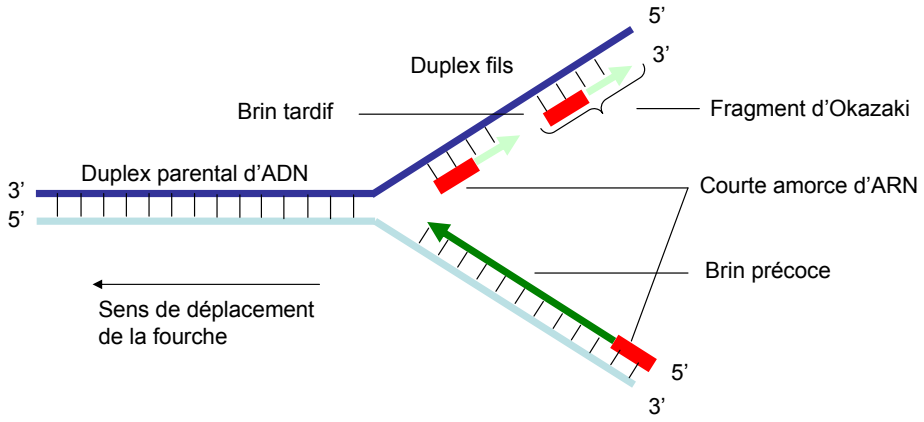


FIG. 1.7 – *La réplication de l'ADN*

1.3.2 Vérification de la réplication

Il est fondamental pour l'organisme que l'information génétique reste inchangée quand elle est transmise d'une cellule à l'autre ou d'un individu à l'autre. En effet, des lésions non réparées de l'ADN peuvent avoir des conséquences importantes. Ainsi, si ces mutations surviennent dans des cellules évoluant en gamètes, *i.e* des cellules sexuelles, elles pourront être transmises à la génération suivante. Des mutations apparaissant dans des cellules non reproductrices peuvent avoir des conséquences sur la transcription de l'ADN en ARNm, sur la traduction en protéine ou sur la réplication. De ce fait, il est essentiel pour les cellules de l'organisme de posséder un système de vérification des erreurs performant. Il est ainsi estimé qu'en moyenne moins d'une erreur sur mille échappe aux systèmes de réparations de la cellule. Les réparations se font à l'aide d'enzymes spécialisées qui parcourent l'ADN à la recherche de lésions sur un des brins. Par complémentarité, le brin non muté peut alors servir de modèle pour la réparation.

Il existe deux principaux systèmes de réparation :

Réparation par excision de nucléotides

Des enzymes coupent le brin d'ADN muté de part et d'autre de la lésion. Le fragment obtenu est alors retiré, puis la lacune est comblée par complémentarité grâce à une ADN polymérase (Fig. 1.8).

Il est intéressant de noter que la réparation par excision de nucléotides intervient en priorité sur des lésions affectant des gènes activement transcrits. En effet, la réparation de l'ADN se déroulant pendant la transcription, ceci assure que les mutations touchant des gènes fondamentaux soient réparées en premier.

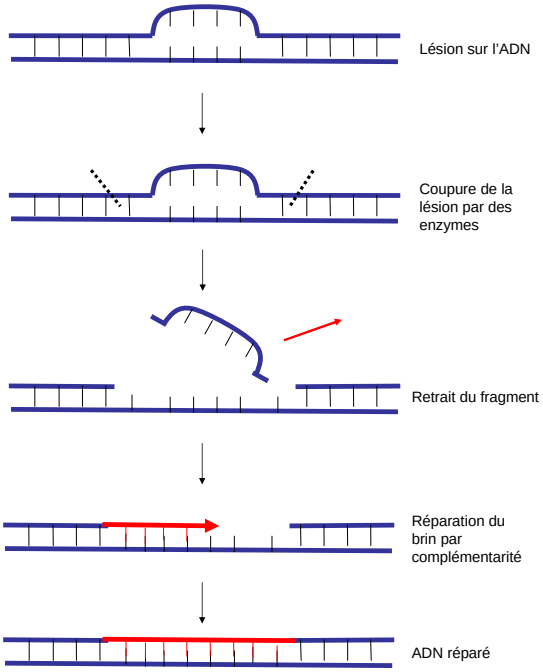


FIG. 1.8 – *Réparation de l'ADN par excision de nucléotides. Les étapes sont décrites dans le texte*

Réparation par excision de bases

Lorsque la mutation observée entraîne des modifications conformationnelles moins importantes, un autre système de réparation est utilisé. Une enzyme repère la base affectée puis l'élimine par clivage de la liaison entre cette base et le sucre correspondant. Une ADN polymérase comble alors le vide avec la base adéquate.

1.4 Les protéines

1.4.1 Les acides aminés

Comme expliqué précédemment, toutes les protéines sont construites par concaténation des **acides aminés**. Les acides aminés ont tous la même structure : ils sont composés d'une portion variable appelée **radical**, d'un **groupement acide** et d'un **groupement amine**.

La nomenclature des vingt acides aminés est donnée dans la Table 1.1.

Acide Aminé	Abré. 3 lettres	Abrév. 1 lettre	Codon(s)
Alanine	Ala	A	GCA, GCC, GCG, GCT
Arginine	Arg	R	$CGA, CGC, CGG, CGT, AGA, AGG$
Acide aspartique	Asp	D	GAC, GAT
Asparagine	Asn	N	AAC, AAT
Cystéine	Cys	C	TGC, TGT
Acide glutamique	Glu	E	GAA, GAG
Glutamine	Gln	Q	CAA, CAG
Glycine	Gly	G	GGA, GGC, GGG, GGT
Histidine	His	H	CAC, CAT
Isoleucine	Ile	I	ATA, ATC, ATT
Leucine	Leu	L	$CTA, CTC, CTG, CTT, TTA, TTG$
Lysine	Lys	K	AAA, AAG
Méthionine	Met	M	ATG
Phénylalanine	Phe	F	TTC, TTT
Proline	Pro	P	CCA, CCC, CCG, CCT
Sérine	Ser	S	$TCA, TCC, TCG, TCT, AGC, AGT$
Thréonine	Thr	T	ACT, ACC, ACG, ACT
Tryptophane	Trp	W	TGG
Tyrosine	Tyr	Y	TAC, TAT
Valine	Val	V	GTA, GTC, GTG, GTT
STOP	-	-	TAG, TAA, TGA

Tab. 1.1 – Nomenclature des acides aminés

1.4.2 Les polypeptides

Les acides aminés peuvent se lier les uns aux autres par une **liaison peptidique**.

L'union de plusieurs acides aminés forme des **polypeptides**, divisés en deux catégories :
- les **peptides**, qui sont constitués de polypeptides de moins de cinquante acides aminés,
- les **protéines**, qui sont des polypeptides de plus de cinquante acides aminés.

Les radicaux des acides aminés possèdent différentes propriétés chimiques. Il s'en suit alors des répulsions, des rapprochements et des formations de liens chimiques entre certains radicaux d'une chaîne d'acides aminés. De plus, la chaîne d'acides aminés a tendance à se replier sur elle-même pour adopter une certaine structure tridimensionnelle. A l'intérieur d'un polypeptide, quatre types d'interactions peuvent intervenir dans le repliement de la chaîne (Fig. 1.9) :

1. **l'hydrophobie :** Les acides aminés dont les radicaux ont plus d'affinité entre eux qu'avec les molécules d'eau les entourant sont dits hydrophobes. Afin d'éviter le contact avec l'eau, le polypeptide se replie sur lui-même pour regrouper les acides aminés hydrophobes en son centre. A l'inverse, les acides aminés hydrophiles ont tendance à se disposer à la périphérie de la chaîne de façon à être en contact avec l'eau.
2. **Les liaisons ioniques**, qui sont les liaisons entre radicaux qui s'ionisent positivement et ceux qui s'ionisent négativement.
3. **Les ponts disulfure**, qui sont des liaisons covalentes (*i.e.* mise en commun d'un doublet d'électrons entre deux atomes) entre des atomes de soufre des radicaux de deux molécules de cystéines.
4. **Les liaisons hydrogène**, qui sont des liaisons électrostatiques de faible intensité.

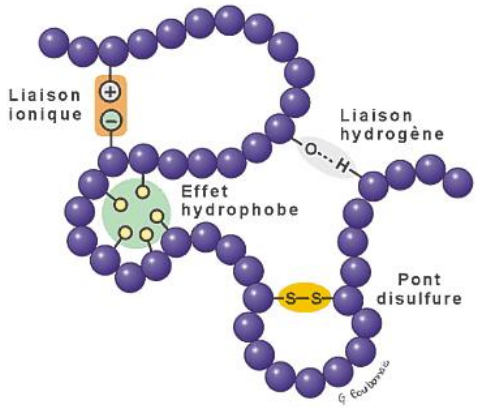


FIG. 1.9 – *Repliements d'un polypeptide (source : internet)*

1.4.3 Structure des protéines

Il existe trois représentations des protéines, de complexité croissante :

La structure primaire est la séquence linéaire des acides aminés constituant la protéine (Section 1.2.2).

La structure secondaire est la structure dans l'espace de chacun des acides aminés qui composent la protéine. Il existe trois types de structures secondaires différentes : en feuillet, en hélice, ou sans aucune structure particulière (Fig. 1.10).

La structure tertiaire de la protéine est la forme qu'adoptent dans l'espace les structures secondaires, qui sont reliées entre elles par des **coudes**. La structure tertiaire est également appelée *sous-unité* (Fig. 1.10).

La structure quaternaire d'une protéine est l'assemblage de plusieurs sous-unités.

1.4.4 Rôles des protéines

Environ cent mille protéines différentes peuvent être fabriquées par notre organisme. La fonction biologique d'une protéine est intimement liée à sa forme. De ce fait, une protéine " déformée " (dénaturée) ne peut donc généralement plus assurer sa fonction.

Les protéines remplissent de nombreux rôles :

- transport de substances dans le sang : comme l'hémoglobine,
- hormones : comme l'insuline,
- transport de substances à travers la membrane cellulaire : par exemple la porine,
- défense du corps humain : les anticorps ou immunoglobulines.

1.4.5 Dégradation des protéines

La demi-vie d'une protéine, c'est-à-dire le temps requis pour que la quantité de cette protéine dans la cellule soit réduite de moitié par rapport à une quantité initiale, varie de quelques minutes à quelques jours et dépend de différents paramètres, par exemple la nature de la protéine, son milieu... Ainsi, la dégradation des protéines est un mécanisme fondamental pour la régulation de leur quantité dans la cellule ou dans le sang. Par exemple, une destruction rapide de certaines protéines est parfois nécessaire pour adapter leur taux en réponse à un stimuli extérieur. De même, des protéines endommagées ou erronées seront repérées et détruites dans la cellule.

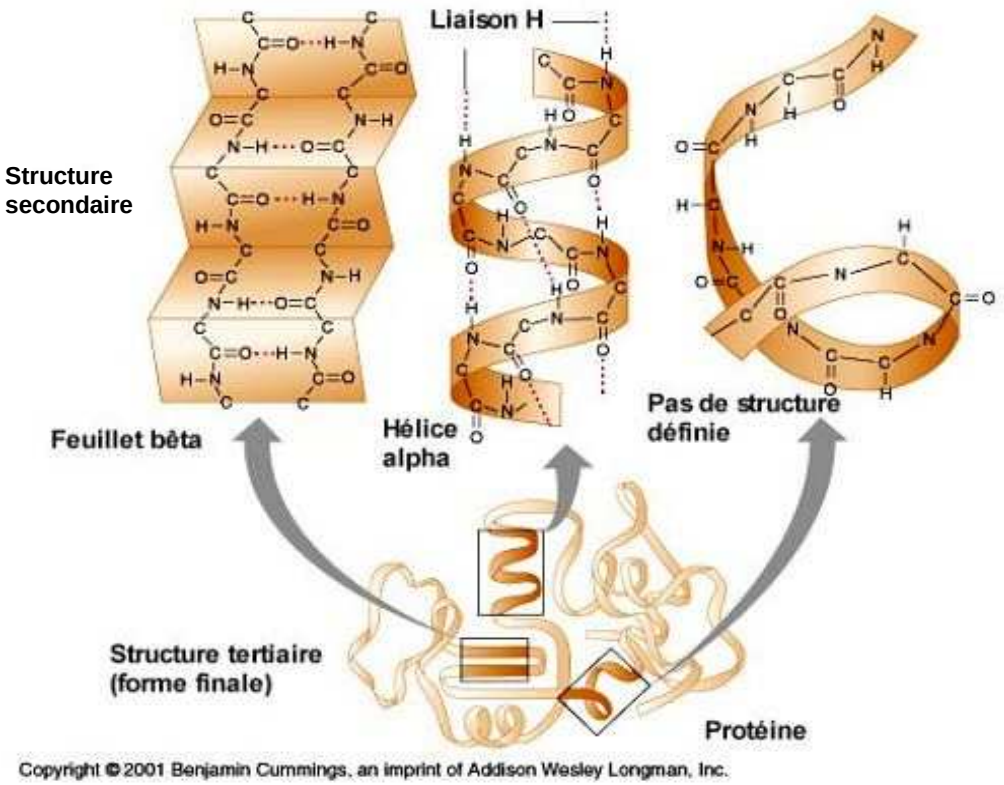


FIG. 1.10 – *Structure des protéines (source : internet)*

La sélection des protéines à détruire se fait principalement via l'utilisation d'un polypeptide appelé **ubiquitine**. Cette molécule est utilisée comme une étiquette apposée sur les protéines à dégrader. Une première molécule d'ubiquitine se fixe à une extrémité de la protéine, puis d'autres viennent se greffer à leur tour, formant ainsi une chaîne d'ubiquitines. Ce mécanisme appelé **ubiquitination**, permet de repérer la protéine qui va être détruite par le **protéasome**, qui est un complexe enzymatique multiprotéique.

Dans la section suivante, nous allons nous intéresser aux anticorps, qui sont des protéines particulières.

1.5 Le système immunitaire

Le système immunitaire permet de discriminer le « soi » du « non-soi » et gère les mécanismes de défense du corps humain [6]. Les substances responsables de la stimulation du système immunitaire sont appelées des **antigènes** (Ag). Ce sont en général des molécules étrangères à l'organisme.

Le système immunitaire est divisé en deux composantes distinctes :

1. le système immunitaire **humoral**,
2. le système immunitaire **cellulaire**.

En cas d'exposition à un antigène, la réponse humorale est assurée par les **lymphocytes B**, qui produisent de grosses protéines véhiculées par le sang appelées **anticorps** ou **immunoglobulines (Ig)**. L'action du système immunitaire cellulaire s'effectue quant-à-elle via les **lymphocytes T** et les **macrophages**. Les lymphocytes T se divisent en deux catégories : les lymphocytes T **cytotoxiques**, qui détruisent les cellules infectées par un virus, et les lymphocytes T **auxiliaires**, qui participent à la régulation des autres lymphocytes.

Remarque : L'ensemble des lymphocytes appartient à la famille des **leucocytes** ou **globules blancs**.

Dans le système immunitaire humoral, les anticorps se divisent en cinq groupes :

1. les **IgE**, propres aux réactions allergiques,
2. les **IgG**, qui interviennent en cas d'infection,
3. les **IgA**, les **IgM** et les **IgD**.

Chaque groupe d'anticorps est divisé en deux catégories :

1. les immunoglobulines **circulantes**, qui sont véhiculées par le sérum,
2. et les immunoglobulines **non circulantes**, qui sont fixées à des cellules.

Structure des anticorps

Les anticorps sont tous constitués de deux chaînes lourdes, notée H pour "Heavy", en violet sur la figure ci-dessous, et deux chaînes légères, notées L pour "Light", colorées en vert. Chaque chaîne légère est constituée d'un domaine constant (noté C) et d'un domaine variable (noté V) ; les chaînes lourdes sont composées d'un fragment variable et de plusieurs fragments constants. Pour un anticorps donné, les deux chaînes lourdes sont identiques, de même pour les deux chaînes légères. Enfin, la partie constante des chaînes lourdes caractérise l'appartenance à un groupe d'anticorps.

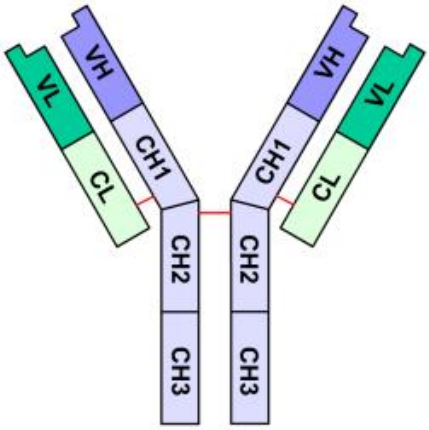


FIG. 1.11 – *La structure des anticorps* (source : Wikipedia)

A l'intérieur d'un même groupe, les anticorps se distinguent les uns des autres par leur domaine variable, qui caractérise la protéine contre laquelle ils sont dirigés. Les régions de la protéine qui stimulent la production d'anticorps sont appelées **épitopes**.

Partie I

Recherche de biomarqueurs du cancer

Chapitre 2

Introduction à la cancérologie

2.1 Les mutations génétiques

Avant d'introduire les mécanismes principaux du cancer, il est nécessaire de définir le concept de mutation.

2.1.1 Définition

En biologie, le terme mutation est utilisé pour désigner la modification de la séquence des bases de l'ADN ou de l'ARN. Les conséquences d'une mutation peuvent être très importantes. En effet, une modification d'un unique nucléotide induisant un changement d'acide aminé peut modifier la structure tridimensionnelle de la protéine correspondante, et par conséquent altérer sa fonction.

Un individu atteint d'une mutation est appelé mutant.

2.1.2 Transmission des mutations

Une mutation affectant les cellules germinales, c'est-à-dire les cellules susceptibles de former des gamètes, peut être transmise aux descendants du sujet mutant. En revanche, une mutation affectant les cellules somatiques, c'est-à-dire les cellules qui ne sont jamais à l'origine des gamètes, comme en cas d'exposition à des radiations ou à des substances chimiques, ne sera pas transmise à la descendance et n'affectera donc que l'individu touché initialement.

2.1.3 Les types de mutations

Les mutations peuvent être classées en fonction de leur impact sur la séquence d'ADN ou d'ARN :

- les substitutions, c'est-à-dire le remplacement d'un nucléotide par un autre ;
- les insertions, c'est-à-dire l'ajout d'un ou plusieurs nucléotides dans une séquence ;
- les délétions ou gaps, qui consistent en le retrait d'un nucléotide.

Ces mutations peuvent avoir des conséquences diverses sur la protéine après traduction de l'ARNm.

En effet, certains acides aminés sont représentés par plusieurs codons, comme par exemple l'*arginine*, qui est peut être codée par AGA ou AGG (Table 1.1). Ainsi une substitution modifiant le deuxième A de AGA en G n'induirait donc pas de changement d'acide aminé. Une telle mutation est dite **silencieuse** ou **muette** (Fig. 2.1).

Par ailleurs, une mutation provoquant un changement d'acide aminé n'implique pas nécessairement de modification fonctionnelle de la protéine résultante. Ces mutations sont dites **neutres**. C'est le cas par exemple d'une substitution de G par A dans le codon AGA, remplaçant ainsi une arginine par une lysine (Fig. 2.1).

Enfin, les délétions et les insertions ont nécessairement un impact important sur la traduction de l'ARNm en protéine, puisqu'elles entraînent un décalage du cadre de lecture du ribosome (Fig. 2.1).

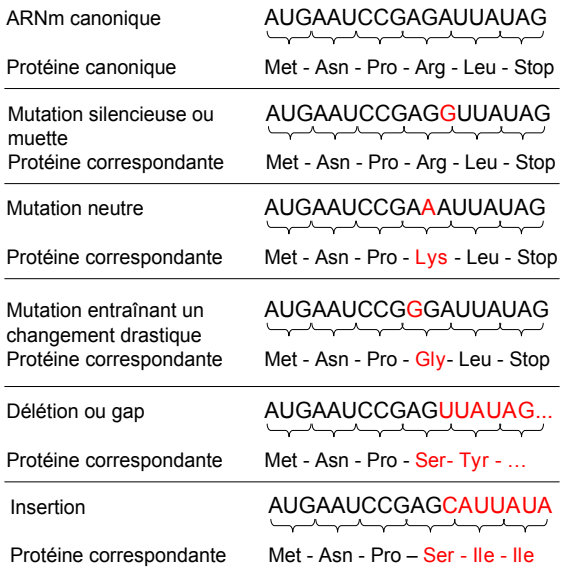


FIG. 2.1 – *Conséquences de mutations sur l'ARNm*

2.2 Quelques notions de cancérologie

La cancérologie ou oncologie est la spécialité médicale de l'étude, du diagnostic et du traitement des cancers.

Le cancer est une maladie caractérisée par la prolifération anarchique de cellules anormales au sein d'un tissu normal de l'organisme. Ces cellules dérivent toutes d'une même cellule initiatrice qui a acquis certaines caractéristiques lui permettant de se diviser indéfiniment et de se disséminer dans l'organisme. Les nouvelles cellules résultantes peuvent former une tumeur maligne (un néoplasme) puis se propager à travers le corps. Une tumeur est une partie du corps d'un animal ou d'un végétal ayant augmenté de volume. Cette grosseur peut être notamment due à un oedème ou à un tissu cancéreux ou précancéreux.

2.2.1 Caractéristiques biologiques du cancer

Le cancer est considéré actuellement comme une " maladie des gènes ".

On distingue :

- les **oncogènes**, qui sont des gènes qui favorisent l'apparition d'un cancer quand ils sont touchés par une mutation,
- les **gènes suppresseurs de tumeurs**, qui sont des gènes qui, à l'état normal, empêchent la création de tumeurs. Si une mutation inhibe leur production ou modifie leur fonction, les cellules peuvent proliférer de façon anarchique jusqu'à former une tumeur. Par exemple, environ la moitié des tumeurs sont déficientes en gène suppresseur de tumeur *TP53*.

2.2.2 Facteurs de risque

La survenue d'un cancer peut avoir différentes causes :

- les radiations ou des produits chimiques qualifiés de carcinogènes,
- des prédispositions héréditaires,
- certains virus, comme le *papillomavirus*, impliqué dans certains cancers de l'utérus.

  Ces virus contiennent habituellement dans leur génome certains oncogènes ou gènes inactivateurs de gènes suppresseurs de tumeur. Les virus semblent jouer un rôle dans environ 15% des cancers ;
- des bactéries, comme *Helicobacter pylori*, peuvent provoquer des carcinogenèses, c'est-à-dire la création d'un cancer, par un processus d'inflammation chronique.

Le nombre total des décès par cancer en France était, en 1997, de 146 705, soit environ 241 décès pour 100 000 habitants. C'est la deuxième cause de mortalité après les maladies cardio-vasculaires. En 2000, 278 000 personnes étaient atteintes d'un cancer et 150 000 personnes en sont mortes. Notons que la France est le pays ayant la plus longue survie après la survenue d'un cancer.

Le pourcentage de cancers causés par certains facteurs de risque est donné Table 2.1. Le pourcentage de décès dus à ces mêmes facteurs de risque est également donné [7].

Facteurs de risque	Cas	Décès
Tabac	18,2	23,9
Alcool	8,1	6,9
Agents infectieux (ex : virus, bactéries...)	3,3	3,7
Inactivité physique	2,3	1,6
Obésité et surpoids	2,2	1,6
Hormones	2,1	0,9
Ultraviolets	2,0	0,7
Expositions diverses (ex : amiante...)	1,6	2,4
Facteurs de reproduction (ex : grossesses multiples...)	0,8	0,4
Polluants	0,1	0,2

TAB. 2.1 – Pourcentages de cancers causés et de décès pour chaque facteur de risque

Remarque : Un même cancer peut être attribué à plusieurs facteurs de risque. De ce fait, les pourcentages de la Table 2.1 ne peuvent être additionnés.

2.2.3 L'évolution de la maladie

De son foyer initial, le cancer se développe tout d'abord localement depuis la tumeur. L'irrigation de la tumeur se fait par la création de nouveaux vaisseaux sanguins et l'envahissement du système circulatoire. Ce phénomène est appelé *angiogenèse*. Les organes voisins sont alors compressés et les tissus adjacents peuvent être détruits. Le cancer se propage ensuite à distance dans d'autres tissus pour former des tumeurs secondaires appelées **métastases**.

L'évolution du cancer dépend du type du cancer et de sa prise en charge : certains ne font que très peu de métastases et sont très sensibles aux traitements permettant d'aboutir dans la grande majorité des cas à une guérison. D'autres sont malheureusement encore très difficilement maîtrisables et peuvent entraîner le décès à court terme.

2.2.4 Conséquences du cancer

La vie dépend de la bonne marche d'un certain nombre de fonctions, dont la respiration, la digestion et l'excrétion. Selon que les cellules cancéreuses altèrent l'un de ces trois systèmes, le patient peut mourir :

- d'insuffisance respiratoire ;
- de dénutrition ;
- d'empoisonnement, par accumulation de substances toxiques normalement filtrées et excrétées par les reins et le foie.

2.3 La médecine du cancer

2.3.1 Diagnostic

Même s'il existe des éléments permettant d'identifier un cancer avec une grande probabilité, le diagnostic de certitude ne se fait que sur analyse au microscope (anatomopathologie) d'un échantillon de la tumeur (éventuellement aidé par d'autres techniques). Cet échantillon vient soit d'une biopsie (prélèvement d'un morceau de la tumeur) qui peut être faite, suivant la localisation, suivant différentes procédures (fibroscopie, ponction à travers la peau...), soit du retrait d'une pièce opératoire (tumeur enlevée par le chirurgien).

Le diagnostic précoce du cancer est à l'heure actuelle un réel enjeu de santé publique. En effet, plus un cancer est détecté tôt, plus il a de chances d'être guéri. De ce fait, la recherche de **biomarqueurs**, c'est-à-dire de protéines plus fréquemment produites chez des patients cancéreux que chez des patients sains, est l'objet de nombreuses études de biologie et de médecine.

2.3.2 Traitements

Le traitement nécessite :

- d'avoir un diagnostic de certitude et de connaître le type de cancer,
- d'évaluer son extension locale, régionale et la présence ou non de métastases,
- d'évaluer l'état général du patient (âge, fonctions cardiaque et rénale, présence d'autres maladies).

Suivant les cas, il repose sur :

- l'exérèse (l'ablation) chirurgicale de la tumeur quand cela est possible ;
- une chimiothérapie, prescription de médicaments s'attaquant au cancer et à ses métastases ;
- une radiothérapie, l'irradiation de la tumeur permettant de la réduire, voire de la faire disparaître.

Certains cancers peuvent bénéficier également :

- d'un traitement hormonal ;
- d'un traitement à visée immunologique.

2.3.3 Prévention

Elle se base sur :

- l'évitement ou la diminution de l'exposition aux carcinogènes de l'environnement ;
- la détection précoce des cancers (cancer du sein par mammographie systématique et par autopalpation, recherche de sang dans les selles pour les cancers colorectaux...).
- le traitement des lésions pré-cancéreuses.

Chapitre 3

Contrôle du risque de première espèce dans un ensemble de tests

Donnons comme références globales sur les tests multiples [8] et [9].

3.1 Introduction

Lorsque l'on est amené à effectuer des milliers, voire des millions de tests simultanés, il se pose le problème du contrôle du risque de première espèce de cet ensemble de tests. Par exemple, dans l'analyse des données du transcriptome, on cherche à savoir quels sont les gènes différentiellement exprimés entre deux conditions expérimentales. On dispose de m gènes. Pour $i = 1, \ldots, m$, on considère le test :

$$\begin{cases} H_{0,i} : \text{le gène } i \text{ n'est pas différentiellement exprimé,} \\ H_{1,i} : \text{le gène } i \text{ est différentiellement exprimé.} \end{cases} \tag{3.1}$$

On note m_0 le nombre de gènes dans H_0 (c'est-à-dire tels que H_0 soit vraie), et m_1 le nombre de gènes dans H_1. Soit une statistique T_i qui permet de tester $H_{0,i}$. On note $t_{i,obs}$ la réalisation observée de T_i. On note p_i la probabilité critique (p-value ou p-valeur) :

$$\begin{aligned} p_i &= P(T_i \geq t_{i,obs} \text{ sous } H_0) \\ &= 1 - P(T_i < t_{i,obs} \text{ sous } H_0) \\ &= 1 - F_{0,i}(t_{i,obs}), \end{aligned} \tag{3.2}$$

où $F_{0,i}$ est la fonction de répartition de T_i sous H_0.

p_i est la réalisation observée de la variable aléatoire réelle (v.a.r) :

$$P_i = 1 - F_{0,i}(T_i). \tag{3.3}$$

Sous certaines hypothèses (par exemple : $F_{0,i}$ continue et strictement croissante), la loi de $F_{0,i}(T_i)$ est $\mathcal{U}(0,1)$ sous $H_{0,i}$; alors, sous $H_{0,i}$, la loi de P_i est $\mathcal{U}(0,1)$.

On fait dans la suite, sauf spécification contraire, les hypothèses :

(1) $P_i \sim \mathcal{U}(0,1)$,

(2) les variables aléatoires réelles $P_1, \ldots, P_m$ sont mutuellement indépendantes.

3.2 Procédures de tests multiples : éléments de base

3.2.1 Règle de décision

Définition *La règle de décision d'un test multiple consiste à définir un seuil t et, pour tout i, à rejeter l'hypothèse $H_{0,i}$ lorsque $p_i \leq t$.*

Problème : Définir t, seuil ou risque de première espèce de chaque test.

3.2.2 Faux positifs et faux négatifs

Définitions *: On dit qu'un test est*

- *positif si l'on rejette H_0,*
- *faux positif si l'on rejette H_0 alors que H_0 est vraie,*
- *vrai positif si l'on rejette H_0 alors que H_0 est fausse,*
- *négatif si l'on ne rejette pas H_0,*
- *vrai négatif si l'on ne rejette pas H_0 alors que H_0 est vraie,*
- *faux négatif si l'on ne rejette pas H_0 alors que H_0 est fausse.*

On définit les variables aléatoires :

1. $V(t)$, *le nombre de faux positifs au seuil t,*
2. $S(t)$, *le nombre de vrais positifs au seuil t,*
3. $R(t) = V(t) + S(t)$, *le nombre de tests positifs au seuil t,*
4. $U(t)$, *le nombre de vrais négatifs au seuil t,*
5. $T(t)$, *le nombre de faux négatifs au seuil t,*
6. $W(t) = U(t) + T(t)$, *le nombre de tests négatifs au seuil t.*

Représentons ces variables dans le tableau suivant :

	H_0 acceptée	H_0 rejetée	Somme
H_0 vraie	$U(t)$	$V(t)$	m_0
H_0 fausse	$T(t)$	$S(t)$	m_1
Somme	$W(t)$	$R(t)$	m

Remarquons que m_0 et m_1 sont des constantes inconnues. Par contre, $W(t)$ et $R(t)$ sont deux variables aléatoires observées.

Loi de $V(t)$

On considère les m_0 tests pour lesquels H_0 est vraie. On suppose sans perte de généralité que ce sont les m_0 premiers tests.
Soit, pour $i = 1, \ldots, m_0$:

$$V_i(t) = \begin{cases} 1 & \text{si l'on rejette } H_0 \text{ au } i^{eme} \text{ test,} \\ 0 & \text{sinon.} \end{cases} \tag{3.4}$$

D'après l'hypothèse (1) :

$$P(V_i(t) = 1) = P(P_i \leq t) = t \text{ (car } H_0 \text{ est vraie).} \tag{3.5}$$

Les v.a.r $V_1(t), \ldots, V_{m_0}(t)$ sont mutuellement indépendantes sous l'hypothèse (2).

Or :

$$V(t) = V_1(t) + \ldots + V_{m_0}(t), \tag{3.6}$$

donc

$$V(t) \sim \mathcal{B}(m_0, t). \tag{3.7}$$

Comme $U(t) = m_0 - V(t)$:

$$U(t) \sim \mathcal{B}(m_0, 1-t). \tag{3.8}$$

Loi conditionnelle de $V(t_1)$ par rapport à $V(t_2)$, $t_2 > t_1$

Comme $t_1 < t_2$, $V(t_1) \leq V(t_2)$, car $P_i \leq t_1$ implique $P_i \leq t_2$. Supposons $V(t_2) = k_2$: il y a k_2 tests parmi les m_0 pour lesquels $P_i \leq t_2$; $V(t_1)$ a alors pour réalisations possibles $0, 1, \ldots, k_2$; la probabilité conditionnelle $P(P_i \leq t_1 | P_i \leq t_2)$ est égale à $\frac{P(P_i \leq t_1)}{P(P_i \leq t_2)} = \frac{t_1}{t_2}$; donc la loi conditionnelle de $V(t_1)$ lorsque $V(t_2) = k_2$ est $\mathcal{B}(k_2, \frac{t_1}{t_2})$.
On a donc :

$$\mathcal{L}(V(t_1)|V(t_2)) = \mathcal{B}(V(t_2), \frac{t_1}{t_2}). \tag{3.9}$$

$$E[V(t_1)|V(t_2)] = V(t_2)\frac{t_1}{t_2}, \text{ pour } t_1 < t_2 \tag{3.10}$$

$$E\left[\frac{V(t_1)}{t_2}|V(t_2)\right] = \frac{V(t_2)}{t_1}, \text{ pour } t_1 < t_2 \tag{3.11}$$

3.2.3 Critères de détermination du seuil t

1. Critères basés sur le nombre de faux positifs

a) On a

$$E[V(t)] = m_0 t \leq mt. \tag{3.12}$$

Critère 1 *On cherche t tel que le nombre moyen de faux positifs $E[V(t)]$ soit majoré par un nombre fixé.*

Définitions

1. *$E[V(t)]$ est appelé le taux d'erreur par famille ($PFER$ "per-family error rate").*
2. *$\frac{E[V(t)]}{m}$ est appelé le taux d'erreur par comparaison ($PCER$ "per-comparison error rate").*

b) ***Critère 2*** *Un nombre maximal v de faux positifs étant fixé, on cherche t tel que $P(V(t) > v) < \alpha$, pour α fixé.*

Définition *$P(V(t) > v)$ est appelé le taux d'erreur global généralisé ($gFWER(v)$* "generalized family-wise error rate").

c) **Définition** *$P(V(t) > 0)$ est appelé le taux global d'erreur ($FWER$* "family-wise error rate") : *c'est la probabilité de rejeter $H_{0,i}$ alors que $H_{0,i}$ est vraie pour au moins un i.*

Critère 3 *α étant fixé, on cherche t tel que $P(V(t) > 0) < \alpha$.*

Sous l'hypothèse (2) :

$$P(V(t) > 0) = 1 - (1-t)^{m_0}. \tag{3.13}$$

$$1 - (1-t)^{m_0} < \alpha \Leftrightarrow t < 1 - (1-\alpha)^{\frac{1}{m_0}}. \tag{3.14}$$

Or pour $\alpha \neq 0$:

$$(1-\alpha)^{\frac{1}{m_0}} < 1 - \frac{\alpha}{m_0} \Leftrightarrow \frac{\alpha}{m_0} < 1 - (1-\alpha)^{\frac{1}{m_0}}. \tag{3.15}$$

Donc la relation (3.14) est vérifiée pour $t = \frac{\alpha}{m_0}$. Si on ne connaît pas m_0, on peut prendre $t = \frac{\alpha}{m} \leq \frac{\alpha}{m_0}$.

2. Critères basés sur la proportion de faux positifs parmi les positifs

a) **Définition** *Le taux de fausses découvertes (ou false discovery rate (FDR)) est défini par :*

$$FDR(t) = E\left[\frac{V(t)}{R(t)}\mathbb{1}_{(R(t)>0)}\right]. \tag{3.16}$$

C'est le taux moyen de faux positifs parmi les positifs.

Critère 4 *α étant fixé, on cherche t tel que $FDR(t) < \alpha$.*

b) ***Critère 5*** *α étant fixé, un taux moyen β de faux positifs parmi les positifs étant fixé, on cherche t tel que*

$$P\left(\frac{V(t)}{R(t)}\mathbb{1}_{(R(t)>0)} > \beta\right) < \alpha. \tag{3.17}$$

3.3 Estimation et contrôle du taux de fausses découvertes

Théorème 3.3.1 *Soit $0 < \beta < 1$. On suppose les p_i ordonnés : $p_1 \leq \ldots \leq p_m$. Soit $r = max\left\{i : \frac{mp_i}{i} \leq \beta\right\}$. En prenant $t = p_r$, sous l'hypothèse (2), on a :*

$$FDR(t) \leq \frac{m_0}{m}\beta. \tag{3.18}$$

Ce résultat a été établi par Benjamini et Hochberg [10].

3.3.1 Estimation du FDR

Storey *et al.* [11] ont établi les deux théorèmes suivants :

a) Cas où m_0 est connu

Théorème 3.3.2 *Soit*

$$\widehat{FDR}(t) = \begin{cases} \frac{m_0 t}{R(t)} & \text{si } R(t) > 0 \\ 0 & \text{si } R(t) = 0. \end{cases} \tag{3.19}$$

On a

$$E[\widehat{FDR}(t)] \geq FDR(t). \tag{3.20}$$

L'estimateur $\widehat{FDR}(t)$ de $FDR(t)$ a un biais positif ou nul.

b) Cas où m_0 est inconnu

Soit $\hat{m}_0(\lambda) = \frac{W(\lambda)}{1-\lambda}$, où $\lambda \in [0, 1[$.

On a

$$W(\lambda) = U(\lambda) + T(\lambda) \geq U(\lambda). \tag{3.21}$$

$$E[W(\lambda)] \geq E[U(\lambda)] = m_0(1-\lambda) \Leftrightarrow \frac{E[W(\lambda)]}{1-\lambda} \geq m_0. \tag{3.22}$$

L'estimateur $\hat{m}_0(\lambda) = \frac{W(\lambda)}{1-\lambda}$ de m_0 a un biais positif ou nul.

Théorème 3.3.3 *Soit*

$$\widehat{FDR}(t, \lambda) = \begin{cases} \frac{\hat{m}_0(\lambda)t}{R(t)} & \text{si } R(t) > 0 \\ 0 & \text{si } R(t) = 0. \end{cases} \tag{3.23}$$

On a

$$E[\widehat{FDR}(t, \lambda)] \geq FDR(t). \tag{3.24}$$

L'estimateur $\widehat{FDR}(t, \lambda)$ de $FDR(t)$ a un biais positif ou nul.

3.3.2 Contrôle du FDR

a) Si l'on connaît m_0, on définit t_α par :

$$t_\alpha = \sup \left\{ t : \widehat{FDR}(t) \leq \alpha \right\}. \tag{3.25}$$

Théorème 3.3.4 *On a* $FDR(t_\alpha) = \alpha$.

b) Si l'on ne connaît pas m_0, on définit t_α par :

$$t_\alpha = \sup \left\{ t : \widehat{FDR}(t, \lambda) \leq \alpha \right\}. \tag{3.26}$$

Théorème 3.3.5 *On a* $FDR(t_\alpha) = (1 - \lambda^{m_0})\alpha \leq \alpha$.

3.4 Estimation de $\pi_0 = \frac{m_0}{m}$

3.4.1 Location Based Estimator pour une v.a absolument continue

La connaissance d'une estimation de π_0 nous permettra par exemple d'avoir une estimation du nombre moyen de faux positifs, $E[V(t)] = m_0 t = m\pi_0 t$. L'estimation de π_0 va être réalisée à partir de la distribution du degré de signification P. Exprimons la densité de P. Notons :

- f_0 la densité de P sous H_0 (densité uniforme sur (0,1)) ;
- f_1 la densité de P sous H_1.

La densité de P s'écrit sous la forme d'un modèle de mélange :

$$f(x) = \pi_0 f_0(x) + (1-\pi_0) f_1(x). \tag{3.27}$$

La méthode LBE (Location Based Estimator) permet d'obtenir aisément un estimateur avec biais positif du nombre moyen de faux positifs. Elle est due à Dalmasso, Broët et Moreau [12].
On a :

$$E[P] = \int_{\mathbb{R}} x f(x) dx = \pi_0 E_0[P] + (1-\pi_0) E_1[P] \tag{3.28}$$

$$\iff \quad \frac{E[P]}{E_0[P]} = \pi_0 + (1-\pi_0)\frac{E_1[P]}{E_0[P]} \geq \pi_0. \tag{3.29}$$

Or $E_0[P] = \frac{1}{2}$, car $\mathcal{L}(P|H_0) = \mathcal{U}(0,1)$. Donc $2E[P] \geq \pi_0$.

Un estimateur sans biais de $2E[P]$ est $2\frac{1}{m}\sum_{i=1}^{m} P_i$; c'est un estimateur de π_0 avec un biais positif.

3.4.2 Location Based Estimator pour une v.a discrète

Supposons que $T_i = T$ soit une variable aléatoire discrète. Dans ce cas, le degré de signification P correspondant ne suit pas la loi $\mathcal{U}_{[0,1]}$.
Soit $\{t_i, i \in I\}$, avec $t_l < t_m$ pour $l < m$, l'ensemble des modalités de la variable aléatoire discrète T. Calculons $E_0[P] = E[P|H_0]$.

$$\begin{aligned} E[P|H_0] &= 1 - E[F_0(T)|H_0] \\ &= 1 - \sum_{i\in I} F_0(t_i) P(T = t_i|H_0) \\ &= 1 - \sum_{i\in I}\sum_{l<i} P(T = t_l|H_0) P(T = t_i|H_0). \end{aligned} \tag{3.30}$$

Or

$$(\sum_{i\in I} P(T = t_i|H_0))^2 = 1 = \sum_{i\in I} P^2(T = t_i|H_0) + 2\sum_{i\in I}\sum_{l<i} P(T = t_l|H_0) P(T = t_i|H_0). \tag{3.31}$$

Donc

$$\begin{aligned} E[P|H_0] &= 1 + \frac{1}{2}(\sum_{i\in I} P^2(T = t_i|H_0) - 1) \\ &= \frac{1}{2} + \frac{1}{2}\sum_{i\in I} P^2(T = t_i|H_0) > \frac{1}{2}. \end{aligned} \tag{3.32}$$

En écrivant que :

$$E[P] = E[P|H_0]P(H_0) + E[P|H_1]P(H_1) \geq E[P|H_0]P(H_0) > \frac{1}{2}\frac{m_0}{m}, \tag{3.33}$$

on obtient que :
- $\frac{1}{m}\sum_{i=1}^{m} P_i$ est un estimateur avec biais positif de $\frac{1}{2}\frac{m_0}{m}$,
- $2\sum_{i=1}^{m} P_i$ est un estimateur avec biais positif de m_0,
- $2t\sum_{i=1}^{m} P_i$ est un estimateur avec biais positif de $m_0 t$, qui est un majorant du nombre moyen de faux positifs.

