Université de Montréal

Algorithme Génétique spécifique à l'analyse de la susceptibilité à l'hypertension de la population du Saguenay-Lac-Saint-Jean

par
Louis-Philippe Lemieux Perreault

Département de biochimie
Faculté de médecine

Mémoire présenté à la Faculté des études supérieures
en vue d'obtention du grade de maîtrise
en bio-informatique

Décembre 2007

Université de Montréal
Faculté des études supérieurs
Ce mémoire intitulé :

Algorithme Génétique spécifique à l'analyse de la susceptibilité

à l'hypertension de la population du Saguenay-Lac-Saint-jean

présenté par :
Louis-Philippe Lemieux Perreault

a été évalué par un jury composé des personnes suivantes :

Damian Labuda, Ph.D.
président-rapporteur

Ettore Merlo, Ph.D.
directeur de recherche

Pavel Hamet, M.D. Ph.D.
codirecteur

Daniel Sinnett, Ph.D.
membre du jury

Abstract

Hypertension is a complex multifactorial disorder. It affects about a quarter of the world's population and is more prevalent in economically developed countries than in developing ones. This disorder is one of the principal risk factors for cardiovascular disease in Canada. There are several risk factors for hypertension falling into two broad categories : environmental and genetic. Genetic factors lead to a certain predisposition to hypertension in an individual given the particulars of his or her environment. The task of identifying these genetic factors should receive high priority within the research community so that we may properly diagnose and cure hypertension.

To study the genetic components causing a rise in blood pressure among a French Canadian population living in the Saguenay-Lac-Saint-Jean region, a genetic algorithm was designed and implemented. It uses haplotypes, derived from single nucleotide polymorphisms in linkage disequilibrium, in order to create a signature (a haplotype group) which could explain the susceptibility to hypertension within this population. Furthermore, in addition to using the genetic data generated by genotyping, the algorithm may use environmental data to enrich the search.

Graphs showing the relation between precision and recall (two values used to demonstrate the quality of the results) have been elaborated upon. It is therefore possible to determine through the use of our algorithm the maximal precision associated with a certain recall. Finally, the results were statistically validated using resampling methods.

Key words : heuristic, genetic algorithm, hypertension, haplotypes, SNPs

Résumé

L'hypertension est un désordre multifactoriel complexe. Elle touche environ un quart de la population mondiale et est beaucoup plus présente dans les pays développés que dans les pays en voie de développement. Cette maladie est l'un des principaux facteurs de risque des maladies cardiovasculaires au Canada. Plusieurs facteurs prédisposant à l'hypertension existent et peuvent être classés en deux grandes catégories : les facteurs environnementaux et les facteurs génétiques. La deuxième catégorie apporte une certaine prédisposition à l'hypertension face à l'environnement immédiat des porteurs. Il est important de déterminer la composante génétique responsable de cette prédisposition afin de bien diagnostiquer et guérir l'hypertension.

Afin d'étudier les composantes génétiques susceptibles d'entraîner une hausse de la pression artérielle chez une population Canadienne française habitant la région du Saguenay-Lac-Saint-Jean, un algorithme génétique fut imaginé et implanté. Celui-ci utilise des haplotypes, créés à partir de polymorphismes à simple nucléotide en déséquilibre de liaison, afin de créer une signature (ensemble d'haplotypes) pouvant expliquer la susceptibilité à l'hypertension des porteurs. Cet algorithme, en plus d'utiliser des données générées par génotypage, peut aussi utiliser des données relatives à l'environnement des sujets à l'étude.

Des graphiques mettant en relation la précision et le rappel, deux valeurs permettant de mettre en évidence la qualité des résultats, furent élaborés. Il est donc possible de savoir, à la suite d'une analyse avec notre algorithme, quelle est la précision maximale que nous pouvons obtenir lorsque nous considérons un rappel donné. Finalement, les résultats furent validés statistiquement à l'aide de la méthode de rééchantillonnage.

Mots clés : heuristique, algorithme génétique, hypertension, haplotypes, SNPs

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

ADN : Acide DésoxyriboNucléique

EA : algorithmes évolutionnaires (Evolutionay Algorithms)

ARN : Acide RiboNucléique

BMI : indice de masse corporel (Body Mass Index)

BP : pression sanguine (Blood Pressure)

CHUM : Centre Hospitalier de l'Université de Montréal

DBP : pression artérielle diastolique (Diastolic Blood Pressure)

GA : algorithme génétique (Genetic Algorithm)

LD : déséquilibre de liaison (Linkage Disequilibrium)

MCMC : Markov Chain Monte Carlo

Pb : Paires de Bases

pc : Probabilité de Croisement

pm : Probabilité de Mutation

QTL : locus à trait quantitatif (Quantitative Trait Loci)

RFLP : Restriction Fragment Length Polymorphism

RI : immigrant aléatoire (Random Immigrant)

RWS : Roulette Wheel Selection

SBP : pression artérielle systolique (Systolic Blood Pressure)

sh-n croisement à n-point avec mélange

SLSJ : Saguenay-Lac-Saint-Jean

SNP : polymorphisme de simple nucléotide (Single Nucleotide Polymorphism)

SQL : Structured Query Language

SUS : Stochastic Universal Selection

TSP : problème du voyageur de commerce (Traveling Salesman Problem)

Dédicaces

« "L'envie de savoir" est le plus puissant moteur humain. »
Bernard Werber

À ma copine, ma famille et mes amis(ies)...

Remerciements

Je voudrais remercier tous ceux et celles qui, dans les moments les plus sombres, n'ont cessé de croire en moi et de m'encourager à continuer et à persévérer. Je voudrais aussi remercier mon directeur et mon codirecteur de maîtrise, Ettore Merlo (*Ph.D.*) et Dr. Pavel Hamet, qui m'ont laissé entreprendre un projet intéressant et utile pour la société. Merci aussi pour leur encouragement et leur soutien. Un remerciement important est aussi adressé à mes collègues de laboratoire, François Gauthier, Audrey Noël, Pierre-Luc Brunelle et Alexandru Gurau. Ils ont toujours été là pour m'encourager et m'offrir du soutien technique (graphique et couleurs dans R, programmation en Python, utilisation de postgreSQL, simulations, etc.).

Merci à mes collègues de baccalauréat et de maîtrise, pour ces agréables moments passés en leur compagnie, qui ont su égayer les nombreux cours obligatoires du programme de bio-informatique. Que de bons souvenirs à garder de ces dîners, assis autour d'une soupe tonkinoise ou d'un plateau de Sushi...

Un merci spécial à ma copine, Marie-Josée, et à toute ma famille pour le soutien moral et psychologique constant, dans les moments de joie et de peine. Il est toujours agréable et stimulant de savoir que vous croyez en mes compétences.

Sans vous, rien n'aurait été possible!

Première partie

Introduction

1. Génétique humaine

Le génome de l'être humain est composé de 23 paires de chromosomes, soit 22 paires d'autosomes et 1 paire de chromosomes sexuels. Au total, le génome est composé d'environ 3×10^9 paires de bases et comporte plus de 22 000 gènes codant pour des protéines[1]. Les gènes sont les déterminants majeurs de la variation humaine. Il a été estimé que toute paire de gènes choisie aléatoirement chez deux individus diffère d'une simple base nucléotidique à tous les 200 à 300 bases [2].

1.1 Désordres génétiques

Pour la détermination d'un phénotype, un ou deux allèles peuvent être suffisants, mais généralement, il faut considérer l'interaction entre gènes ou avec un ou plusieurs facteurs environnementaux [3]. Parmi les différents désordres causés entièrement ou partiellement par des facteurs génétiques, trois grandes catégories sont reconnues [4] :

1. les désordres monogéniques ;
2. les syndromes chromosomiques ;
3. et les désordres multifactoriels.

1.1.1 Les désordres monogéniques

Les désordres monogéniques, de l'anglais « single-gene disorders », résultent d'une mutation sur l'un ou sur les deux allèles d'un gène situé sur un autosome, un chromosome sexuel ou sur de l'ADN mitochondrial [3]. Même s'ils sont rares (lorsque considéré

[1]Données tirées du site internet « Ensembl » [1] (`http://www.ensembl.org/Homo_sapiens/index.html`).

individuellement), les désordres monogéniques sont responsables d'une proportion non négligeable de maladies et de décès (soit environ 2 % de la population) [4]. Le syndrome de Liddle, une forme autosomale dominante de l'hypertension, est un exemple de désordre monogénique, causé par une réabsorption rénale de sodium par le rein [5].

1.1.2 Les syndromes chromosomiques

Les désordres chromosomiques sont présents s'il y a une altération visible dans le nombre ou la structure des chromosomes [3]. Les différentes maladies sont donc causées par un excès ou un déficit des gènes contenu dans certains segments de chromosome, ou sur des chromosomes en entier. Par exemple, une maladie très connue, la trisomie 21, est causée par un chromosome surnuméraire, le vingt-et-unième. Les désordres chromosomiques sont des phénomènes très courants et ont des répercussions importantes sur le fonctionnement du corps humain (phénotypes). En effet, environ 50 % de tous les avortements spontanés sont causés par les désordres chromosomiques [4].

1.1.3 Les désordres multifactoriels

Les désordres multifactoriels découlent de l'interaction d'un gène ou plus avec un ou plusieurs facteurs environnementaux. La contribution génétique prédispose l'individu concerné à l'action des différents agents environnementaux. Dans la plupart des cas de désordres multifactoriels, la nature des agents environnementaux ainsi que la prédisposition génétique ne sont pas connues et sont sujets d'intensives recherches [3].

1.2 Marqueurs génétiques

Afin de permettre l'étude des différents désordres génétiques énumérés précédemment, il est nécessaire de diminuer la quantité de données à considérer. Puisqu'il est difficile d'analyser à la fois les 3×10^9 paires de bases, les différentes équipes de recherche utilisent des marqueurs génétiques. Un marqueur génétique est une séquence

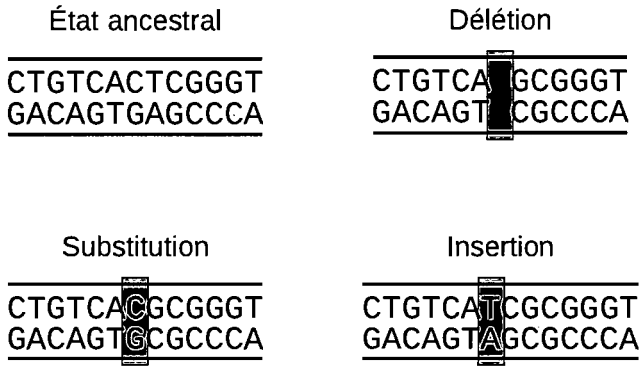


Figure 1.1: Les différents types de SNPs. Il y a trois types de SNPs générés à partir de l'état ancestral. Le premier type est créé par substitution, le deuxième type, par délétion et le troisième l'est par insertion. La figure est tirée de [3].

d'ADN connue permettant un balisage ; un ordonnancement des marqueurs donne une carte utile pour se situer physiquement dans le génome. Il existe plusieurs marqueurs différents, allant de petites régions d'un nucléotide à des régions beaucoup plus grandes. Les marqueurs utilisés dans le cadre de cette recherche sont les polymorphismes d'une seule paire de bases, nommés SNPs.