Le résultat obtenu dans la section précédente reste donc valable dans le cas d'une statistique de test discrète. Notons que ce cas n'a pas encore été développé dans la littérature.

3.4.3 Autres méthodes

D'autres méthodes plus complexes d'estimation de π_0 existent, citons notamment les méthodes de la $Q-$value [13], BUM [14] et SPLOSH [15].

Chapitre 4

Comparaison de la probabilité de survenue d'une substitution sur un ARNm sain et sur un ARNm cancéreux

4.1 Introduction

Le cancer est une maladie génétique causée par une accumulation de mutations dans les oncogènes et les gènes suppresseurs de tumeurs [16] (Chapitre 2). Des travaux récents montrent que des mutations somatiques surviennent plus fréquemment que ce qui était suspecté. Cependant, elles restent relativement rares (3,1 mutations pour 10^6 bases), soit en moyenne 90 substitutions d'acides aminés par tumeur [17]. De ce fait, l'hétérogénéité caractéristique des cellules cancéreuses ainsi que le grand nombre de variants protéiques observés dans certaines études de protéomique [18], ne peuvent être uniquement expliqués par l'occurrence de mutations somatiques.

L'éventuelle implication de la transcription dans l'hétérogénéité moléculaire du cancer n'avait jusqu'alors pas été considérée. En effet, la transcription est supposée être un mécanisme fidèle, contrôlé de plus par un système complexe de vérification. Ainsi, le but de cette étude est de comparer, pour 17 gènes d'intérêt, la variabilité des séquences nucléotidiques des ARNm extraits de tissus cancéreux avec celle des ARNm extraits de tissus sains.

Pour cela, une analyse statistique est mise en place afin de comparer les probabilités de survenue d'une substitution pendant la transcription dans les cas sain et cancéreux, à chaque position de la séquence d'ARNm d'un gène.

Nous introduirons ainsi dans un premier temps les séquences EST (Expressed Sequences Tags) [19], qui jouent un rôle central dans notre étude. Ensuite, nous décrirons l'approche bioinformatique appliquée à la collecte et au formatage des données, puis l'analyse statistique réalisée et les résultats obtenus.

4.2 Les Expressed Sequences Tags (EST)

L'ARNm est une molécule caractérisée par son instabilité. De ce fait, séquencer l'ARNm, *i.e* déterminer la séquence de nucléotides qui le composent, est un processus long et complexe. Ainsi, il n'existe pas à l'heure actuelle de technique de séquençage directe de l'ARNm qui permette de réaliser rapidement et simplement une étude à large échelle. Les études reposant sur l'analyse des séquences d'un grand nombre d'ARNm nécessitent la préparation en laboratoire d'une

autre molécule plus stable : l'**ADN complémentaire (ADNc)**. Les EST sont des séquences de nucléotides correspondant à de courts fragments d'ADNc.

4.2.1 Synthèse de l'ADNc et séquençage des EST

Le protocole commence par l'isolation des ARNm contenus dans un ensemble de cellules. C'est en utilisant la particularité de l'ARNm de se terminer par une **queue poly(A)**, c'est à dire une série de nucléotides A, que la fabrication de l'ADNc peut commencer. A l'inverse de l'ARNm, l'ADNc est une molécule à deux brins, qui sont synthétisés l'un après l'autre, en cinq étapes distinctes (Figure 4.1) :

1. Un **oligo-dT**, c'est à dire une courte suite de nucléotides T, vient s'associer (ou **s'hybrider**) à la queue poly(A) de l'ARN pour amorcer la fabrication du premier brin ;
2. Le premier brin de l'ADNc est d'abord obtenu par **transcription inverse** de l'ARNm : une enzyme appelée *reverse transcriptase* constitue un nouveau brin d'ADN à partir du brin d'ARNm, par la règle suivante : en face de la base A de l'ARNm est placé un T sur l'ADNc, en face de la base U est placé un A, en face de C un G et en face de G un C ;
3. L'ARNm est ensuite extrait du milieu, puis une queue poly(G) est ajoutée à l'extrémité 3' du premier brin d'ADNc ;
4. Une amorce d'oligo-dC est ajoutée pour s'hybrider avec la queue poly(G) du premier brin d'ADNc ;
5. Le deuxième brin est synthétisé par complémentarité du premier brin à partir de cette amorce.

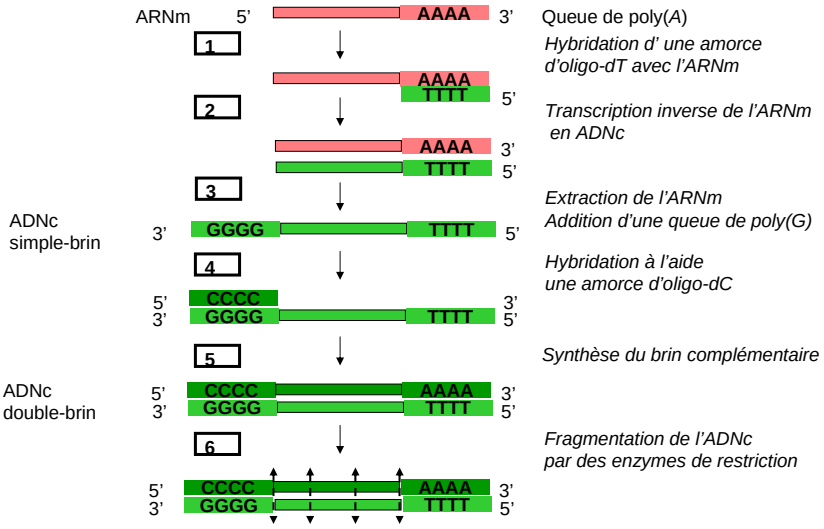


FIG. 4.1 – *La synthèse de l'ADNc*

Une fois l'ADNc double-brin entièrement synthétisé, des enzymes dites de **restriction**, comme par exemple *Alu1* et *EcoRI*, découpent l'ADNc en plusieurs courts fragments. Chaque fragment

d'ADNc est ensuite inséré dans un **plasmide** ou **vecteur de clonage**, qui est une séquence d'ADN circulaire (Figure 4.2). Cette opération permet le séquençage du fragment inséré, *i.e* la lecture de la suite de nucléotides A, T, C, G le constituant. Les **Expressed Sequences Tags (EST)**, ou **marqueurs de séquences exprimées**, sont les séquences de nucléotides obtenues à la fin de l'expérience, d'une longueur allant en général de 200 à 500 bases.

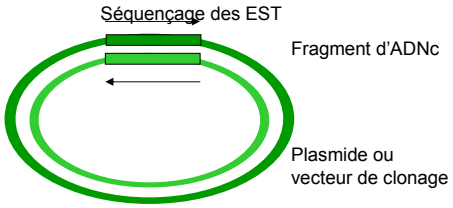


FIG. 4.2 – *Insertion des fragments d'ADNc dans un plasmide*

En résumé, les EST sont des copies partielles des séquences des ARNm d'un gène présents dans un tissu. Bien que ce phénomène soit rare, il est admis que certains nucléotides puissent être sujets à des **erreurs de copie** pendant la transcription inverse.

Remarque : Il existe en fait plusieurs techniques de synthèse des ADNc et de séquençage des EST, qui peuvent différer en fonction du type d'amorce choisi, des enzymes de restriction... Cependant, la trame générale, exposée ici, reste identique.

La base de données **dbEST** du *National Center for Biotechnology Information (NCBI)* , disponible en ligne sur internet (http ://www.ncbi.nlm.nih.gov/), réunit aujourd'hui plus de 6,1 millions d'EST d'origine humaine. La moitié environ provient de tissus d'origine cancéreuse. En pratique, les EST sont contenus dans des fichiers informatiques, dans lesquels se trouvent la séquence nucléotidique ainsi qu'un certain nombre de champs descriptifs (laboratoire dépositaire, espèce, origine du tissu...).

Lors de la création des EST, des ARNm issus de gènes différents se trouvent dans la même préparation. Par conséquent, les séquences d'ADNc obtenues ne sont pas assignées à un gène en particulier. De ce fait, le gène auquel se réfère un EST n'est pas connu a priori dans dbEST. Cependant, une autre banque de données de NCBI appelée *RefSeq* fournit pour chaque gène une **séquence de référence** représentant le brin d'ARNm du gène en question. C'est le processus d'alignement, décrit plus loin, qui permet de regrouper les EST relatifs au même gène. La séquence de référence d'un gène est la même pour tous les individus, qu'ils soient sains ou cancéreux. Les positions d'un gène pouvant être sujettes à des variabilités individuelles connues, appelées *Single Nucleotide Polymorphism (SNP)*, sont répertoriées dans la base de données *dbSNP* et indiquées sur sa séquence de référence du gène [20].

4.2.2 Les erreurs de séquençage

Lors du séquençage, une base d'un ADNc peut être mal lue et remplacée par une autre base. Ainsi, on peut observer en pratique sur l'EST une base différente de celle existant en réalité sur l'ADNc. Ce phénomène est appelé **erreur de séquençage** : à n'importe quelle position, il est admis qu'une telle faute de lecture puisse se produire avec une probabilité ϵ comprise entre 1 et 5%. En d'autres termes, ceci signifie qu'en moyenne 1 à 5 bases sur 100 sont mal lues. Il

n'est pas précisé dans la littérature comment ϵ a été estimé. En effet, compte tenu du mode d'obtention des séquences (laboratoires et procédés différents, amélioration des techniques dans le temps...), il n'est pas possible de connaître la vraie valeur de la probabilité de l'erreur de séquençage. De plus, la probabilité ϵ peut en réalité dépendre de la position du nucléotide sur l'ARNm. Ainsi, les EST ont la réputation d'être peu fiables et de mauvaise qualité [21, 22].

Par abus de langage, dans cette étude, l'erreur de séquençage englobe également les éventuelles erreurs de copies induites par la transcription inverse lors du passage de l'ARNm à l'ADNc. Ainsi :

$$\text{erreur de séquençage} = \begin{cases} \text{erreur de copie pendant la transcription inverse} \\ \text{ou} \\ \text{erreur de lecture} \end{cases}$$

La séquence de référence est en revanche considérée comme très fiable, car son séquençage est réalisé plusieurs fois - à l'inverse des EST dont la lecture est unique - et est d'excellente qualité (l'erreur est estimée à 0,01%).

Le but de cette étude est de tester s'il existe une différence significative entre les fréquences de survenue d'une substitution à une position fixée d'un ARNm cancéreux et d'un ARNm sain. Les EST étant les "miroirs" des ARNm - aux erreurs de séquençage près - , une substitution de certains nucléotides de la séquence d'ARNm d'un gène devrait se visualiser dans les EST.

Nous allons donc comparer les séquences des EST cancéreux de 17 gènes à celles des EST normaux de ces mêmes gènes. La liste des gènes étudiés est donnée ci-dessous :

ALB, ALDOA, ATP5A1, CALM2, ENO1, FTH1, TL, GAPDH, HSPA8, LDHA, RPL7A, RPS4X, RPS6, TMSB4X, TPI1, TPT1, VIM.

Ces gènes ont été retenus pour leur grand nombre d'EST disponibles. Par ailleurs, ils codent pour des protéines aux fonctions variées, n'ayant pas nécessairement de rôle connu dans les mécanismes du cancer.

L'approche bioinformatique mise en place pour collecter et formater ces données est décrite dans la section suivante.

4.3 Approche bioinformatique

Pour chaque gène, le même protocole est suivi :

1. Recherche de la séquence de référence d'ARNm dans la banque RefSeq,
2. Alignement de l'ensemble des EST de dbEST sur la séquence de référence à l'aide du logiciel MegaBLAST 2.2.13 [23] (Figure 4.3),
3. Séparation des EST du gène en deux blocs, en fonction de l'origine du tissu : cancer ou sain. Les EST pour lesquels le champ correspondant à l'origine du tissu n'est pas renseigné sont automatiquement retirés de l'étude.

Remarque : Lors de la deuxième étape, tous les EST de la banque sont alignés sur la séquence de référence du gène étudié. Seuls sont retenus les EST ayant 100% d'idendité sur au moins 16 bases consécutives et plus de 90% d'identité sur au moins 50 bases. Ces critères stricts permettent d'admettre raisonnablement que seuls les EST relatifs au gène d'intérêt, portant ou non des substitutions, sont conservés.

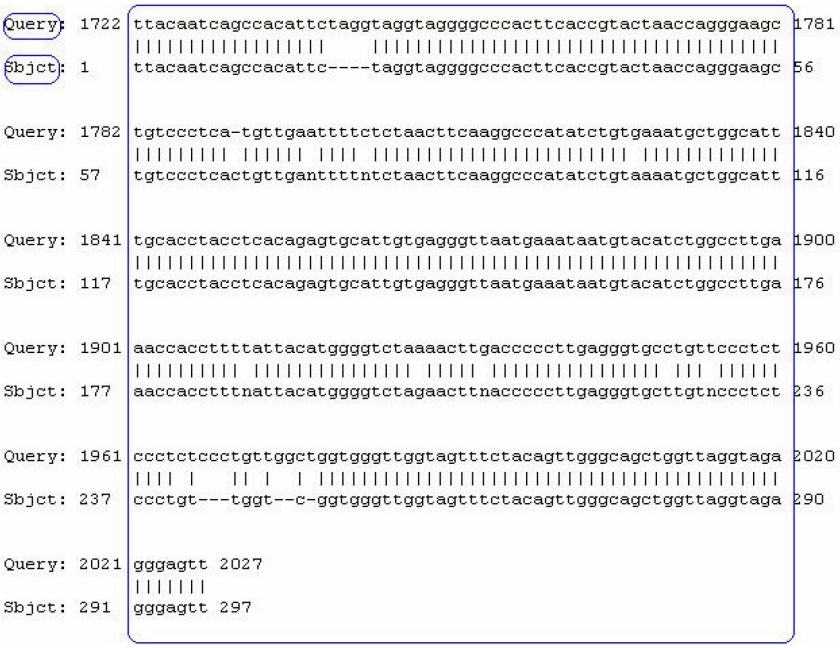


FIG. 4.3 – Un exemple de sortie Blast. La première ligne représente la séquence de référence du gène étudié. La seconde ligne correspond à l'EST aligné sur cette séquence. Deux nucléotides occupant la même position sur le gène sont liés par un tiret vertical. Les parties manquantes de l'EST sont remplacées par une suite de tirets horizontaux. Enfin, un nucléotide non indiqué sur l'EST est identifié par la lettre "n".

4.4 Analyse statistique

4.4.1 Formalisation du problème

Soit un gène fixé. Si les passages de l'ADN à l'ARNm, et de l'ARNm aux EST, sont des mécanismes fidèles, tout individu, cancéreux ou non, devrait porter à une position donnée le même nucléotide que celui de la séquence de référence (aux erreurs de séquençage et aux SNP près). Or en alignant les EST sur la séquence de référence, il apparaît en pratique que certains nucléotides sont touchés par des substitutions, y compris dans le cas normal.

Ainsi, en supposant que des erreurs soient possibles, il semble naturel d'introduire un modèle probabiliste pour les prendre en compte.

Raisonnons à une position i fixée de la séquence de référence du gène.

Soit $B \in \{A, T, C, G\}$ la base sur la séquence de référence à la position i. Soit p_{1B} (resp. p_{2B}) la probabilité qu'un EST cancéreux (resp. sain) ait effectivement la base B sur l'ADNc. Ces probabilités mesurent la variabilité des EST cancéreux et sains.

Le but est de tester si la variabilité des EST cancéreux est la même que celle des EST sains. Plutôt que de comparer la distribution des quatre bases dans le cas normal et le cas cancer, nous allons comparer les probabilités d'avoir la même base que celle de la base de la séquence de référence.

Cela revient à effectuer le test de comparaison de probabilités :

$$\begin{cases} H_0 : p_{1B} = p_{2B} \\ H_1 : p_{1B} \neq p_{2B} \end{cases} \tag{4.1}$$

Notons que ce test est équivalent à :

$$\begin{cases} H_0 : p_{1\bar{B}} = p_{2\bar{B}} \\ H_1 : p_{1\bar{B}} \neq p_{2\bar{B}} \end{cases} \tag{4.2}$$

où $p_{1\bar{B}}$ (resp. $p_{2\bar{B}}$) est la probabilité qu'un EST cancéreux (resp. sain) ait en position i l'une des bases différentes de B.
Par exemple, si $B = T$, alors $\bar{B} = \{A, C, G\}$ et :

$$p_{j\bar{T}} = p_{jA} + p_{jC} + p_{jG},\ j = 1, 2. \tag{4.3}$$

Problème : Le test (4.2) n'est pas réalisable en pratique. En effet, pour l'effectuer il faut lire les bases de l'ADNc, et comme expliqué précédemment, cette lecture peut être entachée d'erreurs de séquençage. De ce fait, il n'est pas possible d'obtenir des estimations "directes" de ces probabilités avec nos données.

L'impact des erreurs de séquençage sur la formulation du test statistique précédent est étudié dans la section suivante.

4.4.2 Impact de l'erreur de séquençage sur le modèle

Soit ϵ la probabilité d'avoir une erreur de séquençage à la position i fixée, *i.e* de lire une base différente de celle de la séquence de référence.

Après concertation avec les biologistes, il semble raisonnable d'introduire les hypothèses suivantes :

1. (E_1) La probabilité ϵ d'avoir une erreur de séquençage ne dépend pas de l'état sain/cancer,
2. (E_2) La probabilité ϵ d'avoir une erreur de séquençage ne dépend pas de la véritable base B à la position d'intérêt,
3. (E_3) Une base B touchée par une erreur de séquençage est indifféremment remplacée par l'une ou l'autre des trois bases de $\bar{B}$ avec la même probabilité $\frac{\epsilon}{3}$.

Soit q_{1B} (resp. q_{2B}) la probabilité de lire la base de référence $B \in \{A, T, C, G\}$ sur un EST cancéreux (resp. sain).
Supposons que $B = T$ et déterminons $q_{jT},\ j = 1, 2$:

$$\begin{aligned} q_{jT} &= P(\text{“ la base lue est T ”}) \\ &= P(\text{“ la base réelle est T et il n'y a pas eu d'erreur de séquençage”}) \\ &+ P(\text{“ la base réelle est A et il y a eu une erreur de séquençage de A vers T ”}) \\ &+ P(\text{“ la base réelle est C et il y a eu une erreur de séquençage de C vers T ”}) \\ &+ P(\text{“ la base réelle est G et il y a eu une erreur de séquençage de G vers T ”}). \end{aligned} \tag{4.4}$$

D'où sous (E1), (E2) et (E3) :

$$\begin{aligned} q_{jT} &= p_{jT}(1-\epsilon) + p_{jA}\frac{\epsilon}{3} + p_{jC}\frac{\epsilon}{3} + p_{jG}\frac{\epsilon}{3} \\ &= p_{jT}(1-\epsilon) + \frac{\epsilon}{3}(1 - p_{jT}) \\ &= (1 - \frac{4}{3}\epsilon)p_{jT} + \frac{\epsilon}{3}. \end{aligned} \tag{4.5}$$

$$\tag{4.6}$$

De façon générale, pour toute base B :

$$q_{jB} = (1 - \frac{4}{3}\epsilon)p_{jB} + \frac{\epsilon}{3}. \tag{4.7}$$

De même :

$$\boxed{q_{j\bar{B}} = (1 - \tfrac{4}{3}\epsilon)p_{j\bar{B}} + \epsilon.} \tag{4.8}$$

Enfin, sous (E_1), le test bilatéral (4.1) s'écrit de façon équivalente :

$$\begin{cases} H_0 : q_{1B} = q_{2B} \\ H_1 : q_{1B} \neq q_{2B} \end{cases} \tag{4.9}$$

ou bien

$$\begin{cases} H_0 : q_{1\bar{B}} = q_{2\bar{B}} \\ H_1 : q_{1\bar{B}} \neq q_{2\bar{B}} \end{cases} \tag{4.10}$$

Remarques :

1. Ce test peut-être réalisé sans connaître la valeur de la probabilité ϵ d'erreur de séquençage.
2. Les résultats précédents restent également valables dans le cas où la probabilité ϵ d'avoir une erreur de séquençage dépend de la position i,
3. Il n'est pas possible de vérifier statistiquement les hypothèses (E_1),(E_2) et (E_3), car si une substitution est lue sur l'EST, on ne peut savoir si elle est due à une erreur de transcription ou à une erreur de séquençage.

Ainsi, sous les hypothèses $(E_1), (E_2)$ et (E_3), il est équivalent de comparer les fréquences d'occurrences de substitution sur les ESTs (*i.e* compte tenu des erreurs de séquençage) et sur les ARNm.

Dans les sections suivantes sont présentées les différentes statistiques utilisées pour réaliser le test (4.10).

4.4.3 Tests de comparaison de deux probabilités

Test bilatéral

De par leur variabilité intrinsèque, les positions SNP ne sont pas prises en compte.

Soit B la base de la séquence de référence à une position donnée, non SNP. A cette position sont alignés sur la séquence de référence n_1 EST cancéreux et n_2 EST sains.

Soit alors le tableau de contingence :

	B	$\overline{B}$	Somme
Cancéreux	$n_1 - k_1$	k_1	n_1
Sains	$n_2 - k_2$	k_2	n_2
Somme	m_1	m_2	n

Soit K_j la variable aléatoire qui a pour réalisation le nombre k_j d'EST pour lesquels on ne lit pas la base B dans le groupe $j = 1, 2$.

k_1 (resp. k_2) est déterminé en comptant le nombre de nucléotides différents de la base B parmi les n_1 (resp. n_2) EST cancéreux (resp. sain) alignés à la position étudiée (Figure 4.4).

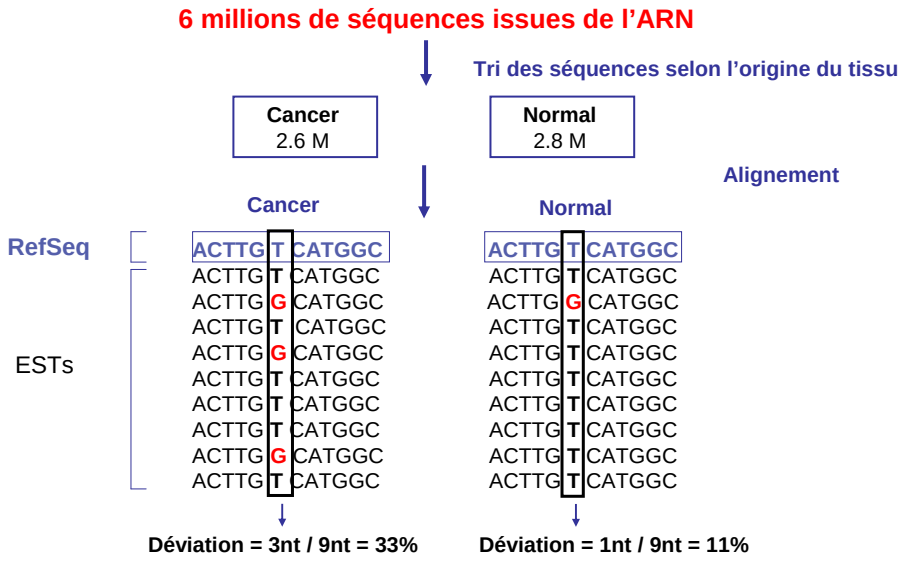


FIG. 4.4 – *Alignement des EST sur la séquence de référence et calcul des pourcentages de déviation dans les EST cancéreux et dans les EST sains.*

La proportion $\frac{k_1}{n_1}$ (resp. $\frac{k_2}{n_2}$) est appelée **pourcentage de déviation observé** dans le tissu cancéreux (resp. sain). La variable aléatoire $\frac{K_j}{n_j}$ (de réalisation $\frac{k_j}{n_j}$) est l'estimateur efficace de $q_{j\bar{B}}$.
On introduit alors la variable aléatoire

$$\hat{P} = \frac{K_1 + K_2}{n_1 + n_2}, \tag{4.11}$$

de réalisation

$$\hat{p} = \frac{k_1 + k_2}{n_1 + n_2}. \tag{4.12}$$

De plus, on pose

$$T = \frac{\frac{K_1}{n_1} - \frac{K_2}{n_2}}{\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \tag{4.13}$$

de réalisation

$$t = \frac{\frac{k_1}{n_1} - \frac{k_2}{n_2}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}. \tag{4.14}$$

Sous les conditions

$$n > 70,\ \frac{n_i m_j}{n} > 5,\ i = 1, 2,\ j = 1, 2, \tag{4.15}$$

et lorsque H_0 est vraie, la loi asymptotique de T est la loi $N(0, 1)$:

$$T \approx N(0, 1). \tag{4.16}$$

Soit également $u(\alpha)$ le quantile d'ordre $1 - \alpha$ de la loi $N(0, 1)$, *i.e* le réel tel que

$$\alpha = P(G > u(\alpha)), \tag{4.17}$$

où $G \sim N(0, 1)$.
On contruit alors la règle de décision du test (4.10) :

Règle de décision :
Si $|t| \geq u(\frac{\alpha}{2})$*, on rejette* H_0 *au seuil* α *; sinon on ne rejette pas* H_0.

Le test (4.10) est alors réalisé à toute position non SNP pour l'ensemble des 17 gènes, lorsque les conditions (4.15) sont vérifiées. Il permet de conclure à une différence significative de variabilité entre le cas cancer et le cas normal lorsque l'hypothèse H_0 est rejetée.

Exemple 1 : Pour le gène *TPT1*, à la position 435, la base de la séquence de référence est T. Soit alors le tableau de contingence :

	T	A, C, G	Somme
Cancéreux	2402	42	2444
Sains	2102	8	2110
Somme	4504	50	4554

Au seuil $\alpha = 5\%$, on a $u(\frac{\alpha}{2}) = u(0,025) = 1,96$. On trouve $t = |t| = 4,33 > 1,96$, donc l'hypothèse $H_0 : q_{1\bar{B}} = q_{2\bar{B}}$ est rejetée au seuil 5%.

Tests unilatéraux

Si le test (4.10) permet de tester si les fréquences de survenue des substitutions sont significativement différentes dans les deux types de tissus, il ne donne pas le sens de la différence. Il est en effet particulièrement important de déterminer si l'hétérogénéité observée est significativement augmentée dans les ARNm cancéreux ou dans les ARNm sains.

Ainsi, puisque l'hypothèse $q_{1\bar{B}} \neq q_{2\bar{B}}$ est équivalente à $q_{1\bar{B}} > q_{2\bar{B}}$ ou $q_{1\bar{B}} < q_{2\bar{B}}$, il semble naturel d'introduire les tests unilatéraux suivants (on conserve les notations introduites précédemment) :

Test unilatéral à droite
Le test

$$\begin{cases} H_0 : q_{1\bar{B}} \leq q_{2\bar{B}} \\ H_1 : q_{1\bar{B}} > q_{2\bar{B}} \end{cases} \tag{4.18}$$

qui est équivalent à

$$\begin{cases} H_0 : p_{1\bar{B}} \leq p_{2\bar{B}} \\ H_1 : p_{1\bar{B}} > p_{2\bar{B}}, \end{cases} \tag{4.19}$$

a la même la règle de décision que le test

$$\begin{cases} H_0 : q_{1\bar{B}} = q_{2\bar{B}} \\ H_1 : q_{1\bar{B}} > q_{2\bar{B}}. \end{cases} \tag{4.20}$$

Règle de décision :

Si $t \geq u(\alpha)$, on rejette H_0 au seuil α ; sinon on ne rejette pas H_0.

Si l'hypothèse H_0 est rejetée, ce test permet de conclure à une variabilité plus forte dans le cas cancer que dans le cas normal. Une telle position sera dite de type $(C > N)$.

Test unilatéral à gauche

Considérons maintenant le test suivant

$$\begin{cases} H_0 : q_{1\bar{B}} \geq q_{2\bar{B}} \\ H_1 : q_{1\bar{B}} < q_{2\bar{B}}. \end{cases} \tag{4.21}$$

Il s'écrit de manière équivalente

$$\begin{cases} H_0 : p_{1\bar{B}} \geq p_{2\bar{B}} \\ H_1 : p_{1\bar{B}} < p_{2\bar{B}}, \end{cases} \tag{4.22}$$

qui a la même règle de décision que le test

$$\begin{cases} H_0 : q_{1\bar{B}} = q_{2\bar{B}} \\ H_1 : q_{1\bar{B}} < q_{2\bar{B}}. \end{cases} \tag{4.23}$$

Règle de décision :

Si $t \leq -u(\alpha)$, on rejette H_0 au seuil α ; sinon on ne rejette pas H_0.

Rejeter l'hypothèse H_0 à une position signifie que la probabilité d'occurence d'une substitution est significativement plus grande dans les ARNm sains que dans les ARNm cancéreux. Une telle position sera dite $(N > C)$.

Exemple 2 : Reprenons l'exemple précédent. On réalise maintenant le test (4.18) au seuil $\alpha = 5\%$. On trouve $t = 4,33 > u(5\%) = 1,645$, donc on rejette l'hypothèse $H_0 : q_{1\bar{B}} \leq q_{2\bar{B}}$ au seuil 5%.

4.4.4 Les p-values

Introduisons la notion de p-value pour les tests (4.10), (4.20) et (4.23).

Soit T la statistique de test de réalisation observée t.

Test bilatéral

On appelle **p-value** (ou **probabilité critique**) du test (4.10) la probabilité :

$$p = P(|T| \geq |t| \text{ sous } H_0). \tag{4.24}$$

Remarquons que p dépend de t et est par conséquent la réalisation de la variable aléatoire $P = G(|T|)$, où $G(u) = P(|T| \geq u \text{ sous } H_0)$. Sous H_0, $P \sim \mathcal{U}_{[0,1]}$.

La règle de décision du test (4.10) peut s'écrire sous la forme équivalente :

Règle de décision :

Si $p \leq \alpha$, on rejette H_0 au seuil α ; sinon, on ne rejette pas H_0.

Exemple 3 : Dans l'exemple 2, nous avions calculé pour le test bilatéral $t = 4,33$. Calculons maintenant p.

Sous H_0, si les conditions (4.15) sont vérifiées, alors T suit asymptotiquement une loi normale centrée réduite :

$$\begin{aligned} p &= P(|T| \geq 4,33 \text{ sous } H_0) \\ &= 1,49.10^{-5} \end{aligned} \tag{4.25}$$

Puisque $p \leq 0,05$, on rejette l'hypothèse H_0 au seuil 5%.

Test unilatéral à droite

On appelle p-value du test (4.20) la probabilité :

$$p = P(T \geq t \text{ sous } H_0). \tag{4.26}$$

On peut alors comme précédemment écrire la règle de décision du test (4.20) sous la forme équivalente :

Règle de décision :
Si $p \leq \alpha$, on rejette H_0 au seuil α ; sinon, on ne rejette pas H_0.

Test unilatéral à gauche

On appelle p-value du test (4.23) la probabilité :

$$p = P(T \leq t \text{ sous } H_0). \tag{4.27}$$

La règle de décision du test (4.23) peut alors s'écrire :

Règle de décision :
Si $p \leq \alpha$, on rejette H_0 au seuil α ; sinon, on ne rejette pas H_0.

Nous venons de définir les statistiques utilisées pour réaliser les tests (4.10), (4.20) et (4.23). Toutefois, les tests précédents sont tous asymptotiques et leur mise en place nécessite que certaines conditions soient vérifiées (4.15). Or on observe que ce n'est pas toujours le cas. En effet, si les gènes étudiés ont en général suffisamment d'EST alignés et vérifient par conséquent la condition $n > 70$, la contrainte sur les effectifs théoriques n'est pas toujours respectée pour certaines positions. De ce fait, les trois tests précédents deviennent inutilisables. Dans ce cas, une alternative est proposée en utilisant le test exact de Fisher, qui ne souffre d'aucune contrainte de validité et qui est présenté dans la section suivante.

4.4.5 Test exact de Fisher

On rappelle ci-dessous le principe du test exact de Fisher, qui sera utilisé dans la suite. Une approche numérique est proposée en fin de section, permettant de faciliter le calcul des $p-$values.

Préliminaire : tableaux 2×2 de marges fixées.

Soit k, n_1 et n trois nombres entiers fixés tels que :

$$0 \leq k \leq n \ , 0 \leq n_1 \leq n. \tag{4.28}$$

Considérons l'ensemble des tableaux à coefficients entiers

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{4.29}$$

tels que

$$\begin{aligned} a + b + c + d &= n \\ b + d &= k \\ a + b &= n_1. \end{aligned} \tag{4.30}$$

k, n_1 et n sont des marges du tableau A.

D'après les relations précédentes, on a

$$\begin{aligned} b &= n_1 - a \\ c &= n - k - a \\ d &= k - n_1 + a. \end{aligned}$$

Puisque k, n_1 et n sont fixés, les coefficients du tableau A ne dépendent que de a. Par ailleurs, comme b et c sont positifs ou nuls, on a $a \leq \min(n - k, n_1)$; de même puisque a et d sont positifs, on a $a \geq \max(0, n_1 - k)$.

Pour k et n_1 compris entre 0 et n, on note alors $\mathcal{M}(n_1, k, n)$ l'ensemble des tableaux A pour lesquels le coefficient a vérifie l'inégalité :

$$\max(0, n_1 - k) \leq a \leq \min(n - k, n_1). \tag{4.31}$$

Reprenons maintenant les notations du paragraphe 4.4.3 et considérons le tableau de contingence :

	B	$\overline{B}$	Somme
Cancéreux	$n_1 - k_1$	k_1	n_1
Sains	$n_2 - k_2$	k_2	n_2
Somme	$n_1 + n_2 - (k_1 + k_2)$	k	n

où $k = k_1 + k_2$ et $n = n_1 + n_2$.

On note comme précédemment $q_{1\bar{B}}$ (resp. $q_{2\bar{B}}$) la probabilité que la base lue ne soit pas la base de la séquence de référence chez un cancéreux (resp. sain).

Test bilatéral

On se propose de réaliser le test exact de Fisher bilatéral suivant [24] :

$$\begin{cases} H_0 : q_{1\bar{B}} = q_{2\bar{B}} \\ H_1 : q_{1\bar{B}} \neq q_{2\bar{B}}. \end{cases} \tag{4.32}$$

Loi du tableau sous l'hypothèse H_0

Remarquons que n_1 et n_2 sont des nombres fixes ; en revanche, k_1 et k_2 sont les réalisations respectives des variables aléatoires indépendantes K_1 et K_2.

On a :
- $K_1 \sim \mathcal{B}(n_1, q_{1\bar{B}})$,
- $K_2 \sim \mathcal{B}(n_2, q_{2\bar{B}})$,
- $n_1 - K_1 \sim \mathcal{B}(n_1, 1 - q_{1\bar{B}}) = \mathcal{B}(n_1, q_{1B})$,
- $n_2 - K_2 \sim \mathcal{B}(n_2, 1 - q_{2\bar{B}}) = \mathcal{B}(n_2, q_{2B})$.

Soit l_1 et l_2 deux entiers vérifiant $0 \leq l_1 \leq n_1$, $0 \leq l_2 \leq n_2$ et $k = l_1 + l_2$. On a d'après l'indépendance de K_1 et K_2 :

$$\begin{aligned} P(K_1 = l_1, K_2 = l_2) &= P(K_1 = l_1)P(K_2 = l_2) \\ &= C_{n_1}^{l_1} q_{1\bar{B}}^{l_1} (1 - q_{1\bar{B}})^{n_1 - l_1} C_{n_2}^{l_2} q_{2\bar{B}}^{l_2} (1 - q_{2\bar{B}})^{n_2 - l_2} \end{aligned} \tag{4.33}$$

Sous H_0, $q_{1\bar{B}} = q_{2\bar{B}} = q$.

Dans ce cas, $K = (K_1 + K_2) \sim \mathcal{B}(n, q)$ et $P(K = k) = C_n^k q^k (1 - q)^{n-k}$.

On a alors :

$$\begin{aligned} P(K_1 = l_1, K_2 = l_2 \mid K = k) &= \frac{C_{n_1}^{l_1} q^{l_1} (1-q)^{n_1 - l_1} C_{n_2}^{l_2} q^{l_2} (1-q)^{n_2 - l_2}}{C_n^k q^k (1-q)^{n-k}} \\ &= \frac{C_{n_1}^{l_1} C_{n_2}^{l_2} q^k (1-q)^{n-k}}{C_n^k q^k (1-q)^{n-k}} \\ &= \frac{C_{n_1}^{l_1} C_{n_2}^{l_2}}{C_n^k} \\ &= \frac{\frac{n_1!}{l_1!\ (n_1 - l_1)!} \frac{n_2!}{l_2!\ (n_2 - l_2)!}}{\frac{n!}{k!\ (n-k)!}} \\ &= \frac{n_1!\ n_2!\ k!\ (n-k)!}{l_1!\ l_2!\ (n_1 - l_1)!\ (n_2 - l_2)!\ n!}. \end{aligned} \tag{4.34}$$

On note alors

$$\pi_{l_1, l_2, k} = \frac{n_1!\ n_2!\ k!\ (n-k)!}{l_1!\ l_2!\ (n_1 - l_1)!\ (n_2 - l_2)!\ n!}. \tag{4.35}$$

En conclusion, sous H_0 :

$$P(K_1 = l_1, K_2 = l_2 \mid K = k) = \pi_{l_1, l_2, k}. \tag{4.36}$$

Pour k fixé, $\pi_{l_1, l_2, k}$ ne dépend que de l_1. On montre que la suite $l_1 \mapsto \pi_{l_1, l_2, k}$ croît au sens large pour $l_1 \leq \frac{(k+1)(n_1+1)}{n+2} - 1$ et décroît strictement après.

Règle de décision

Notons π^0 la probabilité $\pi_{k_1,k_2,k}$ correspondant au tableau de contingence observé, où k_1, k_2 et k sont les réalisations de K_1, K_2 et K.

Pour chaque tableau de $\mathcal{M}(n_1, k, n)$, on calcule la probabilité $\pi_{l_1,l_2,k}$ correspondante ; on considère ensuite la somme p des probabilités $\pi_{l_1,l_2,k}$ qui sont inférieures ou égales à π^0.

Règle de décision :

Si $p \leq \alpha$, on rejette H_0 ; sinon on ne rejette pas H_0.

Tests unilatéraux

Soit le test unilatéral :

$$\begin{cases} H_0 : q_{1\bar{B}} \leq q_{2\bar{B}} \\ H_1 : q_{1\bar{B}} > q_{2\bar{B}}. \end{cases} \tag{4.37}$$

On considère la restriction de $\mathcal{M}(n_1, k, n)$ aux tableaux dont le coefficient b est supérieur ou égal à k_1. On calcule la probabilité $\pi_{k_1,k_2,k}$ pour chacun de ces tableaux ; on appelle p la somme de ces probabilités.

Règle de décision :

Si $p \leq \alpha$, on rejette H_0 ; sinon on ne rejette pas H_0.

De même soit le test unilatéral :

$$\begin{cases} H_0 : q_{1\bar{B}} \geq q_{2\bar{B}} \\ H_1 : q_{1\bar{B}} < q_{2\bar{B}}. \end{cases} \tag{4.38}$$

On considère la restriction de $\mathcal{M}(n_1, k, n)$ aux tableaux dont le coefficient b est inférieur ou égal à k_1. On calcule la probabilité $\pi_{k_1,k_2,k}$ pour chacun de ces tableaux ; on appelle p la somme de ces probabilités.

Règle de décision :

Si $p \leq \alpha$, on rejette H_0 ; sinon on ne rejette pas H_0.

Calculs numériques

Dans la pratique, les factorielles intervenant dans l'expression de $\pi_{l_1,l_2,k}$ peuvent devenir très grandes à mesure que les effectifs augmentent, ce qui pose des problèmes informatiques. La formule (4.35) ne peut donc pas être utilisée directement. Il faut trouver une manière d'arranger les facteurs de l'expression pour calculer ce nombre.

Remarquons que le numérateur et le dénominateur de $\pi_{l_1,l_2,k}$ sont chacun constitués de $2n$ facteurs. L'idée est de construire un algorithme itératif qui à chaque pas divise le plus grand facteur du numérateur par le plus grand facteur du dénominateur. On multiplie ensuite le résultat obtenu par le résultat du pas précédent et on passe au pas suivant en décrémentant les deux maxima trouvés. L'algorithme s'arrête quand tous les facteurs ont été introduits. Le code peut être écrit de la manière suivante :

```
function res=proba_fisher(l1,n1,l2,n2)
  réel res=1;
  k=l1+l2;
  n=n1+n2;
```

```
  tableau t1=[n1,n2,k,n-k];
  tableau t2=[l1,l2,n1-l1,n2-l2,n];

  pour i=1 à 2n faire
    ind1= indice du max de t1;
    ind2= indice du max de t2;
    res=res*t1(ind1)/t2(ind2);
    t1(ind1)=t1(ind1)-1;
    t2(ind2)=t2(ind2)-1;
  fin;

fin;
```

Ainsi, grâce au test exact de Fisher, il est possible de réaliser les tests (4.10), (4.20) et (4.23) à toutes les positions du gène, y compris celles où les conditions (4.15) ne sont pas vérifiées. Nous allons estimer le nombre moyen de tests ayant conduit à rejeter H_0 alors que H_0 est vraie (faux positifs), *i.e* de tests ayant conclu par erreur à une différence significative entre les fréquences de substitution dans les EST sains et dans les EST cancéreux, en utilisant le LBE présenté dans le chapitre précédent.