1.2.1 Polymorphisme de simple nucléotide

Les SNPs, de l'anglais « **S**ingle **N**ucleotide **P**olymorphism » sont des variations d'une seule paire de bases du génome entre individus d'une même espèce. Cette variation est très présente dans le génome humain et est distribuée de façon uniforme à travers celui-ci [4]. En effet, environ 85 % de la variation humaine est basée sur ces SNPs. Il y a trois types de SNPs qui peuvent être créés à partir de l'état ancestral. Le premier est la substitution où un nucléotide est remplacé par un autre. Le deuxième est la délétion, où une paire de bases est enlevée. Finalement, le troisième type représente l'insertion, où une paire de bases est ajoutée à la séquence [3]. La Figure 1.1 présente ces trois types de SNPs. Dans le cadre de notre étude, uniquement les SNPs créés par substitution sont utilisés.

L'intérêt pour l'utilisation des SNPs lors d'études génétiques de maladies humaines complexes ne cesse de grandir en raison de leurs nombreux avantages [6]. Les SNPs sont très abondants ; il existe environ un SNP pour chaque 500 à 1 000 paires de bases [7]. Ils sont donc utiles dans la détection de liaison génétique lorsqu'il y a présence de déséquilibre de liaison [8]. De plus, le génotypage de ces marqueurs est facilement automatisable, facilitant ainsi la cueillette de données [9].

1.2.2 Blocs d'haplotypes

Les différents SNPs peuvent être regroupés afin de former un ensemble nommé blocs d'haplotypes. Ces blocs représentent une région chromosomique où la recombinaison est quasi inexistante. Les différents SNPs inclus dans un bloc particulier sont en déséquilibre de liaison et seront donc hérités ensemble. Les différents allèles des n SNPs formeront les différentes combinaisons d'haplotypes regroupées dans un bloc distinct [10, 11]. Ce type de marqueur génétique est notamment utilisé par de nombreuses études portant sur la génétique des maladies complexes [12], telle l'hypertension.

2. Hypertension

L'hypertension est une maladie complexe multifactorielle touchant environ un milliard d'individus à travers le monde [13], soit environ un quart de la population mondiale [14, 15]. De plus, l'hypertension est plus présente dans les pays économiquement développés (37,3 %) que dans les pays en voie de développement (22,9 %) [15]. Elle est l'un des principaux facteurs de risque pour les maladies cardiovasculaires aux États-Unis ainsi qu'au Canada [16, 17] (ces dernières se classant au dixième rang des causes de mortalité dans le monde [18]). Depuis 1976, un grand bond en avant fut effectué en ce qui concerne le diagnostic, le traitement et le contrôle de l'hypertension. En effet, le pourcentage des personnes connaissant leur condition a augmenté de 20 % pour atteindre 70 % en 2000 selon le « National Health and Nutrition Examination Survey » (NHANES). Malgré tout cet effort, une pression artérielle trop élevée est responsable de 62 % des accidents vasculaires cérébraux et de 49 % des cardiopathies ischémiques à travers le monde [13]. Il est donc important de mieux comprendre les différentes facettes de cette maladie afin de créer un traitement adéquat aux personnes atteintes, puisque le nombre de ces dernières devrait, selon les prédictions statistiques de Kearney *et al.*, augmenter d'ici 2025 pour atteindre un total de 1,56 milliard, soit 29 % de la population mondiale [15]. Ces estimations sont basées sur les changements en taille et en âge des populations, et non sur le changement de l'incidence de l'hypertension.

2.1 Définition de l'hypertension

La pression artérielle, se mesurant en millimètre de mercure (mm Hg), est un trait quantitatif démontrant une distribution continue dans la population (une distribution

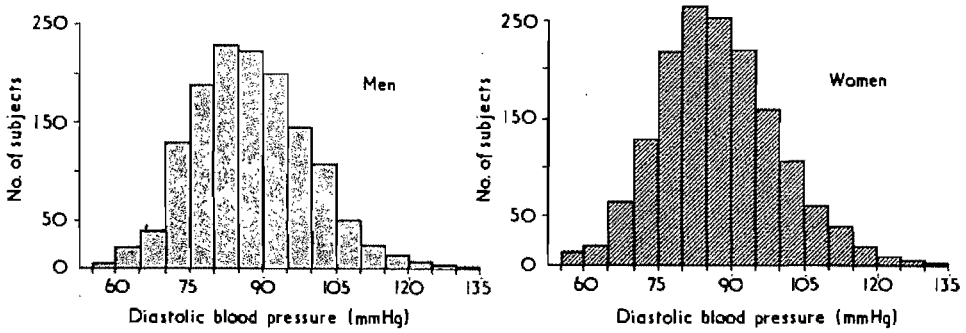


Figure 2.1: Distribution de la pression diastolique. Figure représentant la distribution de la pression artérielle diastolique dans une population [19].

en forme de cloche) (voir Figure 2.1). Il n'y a donc pas deux groupes d'individus (soit ceux avec ou sans hypertension), mais bien une distribution continue de pressions artérielles de la plus basse à la plus haute. Il y a peu de personnes se retrouvant dans les deux extrêmes (haute ou basse pression artérielle) et la majorité des individus se retrouvent au centre de cette même distribution.

La pression artérielle est composée de deux valeurs différentes : la pression systolique (SBP) et la pression diastolique (DBP). Selon le grand dictionnaire terminologique, la SBP est la *valeur de la pression existante dans le système artériel au moment de la systole cardiaque*, soit le moment où le cœur se contracte ou lorsque le sang est éjecté dans les artères. La DBP, quant à elle, est la *valeur de la pression existant dans le système artériel au moment de la diastole*, ou entre deux contractions cardiaques.

Selon le « Seventh Report of the Joint National Committe on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure », une personne est considérée comme souffrant d'hypertension lorsque sa SBP est supérieure à 140 mm Hg et/ou lorsque sa DBP est supérieure à 90 mm Hg (voir Tableau II.I de la page 8).

Il existe trois sous catégories d'hypertension artérielle : (1) l'hypertension systolique isolée (SBP $\geq$ 140 mm Hg et DBP $<$ 90 mm Hg), (2) l'hypertension diastolique

Tableau II.I: Classification de la pression artérielle chez les adultes. Tableau représentant la classification de la pression artérielle chez les adultes de 18 ans et plus[1] [13].

	Pression artérielle (mm Hg)		
Classification de BP	**SBP**		**DBP**
Normal	<120	et	<80
Préhypertension	120-139	ou	80-89
Hypertension stade 1	140-159	ou	90-99
Hypertension stade 2	≤160	ou	≤100

[1]Adulte ne prenant pas de médicament et en santé. Lorsque la SBP et la DBP ne sont pas dans la même catégorie, la plus haute prévaut.

isolée (SBP < 140 mm Hg et DBP ≥ 90 mm Hg) et finalement (3), l'hypertension systolique/diastolique combinée (SBP ≥ 140 mm Hg et DBP ≥ 90 mm Hg).

2.2 Types d'hypertension

La cause de l'augmentation de la pression sanguine chez un individu est inconnue pour près de 95 % des cas [20]. Cette variante de l'hypertension se nomme l'hypertension essentielle. Dans les autres cas, la cause peut être facilement trouvée ; il s'agit d'hypertension secondaire.

2.2.1 Hypertension essentielle

L'hypertension essentielle est la forme d'hypertension la plus rencontrée dans le monde de la médecine (dans 95 % des cas d'hypertension, environ). Elle a été décrite pour la première fois au 19^{e} siècle comme étant une forme d'hypertension qui se développe en l'absence d'évidence clinique de maladie rénale. Elle est présente chez une minorité de jeunes sujets, environ chez 25 % des adultes et chez la majorité de la population âgée de 60 ans et plus [21].

Guyton *et al.* apportent comme hypothèse que l'hypertension essentielle peut être caractérisée par une excrétion anormale du sodium ; les sujets hypertendus excrètent

une quantité donnée de sodium moins facilement que les sujets normotendus lorsque leur pression artérielle est traitée à des niveaux normaux (SBP < 120 et DBP < 80) [22]. Un rôle important du rein dans l'augmentation de la pression artérielle est donc à considérer.

Afin de déterminer les facteurs génétiques expliquant l'hypertension essentielle, des gènes ayant des rôles sur des voies biologiques régissant la pression sanguine dans le corps humain doivent être étudiés ; par exemple, les gènes responsables de l'altération de la régulation de la pression artérielle (dans les stades précoces de la maladie) et ceux qui sont responsables d'un changement dans la sensibilité à l'alimentation ou aux facteurs environnementaux. Plusieurs études tentent de trouver des gènes responsables de la susceptibilité au dommage des organes tels le cœur, les vaisseaux sanguins, le cerveau et les reins. Des dommages à de tels organes sont associés à un stade avancé de l'hypertension [20].

L'hypertension essentielle peut être séparée en trois classes distinctes : l'hypertension monogénique (ou hypertension mendélienne), l'hypertension chromosomique et l'hypertension multifactorielle.

Hypertension monogénique

L'hypertension monogénique (ou hypertension mendélienne) est une forme rare d'hypertension. Elle est causée par une mutation au niveau d'un gène unique ou d'un désordre chromosomique. Elle s'hérite de façon mendélienne de manière autosomique dominante ou autosomique récessive [23]. Une douzaine de formes d'hypertension mendélienne est répertoriée dans la littérature (voir Tableau XII.II de l'annexe, page xv).

La plupart des gènes identifiés comme étant responsables des formes d'hypertension mendéliennes jouent un rôle, de près ou de loin, dans les différentes voies métaboliques

nécessaires au maintien du niveau de sodium dans le sang par les reins.

Hypertension chromosomique

L'hypertension chromosomique résulte d'un désordre chromosomique (altération visible dans le nombre ou la structure des chromosomes). Le syndrome de Turner en est un exemple. Ce syndrome, affectant une naissance femelle sur 2 500 [24, 25], est caractérisé par l'absence complète ou partielle d'un des deux chromosomes X chez la femme [26]. Les maladies cardiovasculaires, telles que l'hypertension, sont la principale cause de décès chez les sujets atteints [27, 28].

Hypertension multifactorielle

L'hypertension multifactorielle est le type d'hypertension le plus rencontré. Il s'agit d'une maladie complexe, polygénique, qui est influencée par une multitude de variations chromosomiques impliquant de nombreux gènes, par l'interaction entre ceux-ci ainsi que par l'environnement immédiat de la personne atteinte, tels l'alimentation, le stress et autres [20]. Il est à noter que les gènes responsables de l'hypertension monogénique peuvent être impliqués dans l'hypertension multifactorielle.

2.2.2 Hypertension secondaire

L'hypertension secondaire est une augmentation de la pression artérielle qui résulte d'un problème sous-jacent, identifiable et parfois même corrigible. Les causes de l'hypertension secondaire sont multiples, allant d'un problème lors de la lecture de la pression par le médecin de famille (brassard trop petit, manches trop serrées, stress d'être chez le médecin) jusqu'à des maladies affectant les reins ou certaines glandes, par exemple (voir Tableau II.II, page 11). La cause principale de l'hypertension secondaire est la sténose de l'artère rénale [29].

Même si cette variante se retrouve chez environ 5 % des cas d'hypertension, cela

Tableau II.II: Distribution des différentes causes de l'hypertension. Tableau représentant les différentes causes de l'hypertension secondaire (tiré de [30], modifié de [31]).