4.5 Résultats

Pour chaque gène sont donnés sa longueur et son nombre de SNP dans le Tableau 4.1.

Gènes	# de positions	# de SNP
ALB	2215	33
ALDOA	2303	31
ATP5A1	1945	14
CALM2	1128	32
ENO1	1812	36
FTH1	1228	33
FTL	870	59
GAPDH	1310	43
HSPA8	2260	21
LDHA	1661	12
RPL7A	890	8
RPS4X	955	27
RPS6	829	19
TMSB4X	627	21
TPI1	1220	14
TPT1	830	36
VIM	1847	28
Ensemble des gènes	23930	467

TAB. 4.1 – Longueur de la séquence de référence de chacun des 17 gènes étudiés et nombre de SNP retirés

Rappelons que dans cette étude les positions SNP ne sont pas traitées et donc retirées des données.

4.5.1 Tests de comparaison de deux probabilités

Test bilatéral

Si les conditions (4.15) sont vérifiées, le test bilatéral de comparaison de deux probabilités (4.10) est réalisé au seuil $\alpha = 5\%$ sur les 23463 positions non SNP restantes de chacun des 17 gènes (lorsque ces conditions ne sont pas vérifiées, on réalise le test exact de Fisher, cf Section 4.5.3) ;

On indique dans le Tableau 4.2 pour chaque gène :

1. le nombre de tests réalisés, *i.e* le nombre de positions du gène pour lesquelles les conditions (4.15) sont vérifiées,
2. le nombre de tests positifs,
3. le pourcentage de tests positifs,
4. la réalisation de l'estimateur avec biais positif du nombre moyen de faux positifs, notée LBE.

Gènes	# de tests réalisés	# de tests positifs	% de tests positifs	LBE
ALB	194	79	40.72	4
ALDOA	336	79	23.51	11
ATP5A1	238	104	43.70	4
CALM2	374	80	21.39	13
ENO1	614	152	24.76	20
FTH1	592	243	41.05	15
FTL	678	150	22.12	24
GAPDH	812	315	38.79	21
HSPA8	513	130	25.34	16
LDHA	181	51	28.18	4
RPL7A	337	106	31.45	9
RPS4X	371	120	32.35	10
RPS6	432	92	21.30	15
TMSB4X	352	118	33.52	10
TPI1	379	106	27.97	10
TPT1	489	123	25.15	15
VIM	752	276	36.70	19
Ensemble des gènes	7644	2324	30.40	220

TAB. 4.2 – Résultats du test bilatéral de comparaison de probabilités pour chacun des 17 gènes.

La dernière ligne du Tableau 4.2 montre que l'hypothèse H_0 a été rejetée pour 30,40% des positions testées. Autrement dit, on conclut pour ces positions à une différence significative entre les probabilités de survenue d'une substitution dans les ARNm sains et dans les ARNm cancéreux. Par ailleurs, le nombre moyen de faux positifs est estimé par le LBE à 220. Le nombre moyen de vrais positifs peut donc être estimé à 2324-220=2104.

Tests unilatéraux

On réalise ensuite au seuil 5% les tests (4.18) et (4.21), sous les conditions (4.15). Les résultats sont présentés dans les Tableaux 4.3 et 4.4.

Test unilatéral à droite

Ce test permet de détecter une augmentation significative des substitutions dans les ARNm cancéreux par rapport aux ARNm sains.

Gènes	# de tests réalisés	# de tests positifs	% de tests positifs	LBE
ALB	194	38	19.59	10
ALDOA	336	81	24.11	11
ATP5A1	238	123	51.68	3
CALM2	374	69	18.45	16
ENO1	614	186	30.29	18
FTH1	592	226	38.18	20
FTL	678	118	17.40	31
GAPDH	812	311	38.30	23
HSPA8	513	163	31.77	13
LDHA	181	70	38.67	4
RPL7A	337	103	30.56	11
RPS4X	371	124	33.42	11
RPS6	432	86	19.91	17
TMSB4X	352	70	19.89	18
TPI1	379	99	26.12	14
TPT1	489	145	29.65	15
VIM	752	269	35.77	24
Ensemble des gènes	7644	2281	29.84	259

TAB. 4.3 – Résultats du test unilatéral à droite de comparaison de probabilités pour chacun des 17 gènes.

Sur l'ensemble des 17 gènes étudiés, 2281 tests sont déclarés positifs, pour un nombre moyen de faux positifs estimé seulement à 259. Autrement dit, les positions correspondantes de la séquence d'ARNm des gènes seraient sujettes à une plus forte hétérogénéité dans les tissus cancéreux que dans les tissus sains.

Test unilatéral à gauche

Gènes	# de tests réalisés	# de tests positifs	% de tests positifs	LBE
ALB	194	56	28.87	8
ALDOA	336	35	10.42	21
ATP5A1	238	6	2.52	20
CALM2	374	40	10.70	21
ENO1	614	31	5.05	42
FTH1	592	57	9.63	38
FTL	678	86	12.68	36
GAPDH	812	61	7.51	57
HSPA8	513	18	3.51	37
LDHA	181	10	5.52	13
RPL7A	337	33	9.79	22
RPS4X	371	27	7.28	25
RPS6	432	40	9.26	25
TMSB4X	352	74	21.02	17
TPI1	379	47	12.40	23
TPT1	489	26	5.32	33
VIM	752	78	10.37	50
Ensemble des gènes	7644	725	9.48	488

TAB. 4.4 – Résultats du test unilatéral à gauche de comparaison de probabilités pour chacun des 17 gènes.

Aux positions des gènes pour lesquelles le test est positif, on conclut à une plus grande probabilité d'erreur dans les tissus sains que dans les tissus cancéreux. Toutefois, on constate que dans de nombreux cas (*ATP5A1, ENO1, ...*), l'estimation du nombre moyen de faux positifs est proche, voire supérieure, au nombre de positifs. Pour ces gènes, il semble que le phénomène de multiplication des erreurs dans les tissus sains ne soit dû qu'au hasard. Pourtant, certains gènes comme *ALB* ou *TMSB4X* ont un nombre de tests positifs beaucoup plus élevé que le LBE.

4.5.2 Etude comparative du test de comparaison de probabilités et du test exact de Fisher

De par les contraintes (4.15), seulement $\frac{7644}{23463} \approx 33\%$ des positions ont pu être testées avec le test de comparaison de probabilités.

Nous allons dans un premier temps effectuer le test exact de Fisher uniquement à ces positions pour en comparer les résultats avec le test de comparaison de probabilités.

Dans les Tableaux 4.5, 4.6 et 4.7 sont présentés respectivement pour les tests (4.10), (4.18) et (4.21), le nombre de positions positives pour le test de comparaison de probabilités, le nombre de positions positives pour le test exact de Fisher et le nombre de positions positives pour les deux tests en même temps (dénommé "intersection").

Pour chacun des tests (4.10), (4.18) et (4.21), on constate qu'une position déclarée positive par le test exact de Fisher l'est également par le test de comparaison de probabilités. Le test de comparaison de probabilités rejette donc plus souvent H_0 que le test de Fisher, fait au même seuil. Ceci traduit le fait que le test de comparaison de probabilités est plus puissant que le test exact de Fisher. Autrement dit, le test de comparaison de probabilités "détecterait" mieux les différences entre les EST cancéreux et les EST sains que le test exact de Fisher.

Gènes	Test de probabilités	Test exact de Fisher	Intersection
ALB	79	77	77
ALDOA	79	76	76
ATP5A1	104	93	93
CALM2	80	74	74
ENO1	152	144	144
FTH1	243	233	233
FTL	150	143	143
GAPDH	315	302	302
HSPA8	130	125	125
LDHA	51	48	48
RPL7A	106	101	101
RPS4X	120	109	109
RPS6	92	83	83
TMSB4X	118	108	108
TPI1	106	98	98
TPT1	123	113	113
VIM	276	255	255
Ensemble des gènes	2324	2182	2182

TAB. 4.5 – Nombre de tests bilatéraux de comparaison de probabilités et de Fisher exact positifs et nombre de tests positifs par les deux méthodes

Gènes	Test de probabilités	Test exact de Fisher	Intersection
ALB	38	35	35
ALDOA	81	70	70
ATP5A1	123	113	113
CALM2	69	57	57
ENO1	186	167	167
FTH1	226	219	219
FTL	118	103	103
GAPDH	311	294	294
HSPA8	163	137	137
LDHA	70	54	54
RPL7A	103	92	92
RPS4X	124	113	113
RPS6	86	72	72
TMSB4X	70	62	62
TPI1	99	81	81
TPT1	145	124	124
VIM	269	234	234
Ensemble des gènes	2281	2027	2027

TAB. 4.6 – Nombre de tests unilatéraux à droite de comparaison de probabilités et de Fisher exact positifs et nombre de tests positifs par les deux méthodes

Par conséquent, pour tester une position d'un gène, le test de comparaison de probabilités est réalisé si les conditions (4.15) sont vérifiées. Sinon le test exact de Fisher est mis en place. Les résultats sont présentés dans la section suivante.

Gènes	Test de probabilités	Test exact de Fisher	Intersection
ALB	56	48	48
ALDOA	35	25	25
ATP5A1	6	3	3
CALM2	40	34	34
ENO1	31	20	20
FTH1	57	44	44
FTL	86	71	71
GAPDH	61	50	50
HSPA8	18	13	13
LDHA	10	9	9
RPL7A	33	28	28
RPS4X	27	18	18
RPS6	40	35	35
TMSB4X	74	64	64
TPI1	47	37	37
TPT1	26	18	18
VIM	78	62	62
Ensemble des gènes	725	579	579

TAB. 4.7 – Nombre de tests unilatéraux à gauche de comparaison de probabilités et de Fisher exact positifs et nombre de tests positifs par les deux méthodes

4.5.3 Résultats du test de comparaison de probabilités ou du test exact de Fisher

On donne les résultats pour toutes les positions testées, avec le test de comparaison de probabilités si c'est possible, sinon avec le test exact de Fisher. La comparaison des probabilités de survenue d'une substitution sur une EST cancéreuse et sur une EST saine est donc effectuée à chaque position non SNP d'un gène.

Test bilatéral

Le nombre de tests positifs passe de 2324 à 3112. Ce gain reste relativement faible par rapport aux 23480 positions testables. De plus, l'estimation du nombre moyen de faux-positifs s'élève désormais à 1221, contre 220 en utilisant uniquement le test de comparaison de probabilités classique. Ceci semble signifier que le peu de tests positifs apportés par le test exact de Fisher pourraient être des faux positifs. Le faible nombre de tests positifs peut s'expliquer par le fait que le test de Fisher est appliqué à des positions pour lesquelles relativement peu d'EST sont alignés ; de ce fait, le test est alors peu puissant et a tendance à ne pas rejeter H_0 lorsque H_0 est fausse. Par ailleurs, on constate que les tests pour lesquels l'hypothèse H_0 n'est pas rejetée présentent des p-values relativement élevées ; puisque le LBE s'écrit à un facteur près comme la somme des p-values, ceci explique pourquoi l'estimation du nombre moyen de faux positifs est aussi élevée.

Gènes	# de tests réalisés	# de tests positifs	% de tests positifs	LBE
ALB	2182	230	10.54	126
ALDOA	2272	131	5.77	164
ATP5A1	1931	175	9.06	114
CALM2	1096	99	9.03	60
ENO1	1776	213	11.99	87
FTH1	1195	271	22.68	58
FTL	811	152	18.74	32
GAPDH	1267	365	28.81	45
HSPA8	2239	200	8.93	121
LDHA	1649	120	7.28	94
RPL7A	882	146	16.55	40
RPS4X	928	143	15.41	45
RPS6	810	115	14.20	38
TMSB4X	606	126	20.79	25
TPI1	1206	151	12.52	58
TPT1	794	133	16.75	33
VIM	1819	342	18.80	81
Ens des gènes	23463	3112	13.26	1221

TAB. 4.8 – Résultats du test bilatéral pour les 17 gènes.

Test unilatéraux

Test unilatéral à droite

Gènes	# de tests réalisés	# de tests positifs	% de tests positifs	LBE
ALB	2182	161	7.38	145
ALDOA	2272	108	4.75	165
ATP5A1	1931	163	8.44	118
CALM2	1096	86	7.85	62
ENO1	1776	216	12.16	92
FTH1	1195	234	19.58	66
FTL	811	121	14.92	39
GAPDH	1267	312	24.63	56
HSPA8	2239	214	9.56	117
LDHA	1649	102	6.19	99
RPL7A	882	111	12.59	50
RPS4X	928	132	14.22	50
RPS6	810	96	11.85	42
TMSB4X	606	73	12.05	36
TPI1	1206	131	10.86	67
TPT1	794	153	19.27	33
VIM	1819	312	17.15	88
Ens des gènes	23463	2725	11.61	1325

TAB. 4.9 – Résultats du test unilatéral à droite pour les 17 gènes.

L'estimation du nombre moyen de vrais positifs s'élève à 2725-1325=1400. Les conclusions sont les mêmes que pour le test bilatéral.

Test unilatéral à gauche

Gènes	# de tests réalisés	# de tests positifs	% de tests positifs	LBE
ALB	2182	114	5.22	141
ALDOA	2272	76	3.35	181
ATP5A1	1931	57	2.95	140
CALM2	1096	47	4.29	71
ENO1	1776	80	4.50	115
FTH1	1195	83	6.95	84
FTL	811	88	10.85	44
GAPDH	1267	116	9.16	81
HSPA8	2239	58	2.59	156
LDHA	1649	68	4.12	114
RPL7A	882	74	8.39	50
RPS4X	928	48	5.17	60
RPS6	810	59	7.28	48
TMSB4X	606	86	14.19	30
TPI1	1206	76	6.30	77
TPT1	794	34	4.28	54
VIM	1819	121	6.65	123
Ens des gènes	23463	1285	5.48	1569

TAB. 4.10 – Résultats du test unilatéral à gauche pour les 17 gènes.

Le LBE est supérieur au nombre de test positifs. De ce fait, il semble que pour la majorité des gènes étudiés, le phénomène de multiplication des erreurs dans les tissus sains ne soit dû qu'au hasard.

4.6 Conclusions

La première conclusion de cette étude est que la probabilité de survenue d'une substitution dans les EST est plus forte dans les tissus cancéreux que dans les tissus sains pour de nombreuses positions des gènes de notre étude. Ainsi, dans le cas du test de comparaison de probabilités classique, l'estimation du nombre moyen de vrais positifs s'élève à 2281-259=2022. Or, dans notre modèle, l'erreur de séquençage ne peut contribuer à cette différence de variabilité. De ce fait, puisque le passage de l'ARNm aux EST ne peut expliquer la différence des probabilités de substitution, la fidélité du mécanisme de transcription se trouve ici remise en cause. En effet, les erreurs observées sur les EST seraient ainsi dues, aux erreurs de séquençage près, à une erreur de transcription des nucléotides de l'ADN. De plus, les cas d'infidélité de la transcription se produisent en général avec une fréquence significativement plus élevée dans les tissus cancéreux que dans les tissus sains.

Cependant, il semble que pour certaines positions des gènes étudiés, la probabilité de survenue d'une substitution puisse être significativement plus grande dans les EST sains que dans les EST cancéreux. Bien que dans le cas du test de comparaison de probabilités unilatéral à gauche, le nombre de tests positifs reste globalement proche du LBE, certains gènes semblent se démarquer, comme *ALB* et *TMSB4X*. Ce phénomène reste cependant marginal.

Il semble que l'infidélité de transcription soit un phénomène pouvant se produire dans les tissus sains et qui serait amplifié dans les tissus cancéreux. Si certaines études avaient déja fait cas par le passé d'erreurs de la transcription *in vivo* chez le rat [25] et le chien [26] et *in vitro* chez

l'homme [27, 28], les conclusions apportées ici montrent que ce phénomène semble également exister *in vivo* chez l'homme, et ce pour un grand nombre de gènes.

Toutefois, notre étude connaît quelques limites :

1. Les seuls événements pris en compte dans l'analyse statistique sont des substitutions de nucléotides dans l'ARNm. Il serait important de comparer par la suite les fréquences de survenue d'insertions ou de délétions de nucléotides dans des ARNm cancéreux et dans des ARNm normaux.
2. Les substitutions observées à des positions différentes d'un ARNm sont considérées comme des événements indépendants. Il faudrait tester cette hypothèse et proposer un modèle tenant compte de la dépendance si celle-ci était avérée.
3. Notre étude a été restreinte à 17 gènes pour lesquels de nombreux EST étaient disponibles. Il faudrait étendre cette étude à toute position de l'ensemble des gènes du génome humain.
4. Les EST sont considérés indépendamment du type de cancer. Un jeu de données de taille plus importante serait nécessaire pour construire une étude unique par type de cancer.
5. Trois hypothèses ont été faites sur l'erreur de séquençage pour construire notre modèle. La première est que la probabilité de survenue d'une erreur de séquençage ne dépend pas de l'état sain/cancer. En ce qui concerne l'erreur de lecture, ceci peut être argumenté par le fait que les banques de données EST sont construites de la même manière, que les ARNm aient été prélévés dans des tissus sains ou dans des tissus cancéreux. Même s'il semble peu probable qu'un séquenceur fasse plus d'erreurs sur un A qu'un T par exemple, ou qu'un G soit plus souvent remplacé par un C, les deux autres hypothèses peuvent être discutées. Le problème est qu'il n'est pas possible de les tester ; en effet, une erreur observée sur un EST ne peut être directement imputée ni à l'erreur de séquençage, ni à l'infidélité de transcription.

Une première vérification biologique de l'existence de l'infidélité de transcription est proposée dans [1], prouvant qu'elle n'est pas un artefact. Toutefois, des séquençeurs de haute qualité seraient nécessaires pour vérifier biologiquement ces substitutions sur les ARNm d'un même individu sans passer par les EST.

Dans notre étude, tous les SNP connus des gènes avaient été retirés préalablement. Bien que certains SNP soient peut-être encore inconnus, il semble peu vraisemblable que l'hétérogénéité des ARNm soit uniquement due à des SNP. Certaines hypothèses sont soulevées dans [1].

Enfin, notons que l'hétérogénéité des ARN cancéreux pourrait expliquer le manque de reproductibilité de certaines études de transcriptomique comme les microarrays [18] et d'études de protéomique [29, 30]. En effet, la traduction des ARNm porteurs de substitutions conduirait à la production de protéines aberrantes dont l'existence était jusqu'alors insoupçonnée. Puisque pour certaines positions des ARNm, les substitutions apparaîssent plus fréquemment dans des tissus cancéreux que dans des tissus sains, ces protéines devraient être présentes en plus grande quantité chez un patient malade que chez un patient sain. De ce fait, mesurer la quantité de ces protéines dans le sang - ou plutôt la quantité d'anticorps dirigés contre ces protéines, car mesurer directement un nombre de protéines est très difficile - permettrait de constituer un ensemble d'éventuels biomarqueurs du cancer. Mener une étude clinique avec des patients sains et cancéreux sur lesquels ces biomarqueurs seraient mesurés, permettrait de réaliser une analyse discriminante. En cas de bons résultats, il pourrait être envisagé de créer un test clinique basé sur une prise de sang, permettant le diagnostic précoce du cancer.

En conclusion, l'étude comparée de la variabilité des ARNm sains et cancéreux semble mettre en évidence un nouveau phénomène biologique, l'infidélité de transcription. Dans un premier temps, la réalité de l'infidélité de transcription devra être prouvée biologiquement. Les travaux suivants devront être tournés vers la détermination de la cause de l'infidélité de transcription,

de ses mécanismes, de son rôle dans le développement des cancers et de ses implications en terme de diagnostic précoce.

Chapitre 5

Modélisation de la probabilité de survenue d'une infidélité de transcription sur un ARNm

5.1 Problème et données

Lors de la transcription, l'ADN est converti en ARNm. Il a été montré au chapitre précédent que certaines erreurs interviennent lors du passage à l'ARNm. Par exemple, un C sur le brin transcrit d'ADN pourrait être transcrit en A au lieu de G.

L'enzyme responsable de la transcription de l'ADN en ARNm est appelé *ARN polymérase* (Chapitre 1). Lorsqu'un nucléotide est transcrit, la polymérase "lit" également plusieurs autres nucléotides de part et d'autre de celui-ci (en particulier les 4 précédents et les 3 suivants). Cet environnement est appelé " contexte nucléotidique". Il est intéressant de déterminer si ces nucléotides ont un rôle dans la survenue d'une substitution. Notre objectif est de prédire si un nucléotide fixé du brin d'ARNm va être sujet à une infidélité de transcription, connaissant son contexte nucléotidique.

Pour cela, on dispose des séquences de référence des 17 gènes étudiés au Chapitre 4. On dispose de plus pour chaque gène des EST correspondants. Rappelons que les EST d'un gène sont des copies partielles de son filament d'ARNm. Pour un gène fixé, on se place à une position i de la séquence de référence et on note n_i le nombre d'EST qui s'alignent à cette position.

Pour chaque position i de la séquence de référence d'un gène, on introduit (Figure 5.1) :

1. le nucléotide x^i sur la séquence de référence à la position i et l'environnement ou contexte nucléotidique $\varphi(i)$ de taille $h_1 + h_2 + 1$ autour de i :

$$\varphi(i) = (x^{i-h_1}, \ldots, x^i, \ldots, x^{i+h_2}), \tag{5.1}$$

   avec $x^j \in \{A, T, C, G\}, \forall j \in \{i - h_1, \ldots, i + h_2\}$. Ainsi $\varphi(i) \in E = \{A, T, C, G\}^{h_1+h_2+1}$

2. la variable aléatoire Z_i ayant pour réalisation le nucléotide sur un EST aligné à la position i,

3. la variable aléatoire Y_i définie par :

$$Y_i = \begin{cases} 1 & \text{si } Z_i \neq x^i \ i.e \text{ s'il y a une substitution sur un EST aligné à la position } i, \\ 0 & \text{si } Z_i = x^i \ i.e \text{ s'il n'y a pas de substitution sur un EST aligné à la position } i \end{cases} \tag{5.2}$$

Supposons que $x^i = A$ et que pour k_i de ces EST, le nucléotide présent n'est pas le même que celui de la séquence de référence. On suppose que ce sont les k_i premiers pour simplifier. Les autres EST alignés ne présentent pas de substitution (Figure 5.1).

On note $z_{i,j}$ le nucléotide sur l'EST j aligné à la position i. Ainsi :

$$z_{i,j} = \begin{cases} T, C \text{ ou } G, \ 1 \leq j \leq k_i \\ A, \ k_{i+1} \leq j \leq n_i. \end{cases} \tag{5.3}$$

Les $z_{i,j}$ constituent un échantillon de Z_i. On pose alors :

$$y_{i,j} = \begin{cases} 1, \ 1 \leq j \leq k_i \\ 0, \ k_{i+1} \leq j \leq n_i. \end{cases} \tag{5.4}$$

Les $y_{i,j}$ constituent un échantillon de de Y_i.

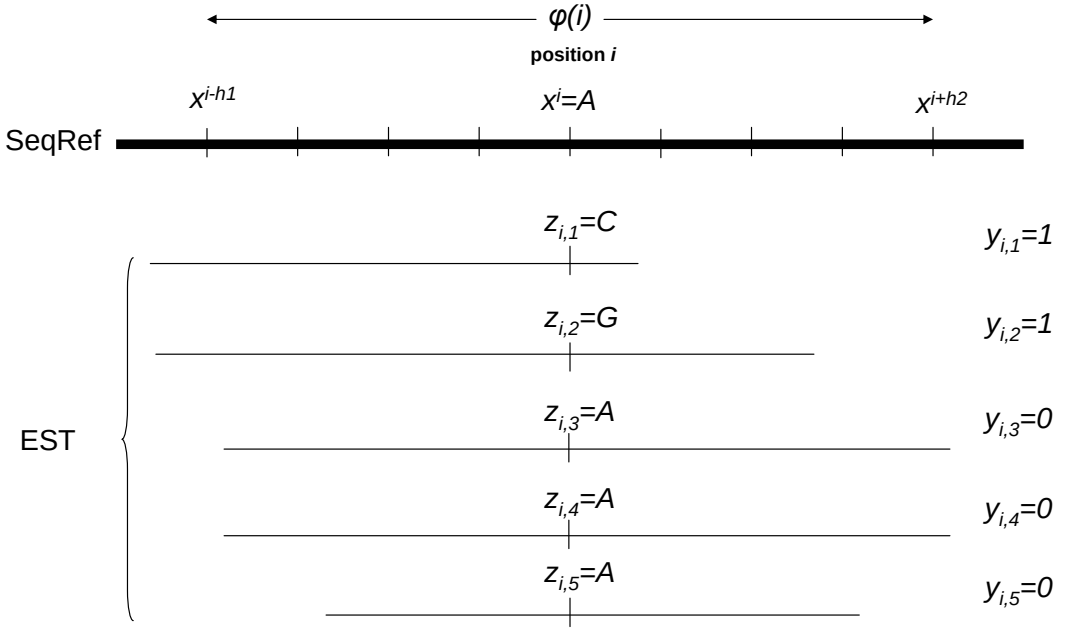


FIG. 5.1 – EST alignés à la position i

Dans la suite, notre objectif sera d'estimer la probabilité $P(Y_i = 1)$ à partir des nucléotides $x^{i-h_1}, \ldots, x^i, \ldots, x^{i+h_2}$, en utilisant les observations faites sur les EST.

Pour cela, nous allons tout d'abord formuler et tester des hypothèses sur le mécanisme d'infidélité de transcription.

5.2 Hypothèses sur le mécanisme d'infidélité de transcription

5.2.1 Hypothèses

On émet les hypothèses suivantes :

- Pour un gène donné de longueur l, les variables aléatoires

$$Y_1, \ldots, Y_l \text{ sont indépendantes.} \tag{5.5}$$

  On suppose également l'indépendance pour des gènes différents.
- Il existe une fonction $f : E \to [0, 1]$ telle que

$$P(Y_i = 1) = f(\varphi(i)) = f(e),\ \forall e \in E. \tag{5.6}$$

  Autrement dit, la survenue d'une infidélité de transcription à une position i dépend du contexte nucléotidique autour de cette position, et deux positions différentes avec le même contexte nucléotidique ont la même probabilité d'être sujettes à une infidélité de transcription. En particulier, la probabilité d'une substitution $P(Y_i = 1)$ ne dépend pas du gène considéré.

On souhaite tester les hypothèses (5.5) et (5.6).

Problème : Rappelons qu'en pratique, certaines erreurs observées sur les EST peuvent être des erreurs de séquençage, qui peuvent intervenir dans les mécanismes permettant de passer des ARNm aux EST. Ainsi, seule une information entachée de certaines erreurs est disponible.

Dans un premier temps, nous allons donc mesurer l'impact de l'erreur de séquençage sur les hypothèses (5.5) et (5.6).

5.2.2 Impact de l'erreur de séquençage sur le modèle

Nous allons reformuler dans cette section le phénomène d'infidélité de transcription en introduisant des variables aléatoires. Plaçons nous à une position i de la séquence de référence d'un gène. Appelons $B = x^i \in \{A, T, C, G\}$ la base à cette position. Considérons un EST aligné à cette position. Soit Z la variable aléatoire à valeurs dans $\{A, T, C, G\}$ ayant pour réalisation la base de l'EST et Z' la variable aléatoire à valeurs dans $\{A, T, C, G\}$ ayant pour réalisation la base lue lors du séquençage. Remarquons que les réalisations de Z et Z' peuvent être toutes deux différentes de B. On introduit également la variable aléatoire :

$$\xi = \begin{cases} 1 & \text{s'il y a une erreur de séquençage à la position } i, \\ 0 & \text{sinon.} \end{cases} \tag{5.7}$$

On note $P(\xi = 1) = \epsilon$, qui est la probabilité d'avoir une erreur de séquençage.
Supposons pour fixer les idées que $B = T$. Ainsi

- si $Z = T$ (pas d'infidélité de transcription), et $\xi = 0$ (pas d'erreur de séquençage), alors $Z' = Z = T$;
- si $Z = T$ et $\xi = 1$, on suppose que Z' prend une valeur $U_Z \in \{A, C, G\}$ uniformément.

De sorte que

$$Z' = Z\mathbb{1}_{(\xi=0)} + U_Z\mathbb{1}_{(\xi=1)}, \tag{5.8}$$

où $(U_B)_{B\in\{A,T,C,G\}}$ est une famille de variables aléatoires indépendantes de (Z, ξ), et telle que U_B suive une loi uniforme sur $\bar{B} = \{A, T, C, G\} - \{B\}$.

On rappelle que

$$Y = \mathbb{1}_{(Z\in\bar{B})} = \begin{cases} 1 & \text{s'il y a une substitution à la position } i, \\ 0 & \text{sinon,} \end{cases} \tag{5.9}$$

et on pose

$$Y' = \mathbb{1}_{(Z'\in\bar{B})} = \begin{cases} 1 & \text{si une substitution est LUE à la position } i, \\ 0 & \text{sinon.} \end{cases} \tag{5.10}$$

Caractériser l'influence de l'erreur de séquençage sur le modèle revient à chercher une relation entre Y et Y'.

Relation entre Y et Y'

On suppose ξ indépendante de Y, c'est-à-dire que les survenues d'une substitution et d'une erreur de séquençage sont des événements indépendants.

a) Supposons que $Z = B$. Dans ce cas, $Y = 0$ et

$$Z' = \begin{cases} Z = B & \text{si } \xi = 0, \\ U_Z = U_B & \text{si } \xi = 1, \end{cases} \tag{5.11}$$

et donc

$$Y' = \begin{cases} 0 & \text{si } \xi = 0, \\ 1 & \text{si } \xi = 1. \end{cases} \tag{5.12}$$

Autrement dit :

$$Y'\mathbb{1}_{(Z=B)} = \mathbb{1}_{(\xi=1)}\mathbb{1}_{(Z=B)} = \mathbb{1}_{(\xi=1)}(1-Y). \tag{5.13}$$

b) Supposons maintenant que $Z \in \bar{B}$. Dans ce cas, $Y = 1$ et

$$Z' = \begin{cases} Z \in \bar{B} & \text{si } \xi = 0, \\ U_Z & \text{si } \xi = 1. \end{cases} \tag{5.14}$$

Notons que si la base de l'EST n'est pas la même que celle de la séquence de référence, c'est-à-dire si $Z \in \bar{B}$, et qu'il y a une erreur de séquençage, c'est-à-dire si $\xi = 1$, la base lue Z' peut être la même que celle de la séquence de référence. On peut retrouver la base B avec une probabilité 1/3. Ce qui conduit à l'expression suivante :

$$\begin{aligned} Y'\mathbb{1}_{(Z\in\bar{B})} &= \mathbb{1}_{(Z\in\bar{B})}(\mathbb{1}_{(\xi=0)} + \mathbb{1}_{(\xi=1)}\theta), & (5.15) \\ &= Y(\mathbb{1}_{(\xi=0)} + \mathbb{1}_{(\xi=1)}\theta), & (5.16) \end{aligned}$$

où θ est une variable aléatoire de loi de Bernoulli de paramètre 2/3 (i.e $P(\theta = 1) = 2/3$), indépendante de (ξ, Z, Y).

En additionnant (5.13) et (5.16), on obtient

$$Y' = \mathbb{1}_{(\xi=1)}(1-Y) + Y(\mathbb{1}_{(\xi=0)} + \mathbb{1}_{(\xi=1)}\theta). \tag{5.17}$$

Comme $\mathbb{1}_{(\xi=0)} = 1 - \mathbb{1}_{(\xi=1)}$, on a finalement :

$$\boxed{Y' = Y + \mathbb{1}_{(\xi=1)}(1 - Y(2-\theta)).} \tag{5.18}$$

Relation entre les lois de Y et Y'

Réécrivons la relation précédente de la manière suivante :

$$Y' = (1 - 2\mathbb{1}_{(\xi=1)} + \theta\mathbb{1}_{(\xi=1)})Y + \mathbb{1}_{(\xi=1)}. \tag{5.19}$$

Comme θ et ξ sont indépendantes de Y, il vient :

$$E[Y'] = E[1 - 2\mathbb{1}_{(\xi=1)} + \theta\mathbb{1}_{(\xi=1)}]E[Y] + P(\xi = 1). \tag{5.20}$$

Puisque θ et ξ sont indépendantes, on a

$$\begin{aligned} E[1 - 2\mathbb{1}_{(\xi=1)} + \theta\mathbb{1}_{(\xi=1)}] &= 1 - 2P(\xi = 1) + P(\theta = 1)P(\xi = 1) \\ &= 1 - 2\epsilon + \frac{2}{3}\epsilon \\ &= 1 - \frac{4}{3}\epsilon\ . \end{aligned} \tag{5.21}$$

Enfin, on obtient

$$P(Y' = 1) = (1 - \frac{4}{3}\epsilon)P(Y = 1) + \epsilon, \tag{5.22}$$

ou encore

$$\boxed{P(Y = 1) = \frac{P(Y'=1)-\epsilon}{1-\frac{4}{3}\epsilon}\ .} \tag{5.23}$$

Notons que l'on retrouve l'expression (4.8) établie au chapitre précédent.

Conséquences

Grâce à la relation (5.23), il est possible de reformuler les hypothèses (5.5) et (5.6).

a) On souhaite réaliser le test (5.5) pour chaque couple de positions de chacun des 17 gènes. Considérons deux positions i_1 et i_2 de la séquence de référence, par exemple $i_1 = 1$ et $i_2 = 2$. Comme précédemment, on introduit les variables aléatoires Y_1, ξ_1, θ_1 et Y_2, ξ_2, θ_2. On suppose que :

1. $\xi_1, \xi_2, \theta_1, \theta_2$ sont indépendantes,
2. $(\xi_1, \xi_2, \theta_1, \theta_2)$ est indépendante de (Y_1, Y_2).

On associe également à ces variables aléatoires :

$$Y_i' = Y_i + \mathbb{1}_{(\xi_i=1)}(1 - Y_i(2 - \theta_i)),\ i = 1, 2. \tag{5.24}$$

Si l'hypothèse (5.5) est vraie, c'est-à-dire, si Y_1 et Y_2 sont indépendantes, alors

$$Y_1' \text{ et } Y_2' \text{ sont indépendantes.} \tag{5.25}$$

En effet, si Y_1 et Y_2 sont indépendantes, alors les variables aléatoires $(Y_1, Y_2, \xi_1, \xi_2, \theta_1, \theta_2)$ sont indépendantes. Par (5.24), Y_1' et Y_2' le sont aussi.

b) L'hypothèse (5.6) est équivalente à : $\exists f' : E \to [0,1]$ telle que

$$P(Y'_i = 1) = f'(\varphi(i)), \tag{5.26}$$

avec

$$f' = (1 - \frac{4}{3}\epsilon)f + \epsilon, \tag{5.27}$$

Conclusion

Grâce à ces relations, nous allons pouvoir tester les hypothèses faites sur Y à partir de Y'. Ainsi, au lieu de tester directement les hypothèses (5.5) et (5.6), nous allons tester dans la suite les hypothèses (5.25) et (5.26).

5.3 Tests des hypothèses

5.3.1 Test de l'hypothèses (5.5)

Plaçons-nous sur un gène donné de longueur l. Pour tout couple de positions $(i,j) \in \{1,\dots,l\}^2$, on souhaite réaliser le test

$$\begin{cases} H_0 : Y_i \textit{ et } Y_j \text{ sont indépendantes} \\ H_1 : Y_i \textit{ et } Y_j \text{ ne sont pas indépendantes.} \end{cases} \tag{5.28}$$

L'hypothèse H_0 du test (5.28) signifie que la survenue d'une infidélité de transcription à la position i est indépendante de la survenue d'une infidélité de transcription à la position j.

Considérons les EST cancéreux ; le cas normal est traité de la même manière. On note $n_{i,j}$ le nombre d'EST qui s'alignent en même temps sur les deux positions i et j.

Soit le tableau de contingence (voir également la Figure 5.2) :

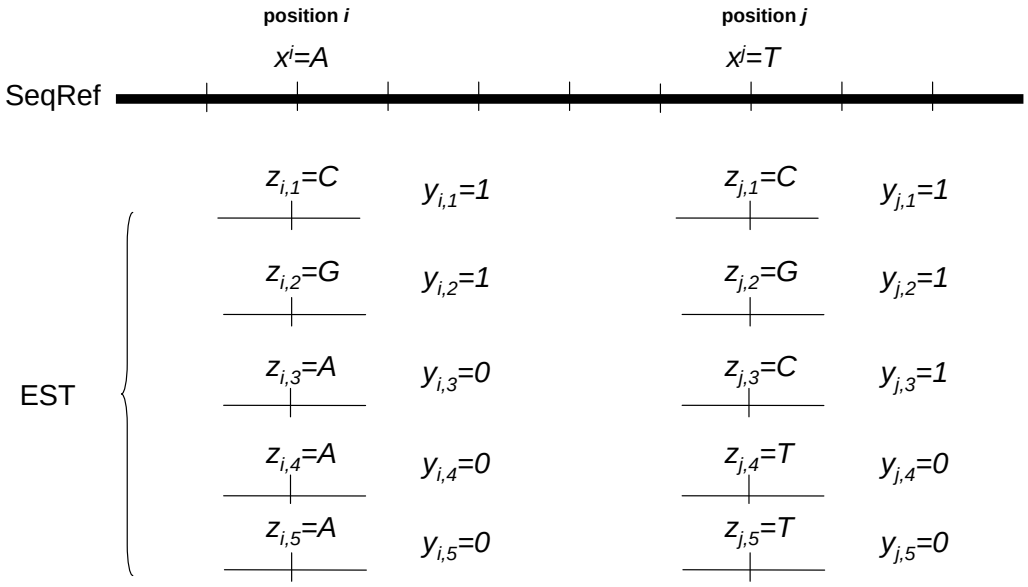
Y_i/Y_j	pas de substitution	substitution
pas de substitution		
substitution		

A l'intersection de la ligne m et de la colonne n, on porte le nombre d'EST ayant la modalité m de Y_i et la modalité n de Y_j parmi les $n_{i,j}$ EST alignés.

On souhaite réaliser le test (5.28) pour chaque couple de positions de chacun des 17 gènes. En pratique, il se peut toutefois qu'aucun EST ne s'aligne aux positions d'intérêt, auquel cas le test est irréalisable. Ce peut être le cas en particulier de deux positions éloignées sur un gène long.

Résultats

Pour chacun des 17 gènes, l'hypothèse H_0 du test (5.28) a été testée pour tout couple de positions où des EST s'alignaient simultanément. C'est le test exact de Fisher qui a été utilisé dans la pratique, car le plus souvent les effectifs théoriques n'étaient pas supérieurs à 5. Les SNPs ont été retirés de l'étude. Les résultats sont présentés dans les Tableaux 5.1 et 5.2, dans le cas cancer et le cas normal.



Yi / Yj	pas de substitution	substitution
pas de substitution	2	1
substitution	0	2

FIG. 5.2 – Exemple d'EST alignés à deux positions i et j, et Tableau 5.3.1 correspondant

Gènes	Couples testés	Couples dépendants	Pourcentage	LBE	Pourcentage
ALB	1685246	18217	1.08%	165961	9.85%
ATP5A1	1537907	38029	2.47%	71071	4.62%
ALDOA	1470301	30623	2.08%	217072	14.76%
CALM2	632717	18784	2.97%	59763	9.45%
ENO1	1488516	53596	3.60%	138310	9.29%
FTH1	690771	41845	6.06%	60941	8.82%
FTL	378015	29670	7.85%	32077	8.49%
GAPDH	848438	66338	7.82%	71577	8.44%
HSPA8	2072260	41684	2.01%	199563	9.63%
LDHA	1183904	24140	2.04%	114487	9.67%
RPL7A	395599	30002	7.58%	34769	8.79%
RPS4X	454963	22218	4.88%	41728	9.17%
RPS6	343206	20718	6.04%	30704	8.95%
TMSB4X	196251	11136	5.67%	17838	9.09%
TPI1	743323	26456	3.56%	69413	9.34%
TPT1	344003	18947	5.51%	30988	9.01%
VIM	1375271	37959	2.76%	131217	9.54%
Ensemble	15840691	530362	3.35%	1487479	9.39%

TAB. 5.1 – Résultats du test (5.28) dans le cas cancer

Gènes	couples testés	couples dépendants	Pourcentage	LBE	Pourcentage
ALB	2168531	50240	2.32%	208663	9.62%
ATP5A1	1420624	13766	0.97%	64116	4.51%
ALDOA	1480063	14163	0.96%	221772	14.98%
CALM2	634644	15665	2.47%	60689	9.56%
ENO1	1527055	26366	1.73%	148773	9.74%
FTH1	624921	22060	3.53%	58720	9.40%
FTL	378015	20864	5.52%	33742	8.93%
GAPDH	836871	30412	3.63%	78589	9.39%
HSPA8	1975575	23812	1.21%	194235	9.83%
LDHA	1171477	12323	1.05%	115498	9.86%
RPL7A	395533	15311	3.87%	37471	9.47%
RPS4X	450163	16123	3.58%	42978	9.55%
RPS6	343206	17226	5.02%	31623	9.21%
TMSB4X	196251	14911	7.60%	17027	8.68%
TPI1	728192	14227	1.95%	70530	9.69%
TPT1	344035	15853	4.61%	31893	9.27%
VIM	1353462	31113	2.30%	131292	9.70%
Ensemble	16028618	354435	2.21%	1547610	9.66%

TAB. 5.2 – Résultats du test (5.28) dans le cas normal

Conclusion : Le faible pourcentage de tests positifs et le LBE très élevé, dans le cas cancer comme dans le cas normal, laissent penser qu'il n'y a globalement pas de dépendance entre deux infidélités de transcription à des positions différentes des séquences de référence des gènes.

5.3.2 Test de l'hypothèses (5.6)

Numérotons de 1 à n les positions des séquences de référence des 17 gènes. Soit $e \in E$ un motif fixé de taille $h_1 + h_2 + 1$. Une manière de tester l'hypothèse (5.6) est de tester l'homogénéité des lois des variables Y_i, où i parcourt l'ensemble des positions des différents gènes où se trouve le motif fixé e. Notons l'ensemble de ces positions $I_e = \{1 \leq k \leq n,\ \varphi(k) = e\} = \{i_1, \ldots, i_l\}$, et supposons que $\text{card}(I_e) \geq 2$; c'est-à-dire que le motif e existe au moins deux fois sur les gènes.