Diagnostique	Pourcentage des Causes
Hypertension rénovasculaire	5–10
Hypertension induite par l'œstrogène	3–5
Aldostéronisme primaire	3–5
Phéochromocytome	<1
Lien avec la drogue (corticostéroïdes, sympathicomimétique, récréative)	<1
Syndrome de Cushing	<0,5
Hyperthyroïdisme	<0,5
Angéite	<0,5

représente un grand nombre de personnes atteintes, en raison de la forte prévalence de l'hypertension en général [32]. Il est donc important de faire la différence entre l'hypertension essentielle et l'hypertension secondaire, puisque cette dernière peut parfois être facilement traitée. Pour plus de détails sur les causes probables de l'hypertension secondaire (en complément au Tableau II.II), voir le Tableau XII.I de l'annexe (page xiv), qui décrit les différentes causes de l'hypertension en détail.

2.3 Facteurs de risques

Selon des études épidémiologiques, 30 % à 60 % des variations au niveau du phénotype de l'hypertension essentielle seraient dues à des facteurs génétiques. Les 40 % à 70 % restant seraient dues à des causes environnementales telles la diète et des facteurs psychoémotionnels tel le stress [33], etc.

Les facteurs de risques peuvent être classifiés en deux catégories : les facteurs qui ne sont pas modifiables, tels l'âge, le sexe, l'ethnicité et la génétique, et les facteurs qui peuvent être modifiés, telle l'alimentation. Une modification de ces derniers peut mener à une baisse significative de la pression artérielle et même prévenir une augmentation de celle-ci [34]. Ils ont donc comme impact l'atténuation de l'effet de la génétique sur le phénotype, par exemple. Voici quelques facteurs de risques pouvant causer l'hyper-

tension ou simplement modifier la pression sanguine de différentes personnes.

2.3.1 Sexe

La pression artérielle (systolique et diastolique) est différente en fonction du sexe de l'individu. En effet, les sujets du sexe féminin semblent avoir une pression moyenne significativement plus basse que les sujets masculins [35]. Selon cette même source, la raison de cette différence de pression peut s'expliquer de façon non-hormonale et hormonale.

De façon hormonale, le niveau d'œstrogène affectant la pression sanguine de façon indirecte [36] fluctue beaucoup chez la femme lorsque l'âge de la ménopause est atteint ou lorsque la femme porte un enfant [37, 38]. Le Tableau II.III de la page 13 présente la prévalence de l'hypertension chez une population canadienne, par âge et sexe. Nous y apercevons une augmentation marquée de la prévalence à l'hypertension chez les femmes ayant atteint l'âge de la ménopause. De façon physique, les femmes sont généralement de stature plus petite que les hommes. Cette caractéristique influence le diamètre de l'artère carotide, la compliance, ainsi que la compliance systémique[1]. Ces derniers influencent la pression sanguine à leur tour.

2.3.2 Âge

La pression artérielle (SBP et DBP) évolue grandement à mesure qu'un individu vieillit ; la moyenne de la pression artérielle tend à augmenter avec l'âge chez les deux sexes. Ce phénomène a été observé dans plusieurs études sur des populations occidentales [40].

Chez l'homme et la femme, la pression systolique semble augmenter de façon quasi linéaire en fonction de l'âge [41]. Elle est généralement plus haute chez l'homme avant

[1]Qui se rapporte à la grande circulation ou aux cavités cardiaques gauches. *Le Grand Dictionnaire terminologique*

Tableau II.III: Prévalence de l'hypertension par âge et sexe. Tableau représentant la prévalence de l'hypertension chez une population canadienne. Celle-ci augmente considérablement pour les femmes lorsque l'âge de la ménopause est atteint [39].

Âge (années)	Sexe	N	Hypertension (%)
18–34	Homme	5 755	11
	Femme	6 041	2
35–64	Homme	3 621	31
	Femme	3 741	21
65–74	Homme	2 000	56
	Femme	1 971	58
Tous	Homme	11 376	26
	Femme	11 753	18
Total		23 129	22

la soixantaine, mais augmente beaucoup chez la femme, pour dépasser celle de l'homme vers soixante-dix ans [17]. La courbe du haut de la Figure 2.2 de la page 13 représente l'évolution de la pression systolique en relation avec l'âge (et le sexe) des individus étudiés.

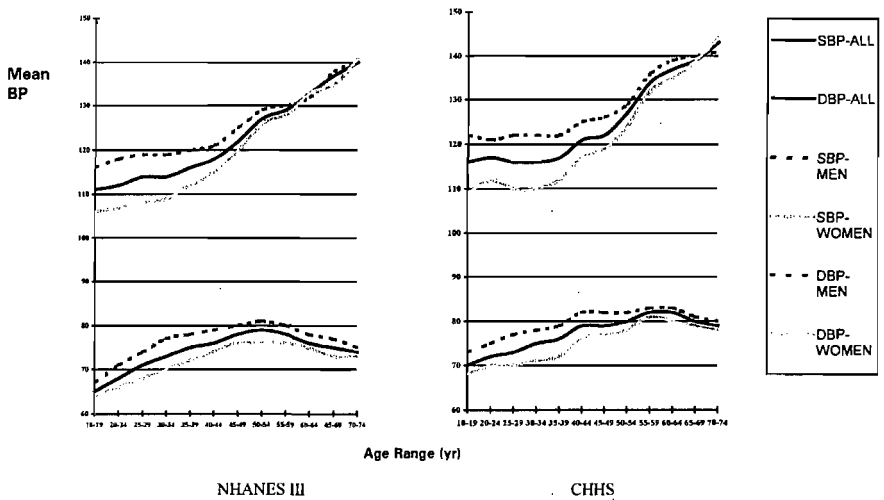


Figure 2.2: Distribution de la SBP et de la DBP selon l'âge et le sexe. Les données sont extraites du « Third National Health and Nutrition Examination Survey » (NHANES III) et du « Canadian Heart Health Surveys » (CHHS). Figure tirée de [17].

La pression diastolique évolue aussi en fonction de l'âge des individus, mais d'une façon différente de l'évolution de la pression systolique. En effet, la DBP augmente

jusqu'à la cinquantaine, mais diminue par la suite pour se stabiliser, et ce, pour les deux sexes [17]. La courbe du bas de la Figure 2.2 représente l'évolution de la pression diastolique en fonction de l'âge (et du sexe).

La prévalence de l'hypertension augmente beaucoup lorsque nous regardons les différentes tranches d'âges des individus. Par exemple, selon une étude de l'hypertension chez une population française [42], la prévalence augmente d'environ 40 % pour atteindre 57 % à l'âge de 50–64 ans (comparativement aux individus âgés de 18 à 49 ans). Cette augmentation de la prévalence chez les personnes âgées peut s'expliquer par la diminution de la compliance[2] artérielle [43], ce qui résulte d'un changement dans la nature et le contenu de collagène et d'élastine dans la paroi artérielle. Une autre statistique importante sur l'hypertension en relation avec l'âge est celle apportée par Vasan *et al.* [44] : le « residual lifetime risk ». Selon ces auteurs, une personne âgée de 55 ans a 90 % de chance de souffrir d'hypertension au cours de sa vie. Le Tableau XII.III de l'annexe à la page xvi représente cette statistique en détail.

2.3.3 Poids

Il est reconnu, depuis de nombreuses années, que l'hypertension est plus commune chez les personnes obèses que chez les personnes de poids normal. Une association existe donc entre le niveau de la pression artérielle et le degré de l'obésité [45]. L'explication de cette affirmation repose sur le fait qu'un gain de poids implique une augmentation du volume des tissus biologiques et, par conséquent, le besoin en apport sanguin (pour l'irrigation de tous les tissus et leur apport en énergie) [46]. Le débit cardiaque[3] augmente donc, ce qui a un effet direct sur la pression artérielle [20, 32, 47]. En effet, une diminution de 1 kg de masse corporelle réduit la SBP et la DBP de 1,2 et 1,0 mm Hg respectivement [45]. Spiegelman *et al.* ont d'ailleurs démontré que l'indice de masse corporelle (BMI) est un bon prédicteur de la pression artérielle [48].

[2]Mesure de la souplesse et des possibilités de distension. *Le Grand Dictionnaire terminologique*
[3]Volume de sang éjecté par chaque ventricule en une minute. *Le Grand Dictionnaire terminologique*

2.3.4 Diète

La diète est aussi un facteur de risque important de l'hypertension. Compte tenu de son apport en minéraux et autres, un individu peut changer sa pression artérielle (l'augmenter ou la diminuer). Il est donc crucial de se soucier de la diète des personnes atteintes de l'hypertension. Trois sels minéraux sont particulièrement à surveiller : le sodium, le calcium et le potassium. La consommation d'alcool peut aussi modifier la pression artérielle [49].

Sels minéraux (sodium, calcium & potassium)

Dès les années soixante, Dahl [50] pose l'hypothèse que, au niveau des populations, la moyenne de sels ingérés corrèle directement avec la prévalence de l'hypertension. Depuis, plusieurs études ont confirmé cette hypothèse en démontrant que l'excrétion du sodium est un facteur déterminant de la pression artérielle [51]. La relation avec la pression sanguine devient plus forte lorsque l'on prend en considération le ratio sodium/potassium urinaire plutôt que le taux de sodium uniquement [52, 53, 54, 55]. Il est important de noter que seulement l'ion de sodium provenant du NaCl provoque une expansion du volume de plasma sanguin et une augmentation de la pression artérielle [56, 57, 58]. La Figure 2.3 à la page 16 présente l'évolution de la pression artérielle en fonction de la quantité de sels ingérés, selon deux diètes spéciales.

Selon Hamet *et al.* [49], le calcium ingéré est aussi un déterminant significatif de la pression sanguine. Il joue le même rôle que le potassium, soit d'empêcher une augmentation de la pression artérielle lors d'une consommation excessive de chlorure de sodium. Il est donc important, dans notre civilisation où le sel fait partie de notre consommation habituelle, d'augmenter l'apport en calcium et en potassium afin de réduire l'impact sur la pression artérielle.

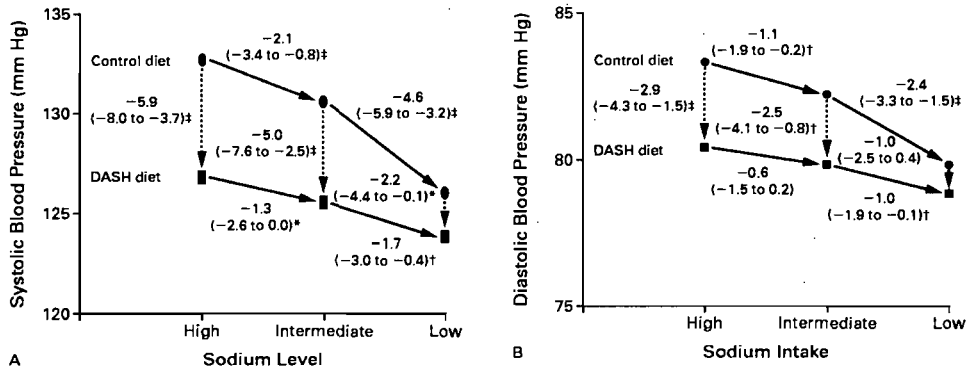


Figure 2.3: Effet sur la SBP (A) et la DBP (B) d'une diminution de sodium ingéré. Les chiffres à côté des lignes représentent le changement moyen dans la pression sanguine. L'intervalle de confiance 95 % est indiqué entre parenthèses. $^{*}P < 0.05$, $^{\dagger}P < 0.01$ et $^{\ddagger}P < 0.001$ [59].