On souhaite alors réaliser le test :

$$\begin{cases} H_0 : \mathcal{L}_{Y_{i_1}} = \cdots = \mathcal{L}_{Y_{i_l}} \\ H_1 : \bar{H}_0. \end{cases} \tag{5.29}$$

Avec la relation (5.22), il est clair que ce test est équivalent au test :

$$\begin{cases} H_0 : \mathcal{L}_{Y'_{i_1}} = \cdots = \mathcal{L}_{Y'_{i_l}} \\ H_1 : \bar{H}_0. \end{cases} \tag{5.30}$$

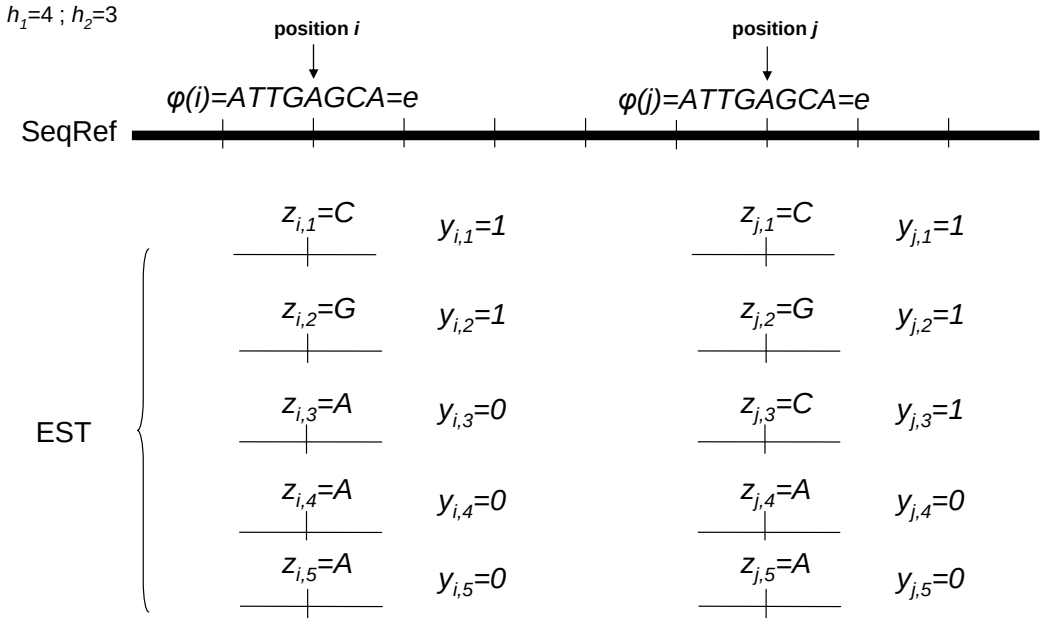
Pour cela, considérons les n_{i_p} EST alignés à la position i_p : on dispose alors d'un n_{i_p}-échantillon de Y_{i_p} : $(Y_{i_p,1}, \ldots, Y_{i_p,n_{i_p}})$.

On remplit le tableau de contingence 5.3 (voir également la Figure 5.3).
En supposant que les l échantillons obtenus pour l'ensemble des positions de I_e soient indépendants, un test d'homogénéité du χ^2 si les conditions de validité sont vérifiées ou un test exact de Fisher sinon permettent de tester H_0.

Toutefois, un problème se pose en pratique si l'on considère des motifs de longue taille. En effet, la majorité des motifs de longue taille n'apparaissent qu'une seule fois sur un unique gène du

position	i_1	...	i_p	...	i_l
pas de substitution			$n_{i_p} - k_{i_p}$		
substitution			k_{i_p}		

TAB. 5.3 – Tableau de contingence des EST alignés aux positions $\{i_1, \ldots, i_l\}$



position	i	j
pas de substitution	3	2
substitution	2	3

FIG. 5.3 – Exemple d'EST alignés à deux positions i et j ayant le même motif $\varphi(i) = \varphi(j) = e$, et Tableau 5.3 correspondant

génome. Il est donc impossible de réaliser ce test pour un motif donné trop long, puisque ce motif n'apparaît qu'à une seule position. On doit donc considérer des motifs plus petits afin d'augmenter le nombre de leurs occurrences.

Puisqu'à une position i fixée, la "bulle de transcription" s'étend des nucléotides $i - 4$ à $i + 3$, on se propose de considérer uniquemement des fenêtres de taille $h_1 + h_2 + 1 = 8$, avec $h_1 = 4$ et $h_2 = 3$.

On dispose dans nos données de 17719 fenêtres de cette taille, dont 6082 sont les répétitions de 2716 motifs différents. Les 11637 autres fenêtres sont des motifs à occurrence unique. Ceci représente $2716 + 11637 = 14353$ motifs différents sur les $4^8 = 65536$ possibles. Le motif le plus représenté apparaît 9 fois.

On souhaite réaliser les tests du χ^2 ou de Fisher pour vérifier si la position correspondant à un motif donné a, à chacune des occurrences de ce motif sur un gène, la même probabilité d'être sujette à une infidélité de transcription. Les motifs apparaissant deux fois et ceux apparaissant

plus de deux fois ont été étudiés séparément. En effet, des problèmes calculatoires ont été rencontrés avec SAS et R pour réaliser le test exact de Fisher sur les tableaux de contingence correspondant aux motifs à plus de deux occurrences. Pour ces derniers, nous nous sommes contentés de réaliser le test du χ^2 lorsque les conditions de validité étaient vérifiées. De ce fait, certains motifs n'ont pu être testés. Les résultats sont présentés dans les Tableaux 5.4, 5.5, 5.6 et 5.7. Une estimation du nombre moyen de faux positifs donnée par le LBE est indiquée dans la légende de chaque tableau.

Cancer	χ^2 +	χ^2 -	χ^2 NR	Sommes
Fisher +	505	0	228	733
Fisher -	13	414	1035	1462
Sommes	518	414	1263	2195

TAB. 5.4 – Résultats du test (5.29) pour les motifs à deux occurrences, dans le cas cancer ; χ^2 + : Nombre de tests du χ^2 positifs, χ^2 - : Nombre de tests du χ^2 négatifs, χ^2 NR : Nombre de tests du χ^2 non réalisables, Fisher + : Nombre de tests de Fisher positifs, Fisher - : Nombre de tests de Fisher négatifs ; LBE χ^2=16, LBE Fisher=85

Normal	χ^2 +	χ^2 -	χ^2 NR	Sommes
Fisher +	252	0	221	473
Fisher -	15	243	1439	1697
Sommes	267	243	1660	2170

TAB. 5.5 – Résultats du test (5.29) pour les motifs à deux occurrences, dans le cas normal ; χ^2 + : Nombre de tests du χ^2 positifs, χ^2 - : Nombre de tests du χ^2 négatifs, χ^2 NR : Nombre de tests du χ^2 non réalisables, Fisher + : Nombre de tests de Fisher positifs, Fisher - : Nombre de tests de Fisher négatifs ; LBE χ^2=9, LBE Fisher=102

Nous pouvons constater grâce aux Tableaux 5.4 et 5.5 que le test du χ^2 est plus puissant que le test exact de Fisher. En effet, dans le cas cancer comme dans le cas normal, le test du χ^2 est positif lorsque le test de Fisher est positif.

Cancer	χ^2 +	χ^2 -	χ^2 NR	Sommes
Nombre de tests	157	37	327	521

TAB. 5.6 – Résultats du test (5.29) pour les motifs à plus de deux occurrences ; LBE=1

Normal	χ^2 +	χ^2 -	χ^2 NR	Sommes
Nombre de tests	67	21	420	508

TAB. 5.7 – Résultats du test (5.29) pour les motifs à plus de deux occurrences ; LBE=0

Remarque : Les nombres globaux de motifs étudiés dans le cas cancer et dans le cas normal sont différents, car il a fallu supprimer les motifs pour lesquels aucun EST n'était aligné ; ce qui ne s'est produit que dans le cas normal.

Il apparaît dans les résultats qu'un nombre non négligeable de tests sont significatifs. L'hypothèse (5.6) étant rejetée dans de nombreux cas, ceci conduit à penser que la probabilité de

survenue d'une infidélité de transcription ne dépend pas que du motif. Ceci signifie que la loi de Y_i peut dépendre de i. De ce fait, la fonction f définie dans l'hypothèse (5.6) peut dépendre de i.
Cette constatation empêche donc de continuer l'étude. Si cette hypothèse avait été vérifiée, on aurait pu modéliser $P(Y_i = 1)$ par exemple en utilisant une fonction de régression logistique.

5.4 Conclusions

Dans ce chapitre, nous avons formulé des hypothèses relatives à la probabilité de survenue d'une infidélité de transcription. Nous avons également montré qu'il était possible de tester ces hypothèses bien que les données disponibles puissent être entachées d'erreurs de séquençage. Il apparaît ainsi qu'à deux positions différentes d'un gène, les infidélités de transcription semblent survenir de manière indépendante. Les résultats du test de l'hypothèse (5.6) semblent indiquer que la survenue d'une infidélité de transcription ne dépend pas que du motif.

Partie II

Recherche de biomarqueurs de l'allergie à l'arachide

Chapitre 6

Introduction aux problématiques de l'allergie à l'arachide

6.1 Introduction

Une **allergie** est une réaction exagérée du système immunitaire vis-à-vis de substances étrangères en principe sans danger pour l'homme, appelées **allergènes**. En particulier, l'allergie à l'arachide, qui touche plus de 0.5% de la population française, est une allergie alimentaire à risque léthal [31] et ne connaît à l'heure actuelle aucun traitement.

Pour un patient allergique, les réactions dues à un contact avec l'arachide peuvent être très graves. Afin d'éviter ces accidents, des mesures d'éviction parfois très strictes sont imposées au patient, ce qui implique une baisse notable de la qualité de vie du patient et de sa famille. La directive européenne 2003/89 du 10 novembre 2003 impose donc un étiquetage des produits alimentaires qui améliore le quotidien des patients, mais d'éventuelles traces d'arachide peuvent malgré tout être présentes dans des produits déclarés sans arachide. Fréquenter les lieux de restauration collective devient également difficile. La détection précoce de l'allergie à l'arachide représente donc un réel enjeu de santé publique.

Pour les personnes à risque (antécédents familiaux, histoire clinique en faveur d'une allergie,...) un TPO (Test de Provocation Orale, en anglais FC pour Food Challenge) en milieu hospitalier est nécessaire pour détecter ou confirmer une allergie à l'arachide. Les patients se voient administrer des doses croissantes d'arachide définies dans un protocole jusqu'à l'apparition d'un symptôme objectif. Le patient est déclaré allergique dès lors qu'une réaction est observée par le clinicien. Actuellement, le TPO est le seul examen qui permette de savoir sans ambiguïté qu'un patient est allergique. Mais le TPO est un test long et coûteux, qui présente toujours un danger potentiel pour le patient car l'intensité de la réaction est imprévisible. Il requiert l'hospitalisation des patients, qui sont souvent des enfants, dans des centres spécialisés.

L'objectif actuel des cliniciens est de mettre en place un test diagnostique avec des variables faciles à mesurer par une prise de sang ou un test cutané et qui soit sans danger pour le patient, afin de remplacer à terme le TPO.

Par ailleurs, le TPO permet d'obtenir de précieuses informations sur la sévérité de l'allergie du patient. La gravité de la réaction est évaluée en attribuant aux symptômes apparus lors du TPO un score allant de 1 à 5 [32], augmentant avec la sévérité et présenté plus loin. La sévérité de l'allergie peut également être évaluée en notant *a posteriori* les symptômes apparus lors du premier accident du patient et qui sont répertoriés dans son dossier médical le cas échéant. Le patient se voit ainsi attribuer un **score de sévérité du premier accident** en utilisant la même graduation que pour le TPO. Connaître la dose provoquant la première réaction, appelée **dose réactogène**, est également utile.

Pour un patient dont l'allergie est déja avérée, être capable de prédire la sévérité des symptômes encourus après ingestion accidentelle d'arachide permettrait d'éviter le TPO. En effet, le TPO est à l'heure actuelle la référence pour évaluer la sévérité d'une allergie [33]. Pouvoir prédire la sévérité d'une allergie sans risque pour le patient avec une simple prise de sang et des tests cutanés serait d'un intérêt clinique évident.

Deux questions se posent alors :

1. Est-il possible de construire un modèle prédictif permettant de détecter l'allergie à l'arachide à l'aide de variables faciles à mesurer ?
2. Est-il possible de prédire la sévérité de la réaction lors d'une ingestion accidentelle d'arachide ?

Pour cela, une étude clinique a été menée sur un échantillon de 243 personnes, dont 129 recrutées à Nancy et 114 à Lille, sur lesquels un certain nombre de variables ont été mesurées. L'échantillon est composé de 179 patients dont l'allergie à l'arachide a été révélée par un TPO et de 64 patients **atopiques**, *i.e.*, allergiques au bouleau ou à des graminées mais pas à l'arachide. Pour les cliniciens, il est en effet plus intéressant de faire la différence entre des allergiques et des atopiques qu'entre des allergiques et des contrôles n'ayant aucune allergie. En effet, la majorité des patients consultant un allergologue ont déja connu des antécédents concernant des allergies diverses.

L'échantillon des données disponibles ainsi que la nature des variables mesurées sont détaillés dans ce chapitre.

Le but de cette étude est de construire des modèles de discrimination permettant de détecter l'allergie à l'arachide et de prédire pour des individus allergiques le score du TPO, le score du premier accident et la dose réactogène. Les méthodes de l'analyse discriminante sont pour cela rappelées dans le chapitre suivant.

Nous nous intéressons dans le Chapitre 8 à la discrimination des individus allergiques et des individus atopiques. L'étude relative à la prédiction des mesures de sévérité est résumée dans le Chapitre 9. L'article correspondant constitue le Chapitre 10. Enfin, des applications de l'algorithme mis au point pour étudier la dose réactogène sont présentées dans le Chapitre 11.

6.2 Description biologique des variables

On mesure la quantité présente de certains anticorps, les *Immunoglobulines de type E (IgE)*, qui sont des protéines produites par le système immunitaire et qui peuvent intervenir dans les mécanismes induisant une réaction immédiate lors d'une allergie.

L'allergie à une substance commence par une phase de **sensibilisation** : lors de la première rencontre avec l'allergène, l'organisme synthétise des IgE qui seront ensuite présentes dans le sérum. A la suite d'un contact accidentel avec l'allergène, un patient sensibilisé est déclaré **allergique** si des symptômes apparaissent dans les minutes ou les heures qui suivent. Certains patients sensibilisés ne deviennent jamais allergiques.

Comme expliqué dans le Chapitre 1, chaque anticorps est codé pour identifier un antigène qui lui est propre et permettre son élimination. Dans le cas de l'allergie, un antigène est appelé un **allergène**. Les IgE sont dites **spécifiques** d'un allergène ou **dirigées contre** un allergène. Par opposition, les IgE sont dites **totales** lorsqu'on s'intéresse à la mesure de l'ensemble de ces anticorps, indépendamment de la substance contre laquelle ils sont dirigés. Par exemple, pour un patient allergique à l'arachide et au lait, les IgE totales sont la réunion de l'ensemble des IgE dirigées contre des protéines contenues dans l'arachide et de l'ensemble des IgE dirigées contre des protéines contenues dans le lait.

Par ailleurs, les IgE sont divisées en deux catégories :

1. les IgE **circulantes**, qui sont véhiculées par le sérum,
2. et les IgE **non circulantes**, qui sont fixées par leur extrémité à des **mastocytes** ou des **basophiles**, qui sont des cellules produisant les médiateurs chimiques induisant les symptômes.

Les IgE circulantes peuvent être mesurées par un dosage immunologique. L'ampleur de la réaction cutanée déclenchée par les IgE non circulantes est mesurée par un **prick test** (test cutané).

6.2.1 Les dosages immunologiques

Cette méthode de dosage permet de quantifier les IgE circulantes spécifiques d'un patient à partir d'un simple prélèvement sanguin. La méthode **ELISA** (*Enzyme-Linked ImmunoSorbent Assay*) permet de doser la concentration de chaque anticorps en mesurant une réaction de fluorescence.

Le protocole expérimental se déroule en six étapes (Figure 6.1) :

1. L'allergène d'intérêt est d'abord immobilisé sur un support.
2. Le sérum du patient est placé dans le récipient contenant l'allergène. Les IgE spécifiques de l'allergène contenues dans le sérum du patient viennent alors se fixer sur l'allergène, tandis que les IgE non spécifiques restent "libres" dans le sérum.
3. Les IgE non fixées sont évacuées par une phase de lavage.
4. Des immunoglobulines de synthèse couplées à une enzyme sont introduites dans le récipient et viennent se fixer aux IgE encore présentes. Ces anticorps, qui sont des protéines de synthèse, sont dits "anti-IgE".
5. L'ajout du substrat de l'enzyme entraîne la libération d'un produit fluorescent quantifiable. La mesure de ce signal est convertie en quantité d'IgE en kilo-unités par litre (1 kU/L = 2.4 mg d'IgE par litre de sérum).
6. Le test est dit positif pour un type d'IgE déterminé si la mesure enregistrée dépasse le seuil de détection fixé par le fournisseur de réactifs, dans cette étude 0.10 kU/L.

Le dosage des IgE par la méthode ELISA peut varier légèrement par manque de répétabilité. Par exemple, des mesures réalisées dans un même laboratoire peuvent présenter un coefficient de variation (l'écart-type des mesures divisé par la moyenne) de 5% .

Par ailleurs, la variabilité inter-laboratoires ne peut être négligée. En effet, une étude menée sur 463 laboratoires a ainsi présenté des résultats de dosages immunologiques de certains anticorps dont le coefficient de variation s'élevait à 10% (Quality Club Phadia d'Octobre 2006, *http ://www.phadia.com*).

En dépit de ces limites, le dosage par ELISA est actuellement la méthode offrant la meilleure répétabilité des mesures. C'est également la méthode la plus simple et la plus précise pour doser les immunoglobulines.

Remarque : Comme expliqué précédemment, certains patients sensibilisés ne deviennent jamais allergiques, bien qu'ils puissent éventuellement avoir des niveaux d'IgE élevés. La présence de ces immunoglobulines dans le sérum n'atteste donc pas nécessairement la présence d'une allergie.

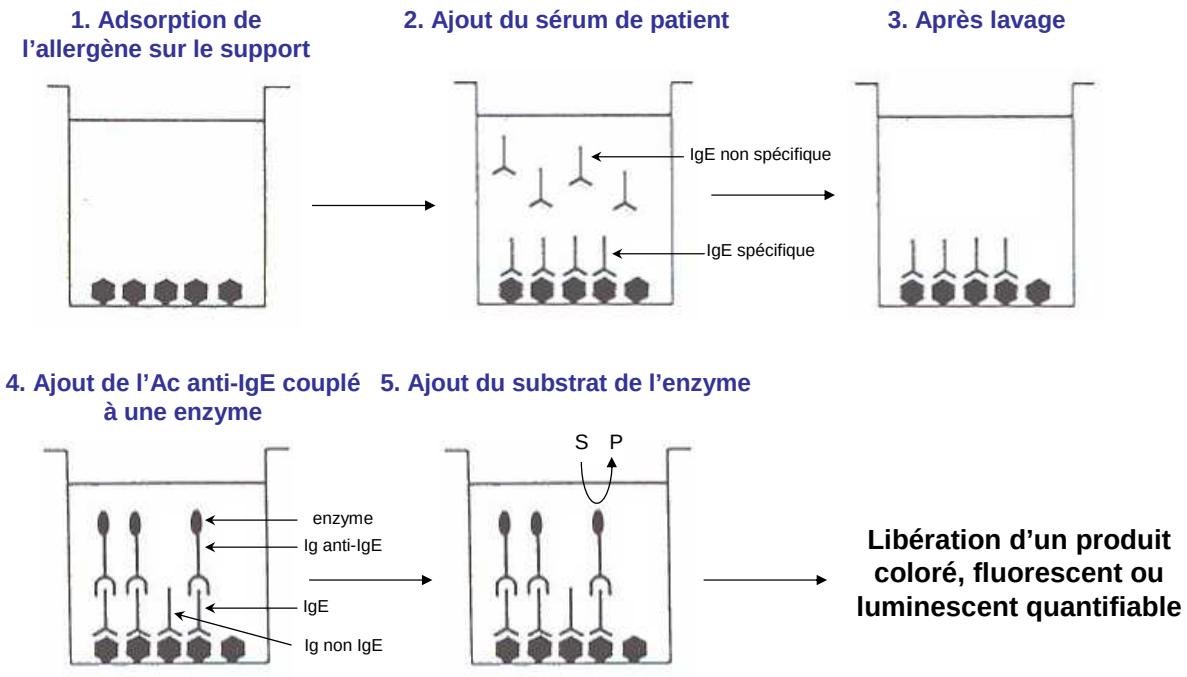


FIG. 6.1 – Dosages des IgE spécifiques

Cas particulier de l'allergie à l'arachide

Pour les cliniciens, diagnostiquer l'allergie à l'arachide par le dosage des IgE dirigées contre les protéines de l'arachide constitue un test très sensible mais peu spécifique [34]. Ceci peut être dû au fait que de multiples protéines sont contenues dans la cacahuète, y compris des protéines n'ayant aucun rôle allergénique, comme celles nécessaires au bon fonctionnement de la plante par exemple. La présence de ces protéines peut ainsi entraîner une réponse positive au test alors que l'individu n'est pas allergique et fausser le résultat en générant des faux positifs. De ce fait, il peut être intéressant de doser des IgE dirigées contre des protéines particulières contenues dans l'arachide. Ces variables seront également utilisées dans l'étude de la prédiction de la sévérité de l'allergie à l'arachide.

Ainsi pour chacun des 243 individus de l'échantillon ont été mesurés 6 dosages immunologiques :

1. les *IgE totales*,
2. les *IgE dirigées contre l'arachide (f13)*, *i.e.*, les anticorps dirigés contre l'ensemble des protéines contenues dans un extrait de cacahuète crue [32],

   et les IgE spécifiques de

   (a) *rAra-h1*,

   (b) *rAra-h2*,

   (c) *rAra-h3*,

   (d) *rAra-h8*,

   qui sont des allergènes majeurs de l'arachide, *i.e.*, des protéines particulières de l'arachide connues pour provoquer des symptômes [32].

Remarque : Notons que *rAra-h1, rAra-h2, rAra-h3* et *rAra-h8* sont des protéines recombinantes, *i.e.*, produites dans la bactérie *E. coli* chez Genclis [32]. Cette technique permet la production d'un grand nombre de protéines hautement purifiées.

Les mesures des dosages immunologiques sont des caractères quantitatifs continus.

Notons que pour les individus allergiques, les échantillons sanguins permettant les dosages immunologiques sont prélevés au moment du TPO.

6.2.2 Mise en évidence des IgE non circulantes par les prick tests

Les **prick tests** (ou SPT pour *Skin Prick Test*) sont des tests cutanés permettant de détecter une sensibilité à un allergène donné en mettant en évidence les IgE non circulantes du patient.

Le protocole pour réaliser des prick tests est très simple : une faible dose de l'allergène d'intérêt est piquée dans l'épiderme, entraînant une éventuelle réaction inflammatoire prenant la forme d'une papule (boursouflure). Son diamètre est alors mesuré en millimètres. Cette réaction atteste d'une libération d'histamine, qui est due à la rencontre des IgE et de l'allergène. En guise de témoin positif, un prick test à la codéine est effectué, provoquant toujours la libération d'histamine, même en l'absence d'IgE. Le prick test à la codéine permet de vérifier la réactivité de la peau à ce type de test. Chez toute personne testée, allergique ou non, le prick test à la codéine induit une boursouflure, dont le diamètre est mesuré. Le rapport des diamètres allergène/codéine est alors utilisé comme une mesure de l'importance de la réaction allergique.

Remarque : Dans certaines études la mesure d'un prick test est parfois discrétisée, en disant que le prick test est positif si le rapport diamètre de la papule / diamètre de la codéine est supérieur à un certain seuil (en général 0.5 ou 0.75), et négatif sinon.

Toutefois, en dépit d'un usage international très répandu, le protocole de réalisation des prick tests n'est pas standardisé. Toutes les papules n'ont pas une forme circulaire et quantifier l'ampleur de la réaction cutanée s'avère dans ce cas plus difficile. Certains auteurs mesurent l'aire de la papule avec un appareil particulier [35].

En conclusion, les prick tests permettent de mettre en évidence les IgE non circulantes du patient, sans pour autant les quantifier exactement.

Cas particulier de l'allergie à l'arachide

A l'instar du dosage immunologique des IgE dirigées contre l'arachide, le prick test à l'arachide est un test très sensible mais peu spécifique [34]. De plus, l'étude des capacités de diagnostic des prick tests à *rAra-h1*,*rAra-h2*,*rAra-h3* et *rAra-h8* a déja été menée dans [32].

Aucun prick test n'a été mesuré sur les individus atopiques. En revanche, certains prick tests pourraient s'avérer utiles pour la prédiction des mesures de la sévérité menée sur les patients allergiques de Nancy et que nous developperons dans les prochains chapitres.

Ainsi, pour les 94 individus allergiques de Nancy, 34 prick tests à différents allergènes, divisés en trois familles, sont effectués :

1. 10 légumineuses : *pois chiche, fève, lentille, haricot sec, petit pois, soja, lupin, pois blond, caroube, néré*, car l'arachide est une légumineuse,
2. 12 fruits à coque : *amande, noix, noisette, noix de cajou, noix du Brésil, noix de Macadamia, noix de pécan, pistache, pignon, châtaigne, noix de Nangaille, arachide grillée*, qui sont souvent associés à une allergie à l'arachide,
3. 12 pneumallergènes : *Dermatophagoïdes Pteronyssinus (acariens), Alternaria, blatte, chat, chien, 12 graminées, bouleau, armoise, plantain, frêne, colza, latex*, qui sont des tests cliniques standards en consultation.

Les diamètres des papules des prick tests sont divisés par celui de la papule du prick test à la codéine, qui est mesuré également. Ce sont donc des caractères quantitatifs continus.

Les prick tests sont effectués au moment du TPO.

Comme expliqué précédemment, un patient sensibilisé peut développer des IgE sans devenir nécessairement allergique. Seul le TPO permet de confirmer ou d'infirmer de manière sûre des présomptions d'allergie.

6.2.3 Le test de provocation orale (TPO)

A l'heure actuelle, le TPO est le seul moyen de confirmer ou d'infirmer une allergie alimentaire. Pour la sécurité des patients, il est réalisé en milieu hospitalier sous haute surveillance. Des doses croissantes d'arachide en poudre sont administrées par paliers toutes les dix minutes, tant qu'aucune réaction n'est observée. Le test est arrêté lorsque des symptômes objectifs apparaissent.

La dose ayant provoqué cette première réaction est appelée **dose réactogène** (DR). Elle apporte une double information : la quantité tolérée par le patient et la quantité déclenchant une réaction allergique. Cependant, comme les doses sont administrées par paliers, on ne connaît pas la valeur exacte de la dose réactogène, mais l'intervalle auquel elle appartient.

La gravité de la réaction observée est ensuite évaluée en lui attribuant un score qui dépend des symptômes apparus, allant de 1 à 5, du moins grave au plus grave [32] :

- **Score 1 :** un ou plusieurs symptômes bénins parmi : douleurs abdominales disparaissant sans traitement en moins de 30 minutes et/ou rhinoconjonctivite et/ou urticaire de moins de 10 papules (boursouflures) et/ou poussée d'eczéma ;
- **Score 2 :** 1 symptôme modéré parmi : douleurs abdominales nécessitant un traitement ou urticaire généralisée ou angioedème non laryngé ou toux ou chute du DEP (Debit Expiratoire de Pointe) allant de 15 à 20% ;

- **Score 3 :** 2 symptômes modérés dans la liste précédente ;
- **Score 4 :** 3 symptômes modérés dans la liste précédente ou asthme nécessitant un traitement ou angioedème laryngé ou hypotension ;
- **Score 5 :** symptômes nécessitant une hospitalisation en soins intensifs, comme un choc anaphylactique.

Notons toutefois que d'autres scores de sévérité existent [36, 37, 38, 39, 40].

Comme certains symptômes peuvent être liés à la peur de la réaction, comme des maux de ventre, un placebo est parfois administré au patient sans qu'il le sache. Si des symptômes apparaissent suite à l'ingestion du placebo, le TPO peut continuer, puisque la réaction ne serait due dans ce cas qu'à l'appréhension du patient. L'arachide et le placebo sont pour cela mélangés à de la purée de pommes de terre ou de la compote de pommes afin que le patient ne puisse les distinguer par le goût ou par la texture. Le TPO est alors dit réalisé "en simple aveugle" (TPOSA), ou *Single-Blind Placebo Controlled Food Challenge* (SBPCFC). De même, le clinicien peut lui-même ne pas savoir non plus si l'extrait donné contient l'allergène ou pas : on parle alors de TPO en double aveugle (TPODA) ou *Double-Blind Placebo Controlled Food Challenge* (DBPCFC). Ce protocole est parfois mis en place pour éviter que le patient ne détecte des signes d'anxiété chez le médecin. Si aucun placebo n'est utilisé, le TPO est dit "ouvert" (open FC). Ce dernier protocole est souvent utilisé avec les jeunes enfants qui appréhendent moins les risques de réaction [33].

Bien qu'étant une méthode très utilisée, le TPO connaît également quelques limites. Tout d'abord la progression des doses d'arachide administrées n'est pas standardisée et deux services d'allergologie différents peuvent utiliser des paliers différents. De même l'utilisation de paliers induit une imprécision de la dose réactogène. Ainsi, le seuil de tolérance réel du patient se situe en fait entre la dose réactogène et la dose précédente. Enfin, la principale limite du TPO réside dans son principe même. En effet le bons sens et l'éthique médicale interdisent de continuer le TPO après l'apparition des premiers symptômes, c'est pourquoi il est impossible de savoir quelle serait la gravité de la réaction du patient s'il avait été exposé à une dose plus forte. Il est raisonnable de penser que certains patients présenteraient des réactions plus graves en ingérant une plus grande dose d'arachide. Le TPO ne mesure donc que le niveau minimal de l'allergie du patient et ne semble par conséquent pas refléter sa gravité "réelle ", d'où l'intérêt de considérer également le score du premier accident quant à la sévérité de la réaction.

Chapitre 7

Méthodes d'analyse discriminante

On donnera comme références générales de ce chapitre [41, 42, 43, 44].

7.1 Formulation

Soit :
- $I = \{1, \cdots, n\}$ un ensemble de n individus ;
- y une variable qualitative à q modalités exclusives, observée sur les individus de I, qui induit une partition de I en classes $I_1, \cdots, I_q$;
- $x^1, \cdots, x^p$ p variables quantitatives mesurées sur les individus de I.

Exemple. Un ensemble d'individus est divisé en deux classes :

I_1 : classe des individus allergiques à l'arachide ;

I_2 : classe des individus atopiques, c'est-à-dire allergiques à une autre substance qu'à l'arachide ;

$x^1, \cdots, x^p$ sont les mesures de différents anticorps.

Le but de l'analyse discriminante est de prédire la classe d'appartenance d'un nouvel individu dont les mesures $x^1, \cdots, x^p$ sont connues.

Une analyse discriminante est effectuée en général en trois étapes :

1. sélection d'un nombre réduit de variables parmi $x^1, \cdots, x^p$ qui discriminent les classes ;
2. construction de différentes règles de classement, qui permettent d'affecter un nouvel individu à une des classes grâce aux mesures des variables sélectionnées à l'étape précédente ;
3. validation de la règle sur un ensemble indépendant d'individus.

On présente dans ce chapitre des méthodes de sélection de variables, de classement et de validation que l'on a utilisées.

7.2 Sélection des variables discriminantes

Avant de procéder à une analyse discriminante, il faut éliminer les variables n'apportant pas d'information significative sur l'appartenance à une classe. En effet, retenir trop de variables non discriminantes dans un modèle peut atténuer les différences entre les classes. En outre, l'utilisation dans un modèle de variables non discriminantes augmente inutilement le nombre de mesures à effectuer. On peut aussi être confronté à un phénomène de surapprentissage lorsqu'il y a trop de variables explicatives par rapport au nombre d'individus : le modèle déterminé s'adapte très bien aux données mais conduit à de mauvaises prédictions pour des individus ne faisant pas partie de l'ensemble d'apprentissage.

On utilise dans la suite deux façons de sélectionner les variables : soit individuellement par des tests d'homogénéité, soit en présence d'autres variables par une sélection pas-à-pas.

7.2.1 Tests d'homogénéité

On peut utiliser des tests de comparaison de moyennes ou de lois.

Comparaison de moyennes

Soit une variable x prise parmi $x^1, \cdots, x^p$.

On note $b_1, \cdots, b_q$ les modalités de y.

Pour $i = 1, \ldots, q$, on note x_{ij} la j^{ieme} mesure de x lorsque y a la modalité b_i, $j = 1, \cdots, n_i$.

On suppose que x_{ij} est la réalisation d'une variable aléatoire réelle (v.a.r.) X_{ij} et qu'il existe des nombres fixes $\mu, \beta_1, \cdots, \beta_q$ et une v.a.r. R_{ij} de loi $\mathcal{N}(0, \sigma)$, les R_{ij} étant mutuellement indépendantes, tels que :

$$X_{ij} = \mu + \beta_i + R_{ij}.$$

On effectue alors le test :

$$\begin{cases} H_0 : \beta_1 = \cdots = \beta_q = 0 \\ H_1 : \overline{H}_0 \end{cases} \tag{7.1}$$

Ce test est un test d'analyse de la variance à un facteur. Dans le cas $q = 2$, ce test équivaut au test de comparaison des moyennes de Student.

Si l'on rejette l'hypothèse H_0, on conclut que le caractère x discrimine de façon significative les classes $I_1, \cdots, I_q$ et il est alors conservé dans la suite de l'étude.

Lorsque les hypothèses de normalité ou d'homoscédasticité ne sont pas ou ne peuvent pas être vérifiées, il est possible d'utiliser des tests non paramétriques d'homogénéité de plusieurs lois.

Test de comparaison de lois

On compare les lois conditionnelles d'une variable x^j dans les classes $I_1, \ldots, I_q$. On dispose de plusieurs tests de comparaison de lois, par exemple :

- pour q quelconque, test d'homogénéité du khi-deux, pour des effectifs n_i suffisamment grands ;
- dans le cas $q = 2$, test de Kolmogorov-Smirnov ;
- dans le cas $q = 2$, tests de Mann-Whitney ou de Wilcoxon ; pour q quelconque, test de Kruskal-Wallis.

Comme nous utilisons le test de Kruskal-Wallis dans la prédiction de la sévérité de l'allergie à l'arachide, nous rappelons brièvement son déroulement.

Soit q variables aléatoires $X_1, \ldots, X_q$. On fait l'hypothèse que leur fonction de répartition est continue. On considère le test d'hypothèses :

$$\begin{cases} H_0 : & \mathcal{L}(X_1) = \ldots = \mathcal{L}(X_q) \\ H_1 : & \bar{H}_0. \end{cases} \tag{7.2}$$

On observe un échantillon $i.i.d$ de taille n_i de chaque v.a.r X_i noté $(x_{i1}, \ldots, x_{in_i})$; on suppose les q échantillons mutuellement indépendants.

a) Statistique de test.

On réunit les q échantillons. On range la suite obtenue par ordre croissant. On note $r(x_{ij})$ le rang de x_{ij} dans la suite ordonnée, réalisation de la v.a.r $r(X_{ij})$. On note $n = \sum_{i=1}^{q} n_i$.

Soit $W_i = \sum_{j=1}^{n_i} r(X_{ij})$. On a $\sum_{i=1}^{q} W_i = \sum_{l=1}^{n} l = \frac{n(n+1)}{2}$.
On considère la statistique de test :

$$T = (n-1)\frac{\sum_{i=1}^{q} n_i(\frac{W_i}{n_i} - \frac{n+1}{2})^2}{\sum_{i=1}^{q}\sum_{j=1}^{n_i}(r(X_{ij}) - \frac{n+1}{2})^2}. \tag{7.3}$$

En utilisant le fait que $\sum_{i=1}^{q}\sum_{j=1}^{n_i}(r(X_{ij}) - \frac{n+1}{2})^2 = \frac{(n-1)n(n+1)}{12}$, la statistique s'écrit :

$$T = \frac{12}{n(n+1)}\sum_{i=1}^{q}\frac{W_i^2}{n_i} - 3(n+1). \tag{7.4}$$

b) Règle de décision.
La région d'acceptation de H_0 est de la forme

$$A_\alpha = \{x \in \mathbb{R}^n : T < c(\alpha)\}. \tag{7.5}$$

Deux cas se présentent concernant la loi de T sous H_0 :
- si $q = 3$ et $n_1, n_2, n_3 \leq 5$, alors la loi de T est tabulée,
- si $q \geq 4$, la loi de T peut être approchée par la χ^2_{q-1}.

Donnons la règle de décision pour $q \geq 4$. Soit $z(\alpha)$ le réel tel que $P(\chi^2_{q-1} \geq z(\alpha)) = \alpha$. Soit t une réalisation de T.
Règle de décision. Si $t \geq z(\alpha)$, on rejette H_0 ; sinon on ne rejette pas H_0.

Mesure de la capacité de discrimination d'une variable et p-value

Dans les tests de comparaison au seuil α que nous avons étudiés, T étant la statistique de test et t_0 la réalisation observée de T, on peut écrire la règle de décision sous la forme suivante :
"si $t \geq t(\alpha)$, on rejette H_0 ; sinon on ne rejette pas H_0".
On définit la p-value (ou *probabilité critique*, ou p-valeur), notée p :

$$p = P(T \geq t_0) \quad \text{sous } H_0.$$

Comme $P(T \geq t(\alpha)) \leq \alpha$ sous H_0, la règle de décision peut s'écrire ainsi :
"si $p \leq \alpha$, on rejette H_0 ; sinon on ne rejette pas H_0".
Dans le cas où l'on rejette H_0, la probabilité critique est une mesure de la capacité individuelle de discrimination de la variable étudiée, sans tenir compte de la présence des autres variables.

7.2.2 Sélection pas-à-pas

On s'intéresse dans ce paragraphe à la mesure de la capacité de discrimination d'une variable en présence d'autres variables. Il se pose alors le problème de la sélection d'un sous-ensemble de r variables discriminantes parmi les p variables de départ. On peut remarquer que chercher parmi les C_p^r choix possibles de sous-ensembles un d'entre eux optimisant un certain critère peut prendre un grand temps de calcul. Des méthodes de sélection pas-à-pas, plus économiques d'un point de vue calculatoire mais qui ne donnent pas de façon générale un sous-ensemble optimal, sont présentées dans la suite.

Donnons trois exemples de critères classiques de la mesure du pouvoir discriminant d'un ensemble de variables. Notons pour cela T la matrice d'inertie totale, W la matrice d'inertie intra et B la matrice d'inertie inter.
- le Lambda de Wilks $\det W / \det T$,

- l'inertie inter classes $tr(T^{-1}B)$,
- le pourcentage de bien-classés obtenu avec un règle de classement à partir des variables choisies (voir plus loin).

On présente les méthodes ascendante, progressive et descendante.

Méthode ascendante

Cette méthode repose à chaque pas sur l'ajout de la variable qui, en présence des variables précédemment introduites, va optimiser un certain critère de discrimination à définir.

Algorithme. On décrit le pas 1 et le pas général l.

1^{er} pas : on détermine la variable $x^{(1)}$ parmi $x^1, \cdots, x^p$ qui a le plus grand pouvoir discriminant relativement à y.

$l^{ième}$ pas : on adjoint aux variables $x^{(1)}, \cdots, x^{(l-1)}$ déjà introduites la variable $x^{(l)}$ telle que le l-uplet $(x^{(1)}, \cdots, x^{(l-1)}, x^{(l)})$ ait le plus grand pouvoir discriminant parmi les l-uplets $(x^{(1)}, \cdots, x^{(l-1)}, x^j)$ pour $j \neq (1), \cdots, (l-1)$.

Règle d'arrêt. Une règle d'arrêt basée sur un critère probabiliste est celle du Lambda de Wilks.

Supposons que les vecteurs des mesures des variables forment un échantillon i.i.d. d'un vecteur aléatoire X dans $\mathbb{R}^p$ défini sur un espace probabilisé $(\Omega, \mathcal{A}, P)$. Ω est partitionné en q classes $\Omega_1, \ldots, \Omega_q$.

On se place sous les hypothèses du modèle probabiliste normal :

$$\mathcal{L}(X \mid \Omega_k) = \mathcal{N}(m_k, \Sigma_k), k = 1, \cdots, q \tag{7.6}$$

$$\Sigma_1 = \ldots = \Sigma_q. \tag{7.7}$$

Au pas l, on teste si l'introduction de la variable $X^{(l)}$ améliore la discrimination effectuée par les variables $X^{(1)}, \cdots, X^{(l-1)}$:

$$\begin{cases} H_0 : E[X^{(l)}|X^{(1)}, \cdots, X^{(l-1)}, \mathbb{1}_{\Omega_1}] = \cdots = E[X^{(l)}|X^{(1)}, \cdots, X^{(l-1)}, \mathbb{1}_{\Omega_q}] \\ H_1 : \overline{H}_0. \end{cases} \tag{7.8}$$

Soit les matrices d'inertie totale T_l (respectivement T_{l-1}) et intra W_l (resp. W_{l-1}) calculées en utilisant les caractères introduits au pas l (respectivement $l-1$).

On introduit la variable $X^{(l)}$ telle que le Lambda de Wilks $\Lambda_l = \frac{\det W_l}{\det T_l}$ soit minimal.

On considère alors la statistique de réalisation

$$\theta_l = \frac{\Lambda_l}{\Lambda_{l-1}} \tag{7.9}$$

Sous H_0, $\dfrac{1-\theta_l}{\theta_l} \dfrac{n-(l-1)-q}{q-1}$ est distribué suivant la loi de Fisher-Snedecor $\mathcal{F}(q-1, n-(l-1)-q)$.