Alcool

La consommation d'alcool est un facteur de risque de l'hypertension indépendant des autres facteurs [60, 61], comme démontré à la Figure 12.1 de l'annexe à la page xvii. Bien que les mécanismes expliquant l'augmentation de la pression sanguine suite à une consommation d'alcool ne soient pas complètement expliqués [62], environ 16 % des maladies liées à l'hypertension sont attribuables à la consommation d'alcool [63].

2.3.5 Génétique

Comme il est mentionné précédemment, environ 30 % à 60 % des variations au niveau du phénotype de l'hypertension essentielle sont dues à des facteurs génétiques. En effet, des études sur des enfants biologiques et adoptés ont clairement montrées l'effet de la génétique sur le phénotype de l'hypertension [64]. Il est donc important de déterminer ces facteurs familiaux. Pour ce faire, il existe quatre stratégies principalement utilisées pour l'étude de la génétique de l'hypertension essentielle :

1. étude de l'hypertension mendélienne ;
2. étude par l'approche de gènes candidats choisis selon leur fonction biochimique ou physiologique pouvant expliquer une augmentation de la pression chez un individu ;

3. étude des régions chromosomiques homologues à celles responsables d'une augmentation de la pression artérielle trouvées chez les modèles d'animaux, ou entourant les gènes reconnus comme étant liés à l'hypertension chez les animaux modèles ;
4. recherche à travers tout le génome pour une association (ou un déséquilibre de liaison).

Toutes ces méthodes ont apporté de nouvelles connaissances dans le domaine de l'hypertension, tel que présenté par Hamet *et al.* [14, 23]. Elles possèdent aussi leur lot d'avantages et d'inconvénients, qui sont bien résumés par Cowley Jr. [20], Ruppert *et al.* [34], Gong *et al.* [47] et Kato [65]. Hamet *et. al.* [66] résument les différentes découvertes réalisées à l'aide de recherches à travers tout le génome pour l'hypertension.

Une étude récente de l'hypertension sur la population du Saguenay-Lac-Saint-Jean a été menée par l'équipe de Hamet *et al.* Elle a permis la découverte de 46 loci pouvant expliquer l'apparition du phénotype de l'hypertension. Fait intéressant, les regroupements les plus importants des QTLs[4] se retrouvent sur les chromosomes 1 et 3 [14]. Ces résultats sont concordants avec d'autres études [67, 68, 69].

[4]Loci à trait quantitatif (*Quantitative Trait Loci*)

3. Algorithmes Génétiques

La complexité des problèmes d'optimisation ne cessant de grandir, l'intérêt pour les méthodes de recherche heuristique devint de plus en plus important durant les vingt dernières années [70]. Cette augmentation en popularité peut s'expliquer par le fait que ces méthodes peuvent trouver des optimums globaux dans des espaces de recherche désordonnés comportant un grand nombre de solutions possibles. Une de ces heuristiques s'est démarquée par ses avantages : sa facilité à s'adapter à une multitude de problèmes aussi différents les uns des autres, sa possibilité à dépasser les optimums locaux, son pouvoir de généralisation, etc. Il s'agit des algorithmes génétiques. Depuis 1989, cette méthode de recherche a fait ses preuves dans des domaines les plus variés allant des mathématiques à la médecine en passant par les sciences politiques et l'ingénierie [71]. Afin de bien comprendre l'importance de ces algorithmes évolutionnaires, certains concepts doivent préalablement être expliqués.

3.1 Optimisation

La théorie de l'optimisation peut se résumer ainsi. Soit l'espace de recherche $\mathcal{V}$ et la fonction

$$g : \mathcal{V} \mapsto \mathbb{R}.$$

Le problème d'optimisation est donc de trouver

$$\arg\max_{v \in \mathcal{V}} g$$

où v est un vecteur de variables de décision et g est la fonction objectif à optimiser [70]. La suite de ce mémoire sera réduite à l'utilisation de la maximisation, puisque tous les concepts peuvent être facilement appliqués à la minimisation. Afin de résoudre un problème d'optimisation de ce genre, plusieurs techniques de recherche existent.

3.2 Techniques de recherche

Un algorithme naïf pour résoudre un problème d'optimisation est la méthode d'essai et erreur. Cette technique repose sur l'énumération de toutes les solutions possibles afin de trouver la solution optimale (voir Figure 3.1). On peut résumer cette mé-

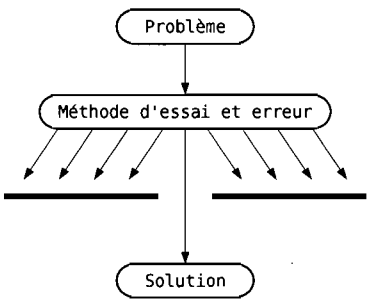


Figure 3.1: Technique d'essai et erreur. La technique de recherche par essai erreur, souvent appelée méthode naïve, peut être représentée par un schéma à trois niveaux [72].

thode ainsi : lorsque la résolution d'un nouveau problème est demandée, il est facile de calculer plusieurs solutions possibles, tout en éliminant les solutions inadéquates jusqu'à l'obtention d'une solution optimale. Par contre, si l'espace de recherche est très grand, le temps d'exécution d'un tel algorithme peut devenir rapidement un tel fardeau que l'énumération de toutes les solutions est impossible. Afin de pallier ce problème, plusieurs techniques de recherche ont été inventées. La Figure 3.2 de la page 20 [73] regroupe ces méthodes en trois grandes classes : les techniques de calculs différentiels, les techniques énumératives et les techniques aléatoires dirigées. Les techniques de calculs différentiels utilisent un ensemble de conditions nécessaires à la résolution d'un problème d'optimisation. Cette classe regroupe les algorithmes se basant sur la notion de recherche d'optimum par la notion de « hill-climbing ». Ils retrouveront donc un