Règle de décision : si $\dfrac{1-\theta_l}{\theta_l} \dfrac{n-(l-1)-q}{q-1} > f_{q-1,\, n-(l-1)-q}(\alpha)$, on rejette H_0 ; sinon on ne rejette pas H_0.

On arrête l'introduction d'une variable lorsque on ne rejette pas H_0.

Méthode progressive

On reprend l'algorithme de la méthode ascendante.

A chaque pas, on teste la capacité de discrimination de chacune des variables déjà introduites en présence des autres. On peut donc éliminer des variables introduites aux pas précédents.

Méthode descendante

On part de l'ensemble des variables $(x^1, \cdots, x^p)$.

A chaque pas, on ôte une variable telle que l'ensemble des variables restantes ait le plus grand pouvoir discriminant.

7.3 Méthodes de classement

On souhaite prédire la modalité de y pour un individu pour lequel on a mesuré les variables $x^1, \cdots, x^p$.

Nous allons présenter dans cette section les méthodes que nous avons utilisées dans la suite.

7.3.1 Règle de classement linéaire *(Linear Discriminant Analysis (LDA))*

Soit un nouvel individu sur lequel on observe $x^1, \cdots, x^p$ et que l'on représente par un point a de $\mathbb{R}^p$. On définit une distance de a à la classe I_k par

$$d(a, I_k) = \sqrt{(a - g_k)' M (a - g_k)}, \tag{7.10}$$

où g_k est le barycentre de la classe I_k et M une métrique. Dans R et SAS, la métrique utilisée par défaut est $(\frac{n}{n-q}W)^{-1}$, W étant la matrice d'inertie intra définie par $W = \Sigma_{k=1}^{q} W_k$, où W_k est la matrice d'inertie de la classe I_k.

L'individu a est alors affecté à la classe I_k telle que $d(a, I_k)$ soit minimale.

7.3.2 Règle de classement quadratique *(Quadratic Discriminant Analysis (QDA))*

On garde les mêmes notations. Le principe de la règle de classement quadratique est le même que celui de la règle linéaire, mais la métrique utilisée est spécifique de la classe considérée. Ceci permet de tenir compte de la dispersion de chacune des classes. On définit dans SAS la distance de a à la classe I_k par

$$d(a, I_k) = \sqrt{(a - g_k)' W_{1,k}^{-1} (a - g_k) + \ln \det W_{1,k}}, \tag{7.11}$$

où $W_{1,k} = \frac{n}{n_k - 1} W_k$.

Il existe d'autres définitions de la distance.

7.3.3 Les k plus proches voisins *(k-Nearest Neighbours (k-NN))*

Soit k un nombre entier positif fixé. Le principe de la méthode des k plus proches voisins est d'affecter l'individu a à la classe la plus représentée parmi les k individus les plus proches de a dans $\mathbb{R}^p$. La distance calculée entre l'individu a et un voisin z est donnée par

$$d(a,z) = \sqrt{(a-z)'M(a-z)}, \tag{7.12}$$

où M est une métrique. Notons que dans le logiciel SAS, la métrique utilisée par défaut est $(\frac{n}{n-q}W)^{-1}$. Dans R, la métrique utilisée est l'identité I_p.

7.3.4 La régression logistique

Supposons que les vecteurs des mesures des p variables quantitatives sur les n individus de I constituent un échantillon i.i.d. d'un vecteur aléatoire X dans $\mathbb{R}^p$ défini sur un espace probabilisé $(\Omega, \mathcal{A}, P)$.

Pour simplifier, plaçons-nous dans le cas $q = 2$. Ω est alors partitionné en deux classes Ω_1 et Ω_2.

Soit un individu dont la réalisation de X est x. On suppose qu'il existe des paramètres inconnus $\beta_0 \in \mathbb{R}$ et $\beta \in \mathbb{R}^p$ tels que :

$$P(\Omega_1|X=x) = \frac{e^{\beta_0+\beta'x}}{1+e^{\beta_0+\beta'x}} \Longleftrightarrow P(\Omega_2|\ X=x) = \frac{1}{1+e^{\beta_0+\beta'x}} \quad .$$

Les paramètres sont estimés par la méthode du maximum de vraisemblance. Des variables peuvent ensuite être retirées de l'étude si, à la suite d'un test, on ne rejette pas l'hypothèse de nullité du paramètre qui leur correspond. Pour cela, plusieurs tests existent, comme le test du rapport de vraisemblance, le test de Wald et le test du score [45].

Ayant estimé les probabilités $P(\Omega_i|X=x)$, on affecte l'individu à la classe de probabilité $P(\Omega_i|X=x)$ maximale (i.e. $P(\Omega_i|X=x) > \frac{1}{2}$ dans le cas de deux classes).

7.3.5 Segmentation par arbres de décision

Le principe de la segmentation est le suivant : pour expliquer y par $x^1, \cdots, x^p$, on partitionne I en classes définies à partir des valeurs de $x^1, \cdots, x^p$ qui sont hétérogènes entre elles relativement à y et homogènes à l'intérieur relativement à y.

On définit pour cela une distance Δ entre deux classes relativement à y. L'algorithme procède pas à pas.

Algorithme de la segmentation

$\underline{1^{er} \text{ pas}}$

a) Pour chaque variable quantitative x^j :

pour un nombre réel quelconque c, on considère la dichotomie $\left\{I_1^j(c), I_2^j(c)\right\}$ de I telle que : $I_1^j(c) = \{x^j < c\}, I_2^j(c) = \{x^j \geq c\}$;

pour chacune de ces dichotomies, on calcule la distance $\Delta(I_1^j(c), I_2^j(c))$;

on retient la dichotomie $\left\{I_1^{j*}, I_2^{j*}\right\}$ qui rend la distance Δ maximale.

b) On détermine ensuite parmi $x^1, \cdots, x^p$ une variable qui donne la plus grande valeur de $\Delta(I_1^{j*}, I_2^{j*})$; on retient la dichotomie $\{I_1, I_2\}$ de I correspondant à cette variable.

$2^{ème}$ pas

Pour chacune des classes I_1 et I_2 obtenues au premier pas, on effectue une dichotomie selon le principe défini précédemment.

$l^{ième}$ pas

On effectue une dichotomie de chacune des classes obtenues au pas précédent selon le même principe.

Règle d'arrêt

Pour les deux classes issues d'une dichotomie, on peut faire un test d'homogénéité des lois conditionnelles de y dans ces classes ; si l'on ne rejette pas l'hypothèse d'homogénéité, on ne procède pas à la dichotomie.

Un deuxième critère d'arrêt est basé sur la taille d'une classe. Un troisième fait appel à la notion de pureté d'une classe ; une classe est dite pure si une seule modalité de y est représentée dans cette classe. On ne peut pas procéder à une dichotomie d'une classe pure.

Distances entre classes

Plusieurs distances entre classes peuvent être utilisées : les distances utilisées en Classification Ascendante Hiérarchique, comme la distance de Ward, des distances basées sur une statistique de comparaison de distributions, comme la distance de Kolmogorov-Smirnov, ou des distances basées sur la notion d'impureté d'un segment, comme l'entropie de Shannon ou l'indice de Gini. Rappelons birèvement cette dernière distance.

On note I_l l'ensemble des individus de I ayant la modalité l de y. Soit un segment S. On note $P(I_l|S)$ la proportion d'éléments de I_l dans le segment S. On dit que le segment S est pur lorsqu'il existe l_0 tel que $P(I_{l_0}|S) = 1$.

L'indice de diversité de GINI est le nombre positif ou nul

$$i(S) = \sum_{l=1}^{q} \sum_{m \neq l} P\left(I_l \mid S\right) P\left(I_m \mid S\right) = 1 - \sum_{l=1}^{q} P^2\left(I_l \mid S\right).$$

Ainsi lorsque le segment S est pur $i(S) = 0$.

Soit la partition (S_1, S_2) de S. On note $p_1 = P(S_1|S)$ la proportion d'éléments de S_1 dans S et $p_2 = P(S_2|S) = 1 - p_1$ la proportion d'éléments de S_2 dans S.

On définit alors la distance Δ entre les classes S_1 et S_2 :

$$\Delta(S_1, S_2) = i(S) - \big(p_1\, i(S_1) + p_2\, i(S_2)\big).$$

Règle de classement

Soit un individu pour lequel on ne connaît pas la modalité de y mais pour lequel on a observé $x^1, \ldots, x^p$.

On détermine le segment terminal dans l'arbre construit auquel appartient cet individu.

Puisque la variable y est qualitative, on peut prédire la modalité de y pour cet individu par la modalité de fréquence maximale dans l'ensemble des individus de I appartenant à ce segment.

7.3.6 Les Support Vector Machine *(SVM)*

Supposons $q = 2$ et $y \in \{-1, 1\}$. Notons $I_1 = \{y = -1\}$ et $I_2 = \{y = 1\}$. On note également $x_i = (x_i^1, ..., x_i^p)$ le vecteur des mesures des p variables sur l'individu i. On dispose donc d'un ensemble de couples $\{(x_i, y_i)\}_{1 \leq i \leq n}$.

Le principe des SVM [46] consiste à chercher un hyperplan de séparation entre les ensembles $\{x_i, i \in I_1\}$ et $\{x_i, i \in I_2\}$, d'équation :

$$\beta' x + \beta_0 = 0, \quad \beta \in \mathbb{R}^p. \tag{7.13}$$

Dans le cas où l'hyperplan sépare parfaitement les deux classes I_1 et I_2, appelé cas séparable, on suppose que

$$\begin{cases} y_i = 1 \text{ si } \beta' x_i + \beta_0 > 0, \\ y_i = -1 \text{ si } \beta' x_i + \beta_0 < 0. \end{cases} \tag{7.14}$$

Soit un nouvel individu sur lequel on a mesuré x. L'individu est classé dans :

$$\begin{cases} I_1 \text{ si } \beta' x_i + \beta_0 < 0, \\ I_2 \text{ si } \beta' x_i + \beta_0 > 0. \end{cases} \tag{7.15}$$

Pour simplifier on se placera uniquement ici dans le cas séparable.
La distance d'un point x_i à l'hyperplan H d'équation $\beta' x + \beta_0 = 0$ est donnée par :

$$\frac{\mid \beta' x_i + \beta_0 \mid}{\parallel \beta \parallel}. \tag{7.16}$$

On détermine l'hyperplan H qui maximise la distance au point x_i le plus proche de H.
Soit d_m la distance minimale à l'hyperplan H ; on a :

$$\frac{\mid \beta' x_i + \beta_0 \mid}{\parallel \beta \parallel} \geq d_m, \quad i = 1, ..., n. \tag{7.17}$$

On remarque que $y_i (\beta' x_i + \beta_0) = \mid \beta' x_i + \beta_0 \mid$.
On cherche (β, β_0) qui forment une solution du problème :

$$\begin{cases} \max d_m \\ \text{sous les contraintes} \dfrac{y_i (\beta' x_i + \beta_0)}{\parallel \beta \parallel} \geq d_m, \quad i = 1, \ldots, n. \end{cases} \tag{7.18}$$

On constate que β et β_0 sont indéterminés à une constante multiplicative près. On va alors poser arbitrairement : $\parallel \beta \parallel = \frac{1}{d_m} \Leftrightarrow d_m = \frac{1}{\parallel \beta \parallel}$.
On cherche donc (β, β_0) qui forment une solution de :

$$\begin{cases} \min \parallel \beta \parallel \\ \text{sous les contraintes } y_i(\beta' x_i + \beta_0) \geq 1, \ i = 1, ..., n \end{cases} \tag{7.19}$$

ce qui est équivalent à :

$$\begin{cases} \min \dfrac{1}{2} \parallel \beta \parallel^2 \\ \text{sous les contraintes } y_i(\beta' x_i + \beta_0) \geq 1, \ i = 1, ..., n. \end{cases} \tag{7.20}$$

Il s'agit d'un problème d'optimisation convexe. On introduit alors les multiplicateurs de Lagrange $\alpha_i \geq 0$. Soit L la fonction de Lagrange définie par :

$$L(\beta, \beta_0; \alpha) = \frac{1}{2} \parallel \beta \parallel^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(\beta' x_i + \beta_0) - 1 \right] \tag{7.21}$$

où $\alpha = (\alpha_1, ..., \alpha_n)$.

On résout le système :

$$\begin{cases} \dfrac{\partial L}{\partial \beta} = \beta - \displaystyle\sum_{i=1}^{n} \alpha_i \, y_i \, x_i = 0 \\ \dfrac{\partial L}{\partial \beta_0} = - \displaystyle\sum_{i=1}^{n} \alpha_i \, y_i = 0 \end{cases} \tag{7.22}$$

En développant l'expression de L, on obtient :

$$\begin{array}{rcl} L(\beta, \beta_0; \alpha) & = & \frac{1}{2} \parallel \beta \parallel^2 - \sum_{i=1}^{n} \alpha_i \, y_i \, \beta' x_i + \sum_{i=1}^{n} \alpha_i \\ & = & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \, \alpha_j \, y_i \, y_j \, x_i' x_j \end{array} \tag{7.23}$$

Le problème dual est donc :

$$\begin{cases} \max \left(\displaystyle\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \, \alpha_j \, y_i \, y_j \, x_i' x_j \right) \\ \text{sous les contraintes} \\ \alpha_i \geq 0, \; i = 1, ..., n, \\ \displaystyle\sum_{i=1}^{n} \alpha_i \, y_i = 0, \\ \beta = \displaystyle\sum_{i=1}^{n} \alpha_i \, y_i \, x_i, \\ \alpha_i \left[y_i \left(\beta' x_i + \beta_0 \right) - 1 \right] = 0, \; i = 1, ..., n. \end{cases} \tag{7.24}$$

On appelle vecteur support tout vecteur x_i tel que $y_i \left(\beta' x_i + \beta_0 \right) = 1$. Remarquons qu'un vecteur support est donc un vecteur à distance minimale de l'hyperplan, c'est-à-dire situé sur un des deux bords de la marge. Pour un vecteur support, on a $\alpha_i \geq 0$, tandis que pour tout autre vecteur, $\alpha_i = 0$. Soit S l'ensemble des indices des vecteurs supports ($S \subset I$).

On est donc ramené à un problème de programmation quadratique. Soit une solution de ce problème (α_1^0, ..., α_n^0). On a alors :

$$\begin{cases} \beta = \displaystyle\sum_{i \in S} \alpha_i^0 \, y_i \, x_i \\ \beta_0 = \dfrac{1}{y_i} - \beta' x_i \text{ pour } i \in S. \end{cases} \tag{7.25}$$

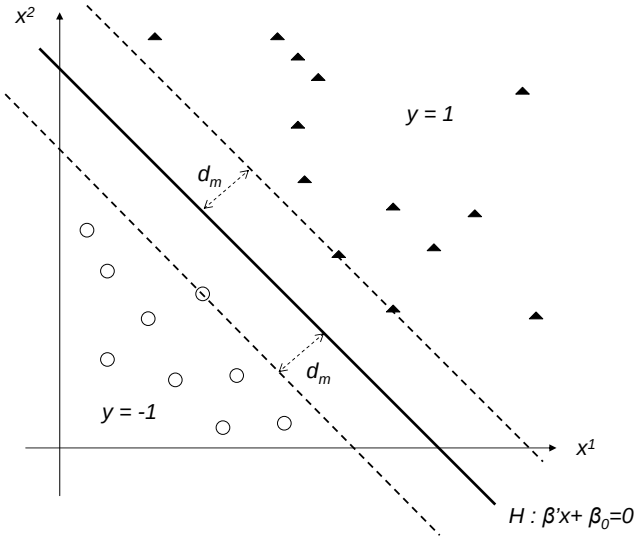


FIG. 7.1 – *Représentation des SVM en deux dimensions*

7.3.7 Les courbes ROC

Supposons que les vecteurs des mesures des variables forment un échantillon i.i.d. d'un vecteur aléatoire X dans $\mathbb{R}^p$ défini sur un espace probabilisé $(\Omega, \mathcal{A}, P)$.

Supposons que Ω est partitionné en deux sous-ensembles Ω_0 et Ω_1. On pose $p_0 = P(\Omega_0)$ et $p_1 = P(\Omega_1)$.

On souhaite affecter un individu ω à Ω_0 ou Ω_1.

Soit $x = X(\omega)$. On suppose dans ce paragraphe que l'on a déterminé une fonction score S définie sur $X(\Omega)$ à valeurs réelles et un seuil s appartenant à un ensemble $\mathcal{S}$, tels que l'individu est déclaré positif et classé dans Ω_1 si $S(x) \geq s$ et déclaré négatif et classé dans Ω_0 si $S(x) < s$.

On introduit donc la fonction de décision :

$$\psi(X(\omega)) = \psi(x) = \begin{cases} 1 \text{ si } S(x) \geq s \\ 0 \text{ si } S(x) < s. \end{cases} \tag{7.26}$$

On introduit également la fonction de perte L_ψ :

$$L_\psi(\omega) = \begin{cases} c_{1,0} \text{ si } \omega \in \Omega_0 \text{ et est classé dans } \Omega_1, \\ c_{0,1} \text{ si } \omega \in \Omega_1 \text{ et est classé dans } \Omega_0. \end{cases} \tag{7.27}$$

Son espérance est appelée fonction de risque,

$$E[L_\psi] = c_{0,1}(1 - E[\psi(X)|\Omega_1])p_1 + c_{1,0}E[\psi(X)|\Omega_0]p_0, \tag{7.28}$$

et représente le coût moyen d'erreur de classement.

On définit alors les grandeurs suivantes :

- la sensibilité $Se(s) = P(S \geq s|\Omega_1) = E[\psi(X)|\Omega_1]$, qui est la probabilité qu'a un individu de Ω_1 d'être déclaré positif,

- la spécificité $Sp(s) = P(S < s|\Omega_0) = 1 - E[\psi(X)|\Omega_0]$, qui est la probabilité qu'a un individu de Ω_0 d'être déclaré négatif,
- la valeur prédictive positive $VPP(s) = P(\Omega_1|S \geq s)$, qui est la probabilité qu'a un individu déclaré positif d'appartenir effectivement à Ω_1,
- la valeur prédictive négative $VPN(s) = P(\Omega_0|S < s)$, qui est la probabilité qu'a un individu déclaré négatif d'appartenir effectivement à Ω_0.

Notons que ces grandeurs dépendent du seuil s fixé a priori. On considère la courbe ensemble des points $(1 - Sp(s), Se(s))$ pour $s \in \mathcal{S}$. On l'appelle la courbe **ROC** (Receiver Operator Characteristic curve) [47]. Cette courbe est croissante et contenue dans le carré unité.

Choix de la fonction score

Donnons deux exemples de fonctions score lorsque X est unidimensionnel :

1. $S(x) = x$,
2. $S(x) = (\frac{x-m_1}{\sigma_1})^2 - (\frac{x-m_0}{\sigma_0})^2$, où m_0 et m_1 sont les moyennes de X dans Ω_0 et Ω_1 respectivement, et σ_0, σ_1 les écarts-types correspondants.

On donne dans la suite trois exemples de fonctions score lorsque X est multidimensionnel :

1. Munissons $\mathbb{R}^p$ d'une métrique M, et appelons a le premier facteur de l'ACP du vecteur X, qui appartient au dual $\mathbb{R}^{p*}$.

   On peut définir comme fonction score

   $$S(x) = x'a. \tag{7.29}$$

   Notons que le facteur a étant défini au signe près, on peut être amené à remplacer a par $-a$ dans l'expression précédente, afin que la fontion score S soit en adéquation avec la règle (7.26).
2. De même, en appelant cette fois a le premier facteur de l'AFD du vecteur X, on peut également utiliser comme fonction score $S(x) = x'a$.
3. Par la méthode des SVM, l'hyperplan de séparation optimal des classes Ω_0 et Ω_1 a pour équation dans le cas séparable :

   $$\beta' x + \beta_0 = 0, \quad \beta \in \mathbb{R}^p. \tag{7.30}$$

   On affecte alors un individu à l'une des deux classes en fonction de sa position par rapport à l'hyperplan :

   $$\psi(x) = \begin{cases} 1 \text{ si } \beta' x_i + \beta_0 > 0, \\ 0 \text{ si } \beta' x_i + \beta_0 < 0. \end{cases} \tag{7.31}$$

   On peut alors définir comme fonction score

   $$S(x) = \beta' x + \beta_0. \tag{7.32}$$

Critères de choix d'un seuil optimal

Une fois la courbe ROC tracée, on cherche un seuil s^* optimal au sens d'un certain critère. Donnons deux exemples.

1. **Minimisation de la fonction de risque**

   On cherche un seuil s^* qui minimise la fonction de risque

   $$E[L_\psi] = c_{0,1}(1 - Se(s))p_1 + c_{1,0}(1 - Sp(s))p_0, \tag{7.33}$$

   ce qui est équivalent à maximiser la fonction

   $$c_{0,1}Se(s)p_1 + c_{1,0}Sp(s)p_0. \tag{7.34}$$

2. **Minimisation de la distance au point (0,1)**

   Le point (0,1) représente le cas idéal où la sensibilité et la spécificité de la méthode valent toutes deux 1. On cherche le seuil pour lequel le point correspondant sur la courbe ROC est le plus proche du point (0,1), ce qui revient à minimiser la fonction

   $$(1 - Sp(s))^2 + (1 - Se(s))^2. \tag{7.35}$$

Estimations de la sensibilité, de la spécificité et des valeurs prédictives

On effectue un tirage au hasard de n individus de Ω.
On construit la matrice de confusion :

	$S < s$	$S \geq s$	Sommes
Ω_0	VN	FP	n_0
Ω_1	FN	VP	n_1
Sommes	m_0	m_1	n

où :
- VP est le nombre de vrais positifs,
- FP est le nombre de faux positifs,
- FN est le nombre de faux négatifs,
- VN est le nombre de vrais négatifs.

On estime :
- la sensibilité $Se(s)$ par $\frac{VP}{n_1}$,
- la spécificité $Sp(s)$ par $\frac{VN}{n_0}$,
- la valeur prédictive positive $VPP(s)$ par $\frac{VP}{m_1}$,
- la valeur prédictive négative VPN par $\frac{VN}{m_0}$.

7.4 Mesures de la qualité d'une règle de classement

On utilise dans la suite les deux méthodes suivantes.

7.4.1 Méthode de l'échantillon-test

On dispose de l'ensemble d'apprentissage I, divisé en classes $I_1, \cdots, I_q$. On suppose que I est un échantillon d'une population Ω divisée en classes $\Omega_1, \cdots, \Omega_q$.

On effectue un sondage par stratification dans I : on tire au hasard de chaque classe I_k des individus, en nombre proportionnel à card Ω_k, ou à défaut à card I_k ; ces individus constituent *l'échantillon-test*. On peut prendre jusqu'à 30% des individus de I.

Les individus non tirés constituent *l'échantillon de base*.

On construit une règle de classement à partir de l'échantillon de base. On applique cette règle aux individus de l'échantillon-test : on classe ces individus. Or on connaît leur classement réel. On calcule alors le *pourcentage d'individus bien classés*, qui est une mesure de la qualité de la règle de classement.

Remarque. On peut utiliser le même principe pour déterminer la valeur optimale de l'entier k dans la méthode des k plus proches voisins : on fait varier la valeur de k et on calcule pour chaque valeur le pourcentage de bien classés.

7.4.2 Validation croisée *(cross-validation)*

On utilise cette méthode lorsque le cardinal de l'ensemble d'apprentissage I est peu élevé.

On divise l'ensemble d'apprentissage I en m parties égales. On détermine la règle de classement à partir de la réunion de $(m-1)$ parties qui constitue l'échantillon de base à partir duquel on construit la règle de classement. L'échantillon-test est constitué de la $m^{\text{ième}}$ partie. On calcule le pourcentage de bien classés (% BC) dans l'échantillon-test. On répète ceci m fois, en changeant à chaque fois d'échantillon-test. On calcule le % BC global.

Cas particulier : le leave-one-out

Soit $n = \text{card } I$. On prend $m = n$.

On enlève un individu de I. On construit la règle de classement à partir des $(n-1)$ individus restants. On l'applique à l'individu ôté. On répète cette analyse n fois. On calcule le % BC.

Chapitre 8

Simplification du diagnostic de l'allergie à l'arachide

Nous allons nous intéresser dans ce chapitre à la discrimination des individus allergiques et des individus atopiques. Dans la première section, nous allons tout d'abord effectuer une analyse descriptive des 243 individus de Nancy et Lille à partir des 6 dosages immunologiques : les IgE totales et les IgE dirigées contre *f13* (*i.e.* l'arachide), *rAra-h1,rAra-h2,rAra-h3* et *rAra-h8*. Le modèle construit pour détecter l'allergie à l'arachide est ensuite exposé.

8.1 Etude descriptive des individus allergiques et des individus atopiques

8.1.1 Données cliniques et mesures de la sévérité

Le tableau des données cliniques est donné ci-dessous :

	Nancy		Lille	
	allergiques	atopiques	allergiques	atopiques
effectif	94	35	85	29
hommes	53	22	61	21
femmes	41	13	24	8
âge moyen ± sd	9 ± 4	9 ± 4	7 ± 3	7 ± 3
étendue	3 à 18	4 à 18	1 à 14	2 à 12

Tab. 8.1 – Tableau des données cliniques

Le score du TPO et la dose réactogène sont également renseignés pour les 179 patients allergiques (Tableaux 8.2 et 8.3). Enfin, le score du premier accident n'est disponible que pour 95 patients (55 à Nancy et 40 à Lille) (Tableau 8.4). L'allergie des 84 autres patients ayant été diagnostiquée lors d'un bilan allergologique et confirmée par un TPO, l'éviction a permis à ces enfants d'éviter les accidents.

Les distributions des scores de sévérité du TPO, du premier accident et de la dose réactogène des deux échantillons de patients sont indiquées respectivement dans les Tableaux 8.2, 8.3 et 8.4.

Notons que les distributions des scores du TPO et du premier accident sont significativement différents dans les centres ($p = 7.65 \times 10^{-5}$ et 0.010 respectivement avec le test exact de Fisher).

TPO	1	2	3	4	5
Nancy	22	32	12	26	2
Lille	43	23	13	6	0

TAB. 8.2 – Distribution du score du TPO dans les centres

DR	1.4	4.4	14	15	17	41	44	46	65	66	95	96
Lille	0	0	0	0	1	1	0	1	0	3	0	1
Nancy	1	2	1	2	0	0	11	0	6	0	1	0
DR	115	116	165	166	191	210	215	216	265	266	300	391
Lille	0	4	0	3	1	0	0	1	0	19	0	2
Nancy	2	0	2	0	0	1	19	0	2	0	1	0
DR	400	466	500	516	766	816	866	965	1000	1016	1266	1700
Lille	0	1	0	3	18	1	1	0	0	1	5	7
Nancy	1	0	23	0	0	0	0	6	1	0	0	0
DR	2000	2110	2450	2700	3500	3610	3700	4110	5200	6700	7000	11700
Lille	0	0	1	2	0	0	5	0	1	1	0	1
Nancy	3	1	0	0	1	1	0	1	0	0	5	0

TAB. 8.3 – Distribution de la dose réactogène dans les centres

premier accident	1	2	3	4	5
Nancy	9	20	7	17	2
Lille	9	22	7	2	0

TAB. 8.4 – Distribution du score du premier accident dans les centres

Les proportions de patients dont l'allergie est sévère, c'est-à-dire de score 4 ou 5, sont en effet plus importantes à Nancy qu'à Lille.

Notons que les populations de Nancy et de Lille ont des "profils allergéniques" comparables, en ce sens que les types d'allergies dans les deux zones géographiques sont équivalents. Ceci ne serait plus vrai si l'on considérait par exemple une population méditerranéenne, dans laquelle beaucoup d'individus sont sensibilisés à l'olivier et au cyprès, alors que dans le nord-est, l'allergène principal est le pollen de bouleau.

Dans la section suivante, nous allons réaliser une Analyse en Composantes Principales afin d'avoir une représentation graphique des données et d'étudier les corrélations entre les variables.

8.1.2 Analyse en Composantes Principales

Une Analyse en Composantes Principales (ACP) est réalisée sur les 243 individus avec les 6 variables quantitatives IgE totales, IgE spécifiques de *f13, rAra-h1,rAra-h2,rAra-h3* et *rAra-h8*, qui sont les seules variables disponibles à la fois pour les allergiques et les atopiques.

Les valeurs propres et les pourcentages d'inertie expliquée sont donnés dans le Tableau 8.5.

Il apparaît que le premier axe principal explique à lui seul 54.25% de l'inertie et le deuxième 19.22%.

Le premier facteur est fortement corrélé positivement avec les mesures des IgE dirigées contre *f13*, *rAra-h1*, *rAra-h2* et *rAra-h3* et très peu avec les autres mesures (Table 8.6). Ceci apparaît clairement sur le cercle des corrélations (Figure 8.2). Ces mesures sont également fortement corrélées entre elles. Le premier facteur est donc un facteur qui résume les IgE spécifiques de

facteur	valeurs propre	inertie	inertie cumulée
1	3.26	54	54
2	1.15	19	73
3	0.83	14	87
4	0.52	9	96
5	0.17	3	99
6	0.07	1	100

TAB. 8.5 – Valeurs et propres et pourcentages d'inertie de l'ACP des 243 individus

	facteur 1	facteur 2
IgE	0.12	0.76
f13	0.92	0.00
rAra-h1	0.93	-0.01
rAra-h2	0.94	-0.01
rAra-h3	0.78	0.02
rAra-h8	-0.12	0.76

TAB. 8.6 – Coefficients de corrélations linéaires entre les caractères et les deux premiers facteurs

f13, rAra-h1, rAra-h2 et *rAra-h3*. On constate que les allergiques, qu'ils soient de Lille ou de Nancy, sont répartis le long du premier axe (Figure 8.1). Les individus allergiques présentent des mesures élevées pour les IgE dirigées contre l'arachide (f13) et ont en général des valeurs relativement élevées des IgE spécifiques de *rAra-h1, rAra-h2* et *rAra-h3*. En revanche, les atopiques de Nancy et de Lille ont des valeurs faibles pour le premier facteur et tous les points sont concentrés dans la même zone. Ceci s'explique par le fait que les atopiques ont des valeurs faibles pour les mesures des IgE spécifiques de *f13, rAra-h1, rAra-h2* et *rAra-h3*. Le premier facteur permet donc de discriminer les allergiques des atopiques.

Le deuxième facteur de l'ACP est quant à lui fortement corrélé aux mesures des IgE spécifiques de *rAra-h8* et des IgE totales (Table 8.6 et Figure 8.2). Ces variables sont de plus très peu corrélées aux IgE spécifiques de *f13, rAra-h1, rAra-h2* et *rAra-h3*. Les individus allergiques ont majoritairement des valeurs faibles pour le deuxième facteur. Les atopiques sont quant à eux répartis tout le long du deuxième axe. Rappelons que de nombreux patients atopiques de notre échantillon sont allergiques au pollen de bouleau, dont un allergène majeur, *Bet-v1*, est homologue à *rAra-h8* (ceci signifie que leurs séquences d'acides aminés partagent un certain pourcentage d'idendité, 66% dans ce cas). De ce fait, un patient allergique au pollen de bouleau et non allergique à l'arachide peut présenter un taux élevé d'IgE spécifiques de *rAra-h8*. De même, *rAra-h8* étant une protéine contenue dans la cacahuète, ces patients peuvent également développer des IgE dirigées contre f13. Inversement, un patient allergique à l'arachide peut produire par réactivité croisée de grandes quantités d'anticorps dirigés contre *rAra-h8*. C'est pourquoi certains allergiques peuvent avoir des valeurs élevées pour le deuxième facteur. Ceci peut également être expliqué par une mesure des IgE totales relativement élevée.

Enfin, aucune différence particulière n'apparaît entre les individus des deux centres (Figure 8.1) sur les deux axes. En effet, les allergiques de Nancy et ceux de Lille ont des répartitions comparables, tout comme les atopiques.

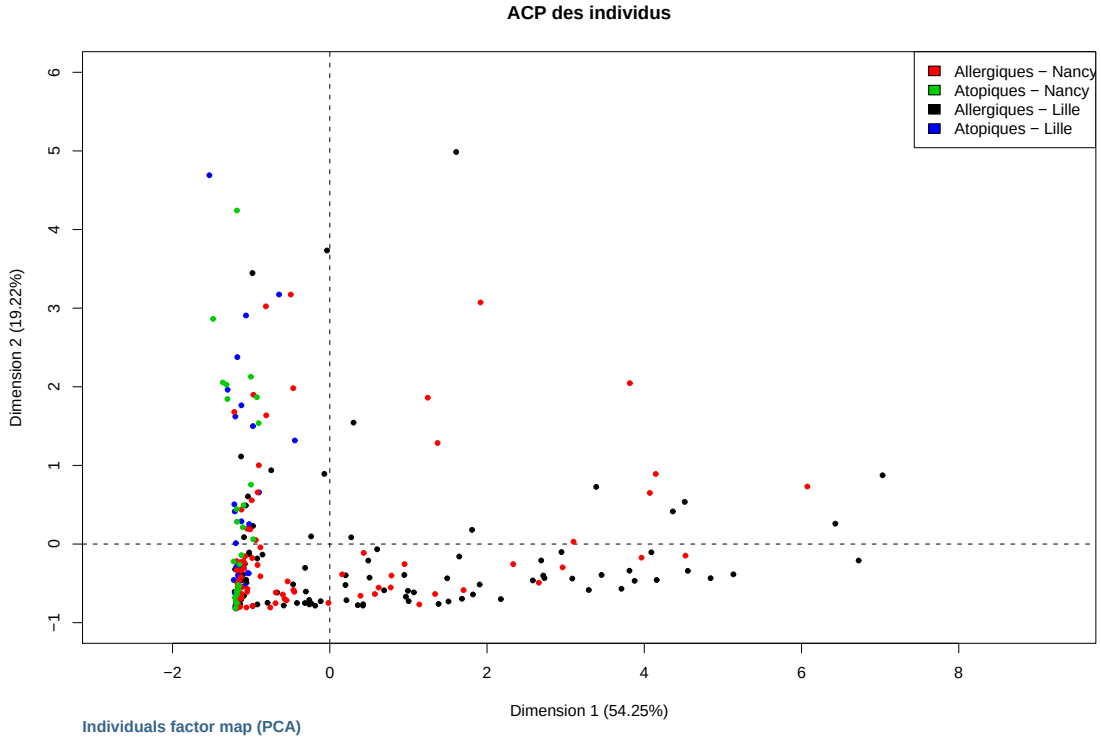


FIG. 8.1 – *Projection des individus sur le premier plan factoriel de l'ACP de tous les individus avec les dosages. Les allergiques de Nancy sont représentés en rouge, ceux de Lille en noir, les atopiques de Nancy en vert et ceux de Lille en bleu.*

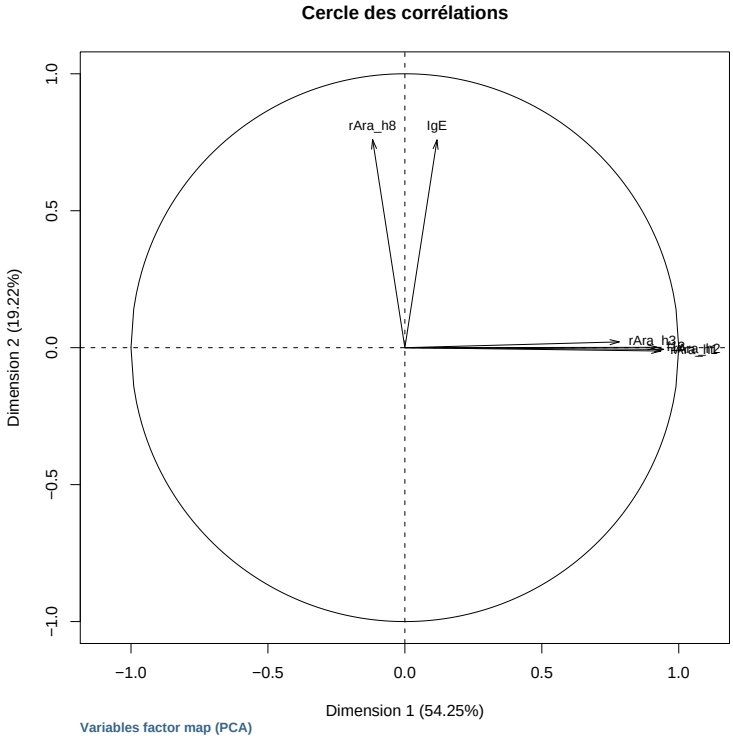


FIG. 8.2 – *Cercle des corrélations des dosages mesurés sur tous les individus*

8.2 Discrimination allergie / atopie à partir des dosages immunologiques

L'objectif de cette étude est de discriminer les 179 patients allergiques des 64 patients atopiques à l'aide des 6 dosages immunologiques mesurés. Les échantillons de Nancy et Lille sont pour cela réunis. La volonté des cliniciens est en effet de mettre en place un test clinique permettant de diagnostiquer l'allergie à l'arachide autrement que par un TPO. Détecter une allergie alimentaire serait ainsi réalisable avec une simple prise de sang et ne présenterait aucun risque pour les patients. Par ailleurs, le test mis en place devrait être le plus sensible et le plus spécifique possible.

8.2.1 Test de comparaison de moyennes

Intéressons-nous tout d'abord aux éventuelles différences de moyennes des IgE spécifiques de *f13* et *rAra-h1,2,3,8* entre les allergiques et les atopiques. Soit pour cela m_1 (resp. m_2) la moyenne de la variable aléatoire X_1 (resp. X_2) ayant pour réalisation la valeur du dosage de l'anticorps considéré chez les allergiques (resp. les atopiques).

Soit le test de comparaison de moyennes suivant :

$$\begin{cases} H_0 : m_1 \leq m_2 \\ H_1 : m_1 > m_2. \end{cases} \tag{8.1}$$

Rappelons brièvement le principe du test. Soit $(X_{1,1}, \ldots, X_{1,n_1})$ un échantillon *iid* de X_1 et $(X_{2,1}, \ldots, X_{2,n_2})$ un échantillon *iid* de X_2, supposés indépendants l'un de l'autre.

On note $\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{i,j}$ et $S_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i}(X_{i,j} - \bar{X}_i)^2$ la moyenne et de la variance de l'échantillon de $X_i, i = 1, 2$.

Alors

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}}}, \tag{8.2}$$

suit asymptotiquement la loi $N(0, 1)$ lorsque $m_1 = m_2$.

Règle de décision Soit t la réalisation observée de T et $p = P(T \geq t | H_0)$ la $p-$value du test. Si $p \leq \alpha$, on rejette H_0 au seuil α ; sinon on ne rejette pas H_0.

Les résultats du test sont donnés Table 8.7.

variable	p-value	allergiques	atopiques
IgE totales	0.91	663 ± 845 [11.5; 5000]	855 ± 1005 [19; 5000]
f13	9.40E-36	39.9 ± 39.4 [0.16; 100]	2.82 ± 4.11 [0; 18.05]
rAra-h1	1.12E-17	18.4 ± 29 [0; 100]	0.04 ± 0.14 [0; 1.02]
rAra-h2	1.80E-30	30.6 ± 35.8 [0; 100]	0.03 ± 0.08 [0; 0.43]
rAra-h3	5.07E-07	8.53 ± 22.4 [0; 100]	0.28 ± 1.86 [0; 14.09]
rAra-h8	0.99	3.43 ± 11.3 [0; 100]	10.9 ± 20.5 [0; 100]

TAB. 8.7 – Tests unilatéral de comparaisons de moyennes de chacun des dosages pour les allergiques et les atopiques. Les résultats sont donnés en termes de $p-$value, moyennes ± écarts-types et étendue dans chaque groupe.

Les IgE dirigées contre *f13* et *rAra-h1, rAra-h2, rAra-h3* ont des moyennes significativement plus élevées dans le groupe des patients allergiques que dans celui des patients atopiques, ce qui n'est pas le cas des IgE totales ni des IgE spécifiques de *rAra-h8*.

Conservons les mêmes notations et considérons maintenant le test

$$\begin{cases} H_0 : m_1 \geq m_2 \\ H_1 : m_1 < m_2. \end{cases} \tag{8.3}$$

On obtient alors pour les IgE dirigées contre *rAra-h8* une $p-$value de 0.003. De ce fait, on conclut que ces anticorps ont une moyenne significativement plus élevée chez les atopiques que chez les allergiques. Avec une p-value de 0.09, aucune différence de moyennes significative n'est observée pour les IgE totales.

Les IgE dirigées contre *f13, rAra-h1, rAra-h2, rAra-h3, rAra-h8* sont donc des variables discriminantes, avec lesquelles nous allons tenter dans la suite de discriminer les allergiques des atopiques.

8.2.2 Discrimination à l'aide du seuil de détection de la méthode

Comme expliqué dans le Chapitre 6, la limite de détection des immunoglobulines spécifiques lors de leur dosage est égale à 0.10 kU/L. Soit x le dosage des IgE dirigées contre *f13,rAra-h1, rAra-h2, rAra-h3* ou *rAra-h8*. On suppose que x est la réalisation d'une variable aléatoire réelle X.

Pour un dosage particulier, un patient est déclaré

$$\begin{cases} \text{positif si } x \geq 0.10 \text{ kU/L} \\ \text{négatif sinon} \end{cases}$$

Pour chaque dosage d'IgE spécifiques, on se propose dans un premier temps de voir si ce seuil de détection permet de discriminer les allergiques des atopiques, en classant comme allergique un patient positif et comme atopique un patient négatif. La sensibilité (Se), la spécificité (Sp), la valeur prédictive positive (VPP) et la valeur prédictive négative (VPN) sont données pour chacun des dosages dans la Table 8.8.

variable	Se	Sp	VPP	VPN
f13	100%	20%	78%	100%
rAra-h1	74%	84%	93%	54%
rAra-h2	95%	84%	94%	86%
rAra-h3	59%	84%	91%	43%
rAra-h8	40%	33%	63%	16%

TAB. 8.8 – Sensibilité (Se), spécificité (Sp), valeur prédictive positive (VPP) et valeur prédictive négative (VPN) obtenue pour chaque dosage avec le seuil canonique 0.10 kU/L

Comme expliqué précédemment tous les patients allergiques produisent des IgE contre l'arachide, ce qui explique la sensibilité de 100%. En revanche, sa spécificité est très faible (20%). La variable offrant le meilleur compromis entre sensibilité et spécificité est la mesure des IgE dirigées contre *rAra-h2*, qui permettent de détecter correctement 95% des allergiques et 84% des atopiques. Bien qu'étant tout aussi spécifiques, les IgE dirigées contre *rAra-h1* et *rAra-h3* sont nettement moins sensibles. Doser les IgE dirigées contre *rAra-h8* n'offre que des résultats très médiocres.

Bien que le dosage des IgE dirigées contre *rAra-h2* donne des résultats tout à fait satisfaisants, la règle de classement utilisée ne repose aucunement sur des estimations réalisées grâce à notre jeu de données. En effet, le seuil de positivité utilisé pour classer un individu comme allergique est le seuil de la méthode. Il est donc raisonnable de penser que ce seuil peut être optimisé en construisant une nouvelle règle de classement grâce à nos données.

On se propose dans la suite de déterminer un seuil optimal pour chaque variable.

8.2.3 Détermination d'un seuil optimal pour chaque variable

Le principe des courbes ROC (Receiving Operator Characteristics) est rappelé dans le Chapitre 7 .

Dans notre cas, Ω_0 est l'ensemble des individus atopiques et Ω_1 l'ensemble des individus allergiques. X a pour réalisation la valeur du dosage des IgE dirigées contre un des 5 allergènes d'intérêt : *f13, rAra-h1, rAra-h2, rAra-h3* ou *rAra-h8*.

Comme ces variables sont discriminantes (cf 8.2.1), la fonction score S utilisée pour chaque anticorps est l'identité. En utilisant le dosage des IgE spécifiques de *f13, rAra-h1, rAra-h2* ou *rAra-h3*, un individu est classé dans le groupe allergique si sa réalisation x est supérieure ou égale au seuil s car la valeur moyenne des allergiques est supérieure à la valeur moyenne des atopiques pour ces variables. Inversement, si les IgE dirigées contre *rAra-h8* sont considérées, l'individu est classé comme allergique si la valeur mesurée est en-dessous du seuil trouvé et atopique sinon. En effet, la quantité moyenne de ces IgE spécifiques est plus petite chez les individus allergiques que chez les atopiques.

La courbe ROC de chaque variable est donnée dans la Figure 8.3.

Le pouvoir discriminant de chaque variable est mesuré par l'aire sous la courbe ROC (AUC pour Area Under Curve). On constate que les dosages des IgE dirigées contre *f13, rAra-h1, rAra-h2, rAra-h3* semblent bien discriminer les deux populations, contrairement aux IgE dirigées contre *rAra-h8*.

Il faut maintenant déterminer un seuil de détection optimal pour chaque variable.

Notre objectif est d'obtenir la meilleure discrimination possible entre les allergiques et les atopiques, mais il est difficile en général de construire un test à la fois sensible et spécifique. Dans notre cas, il ne semble pas acceptable que notre test génère trop de faux négatifs, c'est-à-dire des patients dont on n'aurait pas détecté l'allergie. Ces individus, se pensant ainsi non allergiques, pourraient avoir une conduite à risque, potentiellement fatale pour certains. Pour autant, il n'est pas raisonnable de construire un test générant trop de faux positifs : un cas extrême assurant 100% de sensibilité serait de déclarer allergiques tous les patients testés. Il a donc été décidé de donner la même importance à la sensibilité et à la spécificité.

Le point (0,1) de la courbe ROC représente le cas idéal où la sensibilité et la spécificité du test valent toutes deux 100%. On cherche donc le seuil dont le point correspondant sur la courbe ROC est le plus proche du point (0,1). En notant $s_1, \ldots, s_n$ les différentes réalisations de l'allergène d'intérêt observées sur notre échantillon, cela revient à choisir comme critère de sélection du seuil optimal :

$$s^* = \arg \min_{s \in \{s_1, \ldots, s_n\}} ((1 - Sp(s))^2 + (1 - Se(s))^2), \tag{8.4}$$

Les résultats sont donnés dans la Table 8.9. On constate que l'utilisation du seuil déterminé par la courbe ROC fait passer la spécificité des IgE dirigées contre *rAra-h2* de 84% à 97% et la sensibilité de 95% à 92%.

Toutefois, déterminer le seuil de détection sur un échantillon et reclasser ensuite ce même échantillon peut conduire à une estimation trop optimiste de la sensibilité et de la spécificité. Il

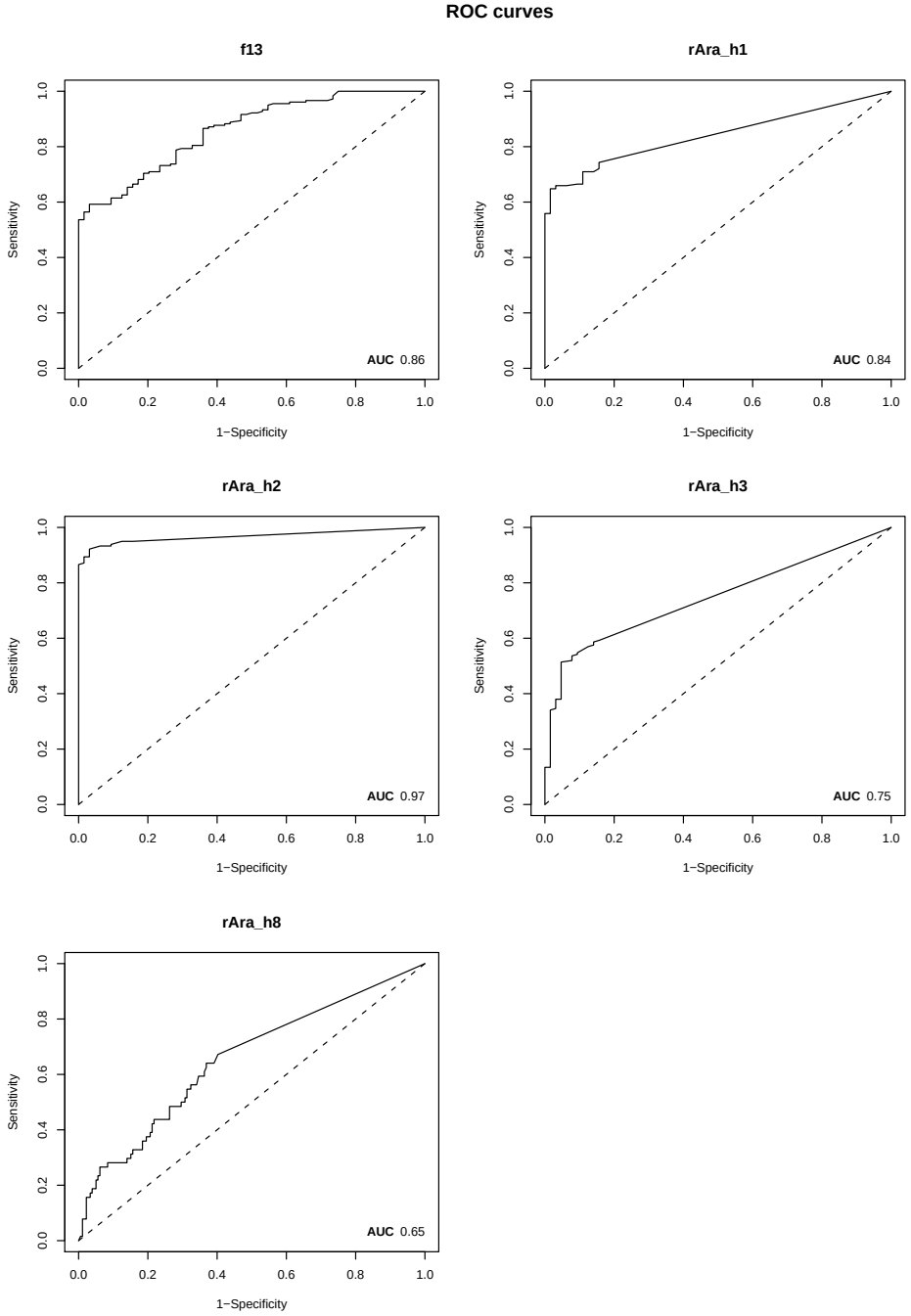


FIG. 8.3 – *Courbes ROC de chacun des dosages immunologiques*

Variable	Seuil	Se	Sp	VPP	VPN
f13	5.68	70%	81%	91%	50%
rAra-h1	0.10	74%	84%	93%	54%
rAra-h2	0.24	92%	97%	99%	82%
rAra-h3	0.11	59%	86%	92%	43%
rAra-h8	0.52	31%	50%	63%	21%

TAB. 8.9 – Sensibilité (Se), spécificité (Sp), valeur prédictive positive (VPP) et valeur prédictive négative (VPN) obtenues pour chaque dosage avec le seuil déterminé sur la courbe ROC

est donc nécessaire de mener une validation sur un échantillon indépendant. 25% des individus de l'échantillon de Nancy et Lille sont tirés aléatoirement afin de constituer un ensemble de test. Les individus restants constituent un ensemble d'apprentissage sur lequel est déterminé pour chaque dosage le seuil optimal défini par la relation (8.4). La sensibilité, la spécificité, la valeur prédictive positive et la valeur prédictive négative sont déterminées ensuite en classant l'échantillon test. Cette opération est répétée pour 100 tirages aléatoires différents de l'ensemble de test.

Les résultats sont présentés dans la Table 8.10. Pour chacun des dosages, la moyenne, l'écart-type et l'étendue de chaque paramètre sont déterminés sur les 100 tirages.

variable	seuil	Se	Sp	VPP	VPN
f13	4.49 ± 1.67	0.71 ± 0.08	0.74 ± 0.11	0.89 ± 0.05	0.48 ± 0.1
	1.45 - 7.18	0.53 - 0.95	0.44 - 1	0.76 - 1	0.25 - 0.86
rAra-h1	0.11 ± 0.01	0.72 ± 0.06	0.85 ± 0.07	0.93 ± 0.04	0.52 ± 0.08
	0.1 - 0.13	0.58 - 0.88	0.67 - 1	0.84 - 1	0.36 - 0.75
rAra-h2	0.23 ± 0.02	0.92 ± 0.03	0.95 ± 0.06	0.98 ± 0.02	0.81 ± 0.08
	0.14 - 0.25	0.83 - 1	0.71 - 1	0.85 - 1	0.6 - 1
rAra-h3	0.11 ± 0.01	0.57 ± 0.06	0.86 ± 0.08	0.92 ± 0.05	0.42 ± 0.07
	0.1 - 0.15	0.36 - 0.73	0.67 - 1	0.75 - 1	0.24 - 0.57
rAra-h8	0.15 ± 0.09	0.39 ± 0.06	0.39 ± 0.13	0.64 ± 0.09	0.18 ± 0.06
	0.1 - 0.93	0.22 - 0.51	0.06 - 0.74	0.45 - 0.87	0.03 - 0.34

TAB. 8.10 – Résultats de la validation par tirage d'échantillons-tests pour chaque dosage. La moyenne ± l'écart-type sont donnés sur la première ligne et le minimum et le maximum sur la seconde ligne pour chaque paramètre mesuré.

Pour les IgE spécifiques de *rAra-h1, rAra-h2* et *rAra-h3*, les seuils moyens trouvés par cette méthode sont proches de ceux déterminés précédemment. Il est également intéressant de noter que les écart-types des seuils de détection calculés sur les 100 tirages sont très faibles. Ceci indique que les seuils trouvés sont peu sensibles aux fluctuations d'échantillonnage.

C'est la mesure des IgE dirigées contre *rAra-h2* qui offre les meilleurs résultats, avec une sensibilité moyenne de 92% et une spécificité moyenne de 95%. En retenant le seuil moyen de 0.23 kU/L, ceci conduirait à une sensibilité de 93% et une spécificité de 95% sur la réunion des échantillons de Lille et Nancy. Par rapport au seuil de détection canonique de 0.10kU/L (Table 8.8), ceci permet d'augmenter la spécificité de 13%, au prix d'une perte de sensibilité de 2%.

Les autres dosages donnent des résultats moyens, voire médiocres pour les IgE dirigées contre *rAra-h8*.

Les résultats sont très satisfaisants et permettent de mettre en évidence l'intérêt de mesurer les IgE dirigées contre *rAra-h2* pour détecter l'allergie à l'arachide. Bien que moins important, le pouvoir discriminant des autres dosages n'est pas nul. Il est donc raisonnable de penser que combiner toutes ces variables permettrait éventuellement d'améliorer les résultats obtenus. Nous allons donc présenter dans la section suivante des modèles prédictifs construits à partir de tous les dosages discriminants.

8.2.4 Discrimination par l'ensemble des prédicteurs

Dans cette section, plusieurs règles de classement vont être mises en compétition afin de tenter d'améliorer encore la discrimination allergie / atopie : les Support Vector Machine (SVM), les arbres de décisions (CART pour Classification And Regression Trees), les règles de classement

linéaire et quadratique (LDA pour Linear Discriminant Analysis et QDA pour Quadratic Discriminant Analysis) et la régression logistique. Le principe de chacune de ces méthodes est rappelé dans le Chapitre 7.

Une validation croisée est réalisée pour chaque règle de classement : l'ensemble de l'échantillon est divisé aléatoirement en quatre quarts. Chaque quart est choisi tour à tour comme ensemble de test, tandis que le modèle prédictif est construit sur les trois autres quarts. Ainsi chaque individu de l'échantillon est classé une fois.

On utilise dans un premier modèle les 5 dosages discriminants comme prédicteurs : les IgE spécifiques de *f13, rAra-h1, rAra-h2, rAra-h3* et *rAra-h8*. On obtient alors les résultats suivants :

Méthode	Se	Sp	VPP	VPN
SVM	89%	97%	99%	77%
CART	91%	94%	98%	79%
LDA	56%	97%	98%	44%
QDA	91%	95%	98%	78%
Logistic	92%	58%	86%	73%

TAB. 8.11 – Résultats des validations croisées construites avec les IgE spécifiques de *f13, rAra-h1, rAra-h2, rAra-h3, rAra-h8*

On construit également un second modèle uniquement avec les IgE spécifiques de *rAra-h1, rAra-h2, rAra-h3* qui sont les variables ayant donné les meilleurs résultats dans les études précédentes.

Méthode	Se	Sp	VPP	VPN
SVM	90%	97%	99%	78%
CART	92%	92%	97%	80%
LDA	47%	100%	100%	40%
QDA	89%	94%	98%	76%
Logistic	78%	97%	99%	61%

TAB. 8.12 – Résultats des validations croisées construites avec les IgE spécifiques de *rAra-h1, rAra-h2, rAra-h3*

Aucun des modèles construits n'améliore sensiblement les résultats obtenus précédemment. Seuls les SVM permettent d'atteindre 97% de spécificité pour les deux listes de variables explicatives, au prix d'une légère perte de sensibilité. Notons par ailleurs la piètre sensibilité obtenue avec la règle de classement linéaire.

Ajouter les autres dosages immunologiques à *rAra-h2* et utiliser ces règles de classement ne permet pas d'améliorer sensiblement le diagnostic de l'allergie à l'arachide. Cette remarque est confortée par le fait que, lorsque l'on utilise la méthode CART, seule la mesure des IgE spécifiques de *rAra-h2* est sélectionnée.

8.3 Conclusion et perspectives

La principale conclusion de cette étude est que le dosage immunologique des anticorps dirigés contre les allergènes recombinants de l'arachide *rAra-h1, rAra-h2, rAra-3* permet d'obtenir une meilleure spécificité qu'en mesurant les IgE dirigées contre *f13*. Comme expliqué précédemment, ce test, qui était le seul dosage immunologique utilisé jusqu'à présent, est très sensible mais

peu spécifique [34]. Or il a été montré dans cette étude que les IgE dirigées contre *rAra-h2* permettaient de discriminer à elles seules les allergiques des atopiques avec une sensibilité de 93% et une spécificité de 95%. D'autres auteurs avaient déja dosé les IgE dirigées contre *rAra-h1, rAra-h2* et *rAra-h3* sur un plus petit échantillon de patients (30 allergiques et 30 atopiques), obtenant des sensibilités de 50%, 100% et 20% respectivement, avec le seuil de 0.10kU/L [32]. Dans cette étude, la spécificité de chacune de ces trois variables était égale à 100%, car aucun individu contrôle n'avait réagi. Tout comme dans notre étude, les patients étaient donc majoritairement sensibilisés à *rAra-h2* puis à *rAra-h1* et *rAra-h3*. Bien que dosés par ELISA, les différences de sensibilité observées des anticorps peuvent être dues au choix du substrat, qui n'est pas le même dans les deux études. De plus l'échantillon étudié ici contient plus de patients dont l'allergie est grave (22% contre 10%). Or nous montrerons dans la suite que la quantité d'anticorps spécifiques présente dans le sang peut augmenter avec la sévérité de l'allergie.

Doser les IgE spécifiques de *rAra-h2* va donc permettre de détecter de manière simple, efficace et sans danger une allergie à l'arachide. Il est conseillé de continuer à mesurer les IgE dirigées contre *f13* en raison de leur grande sensibilité. Le TPO ne devrait dès lors plus être utilisé comme un outil diagnostique. En revanche, il reste utile pour vérifier si l'allergie d'un patient a disparu avec le temps ou pour connaître la dose réactogène d'un individu allergique que l'on veut désensibiliser. De plus, pour les rares patients ne présentant aucun dosage immunologique positif mais dont l'histoire clinique laisse présager un risque d'allergie à l'arachide, un TPO peut être envisagé. Enfin, notons que d'autres allergènes de l'arachide existent, comme *rArah-6, rArah-7* et *rArah-9*, qui pourraient contribuer au diagnostic de l'allergie.

Chapitre 9

Prédiction de la sévérité de l'allergie à l'arachide : résumé

9.1 Introduction

La sévérité de l'allergie à l'arachide peut varier fortement d'un individu à l'autre. En cas d'ingestion accidentelle d'arachide, certains patients allergiques peuvent par exemple développer un léger urticaire, tandis que d'autres peuvent risquer la mort. De ce fait, il est fondamental de connaître la sévérité de l'allergie d'un patient afin d'évaluer les risques qu'il encourt. Comme expliqué dans le Chapitre 6, deux précieuses informations sont apportées par le TPO : le score de sévérité de la réaction ainsi que l'intervalle contenant la dose réactogène. De plus, le score du premier accident, évalué *a posteriori* selon la même grille de symptômes que le TPO à partir du dossier médical du patient, est également un renseignement utile. Le TPO pouvant s'avérer dangereux pour le patient, il serait utile de concevoir un test prédisant la sévérité de l'allergie du patient à l'aide de variables faciles à mesurer sans TPO. Dans cette partie sera présentée l'étude prédictive de la sévérité de l'allergie à l'arachide. Pour cela différentes analyses du score du TPO, du score du premier accident et de la dose réactogène ont été réalisées.

Cette étude a été réalisée uniquement sur les patients allergiques de Nancy afin d'utiliser également les prick tests comme variables explicatives.

On dispose ainsi de 36 variables explicatives, 6 dosages immunologiques et 30 prick tests, mesurées sur 93 individus allergiques (vior la remarque 1).

NB : L'échantillon de données disponibles ainsi que la nature des variables mesurées ont été présentés en détail dans le chapitre 6.

Remarques :

1. Un patient "aberrant" a été retiré des 94 patients de départ en raison de son profil atypique. En effet, cet individu était l'unique patient ayant un dosage immunologique des IgE dirigées contre *f13* positif, mais sans autres IgE spécifiques détectées. Or seuls des patients atopiques ont ces caractéristiques dans notre échantillon. Plusieurs dosages des anticorps de cet individu ont conduit aux mêmes observations et ont écarté les erreurs de mesures. L'allergie de cet individu peut être due à une sensibilité à d'autres allergènes de l'arachide que ceux mesurés dans cette étude.
2. Sur le conseil des cliniciens, les prick tests au *pois blond*, au *caroube*, au *néré* et à la *noix de Nangaille* n'ont pas été pris en compte, car ils sont très peu utilisés dans un bilan allergologique.

9.2 Approche statistique

Notre objectif est de prédire à l'aide des 36 variables mesurées la modalité de chacune des trois variables suivantes : le score du TPO, le score du premier accident, l'intervalle de la dose réactogène. La dose réactogène étant au départ un caractère quantitatif dont les valeurs administrées sont fixées par paliers, un prétraitement - détaillé plus loin - est proposé afin de regrouper les intervalles de valeurs en classes et de sélectionner des variables discriminantes.

9.2.1 Les scores du TPO et du premier accident

Les scores du TPO et du premier accident sont répartis soit en quatre classes de sévérité 1,2,3, $\{4, 5\}$ (à cause du faible effectif de la classe 5), soit en deux classes $\{1, 2, 3\}$ et $\{4, 5\}$. Notre approche statistique est identique pour les scores du TPO et du premier accident et se décompose, dans le cas de quatre classes ou de deux, en deux phases :

1. sélection de caractères discriminants parmi les 36 mesurés, grâce au test de Kruskal-Wallis [24] et à la sélection pas-à-pas progressive par le critère du lambda de Wilks [48],
2. mise en compétition de plusieurs méthodes de classement testées par validation croisée : LDA, QDA, k-NN, CART, AdaBoost [49, 50] (dans le cas de deux classes uniquement).

Remarque : Le principe de chacune de ces méthodes a été rappelé dans le Chapitre 7.

Pour les scores du TPO et du premier accident, le nombre de prick tests discriminants peut être beaucoup plus grand que le nombre de dosages retenus. Dans ce cas, la contribution des dosages au modèle risque d'être supplantée par celle des prick tests. Or les cliniciens souhaitent que les dosages gardent une contribution importante au modèle car :

1. les dosages sont des mesures plus précises et plus simples à réaliser que les prick tests,
2. nous avons montré que les dosages immunologiques dont nous disposons permettent un bon diagnostic de l'allergie à l'arachide (Chapitre 8) ; il est donc raisonnable de penser qu'ils puissent également jouer un rôle dans la prédiction de la sévérité.

Pour remédier à ceci, nous proposons de travailler sur un nombre réduit de facteurs, déterminés à partir des variables discriminantes du score du premier accident ou du TPO par une Analyse Factorielle Multiple (AFM) [51] ; cette méthode d'analyse des données permet d'équilibrer l'influence de groupes de variables dans la détermination des facteurs.

Deux études sont réalisées : l'une en utilisant les variables, l'autre en utilisant les facteurs. La démarche statistique suivie peut être résumée par le schéma de la Figure 9.1.

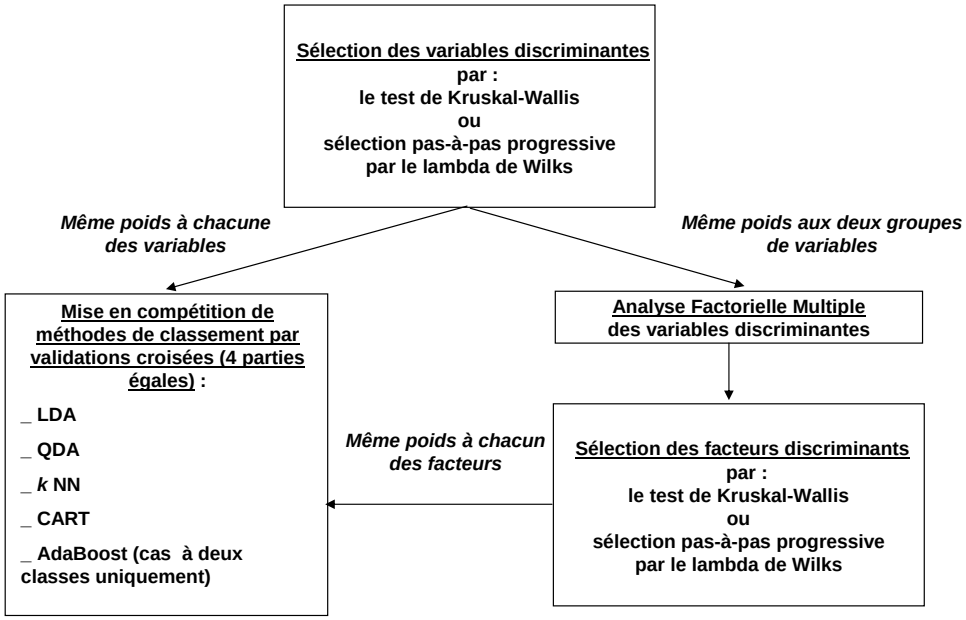


FIG. 9.1 – *Approche statistique pour la prédictions des scores du TPO et du premier accident*

9.2.2 La dose réactogène

La dose réactogène est un caractère quantitatif dont les valeurs administrées sont fixées par paliers. On souhaite regrouper les intervalles de la dose réactogène en un nombre réduit d'intervalles, tout en sélectionnant les variables qui discriminent au mieux ces classes. Pour cela, nous proposons un algorithme qui procède par optimisation alternée. Le critère d'optimalité choisi est le Λ de Wilks.

L'algorithme procède ainsi :

- Pas 1 :
  1. on cherche la partition en classes $\mathcal{C}^1$ des intervalles de la dose réactogène qui minimise Λ, calculé avec tous les prédicteurs (*i.e.* variables ou facteurs) disponibles ;
  2. on choisit le prédicteur v^1 qui minimise Λ correspondant à la partition $\mathcal{C}^1$ précédemment trouvée ;
- Pas 2 :
  1. on cherche la partition en classes $\mathcal{C}^2$ des intervalles de la dose réactogène qui minimise Λ, calculé avec le prédicteur v^1 précédemment trouvé ;
  2. on choisit le prédicteur v^2 tel que le couple de prédicteurs (v^1, v^2) minimise Λ calculé avec la partition $\mathcal{C}^2$;
- et ainsi de suite ... .

– la procédure s'arrête si aucun des prédicteurs restants ne peut améliorer le pouvoir discriminant du modèle, *i.e.*, si la $p-$value de la statistique F d'entrée est plus grande que 0.15 [48], ou si tous les prédicteurs ont déja été sélectionnés.

Des exemples d'utilisation de cet algorithme sont donnés dans le Chapitre 11.

La taille relativement faible de l'échantillon ($n = 93$) ne permet pas de répartir les individus en un grand nombre de classes. De plus, si le nombre de classes est élevé, le nombre de partitions à étudier à chaque pas peut être grand ; en outre, la qualité de la discrimination peut être mauvaise. L'algorithme a été utilisé en fixant le nombre de classes à 4 puis à 2, par analogie avec les scores du TPO et du premier accident. L'algorithme a été appliqué sur les 36 variables puis sur les 36 facteurs de l'AFM correspondants, calculés à partir de toutes les variables disponibles. En effet, il n'existe pas un ensemble "canonique" de caractères discriminants à partir duquel réaliser l'AFM, comme pour les scores du TPO et du premier accident, car le choix des variables discriminantes est lié au choix de la partition et inversement.

Une fois les variables et la partition déterminées, on procède à la comparaison des mêmes méthodes de classement que précédemment par validation croisée. Le schéma de l'étude est représenté dans la Figure 9.2 :

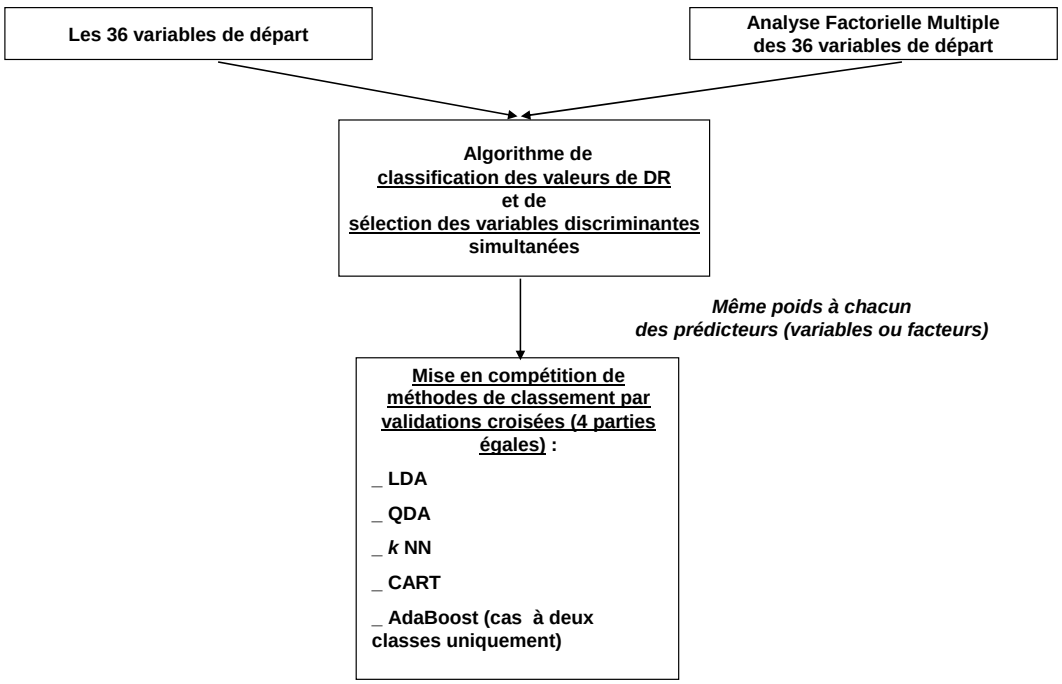


FIG. 9.2 – *Approche statistique pour la prédiction de la dose réactogène*

9.3 Résultats

Dans le tableau suivant est présentée pour chacune des études réalisées la règle de classement offrant le meilleur pourcentage de bien classés. Le pourcentage d'individus dont l'allergie est sévère qui sont bien classés est également indiqué : il s'agit des individus ayant un score 4 ou 5 pour le TPO ou le premier accident, ou les patients appartenant à la classe des doses réactogènes les plus faibles. En effet, un test efficace doit détecter un maximum de patients dont l'allergie est grave, afin d'éviter à ces derniers de suivre une conduite à risque.

mesure de la sévérité	nb de classes	variables	facteurs
premier accident	4	LDA : 50%-26%	3NN : 55%-23%
	2	LDA : 83%-74%	3NN : 82%-64%
TPO	4	LDA : 46%-61%	x
	2	1NN : 72%-66%	x
dose réactogène	4	LDA : 38%-38%	LDA : 40%-73%
	2	5NN : 66%-73%	CART : 82%-85%

TAB. 9.1 – Résumé des résultats obtenus par les règles de classement

On constate tout d'abord que les résultats obtenus pour l'étude en 4 classes sont médiocres en général.

En deux classes, les variables retenues sont pour le 1er accident : les IgE dirigées contre *rAra-h1,2,3* et *poils de chien, noix, pécan, arachide, lupin*. Utiliser ces variables avec la LDA conduit à 83% de bien classés en deux classes.

Pour le TPO, le meilleur pourcentage de bien classés (72%) est obtenu en deux classes avec la méthode du $1 - NN$, en utilisant les variables : *lupin, lentille, blatte, 12 graminées, frêne*. Notons que l'AFM n'a pas été réalisée dans le cas du TPO. La raison en est donnée dans le Chapitre 10.

Enfin, discriminer la dose réactogène avec les facteurs *13, 10, 32, 18, 24* calculés sur les 36 variables disponibles, conduit à 82% de bien-classés en deux classes (dose réactogène ≤ 500 *mg* et dose réactogène > 500 *mg*). Notons également que 85% des individus ayant une allergie sévère (c'est-à-dire réagissant au plus à 500 *mg*) sont bien classés.

9.4 Conclusions et perspectives

Les analyses discriminantes de trois mesures de la sévérité de l'allergie à l'arachide proposées ici sont les premières décrites dans la littérature à notre connaissance. En effet, bien que le but d'études précédentes était de trouver des liens entre les dosages immunologiques ou les prick tests et la sévérité de l'allergie à l'arachide, les analyses statistiques étaient principalement réduites à des tests de comparaison de lois des variables dans deux groupes de patients (en utilisant par exemple le test de Mann-Whitney), ou à des calculs de coefficients de corrélation linéaire [35, 36]. L'utilisation de plusieurs règles de classement nous a permis de comparer les résultats et de choisir la méthode la plus efficace (Table 9.1). De plus, l'utilisation des facteurs d'une AFM comme prédicteurs est une solution pour équilibrer les poids des groupes de variables. Par ailleurs, nous proposons un nouvel algorithme permettant de regrouper les modalités d'une variable qualitative ordinale (dans notre cas, la dose réactogène) et de sélectionner simultanément les variables qui discriminent cette partition.

Les études [35, 36] avaient été réalisées sur des échantillons de faible effectif, *i.e.*, de 30 à 40 patients. Par ailleurs, le nombre de variables mesurées était plus petit que dans notre étude,

limitant de ce fait le choix des prédicteurs. Dans [35], les IgE spécifiques de *rAra-h1,2,3* avaient été mesurées par prick tests plutôt que par dosages immunologiques. Bien qu'une réponse positive à un prick test permette de prouver la sensibilité à un allergène donné, sa mesure est moins précise que celle d'un dosage immunologique.

D'autres scores existent dans la littérature pour évaluer la sévérité de l'allergie à l'arachide. Par exemple, Hourihane *et al.* ont mis au point un score complexe combinant la gravité des symptômes observés avec la dose réactogène et comprenant plus de 25 classes [36]. Une graduation des symptômes a également été proposée par Müller [37] et utilisée dans plusieurs articles, mais ce score est basé sur des réactions induites par une sensibilisation au venin de guêpe ou d'abeille, ce qui ne nous a pas semblé approprié à l'allergie à l'arachide. Une standardisation du score utilisé est donc nécessaire et faciliterait les études multi-centriques et les comparaisons de résultats. Ce problème pourrait être résolu en comparant les résultats d'analyses discriminantes du score proposé par Hourihane avec ceux du score utilisé dans cette étude, réalisées sur les mêmes patients. De plus, le nombre de classes de ce score pourrait être réduit préalablement par notre algorithme. Donnons encore trois autres références de scores de sévérité [38, 39, 40].

Néanmoins, un certain nombre de biais doivent être notés. Tout d'abord, les diamètres des prick tests sont des données relativement imprécises. Il n'existe pas de standardisation de la mesure de la réaction observée, qui est également parfois représentée par l'aire de la papule [35]. De plus, le score du premier accident peut lui même être incomplet ou imprécis. En effet, le dossier médical est conçu lorsque le patient consulte pour la première fois l'allergologue et lui fait part des réactions apparues après l'ingestion accidentelle d'arachide. Bien que certains patients aient parfois été soignés au moment de l'accident par un médecin et connaissent donc précisément les symptômes apparus, le score du premier accident dépend majoritairement des souvenirs du patient, qui peut minimiser ou oublier certains symptômes. Des cofacteurs comme la prise de médicaments peuvent aggraver les symptômes au moment de l'accident [52, 53]. Dans ce cas, le score de sévérité attribué au patient est plus élevé que ce qu'il devrait être. Enfin, le premier accident est un événement passé dont le score est prédit par des variables mesurées plus tard au moment du TPO. Hourihane et al. avaient cependant également établi un *community score*, qui est évalué *a posteriori* avec le dossier médical du patient [36], selon la même grille de symptômes que leur score de sévérité du TPO.

Les modèles prédictifs des scores du premier accident et du TPO définis dans notre étude atteignent respectivement 83% et 72% de bien-classés en deux classes. De plus, notre algorithme nous a permis de grouper les intervalles de la dose réactogène en classes et de choisir simultanément les variables qui les discriminent le mieux, offrant ainsi 82% de bien-classés en deux classes. Comme tous les dosages des *IgE* spécifiques ont été sélectionnés au moins une fois dans un modèle, il est raisonnable de penser que mesurer de nouveaux anticorps d'allergènes de l'arachide, comme ceux dirigés contre *rArah-6, rArah-7* et *rArah-9*, pourrait améliorer le pouvoir discriminant de nos modèles. Ceci renforce également l'argument que les anticorps mesurés dans cette étude jouent un grand rôle dans le diagnostic de la sévérité de l'allergie à l'arachide. Par ailleurs, des tests cutanés comme le *lupin* dont des réactivités croisées avec l'arachide ont été précédemment mises en évidence [54], ont été retenus dans nos modèles, indiquant que nos résultats sont en accord avec les récentes découvertes médicales. Notre sélection de variables offre de nouvelles perspectives pour les bilans allergologiques. En effet, des variables inattendues ont été sélectionnées plusieurs fois, comme les prick tests à la *blatte* ou au *frêne*. Si des expériences biologiques venaient à distinguer le lien réel de ces tests cutanés avec l'allergie à l'arachide d'une réactivité croisée, l'importance de ces variables dans le diagnostic de la sévérité de l'allergie à l'arachide serait ainsi confirmée. Certains autres prick tests ne devraient d'ailleurs plus être réalisés lors du TPO pour prédire la sévérité de l'allergie à l'arachide, car n'ayant jamais été sélectionnés, ils semblent ne présenter aucun lien avec l'allergie à l'arachide, comme le *latex* ou la *noix du Brésil*.

Les modèles discriminants décrits dans cette étude sont un premier pas vers un outil de diagnostic simple, efficace et sans danger de la sévérité de l'allergie à l'arachide par la mise en évidence de la présence d'anticorps. Avant d'être appliqués par les cliniciens, ils devront d'abord être validés sur un ensemble indépendant de patients. De nouvelles variables explicatives devraient également être ajoutées afin d'améliorer les pourcentages de bien-classés. Ces modèles pourraient ainsi devenir des outils pratiques pour les cliniciens. Pour prédire la sévérité, les résultats des tests cliniques pourraient par exemple être soumis en ligne sur le site du réseau d'allergovigilance *http ://www.cicbaa.com/*), ou bien un logiciel simple pourrait être programmé et distribué gratuitement.

Chapitre 10

Discriminant analyses of peanut allergy severity scores

O.Collignon[(1),(2)], J.-M.Monnez[(2)], P.Vallois[(2)], F.Codreanu[(3)], J.-M.Renaudin[(1)], G.Kanny[(3)], M.Brulliard[(1)], V.Harter[(1)], B.E.Bihain[(1)], S.Jacquenet[(1)], D.Moneret-Vautrin[(3)]

(1) Genclis SAS, 15 rue du Bois de la Champelle, 54500 Vandoeuvre-lès-Nancy, France
(2) Institut Elie Cartan, UMR 7502, Nancy Université, CNRS, INRIA, BP239, 54506, Vandoeuvre-lès-Nancy, France
(3) Centre Hospitalier Universitaire, Service d'allergologie, 29 av. Mar De Lattre de Tassigny, 54000 Nancy, France

Abstract : Peanut allergy is one of the most prevalent food allergies. The possibility of a lethal accidental exposure and the persistence of the disease make it a public health problem. Evaluating the intensity of symptoms is accomplished with a double blind placebo controlled food challenge (DBPCFC), which scores the severity of reactions and measures the dose of peanut that elicits the first reaction. Since DBPCFC can result in life-threatening responses, we propose an alternate procedure with the long term goal of replacing invasive allergy tests. Discriminant analyses of DBPCFC score, the eliciting dose and the first accidental exposure score were performed in 93 allergic patients using 6 immunoassays and 30 skin prick tests. A Multiple Factorial Analysis was performed to assign equal weights to both groups of variables, predictive models were built by cross-validation with linear and quadratic discriminant analyses, k-NN, CART, and AdaBoost methods. We also developed an algorithm for simultaneously clustering eliciting dose values and selecting discriminant variables. Our main conclusion is that antibody measurements do offer information on the allergy severity, especially those directed against *rAra-h1* and *rAra-h3*. Further independent validation of these results and the use of new predictors will help extend this study to clinical practices.

KEY WORDS : discriminant analysis, peanut allergy, DBPCFC, Multiple Factorial Analysis, classification, variable selection.

10.1 Introduction

An allergy is an abnormal reaction of the immune system towards foreign substances (allergens) that are normally harmless. Peanut allergies in particular affect more than 0.5% of the entire French population, and its increasing prevalence and potentially severe clinical reactions make it a public health problem. It is also the most lethal food allergy [31]. Following a strict avoidance diet is currently the only effective treatment that minimises potentially lethal accidents.

Diagnosing and scoring peanut allergies is currently performed with a *double blind placebo controlled food challenge (DBPCFC)* [33]. Patients are given increasing peanut doses until the first clinical reaction appears. Those showing specific allergy symptoms are declared allergic, and a particular avoidance treatment is then initiated. DBPCFC is also used to judge the severity of an established peanut allergy by determining the cumulative dose that triggers the first reaction, known as the *eliciting dose* [in *milligrams (mg)*]. However, these tests require patient hospitalisation in specialised centers and can potentially result in life-threatening reactions from patients with severe allergies. The DBPCFC is also a costly and time consuming test to conduct.

The severity of peanut allergies is usually scored using the following scale [32] :

- **Score 1 :** Mild symptoms among : abdominal pains that spontaneously resolve under 30 minutes and/or rhinocunjunctivitis and/or urticaria < 10 papulas and/or a rash (eczema onset) ;
- **Score 2 :** One moderate symptom among : abdominal pain requiring treatment or generalized urticaria or non-laryngeal angioedema or cough or fall of Peak Expiratory Flow between 15 and 20% ;
- **Score 3 :** Two moderate symptoms in the preceding list ;
- **Score 4 :** Three moderate symptoms in the preceding list or laryngeal oedema or hypotension or asthma requiring treatment ;
- **Score 5 :** Any symptom requiring hospitalisation in intensive care.

For an already diagnosed allergy, it would be much more advantageous to predict the severity of the reaction from accidental exposure by using a blood sample or cutaneous test. This would replace the DBPCFC test with a simple statistical tool that can still evaluate potential risk without exposing the patient to a life-threatening allergic situation. Such a diagnostic method would be a major advance in food allergies and be beneficial to both patients and clinicians.

The main goal of this study was to predict the DBPCFC score, the eliciting dose and the first accidental exposure score, evaluated *a posteriori* with the patient's medical record according to the same scale as the DBPCFC score. The first accidental exposure score would then reveal the "real" severity of the allergy. Compare this to using the DBPCFC, which only offers a minimal view of the severity since the procedure is terminated once the first symptoms appear.

10.2 Experimental Procedure and Data

A clinical study was performed using 93 allergic patients with ages from 3 to 18 years. Tables 10.1 and 10.2 describe the frequencies observed for DBPCFC score, the first accidental exposure score and the eliciting dose. Note that only 54 out of the 93 patients experienced a first accident. The remaining 39 patients were diagnosed during an allergy check-up and subsequently confirmed by DBPCFC, thus avoiding further accidents. Patients were homogeneously distributed in age and sex across severity scores.

Thirty-six variables were measured to reveal the presence of *Immunoglobulins of type E (IgE)* antibodies. These are proteins produced by the immune system that can elicit allergic reactions [55]. Each antibody is specific to an allergen, *i.e.*, it is coded to identify a particular protein for elimination. We measured the levels of *IgE* for the proteins of interest with the goal of building a predictive models of allergy severity. The variables used to test for *IgE* were measured either by immunoassays or by Skin Prick Tests (SPTs).

10.2.1 Immunoassays

Immunoassays are biochemical tests that quantify the level of antibodies in a blood sample (in *kilo-units per liter*). We performed six immunoassays aimed at measuring the following : the *total IgE*, the *specific IgE to peanut (f13)*, and the *specific IgE to recombinant (r)Ara-h1, rAra-h2, rAra-h3, rAra-h8*, which are *IgE* especially directed against peanut recombinant major allergens [32]. Note that *rAra-h1, rAra-h2, rAra-h3* and *rAra-h8* are recombinant proteins, meaning they are produced in *E. coli* bacteria in our laboratory [32]. This technique allows for large quantities of highly purified proteins.

10.2.2 Skin Prick Tests (SPTs)

SPTs are used to detect an immunological sensitivity to a particular substance. They show the functional aspect of cellular *IgE*, which are linked to mast cells releasing chemical mediators that elicit symptoms [55]. A small dose of allergen is applied under the skin by pricking with a needle, and the diameter of the resulting wheal is measured in millimeters. We also measured the diameter of prick-tests to codeine as a positive control showing the basal reactivity of the skin. The ratio of the two diameters is used to measure the allergen reaction.

We performed prick-tests for 30 allergens divided into three families :

1. **7 legumes** : *chickpea, lentil, green pea, broad bean, dried bean, soybean, lupine flour*, since peanuts are legume ;
2. **11 tree nuts** : *peanut, almond, walnut, hazelnut, cashew nut, Brazil nut, Queensland nut, pecan nut, pistachio, pine nut, chestnut*, which are often related to peanut allergies by cross-reactivity ;
3. **12 aeroallergens** : *Dermatophagoïdes pteronyssinus (Dpte), Alternaria, cockroach, cat epithelia, dog epithelia, 12 grass pollens, birch, mugwort, ribwort, ash, rape seed, latex*, which are common clinical SPTs.

Immunoassays and SPTs were measured immediately before DBPCFC.

10.3 Statistical approach

All computations were performed using the SAS Enterprise Guide 4.1.0.471 ® or R 2.7.0 [56]. The few missing values were replaced by the class mean of the variable.

10.3.1 Design of the study

We first performed a Principal Component Analysis (PCA) to gain an overview of the data. Then the statistical approach proposed in this paper was split into two sections :

1. variable selection by Kruskal-Wallis test [24] and stepwise selection by Wilks' lambda (Λ) criterion [48],
2. comparison of 5 classification rules by cross-validation : linear and quadratic classification, k-NN, CART, and AdaBoost with CART.

Note that nonparametric Kruskal-Wallis test was preferred to ANOVA, because variables were not always normally distributed in the classes induced by the scores. For the DBPCFC and first accident scores, the number of discriminant SPTs can far exceed the number of selected immunoassays, as will be described later. To equalize the influence of both groups of retained

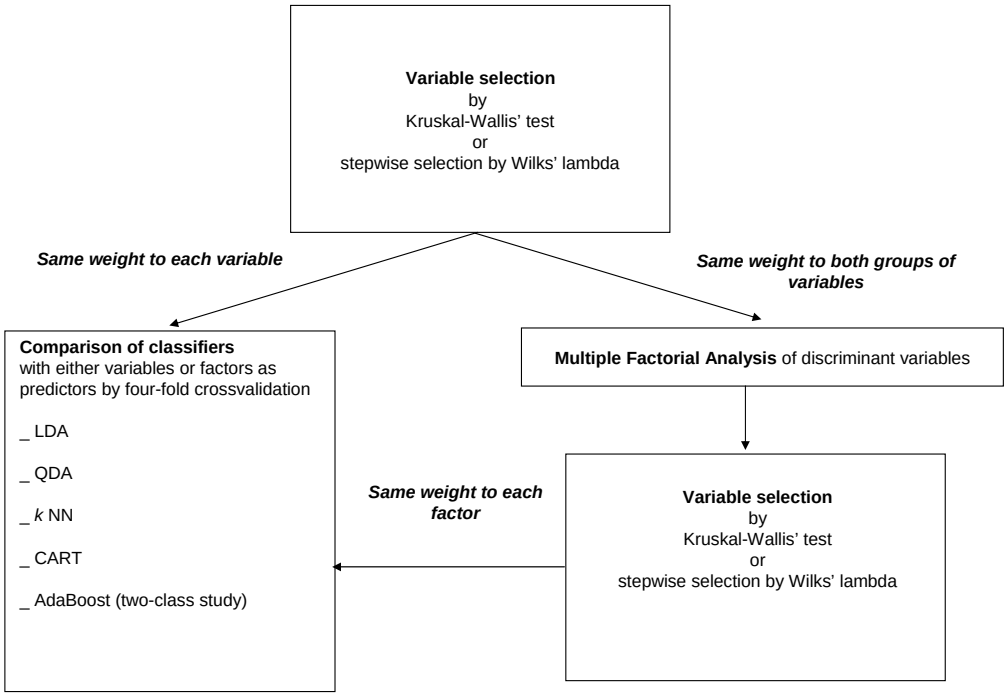


FIG. 10.1 – Statistical Analysis

variables, a Multiple Factorial Analysis (MFA) [57] was performed which enables the use of factors as new predictors of severity.

We also performed two discriminant analyses :

1. by using the discriminant variables as predictors for classification, or
2. by computing MFA factors on these discriminant variables and using a limited number of discriminant factors as predictors.

Our overall statistical approach is summarized in Figure 10.1.

Eliciting dose is a character whose values are fixed by levels by the clinician [32]. To discriminate this variable, we devised an algorithm that simultaneously clusters the eliciting dose values and selects predictors by minimizing the Wilks' lambda. This algorithm was applied by using variables or factors computed by MFA with the 36 available variables as predictors. Once the predictors were chosen and the clusters built, the same classification methods as for DBPCFC and first accident scores were used.

Finally, four studies were performed for each measure of severity by treating it as a four-class variable, and then as a two-class variable, and by using successively variables and factors as predictors, as described below.

10.3.2 Multiple Factorial Analysis (MFA)

MFA was introduced by B.Escofier and J.Pagès [57, 51] for sensory analysis, and is a PCA with a particular choice of metric. The aim of this method is to give a similar part to several groups of variables when determining factors, *i.e.*, uncorrelated linear combinations of the initial

variables. This procedure is useful for avoiding models that are fully influenced by a single group of numerous variables which could partially cancel the effect of the other groups. Briefly, an MFA is performed as follows :

Suppose p variables are measured on n subjects and divided in q groups :

$$x^{1,1}, \quad \dots, \quad x^{1,m_1} \tag{10.1}$$

$$\vdots \tag{10.2}$$

$$x^{q,1}, \quad \dots, \quad x^{q,m_q} \tag{10.3}$$

with $\sum_{k=1}^{q} m_k{=}p$, where m_k is the number of variables in group k.

Denote $\mathbf{X}^k$ the matrix of data of size $n \times p$ corresponding to the k^{th} group of variables, namely :

$$\mathbf{X}^k{=}\begin{array}{c|ccccc} & x^{k,1} & \dots & x^{k,j} & \dots & x^{k,m_k} \\ \hline 1 & & & & & \\ \vdots & & & \vdots & & \\ i & & \dots & x_i^{k,j} & \dots & \\ \vdots & & & \vdots & & \\ n & & & & & \end{array}$$

where the generic element $x_i^{k,j}$ denotes the measure of the variable $x^{k,j}$ for the sample point i.

Let also $\mathbf{X} = (\mathbf{X}^1|\dots|\mathbf{X}^q)$ be the matrix corresponding to the whole set of variables.

For the k^{th} group of variables, let $\mathbf{M}^k$ be a metric matrix in $\mathbb{R}^{m_k}$, $k{=}1,\dots,q$.

Let $\mathbf{D}$ be the diagonal matrix of the weights assigned to the sample points.

The MFA algorithm is then as follows :

- Step 1 : For any $1 \leq k \leq q$, perform PCA($\mathbf{X}^k$,$\mathbf{M}^k$,$\mathbf{D}$), and denote λ_1^k the greatest eigenvalue corresponding to the first factor ;
- Step 2 : consider the metric matrix in $\mathbb{R}^p$:

$$\mathbf{M}{=}\begin{pmatrix} \mathbf{M}^1/\lambda_1^1 & & \\ & \ddots & \\ & & \mathbf{M}^{\mathrm{q}}/\lambda_1^{\mathrm{q}} \end{pmatrix}$$

  and perform PCA($\mathbf{X}$,$\mathbf{M}$,$\mathbf{D}$).

Note that in our case, $q = 2$ with the immunoassays as the first group of variables and the SPTs as the second. Variables were first centered and scaled to unity, and the metric matrix $\mathbf{M}^k$ was then set to the identity matrix $\mathbf{I}^k$ in $\mathbb{R}^{m_k}$.

For the DBPCFC and first accident scores, we thought it made more sense to compute the factors using only the discriminant variables rather than using all available variables. Indeed, the predictive model needed to be built with a reasonable number of characters. Even if a limited number of factors were chosen afterwards, all variables would still have to be measured to compute the factors. Moreover, a non-discriminant variable could have a large coefficient for some retained factors even though it would not improve the overall discriminative power of the model. Nonetheless, we did compute factors using all variables as well, but the results did not improve the discrimination of the first accident score. For the eliciting dose, factors were computed using all 36 available variables, not only the discriminant ones. As described in Section 10.3.4, the set of discriminant variables depends on the choice of the clustering of eliciting dose values and vice versa. Thus it did not seem appropriate to replace the optimal set of variables by factors.

10.3.3 Variable selection

The variable selection methods used here are standard approaches, such as the Kruskal-Wallis' test [24] and the stepwise Wilks' lambda selection [48]. Variables or factors were retained in the model if either the corresponding p-value of the Kruskal-Wallis test was smaller than 0.10, or if the variable was selected by the stepwise Wilks' lambda criterion. For the latter, the maximal $F-to-enter$ p-values used as entry and removal criteria were set by default to 0.15, as recommended by [58].

Note that the Wilks' lambda selection is based on the hypotheses of multi-normality of the variable distribution and equality of the within-class covariance matrices. In 1975, Lachenbruch [59] asserted that the $F-$test is robust to small deviations of these hypotheses. We therefore decided to use the Wilks' lambda selection even though variables were not always normally distributed in the classes included by the measures of severity.

In another test, variables were retained if their distributions were significantly different in the score classes up to level 0.15 for Kruskal-Wallis, but it did not produce better results than limiting the level to 0.10.

10.3.4 An algorithm for simultaneously clustering the response variable and selecting discriminant variables

Eliciting dose is a variable whose values are taken according to an increasing scale of fixed doses of peanut. For a given patient, only an interval including the eliciting dose is actually known. We wished first to group eliciting dose values in a limited number of intervals, and second to select the most discriminant variables for these categories. Here we propose an algorithm to perform these two steps simultaneously using alternate optimization.

Principle

For a given partition in intervals, a set of discriminant variables of fixed cardinal is selected using a certain optimality criterion. A new partition in intervals is then determined to optimize this criterion with the chosen variables, and so on. In order to not repeat this algorithm for different values of the cardinal, the number of variables to include could be increased one-by-one at each step of the procedure. The optimal set of variables could then be searched into all the possible subsets of variables of fixed cardinal corresponding to this step. To avoid heavy computations, variables were included forward in the model. At each step, a new variable was chosen according to the optimality criterion and added to the preceding variables.

Since Wilks' lambda provides a non-empirical stopping rule by testing its significance, this approach was preferred over using within-class inertia computation as the optimality criterion. The p-value of $F-to-enter$ was set to 0.15.

Constructing the clusters

Let y be an ordinal categorial variable of levels $\{m_1, \ldots, m_l\}$, $m_1 < m_2 < \ldots < m_l$. Suppose that we want to cluster the levels of y into a limited number r of intervals $]m_i, m_j]$. Only consecutive levels can be gathered. We build consecutive left-opened and right-closed intervals by selecting the upper bounds. Thus the number of possible clusterings is C_{l-1}^{r-1}.

Algorithm

The specific algorithm for computing intervals and selecting discriminant variables is as follows :

- Step 1 :
  1. choose the clustering $\mathcal{C}^1$ of eliciting dose values that minimises Λ, computed using all 36 available predictors ;
  2. select the predictor v^1 that minimises Λ with the obtained clustering $\mathcal{C}^1$;
- Step 2 :
  1. choose the clustering $\mathcal{C}^2$ of eliciting dose values that minimises Λ, computed using the previously selected predictor v^1 ;
  2. select the predictor v^2 such that the paired predictors (v^1, v^2) minimise Λ with the new clustering $\mathcal{C}^2$;
- and so on ...
- procedure stops if either no left predictor can improve the discriminant power of the model, *i.e.*, if $F - to - enter$ $p-$value is greater than 0.15 [48], or if every predictor is already entered.

Note that the $F - to - enter$ value is only computed when a new variable is entered into the model. This algorithm was used with both the variables and the MFA factors computed using all 36 available variables. Since new discriminant variables could have been chosen at each step of the algorithm, there was no default starting set of discriminant variables. Moreover, it did not seem appropriate to perform the MFA at the end of the algorithm, since the selected variables were specifically chosen to discriminate the found clusters.

10.3.5 Discriminant analysis

Linear and quadratic classification, k-NN and CART are classic methods [49]. For two-class discrimination we also used the AdaBoost algorithm with CART [50]. Since this is still a recently developed algorithm, we briefly summarize its concepts below.

Let $\left\{(\boldsymbol{x}_i, y_i)_{1\leq i\leq n}\right\}$ a dataset, where $\boldsymbol{x}_i \in \mathbb{R}^p$ is the vector of predictors, and $y_i \in \{0, 1\}$ is a binary response variable to discriminate. The principle of the AdaBoost algorithm is to re-weight observations that were misclassified by a base classifier (CART in our case). At each step of the procedure, a new classification tree is randomly built, inducing new misclassified sample points whose weights are updated before the following step starts. The method proceeds according to the following algorithm :

- Step 1 : assign equal weigths to all sample points $w_i^{[0]}=1/n, \forall i = 1, \ldots, n$;
- Step 2 : for $m=1, \ldots, M$ do :
  1. build a classifier $\hat{g}^{[m]}$ trained on data weighted by $w_i^{[m-1]}, \forall i = 1, \ldots, n$;
  2. classify the data by resubstitution : determine $\hat{g}^{[m]}(\boldsymbol{x}_i), i = 1, \ldots, n$;
  3. compute the misclassification rate :

$$err^{[m]} = \frac{\sum_{i=1}^{n} w_i^{[m-1]} \mathbb{1}_{(y_i \neq \hat{g}^{[m]}(\boldsymbol{x}_i))}}{\sum_{i=1}^{n} w_i^{[m-1]}}, \tag{10.4}$$

$$\alpha^{[m]} = \log\left(\frac{1 - err^{[m]}}{err^{[m]}}\right), \tag{10.5}$$

4. update the weights

$$\tilde{w_i} = w_i^{[m-1]} \exp(\alpha^{[m]} \mathbb{1}_{(y_i \neq \hat{g}^{[m]}(\boldsymbol{x_i}))}), \tag{10.6}$$

$$w_i^{[m]} = \frac{\tilde{w_i}}{\sum_{j=1}^{n} \tilde{w_j}}; \tag{10.7}$$

– Step 3 : build the aggregated classifier

$$\hat{f}_{\text{AdaBoost}}(\boldsymbol{x}) = \arg\max_{y \in \{0,1\}} \sum_{m=1}^{M} \alpha^{[m]} \mathbb{1}_{(\hat{g}^{[m]}(\boldsymbol{x})=y)}. \tag{10.8}$$

A novel observation $\boldsymbol{x}$ is classified by the majority vote $\hat{f}_{\text{AdaBoost}}(\boldsymbol{x})$, where vote m is weighted by $\alpha^{[m]}$. Note that the default number M of iterations is set to 50.

10.4 Results

10.4.1 Principal Component Analysis

The PCA representation of the variables was relevant. As seen on the correlation circle of Figure 10.2, intra-family correlations between variables was rather high but inter-family correlations were quite low (with the notable exception of tree nuts and aeroallergens). Moreover, the *total IgE* and *specific IgE to rAra-h8* did not seem closely related to the other immunoassays, an observation fully validated by clinicians. Indeed, as explained earlier, the level of *total IgE* is the global measure of this antibody subclass, whereas the *specific IgE to rAra-h1, rAra-h2* and *rAra-h3* are directed against peanut allergens alone. Also, *rAra-h8* is a homologous protein to the birch pollen allergen *Bet-v1*, sharing about 66% of their amino acid sequences. Thus patients sensitive to both peanut and birch pollen could present high values of *specific IgE to rAra-h8* without being allergic to peanuts.

These results confirmed clinical observations. Also, the individual representation did not provide supplementary information (data not shown), and no particular interpretation was evident for the PCA axes.

10.4.2 First accidental exposure score

Here we show the results for predicting the first accidental exposure score.

Four-class study

Note that in what follows, we combined scores 4 and 5 because of the low frequency of score 5. Thus the four classes considered here are for scores of $\{1\}$, $\{2\}$, $\{3\}$, and $\{4, 5\}$.

Variable selection

Table 10.3 shows the p-value of the Kruskal-Wallis test for each variable, and Table 10.4 shows the stepwise variable selection using the Wilks' lambda criterion. The Λ statistic was computed at each step with all variables already present in the model, whereas the $F-to-enter$ statistic and corresponding p-value measure the discriminant power of a variable added to the preceding ones.

In total, 18 variables (3 immunoassays and 15 prick tests) had either a Kruskal-Wallis p-value lower than 0.10 or were retained in the Wilks' lambda selection. These were *Specific IgE to rAra-h1, rAra-h3, rAra-h8* and *SPT to dog epithelia, walnut, cashew nut, pecan nut, pistachio,*

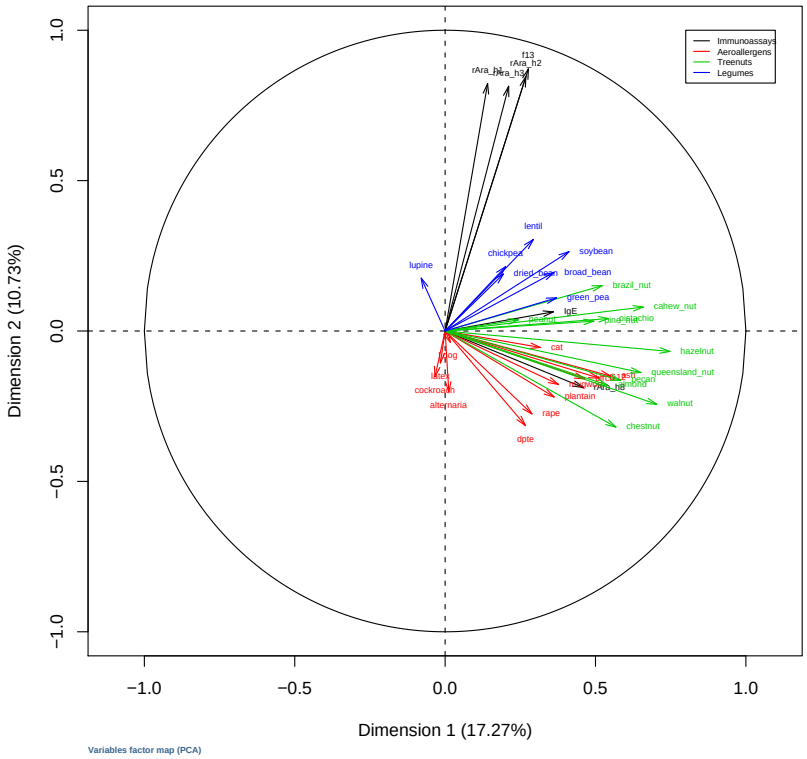


FIG. 10.2 – Circle of correlations

peanut, chick pea, broad bean, green pea, Dpte, ash, 12 grass pollens, Alternaria, almond and *dried bean.*

These immunoassays, as well as SPT to tree nuts and legumes, are variables expected by clinicians. This indicates that our selection process seems to detect "useful" variables. More surprisingly, a few SPTs to aeroallergens were also selected. Indeed, these variables are not known to cross-react with peanuts.

Note that although the variable *soybean* had a p-value lower than 0.10, it was not retained in the model, because its significance was only due to the replacement of missing values. Furthermore, the number of SPTs selected is higher than the number of immunoassays selected. Thus an MFA was performed to equalize the influence of both groups of variables by replacing variables by factors.

Multiple Factorial Analysis of selected variables

Table 10.5 presents factor computations and the corresponding eigenvalues. Factors are numbered by decreasing eigenvalues.

The regular decrease of eigenvalues indicates that the available information cannot be reduced to a limited number of factors. Factors were then chosen in the same manner as the initial characters (Tables 10.6 and 10.7). Note that the five factors selected by the Kruskal-Wallis test are included in the 11 factors chosen with Wilks' lambda. Thus the 11 factors retained to build the predictive models were those selected by the last criterion.

Discriminant analysis

The percentages of good predictions from each classification method (LDA, QDA, k-NN, CART) were determined by fourth-fold cross-validation. The percentage of detected patients with high severity was also computed, *i.e.*, patients of score 4 who were correctly classified in class 4.

According to clinicians, failing to detect patients with severe allergies could lead to inappropriate food intake by the patient.

The use of factors as predictors always improved the classification rates compared to the direct use of variables. The 3-NN classification with factors appeared to be the best method for predicting first accidental exposure score in the four-class study, with 55% of sample points classified successfully. However only 23% of the severe subjects were correctly classified. Using LDA and QDA with factors resulted in a similar percentage of successfully classified sample points, ranging from 50% to 54%. The other models yielded poor results.

Note that the factors numbered 15, 16 and 18 had a low explained proportion of inertia, lower than 1.05. Therefore predictive models were also built without these factors, even though they were selected by Wilks' lambda. It appears, however, that the percentages of good classifications decreased in general, as if complementary information could be brought by these "high background" characters to the discriminative model.

Two-class study

For the two-class discrimination, groups were formed by scores of either $\{1, 2, 3\}$ or $\{4, 5\}$, as recommended by clinicians. As the methodology used was the same as for the four-class study, we only give the list of the selected variables and factors, and the results of the best classification rule.

The eight selected variables were *Specific IgE to rAra-h1, rAra-h2, rAra-h3* for immunoassays and *dog ephitelia, walnut, pecan nut, peanut, lupine flour* for SPT. Factors were then computed with these variables and ranked by decreasing eigenvalues. The indices of the discriminant factors were 3, 1, 6 and 8.

The method with the best results for two-class discrimination was LDA, with variables directly used as predictors (83% of well-classified sample points). Of the score 4 patients, 74% were correctly classified. Contrary to the four-class study, the use of factors in the two-class study did not improve the results of each method. This could be due to the number of immuoassays and SPTs selected not being drastically different from each other (three versus five, respectively).

Note that the CART method can also be seen as a stepwise variable selection method. For that reason, we also tried to use it to model decision trees without any variable pre-selection phase in order not to limit the possible choice of predictors. However, we found that successful classification rates were higher in general when predictors were first selected by the Kruskal-Wallis' test and Wilks' criterion, rather than letting CART choose the predictors.

10.4.3 DBPCFC score

The statistical approach used to predict the DBPCFC score was the same as for the first accident score. Here we show only the variables selected and the results for each classification rule.

Four-class study

Scores of 4 and 5 were again grouped together due to the low frequency of score 5. The nine selected variables were *Total IgE* and *ash, lentil, lupine flour, green pea, dried bean, cockroach, Queensland nut, rape seed*. *Total IgE* was the only immunoassay retained in the model. Thus, no factor was computed. Indeed, assigning the same weight to this single variable as to the eight other variables did not seem appropriate, because *total IgE* are not antibodies specifically involved in peanut allergies, but are involved in all allergic reactions.

Although LDA gave the best results with a 46% successful classification rate, the overall misclassification error remained high.

Two-class study

DBPCFC scores can be gathered into two classes in the same manner as for the first accidental exposure score. This resulted in selected variables of *lupine flour, lentil, cockroach, 12 grass pollens, ash*. Note that only SPT were selected. Overall, using 1-NN classification offered the best results, with 72% of successfully classified sample points and 66% of successfully classified severe patients.

Note that the first accidental exposure score could not be used as a predictor for DBPCFC. Indeed, a p-value of 0.58 for Fisher's exact test [24] did not reject the hypothesis of independence for these two measures of severity. Also note that Hourihane et al. proposed a different classification of reaction severity [36]. In this scheme, DBPCFC and reactions recorded in the patient's medical file (*community score*) were both scored. Although no predictive model was proposed in this paper, no dependence was found between either of their two measures of severity.

10.4.4 Eliciting dose

As explained in Section 10.3.4, an algorithm was developed to stepwise select the best clustering of eliciting dose values and the corresponding best discriminant predictors. This was performed for both four-class ($r = 4$) and two-class ($r = 2$) studies. The minimal number of sample points in each class was set to 10, in order to have class frequencies large enough to perform cross-validation.

Four-class study

Note that the number of possible partitions in four intervals is $C_{22}^3 = 1540$, but partitions with class frequencies lower than 10 were not considered.

Algorithm with variables

Table 10.8 shows the bounds x,y,z for clustering $]-\infty, x]\,,]x, y]\,,]y, z]\,,]z, +\infty[$, and the variables entered at each step. The algorithm stopped at step 7 because no additional improvement resulted afterwards. The selected variables were *hazelnut, birch, pistachio, cashew nut, green pea, total IgE, pine nut* and the bounds were 95, 215, and 500 *mg*. The eliciting dose was correctly predicted for 38% of the patients using LDA. Moreover, only 38% of the patients with eliciting dose lower than 95 *mg* were correctly classified.

Algorithm with factors

The same study was performed with factors as predictors. The final bounds obtained were 165, 300, and 500 *mg* with factors of 10,16,13,20,12, and 18 ($\Lambda = 0.56$). Although predictors were chosen to maximise the discrepancy between the classes, successful predictions for the eliciting dose in the four-class study was only 40% using LDA. Furthermore, the use of factors did not noticeably enhance the model (except for k-NN).

Two-class study

For the two-class study, instead of using three bounds x, y, z a single bound x was searched for that included the two classes $]-\infty, x]$ and $]x, +\infty[$. Also, the number of possible intervals was at most $C_{22}^1 = 22$.

Algorithm with variables

The procedure stopped at $\Lambda = 0.72$ with the variables *walnut, ash, cockroach, lupine flour, plantain* and *hazelnut* as predictors and a bound of 300 *mg*. Using 5-NN, this gave a successful classification rate of 66% with only 27% of highly reactive patients misclassified.

Algorithm with factors

The threshold of eliciting dose was 500 *mg* and the factors selected were 13,10,32,18, and 24 ($\Lambda = 0.75$). Computing factors to discriminate the eliciting dose in the two-class study was the most discriminant model (82% were successfully classified with CART). It also successfully classified 85% of the highly reactive patients. The threshold of 500 *mg* of peanut was also judged relevant by clinicians. Nevertheless, this model cannot easily be used in practice. Indeed, as mentioned earlier, MFA was performed on all 36 variables, not just the discriminant variables. This means that to correctly predict the eliciting dose, all variables would need to be measured to compute the factors ; this does not seem feasible.

Note that in order to evaluate the efficiency of our method, we compared our results of cross-validation to those obtained by discriminating the first clustering displayed by the algorithm. This first clustering has the lowest Wilks' lambda computed using all variables or factors. Predictors were then chosen using the Wilks lambda criterion, and the classification methods were applied. For every measure of severity, our algorithm yielded better results in general.

FIG. 10.3 – Summary of discriminant variables. Selected variables are colored in black for each measure of severity

10.5 Conclusions

To the best of our knowledge, this paper presents the first discriminant analysis of DBPCFC score, first accident score, and eliciting dose measurement of peanut allergy severity. Previous studies were aimed at finding links between immunoassays or SPTs and allergy severity, but their statistical analyses were limited to either comparing distributions between groups of patients (using, for instance, the Mann-Whitney test) or to computing linear correlation coefficients [36, 35]. In our approach, we used several classification rules to aid in comparing and choosing the optimal and most efficient method (see Table 10.9). In addition, we found that using MFA to compute new predictors was an attractive solution when equalizing the weights of group variables, and we proposed a novel algorithm for simultaneously clustering the levels of ordinal qualitative variables and for selecting discriminant variables.

Our work differs from earlier studies in several respects. Previous studies were performed on small sample sizes of 30 to 40 patients [36, 35] using a small number of measured variables, which only permitted a limited choice of discriminating predictors. Also, *specific IgE to Ara-h1,2,3* were measured by SPTs instead of immunoassays [35] and although a positive response to an SPT does indeed indicate allergen sensitivity, it is still less accurate than immunoassays.

There are several scoring methods in the literature to evaluate peanut allergy severity. For example, Hourihane et al. devised a complex 25-class scoring system combining observed reactions and eliciting doses [36]. A graduation of symptoms was also proposed by Müller [37], but this score is based on allergic reactions in response to bee or wasp venom (and not peanut allergens). Thus, developing a standardized scoring method is still necessary and would facilitate comparison studies from different centers. One possible solution would be comparing the results of Hourihane et al.'s score with the one used in this study using the same cohort of patients. Moreover, the large number of scoring levels in Hourihane et al.'s approach could be reduced by our algorithm.

Nevertheless, several potential biases in our study must be noted. First, SPT diameters are relatively imprecise. There is no standard method of measuring SPT reaction since both wheal diameters and areas are popular metrics [35]. Additionally, the score of the first accident can be inaccurate or imprecise since it depends on the medical history of the patient, and hence subject to inaccuracies in the patient's memory which may underestimate symptom severity. Other factors such as alcohol or medication can affect the first reaction symptoms as well [52, 53] yielding a severity score higher than it should be. Finally, the first exposure score is a past event predicted using variables measured during a subsequent DBPCFC. As mentioned in Section 10.4.3, Hourihane et al. also used such a reverse prediction with a community score that was evaluated *a posteriori* using the record file of the patient [36].

The predictive models used in our study yielded correct classifications for the first accidental exposure and DBPCFC of up to 83% and 72% for the two-class study. Our algorithm also allowed us to group eliciting dose values and to select discriminant predictors, leading to an 82% classification rate for the two-class study. This indicates that it is indeed possible to correctly predict peanut allergy severity by measuring well-chosen variables. Considering that all immunoassays of specific *IgE* were selected once, we also hypothesize that measuring new antibodies to peanut allergens, such as those directed against *rArah-6, rArah-7* and *rArah-9*, will further improve the discriminative power of our models. This also argues for these antibodies playing key roles in the diagnosis of peanut allergy severity. Furthermore, some SPTs with proven cross-reactivity to peanuts were retained in our models, such as *lupine flour* [54], indicating that our results were in line with other medical discoveries.

Our variable selection process also offers a new perspective on conducting allergy check-ups. Indeed, some unexpected variables appeared several times in our models, such as SPTs to *coa-*

ckroack and *ash* as shown in Figure 10.3. If future experiments could distinguish the medical relevance of this observation from cross-reactivity, then the importance of these SPTs in discriminating severity would be confirmed. Besides, some SPTs never appeared in our models, such as SPTs to *latex* or *Brazil nut*, and thus should no longer be performed in practice when diagnosing peanut allergy severity.

The discriminating models described in this paper are a first step towards a simple, safe and efficient diagnosis of peanut allergy severity by quantifying antibodies. Before being applied in clinical practices, they must first be validated on an independent set of patients. New variables must also be added as additional predictors toward improving successful classification rates. These models could then become practical tools for clinicians. When scoring severity, the clinical test results could be reported online at the allergy vigilance network (*Réseau d'allergovigilance, http ://www.cicbaa.com/*), or a simple statistical software could be programmed and provided as freeware.

Acknowledgements. The authors would like to thank Frances T. Yen for her critical review of the manuscript.

scores	1	2	3	4	5	sum
DBPCFC	22	31	12	26	2	93
first accident	9	19	7	17	2	54

TAB. 10.1 – Frequencies of DBPCFC and first accidental exposure severity score

eliciting doses (mg)	n	eliciting doses (mg)	n
1.4	1	300	1
4.4	2	400	1
14	1	500	22
15	2	965	6
44	11	1000	1
65	6	2000	3
95	1	2110	1
115	2	3500	1
165	2	3610	1
210	1	4110	1
215	19	7000	5
265	2	Sum	93

TAB. 10.2 – Frequencies of eliciting dose values

variable	p-value	variable	p-value
chick pea	0.003	coackroach	0.289
broad bean	0.007	cat epithelia	0.294
green pea	0.015	lentil	0.295
sp IgE to rAra-h3	0.029	total IgE	0.351
walnut	0.031	Queensland nut	0.359
soybean	0.033	almond	0.377
sp IgE to rAra-h1	0.038	lupine flour	0.406
cashew nut	0.049	ash	0.412
arachide	0.061	Dpte	0.466
pecan nut	0.075	rape seed	0.496
pistachio	0.089	12 grass pollens	0.553
dog epithelia	0.092	sp IgE to rAra-h8	0.555
dried bean	0.115	mugwort	0.564
Alternaria	0.154	pine nut	0.571
sp.IgE to f13	0.155	hazelnut	0.680
Brazil nut	0.248	birch	0.697
chest nut	0.267	plantain	0.776
sp IgE to rAra-h2	0.272	latex	0.794

TAB. 10.3 – p-values of Kruskal-Wallis tests for first accidental exposure score (four-class study). Variables are ordered by increasing p-values.

step	entered	removed	Wilks' Λ	$F-to-enter$	$F-to-enter$ p-value
1	chickpea		0.67	8.11	1.68E-04
2	green pea		0.55	3.58	2.04E-02
3	Sp.IgE to rAra-h8		0.43	4.69	5.97E-03
4	Sp.IgE to rAra-h1		0.35	3.38	2.57E-02
5	dog		0.30	2.76	5.29E-02
6	peanut		0.26	2.19	1.03E-01
7	Dpte		0.23	2.00	1.27E-01
8	ash		0.19	2.52	7.04E-02
9		peanut	0.22	1.85	1.52E-01
10	almond		0.19	2.16	1.06E-01
11	dried bean		0.16	2.75	5.44E-02
12	12 grass pollens		0.14	1.93	1.39E-01
13	Alternaria		0.12	2.49	7.38E-02

TAB. 10.4 – Wilks' lambda stepwise selection for first accidental exposure score (four-class study).

factor	eigenvalue	inertia	cumulated	factor	eigenvalue	inertia	cumulated
1	1.15	18.77	18.77	10	0.20	3.27	88.92
2	1.01	16.43	35.20	11	0.17	2.75	91.68
3	0.76	12.46	47.66	12	0.14	2.24	93.92
4	0.55	8.98	56.63	13	0.10	1.69	95.60
5	0.51	8.38	65.01	14	0.08	1.28	96.89
6	0.45	7.42	72.44	15	0.06	1.05	97.94
7	0.31	5.05	77.49	16	0.06	0.94	98.88
8	0.29	4.71	82.19	17	0.05	0.83	99.70
9	0.21	3.46	85.66	18	0.02	0.30	100.00

TAB. 10.5 – Eigenvalues and corresponding inertia for factors.

factor	p-value	factor	p-value
1	0.0010	18	0.3597
3	0.0436	14	0.4020
5	0.0471	8	0.4809
10	0.0779	13	0.5054
7	0.0960	12	0.6784
9	0.1148	11	0.7386
16	0.1627	17	0.7842
15	0.2775	4	0.8999
6	0.3339	2	0.9277

TAB. 10.6 – p-values of factors for the Kruskal-Wallis test (four-class study). Factors are ordered by increasing p-values.

step	entered	Wilks Λ	$F-to-enter$	$F-to-enter$ p-value
1	1	0.72	6.60	0.0008
2	3	0.56	4.62	0.0063
3	10	0.44	4.15	0.0108
4	5	0.37	3.06	0.0370
5	9	0.31	2.86	0.0471
6	11	0.27	2.56	0.0665
7	7	0.23	2.60	0.0643
8	6	0.20	2.27	0.0937
9	15	0.17	2.54	0.0696
10	16	0.14	2.24	0.0978
11	18	0.12	2.21	0.1023

TAB. 10.7 – Wilks' lambda stepwise selection of factors for first accidental exposure score (four-class study).

step	bounds / predictors	Wilks Λ	$F-to-enter$	$F-to-enter$ p-value
1	95-215-500	0.05		
1	hazelnut	0.73	8.51	6.96E-05
2	95-215-500	0.73		
2	birch	0.64	3.15	0.031
3	95-215-1000	0.63		
3	pistachio	0.56	2.85	0.044
4	95-215-1000	0.56		
4	cashew nut	0.50	2.74	0.050
5	95-215-1000	0.50		
5	green pea	0.45	2.24	0.092
6	95-215-1000	0.45		
6	total IgE	0.41	2.20	0.096
7	95-215-500	0.41		
7	pine nut	0.35	3.24	0.028

TAB. 10.8 – Clusters and variables selected to discriminate eliciting doses and corresponding Λ and $F-to-enter$ statistics (four-class study).

measure of severity	predictors	variables	factors
first accident	4 classes	LDA : 50%-26%	3NN : 55%-23%
	2 classes	LDA : 83%-74%	3NN : 82%-64%
DBPCFC	4 classes	LDA : 46%-61%	x
	2 classes	1NN : 72%-66%	x
eliciting dose	4 classes	LDA : 38%-38%	LDA : 40%-73%
	2 classes	5NN : 66%-73%	CART : 82%-85%

TAB. 10.9 – Summary of the discriminant analysis. For each measure of peanut allergy severity, the best classification method is given for the four-class and two-class studies, with both variables and factors as predictors. Results are expressed as successful classification rates and as severe patient detection rates.

Chapitre 11

Un algorithme de classification et de sélection simultanée de variables discriminantes

11.1 Introduction

Dans l'étude de l'allergie à l'arachide, un des problèmes posés est celui de la prédiction de la dose réactogène à l'aide des mesures d'anticorps spécifiques de certains allergènes (Chapitres 9 et 10). Les valeurs administrées de la dose réactogène sont fixées au préalable par un protocole clinique : des doses croissantes d'arachide sont admininistrées au patient jusqu'à l'apparition de symptômes objectifs. De ce fait, la dose précise induisant la première réaction n'est pas connue ; seul l'intervalle contenant la dose réactogène est connu. Prédire la valeur de la dose réactogène à l'aide d'une analyse de régression n'est donc pas possible. La dose réactogène peut en revanche être assimilée à une variable qualitative ordinale. Mais son grand nombre de modalités, 23 dans notre cas, et le faible effectif de certaines classes empêchent de mener une analyse discriminante classique. Notre objectif consiste à définir un regroupement, optimal au sens d'un certain critère, des intervalles de la dose réactogène en un nombre réduit d'intervalles. On choisit ensuite les variables qui expliquent au mieux ce découpage en classes. Mais ces variables pourraient mieux discriminer un autre découpage en intervalles, qui pourrait lui même induire un choix différent de variables discriminantes. Dans le chapitre précédent, un algorithme permettant de construire simultanément le regroupement des intervalles de la dose réactogène et de sélectionner les variables qui le discriminent le mieux a été présenté. Afin de valider cette procédure, on se propose d'appliquer l'algorithme à trois jeux de données différents.

11.2 Descriptif de l'algorithme

11.2.1 Principe de l'algorithme ascendant

Soit y une variable qualitative ordinale de modalités $\{m_1, \ldots, m_l\}$, $m_1 < m_2 < \ldots < m_l$ et $x^1, \ldots, x^p$ des variables explicatives continues.

Pour une partition en r intervalles fixée, un ensemble de variables discriminantes de cardinal fixé q est cherché parmi les p variables explicatives en utilisant un certain critère d'optimalité. Une nouvelle partition en intervalles est ensuite cherchée pour optimiser ce critère à partir des variables choisies et ainsi de suite.

Pour ne pas répéter cet algorithme pour différentes valeurs de q, le nombre de variables à retenir peut être incrémenté avec le pas de l'algorithme : on cherche la variable la plus discriminante

au pas 1, le couple de variables le plus discriminant au pas 2 , puis le triplet, etc.

Afin de limiter encore les calculs, les variables sont inclues pas-à-pas dans le modèle. A chaque pas, une nouvelle variable est choisie grâce au critère d'optimalité et ajoutée aux caractères précédemment sélectionnés.

Le critère d'optimalité choisi est le Lambda de Wilks, noté Λ, qui offre une règle d'arrêt qui n'est pas empirique, contrairement notamment à l'inertie inter-classes.

11.2.2 Construction des classes

On souhaite regrouper les modalités de y en un nombre limité r d'intervalles $]m_i, m_j]$ ($m_i < m_j$; le premier intervalle est du type $[m_1, m_j]$). Puisque y est une variable qualitative ordinale, seules des modalités consécutives peuvent être regroupées. Ces intervalles ouverts à gauche et fermés à droite sont définis en choisissant les bornes supérieures de $(r-1)$ intervalles parmi $\{m_1, \ldots, m_{l-1}\}$, le dernier intervalle ayant m_l pour borne supérieure. De ce fait, le nombre de regroupements en intervalles possibles est égal à C_{l-1}^{r-1}.

11.2.3 Déroulement de la procédure

L'algorithme procède ainsi :

- Pas 1 :
  1. on choisit le regroupement en r classes (intervalles) $\mathcal{C}^1$ des modalités de y qui minimise Λ, calculé avec les p variables explicatives ;
  2. on sélectionne la variable $x^{(1)}$ parmi $x^1, \ldots, x^p$ qui minimise Λ avec le regroupement $\mathcal{C}^1$;
- Pas 2 :
  1. on choisit le regroupement en r classes $\mathcal{C}^2$ des modalités de y qui minimise Λ, calculé avec la variable $x^{(1)}$ précédemment introduite ;
  2. on sélectionne la variable $x^{(2)}$ telle que le couple de variables $\left(x^{(1)}, x^{(2)}\right)$ minimise Λ avec le nouveau regroupement en classes $\mathcal{C}^2$;
- et ainsi de suite ...
- la procédure s'arrête lorsqu'aucune des variables restantes ne peut améliorer le pouvoir discriminant du modèle, *i.e.*, lorsque la $p-$value du $F - pour - entrer$ [48] est plus grande que 0.15, ou lorsque toutes les variables ont déja été introduites dans le modèle.

Remarques :

- La p-value du $F - pour - entrer$ est calculée uniquement lorsqu'une nouvelle variable est introduite dans le modèle.
- On limite l'étude aux partitions où chaque classe contient au minimum un nombre fixé d'individus afin d'avoir ensuite des effectifs suffisamment grands pour réaliser des validations croisées. L'algorithme a été programmé dans R [56] en interdisant les partitions contenant des intervalles ayant moins d'un nombre à fixer d'individus, par défaut 3.

11.3 Validation de l'algorithme sur divers jeux de données

11.3.1 Modèle probabiliste

Afin de tester la validité de notre procédure, il faut des données :

1. dont on connaît *a priori* le regroupement des valeurs de la variable y qu'il faut retrouver,
2. qui vérifient les hypothèses de la sélection par le Lambda de Wilks (Chapitre 7), à savoir la multinormalité conditionnelle dans chaque classe du vecteur aléatoire ayant pour réalisation le vecteur des valeurs des variables explicatives $x^1, \ldots, x^p$, ainsi que l'égalité des matrices de covariance dans toutes les classes,
3. dont on connaît *a priori* un sous-ensemble de variables $\left\{x^{(1)}, \ldots, x^{(s)}\right\}$ de $\{x^1, \ldots, x^p\}$ qui soit "nécessaire et suffisant" pour bien discriminer le regroupement à retrouver. Autrement dit, aucun autre sous-ensemble de variables de $\{x^1, \ldots, x^p\}$ ne doit mieux discriminer le regroupement à retrouver et l'ajout d'une ou plusieurs autres variables à $\left\{x^{(1)}, \ldots, x^{(s)}\right\}$ ne doit pas non plus améliorer la discrimination.

On définit le modèle probabiliste suivant.

Soit Y une variable aléatoire définie sur un espace probabilisé $(\Omega, \mathcal{A}, P)$ à valeurs dans l'ensemble d'entiers $\{1, \ldots, kr\}$, partitionné en r classes de cardinal k, $I_i = \{k(i-1)+1, \ldots, ki\}$, $i = 1, \ldots, r$.

Soit $\hat{Y}$ la variable aléatoire définie sur $(\Omega, \mathcal{A}, P)$ qui a pour réalisation le numéro de la classe d'un individu :

$$\hat{Y} = \sum_{i=1}^{r} i \times \mathbb{1}_{\{Y \in I_i\}}. \tag{11.1}$$

Soit $X^1, \ldots, X^p$ p variables aléatoires réelles, $p \geq r$, définies sur $(\Omega, \mathcal{A}, P)$.

On suppose qu'un sous-ensemble $\left\{X^{(1)}, \ldots, X^{(s)}\right\}$ de ces variables suffit à effectuer une "bonne" discrimination des classes $I_1, \ldots, I_r$ et que les autres variables ne sont pas discriminantes.

Nous allons tester sur trois exemples si, à partir d'un échantillon $(x_i^1, \ldots, x_i^p, y_i), i = 1, \ldots, n$ de $(X^1, \ldots, X^p, Y)$, on peut retrouver par notre algorithme la partition $(I_1, \ldots, I_r)$ et le sous-ensemble de variables discriminantes $\left\{X^{(1)}, \ldots, X^{(s)}\right\}$.

11.3.2 Données simulées

On suppose que $(X^1, \ldots, X^p, Y)$ vérifient :

$$\mathcal{L}(X|\hat{Y} = i) = \quad N_p(m_i, \sigma^2 I_p) \tag{11.2}$$

$$\mathcal{L}(Y|\hat{Y} = i) = \qquad \mathcal{U}_{I_i} \tag{11.3}$$

où I_p est la matrice identité d'ordre p et $\mathcal{U}_{I_i}$ la loi uniforme discrète dans I_i.

L'hypothèse (11.2) permet d'utiliser la règle d'arrêt du Lambda de Wilks.

Remarque : En prenant $\Sigma = \sigma^2 I_p$ comme matrice de covariance, les variables aléatoires $(X^1, \ldots, X^p)$ sont deux à deux indépendantes conditionnellement à $\left\{\hat{Y} = i\right\}$.

En définissant de façon adéquate les moyennes $m_i, i = 1, \ldots, r$ et l'écart-type σ, on construit un sous-ensemble de variables aléatoires $\left\{X^{(1)}, \ldots, X^{(r-1)}\right\}$ de $\{X^1, \ldots, X^p\}$, tel que :

1. $\left\{X^{(1)}, \ldots, X^{(r-1)}\right\}$ permette de bien discriminer $\hat{Y}$,
2. aucun autre sous-ensemble de variables de $\{X^1, \ldots, X^p\}$ ne discrimine mieux $\hat{Y}$,
3. l'ajout d'une ou plusieurs variables à $\left\{X^{(1)}, \ldots, X^{(r-1)}\right\}$ n'améliore pas la discrimination de $\hat{Y}$.

Sans perte de généralité, supposons que $\left\{X^{(1)}, \ldots, X^{(r-1)}\right\}$=$\{X^1, \ldots, X^{r-1}\}$.
Pour $i = 1, \ldots, r-1$, on définit m_i tel que sa j^{ieme} composante m_i^j soit égale à :

$$m_i^j = \delta_i^j, \ j = 1, \ldots, p \tag{11.4}$$

où δ désigne le symbole de Kronecker,
et on définit

$$m_r = 0_p. \tag{11.5}$$

Ainsi

$$\begin{aligned} m_1 &= (1, 0, 0, \ldots, \overset{r-1}{0}, 0, \ldots, 0) && (11.6)\\ &\vdots \\ m_{r-1} &= (0, 0, 0, \ldots, \overset{r-1}{1}, 0, \ldots, 0) && (11.7)\\ m_r &= (0, 0, 0, \ldots, \ 0 \ , 0, \ldots, 0), && (11.8) \end{aligned}$$

On construit de la sorte $r-1$ variables $X^1, \ldots, X^{r-1}$ telles que conditionnellement à $\hat{Y} = i$, seule X^i a une moyenne non nulle. La variable X^r a une moyenne nulle dans toutes les classes.
On définit σ tel qu'il y ait une probabilité égale à 0.95 que la réalisation de X lorsque $\hat{Y} = i$ appartienne à la boule B_i de centre m_i et de rayon $\epsilon = \frac{1}{4}$:

$$P(||X - m_i|| < \epsilon \mid \hat{Y} = i) = 0.95, i = 1, \ldots, r. \tag{11.9}$$

où $||.||$ est la norme euclidienne usuelle.
Lorsque $\hat{Y} = i$,

$$X = m_i + \sigma G, \ G \sim N_p(0, I_p). \tag{11.10}$$

La relation (11.9) est alors équivalente à

$$\begin{aligned} P(\sigma||G|| < \epsilon) &= 0.95 && (11.11)\\ P(||G||^2 < \frac{\epsilon^2}{\sigma^2}) &= 0.95. && (11.12) \end{aligned}$$

Comme $||G||^2 \sim \chi_p^2$, on en déduit $\frac{\epsilon^2}{\sigma^2}$ et donc σ^2.

Dans le cas $r = 3$, si l'on représente ces boules par projection dans le plan (X^1, X^2), on a le dessin de la Figure 11.1.

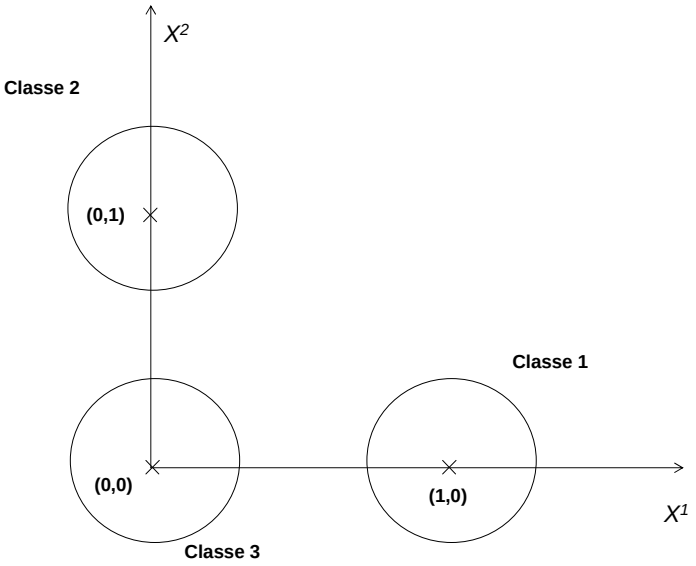


FIG. 11.1 – Représentation des réalisations de $(X^1, \ldots, X^p, Y)$ dans le plan (X^1, X^2)

Les deux variables X^1 et X^2 suffisent pour "bien" discriminer les classes I_1, I_2 et I_3. Par contre, une seule des deux variables ne suffit pas.

Simulation

Dans cette application, on choisit :
- le nombre de classes de $\hat{Y}$: $r = 4$,
- le nombre de valeurs de Y par classe : $k = 10$,
- le nombre de variables explicatives : $p = 10$,
- le nombre de variables discriminantes : $s = 3$.

Ainsi $I_1 = \{1, \ldots, 10\}$, ..., $I_4 = \{31, \ldots, 40\}$.

Pour $i = 1, 2, 3, 4$, on répète 250 fois la simulation suivante :
- on fixe la valeur de $\hat{Y}$ à i ;
- on simule une réalisation de $(X^1, \ldots, X^p, Y)$ sachant que $\hat{Y} = i$ par (11.2).

En projetant les points $(x_i^1, \ldots, x_i^{10})$ sur le sous-espace engendré par les variables X^1, X^2, X^3, les 4 classes sont bien discriminées (Figure 11.2). Deux des variables ne suffisent pas pour discriminer les quatre classes.

En résumé, si notre algorithme est efficace, les classes $I_1 = \{1, \ldots, 10\}$, $I_2 = \{11, \ldots, 20\}$, $I_3 = \{21, \ldots, 30\}$, $I_4 = \{31, \ldots, 40\}$ devraient être retrouvées et les variables X^1, X^2 et X^3 devraient être sélectionnées.

Les résultats obtenus sont présentés dans la Table 11.1 en affichant à chaque pas la borne supérieure des intervalles (fermés à droite), ainsi que la variable ajoutée aux précédentes. Le Lambda de Wilks, la statistique F d'entrée des variables ainsi que la p-value correspondante sont également donnés.

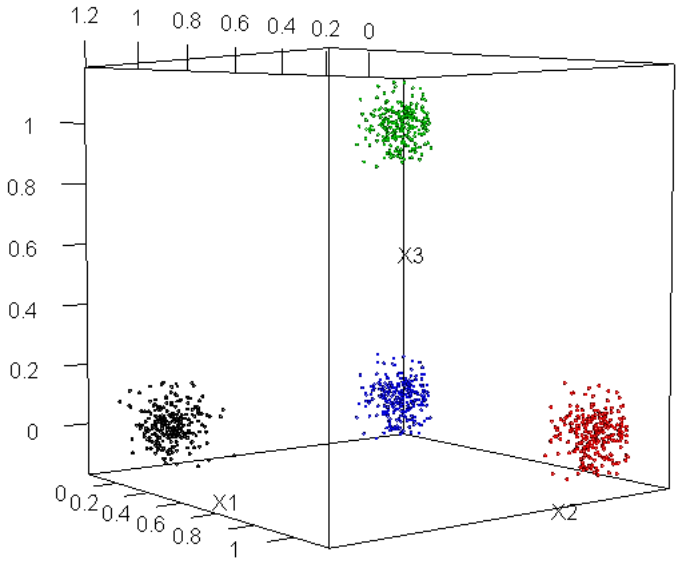


FIG. 11.2 – Représentation du jeu de données simulées sur les variables X^1, X^2, X^3

Pas	Bornes / variable	Lambda de Wilks	statistique F	p-value de F
1	10-20-30	9.32E-06		
1	X^1	0.0175	1380.70	0
2	10-33-35	0.0179		
2	X^2	0.0003	1320.36	0
3	09-10-20	0.0003		
3	X^3	8.69E-06	880.41	0
4	10-20-30	9.68E-06		
4	X^7	8.23E-06	1.38	0.06
5	10-20-30	9.54E-06		
5	X^4	7.80E-06	1.36	0.07
6	10-20-30	9.50E-06		
6	X^8	7.37E-06	1.40	0.05

TAB. 11.1 – Résultats obtenus sur des données simulées

Remarque : La p-value affichée est nulle pour les premiers pas, ce qui signifie qu'elle est inférieure au seuil de précision de R, de l'ordre de 10^{-300}.

Le regroupement attendu a été retrouvé dès le premier pas. Il change ensuite plusieurs fois jusqu'à ce que les variables X^1, X^2 et X^3, qui sont les premières à être intégrées dans le modèle, soient toutes trois sélectionnées. La partition en intervalles reste alors la même jusqu'à la fin de la procédure. On peut noter toutefois que les variables X^7, X^4 et X^8 ont été introduites, alors que d'après notre construction elles ne sont pas discriminantes. Dans cet exemple où les 4

classes sont très bien discriminées par 3 variables, alors que les autres prédicteurs sont inutiles, le seuil de 0.15 pour la p-value du test F est probablement trop grand. En arrêtant la sélection des variables dès que la p-value dépasse 0.05, l'algorithme s'arrêterait alors sur la sélection de X^1, X^2, X^3 et la partition donnée par les bornes 10,20 et 30.

Dans les sections suivantes, l'algorithme est appliqué à deux jeux de données réelles, dont les individus sont naturellement répartis en classes et dont les variables permettant de les discriminer sont connues. La variable $\hat{Y}$ est donc déja définie ; comme dans l'exemple des données simulées, une valeur tirée au hasard est attribuée à chaque individu de la classe $\left\{\hat{Y} = i\right\}, i = 1, \ldots, r$, pour définir la valeur de Y. Des variables non discriminantes sont simulées et ajoutées au jeu de données réelles. L'objectif est de voir si notre algorithme peut retrouver les classes naturelles du jeu de données et les variables discriminantes des classes naturelles, sélectionnées par le Lambda de Wilks. Les données utilisées sont les iris de Fisher et les poissons du lac Laengelmavesi, fournis avec le logiciel SAS ®(procédure STEPDISC).

11.3.3 Les iris de Fisher

Les données sur les iris ont été publiées par Fisher en 1936. La longueur et la largeur du sépale, ainsi que la longueur et la largeur du pétale ont été mesurées en millimètres sur 150 fleurs réparties en trois espèces : *setosa*, *versicolor* et *virginica*. Ces quatre variables permettent de bien discriminer les trois espèces.

Dans la Table 11.2 sont présentés les résultats de la sélection pas-à-pas du Lambda de Wilks sur le jeu de données initial, c'est-à-dire sur les classes *setosa*, *versicolor* et *virginica* :

Pas	Variable	Lambda de Wilks	Statistique F	p-value de F
1	longueur du pétale	0.06	1180.16	2.86E-91
2	largeur du sépale	0.04	43.04	2.03E-15
3	largeur du pétale	0.02	34.57	5.30E-13
4	longueur du sépale	0.02	4.72	0.01

TAB. 11.2 – Résultats de la sélection pas-à-pas sur les iris de Fisher

Dans cet exemple, $\hat{Y}$ prend donc $r = 3$ valeurs. On définit Y en attribuant par un tirage aléatoire uniforme une valeur entre 1 et 5 aux *setosa*, 6 et 10 aux *versicolor* et entre 11 et 15 aux *virginica*. Ainsi $k = 5$ et les classes à retrouver sont $I_1 = \{1, \ldots, 5\}$, $I_2 = \{6, \ldots, 10\}$ et $I_3 = \{11, \ldots, 15\}$.

On construit $p = 30$ variables explicatives en simulant 26 variables non discriminantes qu'on ajoute aux variables de départ. On suppose que le vecteur de ces 26 variables est distribué dans chaque classe selon la même loi multinormale $N_{26}(\mu, V)$.

On fixe un vecteur μ en tirant un échantillon de taille 26 d'une loi $N(m_0, \sigma_0)$, en choisissant m_0 et σ_0 de telle sorte que les variables non discriminantes aient une moyenne comparable à celle des 4 variables discriminantes. On prend pour m_0 la moyenne des moyennes empiriques des 4 variables discriminantes et σ_0 l'écart-type des moyennes empiriques des 4 variables discriminantes. On construit également une matrice de covariance V en tirant uniformément dans $[0, 1]$ les composantes de la matrice R de taille (26,26) et en posant ensuite $V = RR'$.

On simule alors 150 réalisations de la loi multinormale $N_{26}(\mu, V)$ indépendamment de la valeur de $\hat{Y}$.

Remarques :

- Le nombre minimal d'individus dans chaque classe générée par l'algorithme a été fixé à 5.
- Les variables fictives ajoutées sont totalement indépendantes des variables discriminantes.

Les résultats sont présentés dans la Table 11.3.

Pas	Bornes / Variable	Lambda de Wilks	Statistique F	p-value de F
1	05-10	0.01		
1	longueur du pétale	0.06	1180.16	2.86E-91
2	05-10	0.06		
2	largeur du sépale	0.04	43.04	2.03E-15
3	05-10	0.04		
3	largeur du pétale	0.02	34.57	5.30E-13
4	05-10	0.02		
4	longueur du sépale	0.02	4.72	0.01

TAB. 11.3 – Résultats de l'algorithme obtenus sur les iris de Fisher

On constate d'après la Table 11.3 que notre algorithme détecte dès le premier pas la bonne partition en intervalles, c'est-à-dire celle correspondant aux trois espèces d'iris. Cette partition reste la même tout au long de la procédure, les résultats sont identiques à ceux de la sélection pas-à-pas classique. Aucune variable fictive ajoutée au jeu de données réelles n'a été entrée dans le modèle. Si l'on considérait le pas suivant, le regroupement resterait le même et la variable non discriminante X^{25} serait introduite, donnant lieu à une $p-$value de F de 0.16.

11.3.4 Les poissons du lac Laengelmavesi

Cent cinquante neuf poissons, répartis en 7 espèces (*brême, parkki, brochet, perche, gardon, éperlan* et *corégone*), ont été prélevés dans le lac Laengelmavesi en Finlande. Les effectifs de chaque espèce sont donnés dans la Table 11.4.

brême	parkki	perche	brochet	gardon	éperlan	corégone
35	11	56	17	20	14	6

TAB. 11.4 – Effectifs des neuf espèces de poissons

Pour chaque poisson sont mesurés son poids, sa hauteur, sa largeur et trois différentes mesures de sa longueur : du nez du poisson jusqu'au début de la queue, du nez jusqu'à l'entaille de la nageoire caudale, du nez jusqu'au bout de la queue. La hauteur et la largeur du poisson sont exprimées en pourcentage de la troisième variable de longueur.

La Table 11.5 donne la liste des variables sélectionnées pas-à-pas pour discriminer les 7 espèces de poissons.

variable	Lambda de Wilks	statistique F	p-value de F
hauteur	0.245	77.69	1.17E-43
longueur2	0.019	299.31	8.33E-81
longueur3	2.21E-03	186.77	1.08E-66
largeur	9.35E-04	33.72	1.97E-25
poids	5.18E-04	19.73	8.02E-17
longueur1	3.63E-04	10.36	1.49E-09

TAB. 11.5 – Résultats de la sélection pas-à-pas sur les données des poissons

On dispose donc de 6 variables explicatives quantitatives, auxquelles on ajoute 26 variables fictives non discriminantes, simulées de la même manière que pour les iris de Fisher.

Dans cet exemple, la variable $\hat{Y}$ correspond à l'espèce du poisson considéré et prend donc $r = 7$ valeurs différentes. On construit Y comme précédemment en attribuant une valeur au hasard parmi $k = 3$ valeurs possibles à chaque poisson d'une classe. Il faut retrouver la partition de bornes 3-6-9-12-15-18.

Remarque : Le nombre minimal d'individus dans chaque classe a été fixé à 5.

La Table 11.6 présente les résultats obtenus en appliquant notre procédure.

Pas	Bornes / variable	Lambda de Wilks	Statistique F	p-value de F
1	3-6-9-12-15-18	2.02E-04		
1	hauteur	0.245	77.69	1.17E-43
2	3-7-8-13-15-18	0.232		
2	largeur	0.084	43.95	1.08E-30
3	3-6-9-12-15-18	0.030		
3	longueur3	0.004	151.48	8.18E-61
4	3-6-9-12-15-18	0.004		
4	longueur2	9.35E-04	85.96	1.02E-45
5	3-6-9-12-15-18	9.35E-04		
5	longueur1	6.63E-04	10.06	2.67E-09

TAB. 11.6 – Résultats de l'algorithme sur les données des poissons

On observe à nouveau que notre procédure a permis de retrouver la partition en intervalles correspondant aux 7 espèces de poissons. Remarquons que notre procédure n'a pas sélectionné de variables simulées. Cependant, notre algorithme ne retient que cinq variables discriminantes, alors que la méthode classique introduit en plus la variable *poids*. Si on considérait le pas 6, la partition serait inchangée et la variable non discriminante X^{20} serait introduite, donnant lieu à une p-value de F de 0.22. Afin de savoir quel sous-ensemble de variables est le plus discriminant, des analyses discriminantes linéaire (LDA) et quadratique (QDA), les méthodes des $k - NN$ et CART ont été réalisées, permettant ainsi de comparer le pourcentage de bien-classés en leave-one-out (11.7).

Méthode	LDA (%BC)	QDA (%BC)	1-NN (%BC)	CART (%BC)
actuelle	94	99	98	67
classique	94	97	99	66

TAB. 11.7 – Comparaison des pouvoirs discriminants des deux sous-ensembles de variables par LDA et QDA en leave-one-out

La règle de classement linéaire donne le même résultat pour les deux sélections de variables (94% de bien-classés). La règle de classement quadratique en revanche est légèrement meilleure avec notre sélection de variables (99% de bien-classés contre 97%) et ce avec la variable *poids* en moins. Les méthodes du $1 - NN$ et CART donnent des résultats voisins. Notons que les résultats de CART sont mauvais dans les deux cas. Nous pouvons donc conclure que notre algorithme fournit une liste de prédicteurs dont le pouvoir discriminant est équivalent à celui obtenu par la méthode classique.

11.4 Conclusion et perspectives

Dans ce chapitre est présenté un algorithme permettant de regrouper en intervalles les modalités d'une variable qualitative ordinale et de sélectionner des variables discriminantes de cette partition. Son application à trois jeux de données a donné des résultats satisfaisants quant à son utilité. Ceci justifie donc son application à notre problème initial de prédiction de la dose réactogène. Il serait intéressant de considérer une sélection progressive, permettant de retirer des prédicteurs devenus inutiles au cours de la procédure, ainsi qu'une sélection descendante, en retirant une à une les variables de la liste de départ permettant l'initialisation. Enfin notons que le critère du Lambda de Wilks est particulièrement adapté dans le cas où l'on souhaite ensuite appliquer l'analyse discriminante linéaire. En effet, plus le Lambda de Wilks est petit, plus les moyennes conditionnelles du vecteur des variables explicatives sont éloignées. D'autres critères d'optimalité sont ainsi envisagés, notamment la maximalisation de la statistique du score, qui pourrait être utile pour utiliser ensuite une régression logistique.

Conclusion

Les buts principaux de cette thèse de doctorat étaient de mettre en évidence des variables permettant de faciliter le diagnostic précoce du cancer et de l'allergie à l'arachide.

Suite à l'étude présentée dans la partie cancer, il est envisagé de mener une étude clinique dans laquelle des anticorps dirigés contre des protéines aberrantes issues de l'infidélité de transcription seront mesurés sur des patients sains et cancéreux. Ainsi, en réalisant des analyses discriminantes, il sera possible de vérifier si les mesures de telles protéines permettent d'améliorer et de faciliter le diagnostic précoce des cancers.

De plus, certains marqueurs de l'allergie à l'arachide et de sa sévérité, comme les IgE dirigées contre *rAra-h1, rAra-h2* et *rAra-h3*, sont mis en évidence dans cette thèse, et des modèles prédictifs ont été construits à partir de ces variables. Notre objectif est maintenant de valider ces modèles sur d'autres patients et de rechercher de nouveaux prédicteurs. Par ailleurs, des développements de l'algorithme construit pour traiter la dose réactogène devraient être proposés par la suite.

En tout état de cause, une des clefs de ces projets reste à mon sens la collaboration étroite entre médecins, biologistes et statisticiens. En effet, les médecins, confrontés à des cas cliniques, posent les problèmes et fixent les enjeux de l'étude. Les biologistes doivent ensuite comprendre les phénomènes cellulaires ou moléculaires qui induisent la maladie et proposer de nouvelles variables d'intérêt. Les statisticiens conçoivent l'étude clinique, pour traiter ensuite les mesures faites sur les patients. Les résultats doivent ensuite être clairement transmis aux biologistes et aux médecins, en mettant l'accent sur le choix des méthodes utilisées. Il est en effet fondamental que les modèles proposés n'apparaissent pas à leurs yeux comme des boîtes noires. Les résultats obtenus peuvent ainsi amener les biologistes et les cliniciens à se poser des questions et à éventuellement remettre en cause certains fondements. De ce fait, de nouveaux projets peuvent être initiés et la boucle peut recommencer. Ceci repose naturellement en grande partie sur la qualité des rapports humains qui tissent les relations entre les représentants de chacune des trois communautés.

Je tiens ainsi à souligner que mener la collaboration entre l'Institut Elie Cartan de Nancy, le laboratoire Genclis et le Centre Universitaire Hospitalier de Nancy a constitué une part conséquente du travail du thèse, qui peut-être n'apparaît pas directement à la lecture de ce mémoire. Etre immergé dans un laboratoire de biotechnologies m'a permis d'acquérir quelques modestes connaissances en biologie et en médecine, qui sont par ailleurs essentielles pour formaliser le problème en termes mathématiques. De plus, un réel effort de vulgarisation des techniques de statistiques utilisées et de transfert des connaissances a été fait, afin que ces méthodes d'analyse puissent être réutilisées par les collaborateurs.

Appliquer des techniques de statistiques au diagnostic de certaines maladies permet au mathématicien d'apporter une modeste contribution aux recherches médicales dédiées à la guérison des malades. Pour cela, il est évidemment nécessaire que le statisticien puisse s'intéresser en parallèle à l'amélioration des techniques existantes et à la conception de nouveaux outils.

Bibliographie

[1] Brulliard M, Lorphelin D, Collignon O, Lorphelin W, Thouvenot B, Gothie E, Jacquenet S, Ogier V, Roitel O, Monnez J, *et al.*. Nonrandom variations in human cancer ESTs indicate that mRNA heterogeneity increases during carcinogenesis. *Proceedings of the National Academy of Sciences* 2007 ; **104**(18) :7522.

[2] Collignon O, Monnez JM, Vallois P, Codreanu F, Renaudin JM, Kanny G, Brulliard M, Harter V, Bihain B, Jacquenet S, *et al.*. Discriminant analyses of peanut allergy severity scores. *Journal of Applied Statistics* 2009 ; **soumis**.

[3] Karp G. *Biologie cellulaire et moléculaire*. De Boeck, 2004.

[4] Lodish H. *Biologie moléculaire de la cellule*. De Boeck Université, 1997.

[5] Cooper G. *La cellule : Une approche moléculaire*. De Boeck Université, 1999.

[6] Pollard T, Earnshaw W. *Biologie cellulaire*. Elsevier, 2004.

[7] World-Health-Organization. *Attributable Causes of Cancer in France in the Year 2000*. IARC, 2007.

[8] Bar-Hen A, Daudin J, Robin S. Comparaisons multiples pour les microarrays. *Journal de la Société française de statistique* 2005 ; **146**(1-2) :45–62.

[9] Dalmasso C, Broët P. Procédures d'estimation du false discovery rate basées sur la distribution des degrés de signification. *Journal de la Société française de statistique* 2005 ; **146**(1-2) :63–75.

[10] Benjamini Y, Hochberg Y. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995 ; **57**(1) :289–300.

[11] Storey J, Taylor J, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates : a unified approach. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 2004 ; **66** :187–205.

[12] Dalmasso C, Broet P, Moreau T. A simple procedure for estimating the false discovery rate. *Bioinformatics* 2005 ; **21**(5) :660–668.

[13] Storey J, Tibshirani R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* 2003 ; **100**(16) :9440–9445.

[14] Pounds S, Cheng C. Improving false discovery rate estimation. *Bioinformatics* 2004 ; **20**(11) :1737–1745.

[15] Pounds S, Morris S. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics* 2003 ; **19**(10) :1236–1242.

[16] Vogelstein B, Kinzler K. Cancer genes and the pathways they control. *Nature medicine* 2004 ; **10**(8) :789–799.

[17] Sjoblom T, Jones S, Wood L, Parsons D, Lin J, Barber T, Mandelker D, Leary R, Ptak J, Silliman N, *et al.*. The consensus coding sequences of human breast and colorectal cancers. *Science* 2006 ; **314**(5797) :268–274.

[18] Marshall E. Getting the noise out of gene arrays. *Science* 2004 ; **306**(5696) :630.

[19] Lodish H, Berk A, Zipursky S, Matsudaira P, Baltimore D, Darnell J. Molecular Cell Biology. 2000.

[20] Sherry S, Ward M, Kholodov M, Baker J, Phan L, Smigielski E, Sirotkin K. dbSNP : the NCBI database of genetic variation. *Nucleic acids research* 2001 ; **29**(1) :308.

[21] Hillier L, Lennon G, Becker M, Bonaldo M, Chiapelli B, Chissoe S, Dietrich N, DuBuque T, Favello A, Gish W, *et al.*. Generation and analysis of 280,000 human expressed sequence tags. *Genome Research* 1996 ; **6**(9) :807–828.

[22] Boguski M, Lowe T, Tolstoshev C. dbEST : database for expressed sequence tags. *Nature genetics* 1993 ; **4**(4) :332–333.

[23] Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology* 2000 ; **7**(1-2) :203–214.

[24] Conover W. *Practical Nonparametric Statistics*. Wiley & Sons, 1980.

[25] Evans D, Van der Kleij A, Sonnemans M, Burbach J, Leeuwen F. Frameshift mutations at two hotspots in vasopressin transcripts in post-mitotic neurons. *Proceedings of the National Academy of Sciences* 1994 ; **91**(13) :6059–6063.

[26] Benson D, Karsch-Mizrachi I, Lipman D, Ostell J, Wheeler D. GenBank : update. *Nucleic Acids Research* 2004 ; **32**(Database Issue) :D23.

[27] Pomerantz R, Temiakov D, Anikin M, Vassylyev D, McAllister W. A mechanism of nucleotide misincorporation during transcription due to template-strand misalignment. *Molecular cell* 2006 ; **24**(2) :245–255.

[28] Kashkina E, Anikin M, Brueckner F, Pomerantz R, McAllister W, Cramer P, Temiakov D. Template misalignment in multisubunit RNA polymerases and transcription fidelity. *Molecular cell* 2006 ; **24**(2) :257–266.

[29] Baggerly K, Morris J, Coombes K. Reproducibility of SELDI-TOF protein patterns in serum : comparing datasets from different experiments. *Bioinformatics* 2004 ; **20**(5) :777–785.

[30] Petricoin E, Ardekani A, Hitt B, Levine P, Fusaro V, Steinberg S, Mills G, Simone C, Fishman D, Kohn E, *et al.*. Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet* 2002 ; **359**(9306) :572–577.

[31] Bock S, Munoz-Furlong A, Sampson H, *et al.*. Fatalities due to anaphylactic reactions to foods. *Journal of Allergy and Clinical Immunology* 2001 ; **107**(1) :191–3.

[32] Astier C, Morisset M, Roitel O, Codreanu F, Jacquenet S, Franck P, Ogier V, Petit N, Proust B, Moneret-Vautrin D, *et al.*. Predictive value of skin prick tests using recombinant allergens for diagnosis of peanut allergy. *Journal of Allergy and Clinical Immunology* 2006 ; **118**(1) :250–256.

[33] Bindslev-Jensen C, Ballmer-Weber B, Bengtsson U, Blanco C, Ebner C, Hourihane J, Knulst A, Moneret-Vautrin D, Nekam K, Niggemann B, *et al.*. Standardization of food challenges in patients with immediate reactions to foods-position paper from the European Academy of Allergology and Clinical Immunology. *Allergy* 2004 ; **59**(7) :690–697.

[34] Roberts G, Lack G. Diagnosing peanut allergy with skin prick and specific IgE testing. *The Journal of Allergy and Clinical Immunology* 2005 ; **115**(6) :1291–1296.

[35] Peeters K, Koppelman S, van Hoffen E, van der Tas C, den Hartog Jager C, Penninks A, Hefle S, Bruijnzeel-Koomen C, Knol E, Knulst A. Does skin prick test reactivity to purified allergens correlate with clinical severity of peanut allergy ? *Clinical & Experimental Allergy* 2007 ; **37**(1) :108–115.

[36] Hourihane J, Grimshaw K, Lewis S, Briggs R, Trewin J, King R, Kilburn S, Warner J. Does severity of low-dose, double-blind, placebo-controlled food challenges reflect severity of allergic reactions to peanut in the community ? *Clinical & Experimental Allergy* 2005 ; **35**(9) :1227–1233.

[37] Müller H. Diagnosis and Treatment of Insect Sensitivity. *Journal of Asthma* 1966 ; **3**(4) :331–333.

[38] Sicherer S, Furlong T, DeSimone J, Sampson H. Self-reported allergic reactions to peanut on commercial airliners. *The Journal of allergy and clinical immunology* 1999 ; **104**(1) :186.

[39] Ewan P, Clark A. Long-term prospective observational study of patients with peanut and nut allergy after participation in a management plan. *The Lancet* 2001 ; **357**(9250) :111–115.

[40] Lewis S, Grimshaw K, Warner J, Hourihane J. The promiscuity of immunoglobulin E binding to peanut allergens, as determined by Western blotting, correlates with the severity of clinical symptoms. *Clinical & Experimental Allergy* 2005 ; **35**(6) :767–773.

[41] Friedman J, Hastie T, Tibshirani R. *The elements of statistical learning : data mining, inference and prediction*. Springer Series in Statistics, 2001.

[42] Nakache J, Confais J, Cazes P. *Statistique explicative appliquée*. Technip Paris, 2003.

[43] Monnez JM. *Cours d'Analyse des données MASTER 2 IMOI*. IECN, Nancy, France 2007.

[44] Celeux G. Analyse Discriminante sur variables continues. *Collection Didactique INRIA* 1990 ; **7**.

[45] Droesbeke J, Lejeune M, Saporta G. *Modèles statistiques pour données qualitatives*. Editions TECHNIP, 2005.

[46] Cristianini N, Shawe-Taylor J. *An introduction to support vector machines*. Cambridge university press, 2000.

[47] Zweig M, Campbell G. Receiver-operating characteristic (ROC) plots : a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993 ; **39**(4) :561–577.

[48] Jennrich R, Sampson P. Stepwise Discriminant Analysis. *Mathematical Methods for Digital Computers* 1960 ; .

[49] Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning : data mining, inference and prediction. *The Mathematical Intelligencer* 2005 ; **27**(2) :83–85.

[50] Buhlmann P, Hothorn T. Boosting Algorithms : Regularization, Prediction and Model Fitting. *Statistical Science* 2007 ; **22**(4) :477.

[51] Escofier B, Pages J. *Analyses factorielles simples et multiples, 3 ème édition*. Dunod, Paris, 1998.

[52] Lidholm J, Ballmer-Weber B, Mari A, Vieths S. Component-resolved diagnostics in food allergy. *Current Opinion in Allergy and Clinical Immunology* 2006 ; **6**(3) :234.

[53] Moneret-Vautrin D, Morisset M, Flabbee J, Beaudouin E, Kanny G. Epidemiology of life-threatening and lethal anaphylaxis : a review. *Allergy* 2005 ; **60**(4) :443–451.

[54] Moneret-Vautrin D, Guerin L, Kanny G, Flabbee J, Frémont S, Morisset M. Cross-allergenicity of peanut and lupin : the risk of lupin allergy in patients allergic to peanuts. *Journal of Allergy and Clinical Immunology* 1999 ; **104**(1) :883–888.

[55] Pollard T, Earnshaw W, Johnson G. *Cell biology*. Saunders Philadelphia, 2004.

[56] R Development Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2007. URL `http ://www.R-project.org`, ISBN 3-900051-07-0.

[57] Escofier B, Pagès J. Multiple factor analysis (AFMULT package). *Computational Statistics and Data Analysis* 1994 ; **18** :121–140.

[58] Costanza M, Afifi A. Comparison of stopping rules in forward stepwise discriminant analysis. *Journal of the American Statistical Association* 1979 ; **74**(368) :777–785.

[59] Lachenbruch P. *Discriminant analysis*. Hafner, New-York, 1975.

Recherche statistique de biomarqueurs du cancer et de l'allergie à l'arachide

Résumé

La première partie de la thèse traite de la recherche de biomarqueurs du cancer. Lors de la transcription, il apparaît que certains nucléotides peuvent être remplacés par un autre nucléotide. On s'intéresse alors à la comparaison des probabilités de survenue de ces infidélités de transcription dans des ARNm cancéreux et dans des ARNm sains. Pour cela, une procédure de tests multiples menée sur les positions des séquences de référence de 17 gènes est réalisée via les EST (Expressed Sequence Tag). On constate alors que ces erreurs de transcription sont majoritairement plus fréquentes dans les tissus cancéreux que dans les tissus sains. Ce phénomène conduirait ainsi à la production de protéines dites aberrantes, dont la mesure permettrait par la suite de détecter les patients atteints de formes précoces de cancer.

La deuxième partie de la thèse s'attache à l'étude de l'allergie à l'arachide. Afin de diagnostiquer l'allergie à l'arachide et de mesurer la sévérité des symptômes, un TPO (Test de Provocation Orale) est réalisé en clinique. Le protocole consiste à faire ingérer des doses croissantes d'arachide au patient jusqu'à l'apparition de symptômes objectifs. Le TPO pouvant se révéler dangereux pour le patient, des analyses discriminantes de l'allergie à l'arachide, du score du TPO, du score du premier accident et de la dose réactogène sont menées à partir d'un échantillon de 243 patients, recrutés dans deux centres différents, et sur lesquels sont mesurés 6 dosages immunologiques et 30 tests cutanés. Les facteurs issus d'une Analyse Factorielle Multiple sont également utilisés comme prédicteurs. De plus, un algorithme regroupant simultanément en classes des intervalles comprenant les doses réactogènes et sélectionnant des variables explicatives est proposé, afin de mettre ensuite en compétition des règles de classement. La principale conclusion de cette étude est que les mesures de certains anticorps peuvent apporter de l'information sur l'allergie à l'arachide et sa sévérité, en particulier ceux dirigés contre rAra-h1, rAra-h2 et rAra-h3.

Mots-clés : tests multiples, ARNm, cancer, Expressed Sequence Tag, substitution de nucléotides, infidélité de transcription, Analyse Factorielle Multiple, analyse discriminante, apprentissage statistique, classification supervisée, sélection de variables, allergie à l'arachide, test de provocation orale,immunologie

Development of statistical methods for the discovery of novel biomarkers for cancer or peanut allergy

Abstract

The first part of this doctoral dissertation deals with the research of cancer biomarkers. During transcription it was observed that some nucleotides are replaced mistakenly by others. We sought to compare the probabilities of these transcription infidelities in mRNA originating from normal and cancerous tissues. To do this, a multiple testing procedure was performed on the positions of 17 genes by considering their ESTs (Expressed Sequence Tag). The conclusion was reached that the proportions of these transcription errors are mainly increased in cancer tissues as compared to normal ones. This phenomenon would lead to the translation of aberrant proteins, whose detection could help in identifying patients with cancer.

The main goals of the second part are the diagnosis of peanut allergy and the prediction of its severity. Diagnosing peanut allergy and evaluating the intensity of the symptoms are currently accomplished with a double blind placebo controlled food challenge (DBPCFC). Patients are given increasing peanut doses until the first clinical reaction appears. Since DBPCFC can result in life-threatening responses, we propose an alternate procedure with the long term goal of replacing invasive allergy tests. Discriminant analyses of peanut allergy, DBPCFC score, the eliciting dose and the first accidental exposure score were performed in 243 allergic patients using 6 immunoassays and 30 skin prick tests. A Multiple Factorial Analysis was performed to use new factors as predictors. We also developed an algorithm for simultaneously clustering eliciting dose values and selecting discriminant variables. Our main conclusion is that antibody measurements provide information on the allergy and its severity, especially those directed against the peanut allergens rAra-h1, rAra-h2 and rAra-h3.

Keywords : multiple testing, mRNA, cancer, Expressed Sequence Tag, nucleotides substitution, transcription infidelity, Multiple Factorial Analysis, discriminant analysis, statistical learning, supervised classification, variable selection, peanut allergy, double blind placebo controlled food challenge,immunology