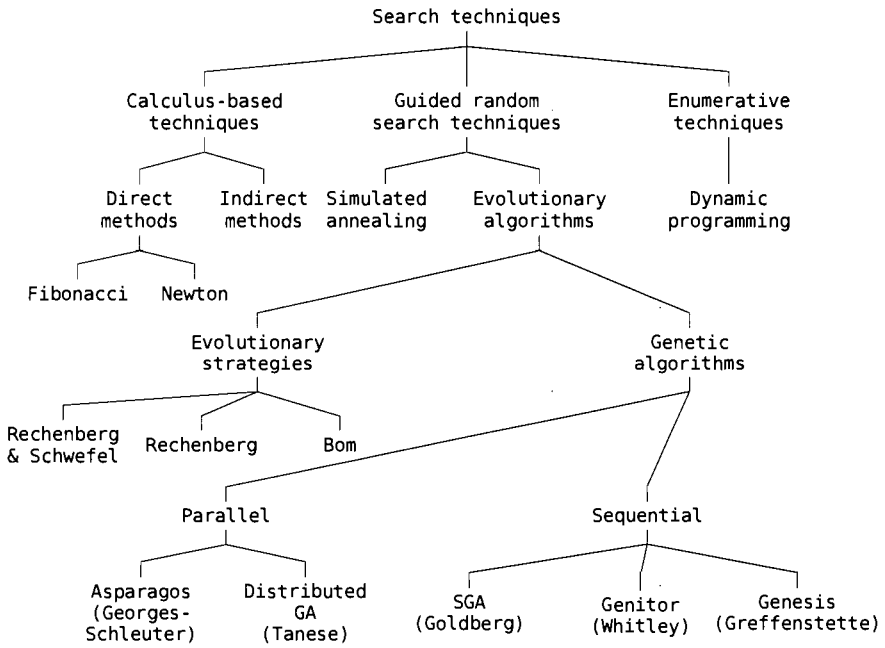


Figure 3.2: Classes de techniques de recherche. Représentation sous forme de trois classes des certaines techniques de recherche dont les algorithmes évolutionnaire [73].

optimum local en considérant le voisinage d'une solution de départ. L'utilisation de solutions initiales différentes vont permettre de trouver différents optimums locaux et, possiblement, l'optimum global. Les techniques énumératives sont des techniques ressemblant beaucoup à la méthode d'essai et erreur décrite précédemment. En effet, elles considèrent tous les points de l'espace de recherche. Encore une fois, un domaine de recherche trop vaste peut rendre impossible l'utilisation de cette catégorie de techniques de recherche. Les techniques aléatoires dirigées se basent sur les techniques énumératives. Elles se distinguent de ces dernières en ne considérant qu'un sous-ensemble de points de l'espace de recherche. Elles sont guidées par de l'information supplémentaire et par un générateur de nombres aléatoires. Parmi ces techniques, les « algorithmes évolutionnaires » sont basés sur les principes de la sélection naturelle. Proposées par Rechenberg et Schwefel dans les années 1970, elles s'adaptent à la majorité des problèmes d'optimisation. Un des algorithmes les plus performants de cette catégorie est l'algorithme génétique. Ce dernier présente de nombreux avantages :

- l'optimisation se fait à l'aide de paramètres continus ou discrets ;

- l'information sur le problème en question n'est pas nécessaire puisque l'algorithme utilise une fonction objectif à optimiser ;
- l'algorithme exécute une recherche simultanée de façon parallèle à travers l'espace de recherche et il s'agit donc d'une recherche globale où le temps de recherche est optimisé ;
- l'algorithme génétique est capable de contourner les optimums locaux afin de trouver le ou les optimums globaux ;
- un GA retournera une liste de solutions optimales au lieu d'une seule solution, permettant ainsi d'étudier d'autres alternatives à l'aide d'une seule recherche ;
- les GA fonctionnent sur des données numériques, expérimentales ou analytiques.

Par contre, un GA est coûteux en temps et en mémoire, puisqu'il fait évoluer une population de solutions (parfois très grande) et doit évaluer chaque solution à l'aide de la fonction objectif. Il y a aussi un risque de surapprentissage (« overfitting »). Le fonctionnement d'un algorithme génétique simple est expliqué au chapitre 6.

3.3 Historique des GA

L'idée d'utiliser des concepts biologiques pour le développement d'algorithmes, appelés algorithmes évolutionnaires (EA), est décrite pour la première fois dans l'ouvrage de Holland [74] paru en 1975. Holland s'est basé sur le concept de la sélection naturelle élaboré par Charles Darwin dans *The Origin of Species* de 1859. Le vocabulaire employé est directement calqué sur celui de l'évolution et de la génétique. En effet, il est question de population, de croisement, de mutation, de sélection, etc. Il existe trois types d'AE qui ont été développés dans les années soixante : les algorithmes génétiques, les stratégies d'évolution et la programmation génétique.

Les algorithmes génétiques manipulent une population de solutions possibles à un problème d'optimisation tout en les faisant évoluer à chaque itération. Pour ce faire, l'algorithme doit encoder les solutions en chaînes de caractères auxquelles des opé-

rations les modifiant et les mélangeant sont effectuées. Chacune de ces solutions est associée à une fonction objectif unique (« fitness function ») permettant de les classer selon leur performance [75]. L'algorithme génétique tentera donc de maximiser (ou minimiser) cette fonction objectif afin de résoudre le problème. Avant d'approfondir la théorie des GA, une certaine nomenclature doit être connue.

La nomenclature des GA est généralement basée sur la génétique. Les solutions (ou génotypes) parcourues par l'algorithme génétique sont encodées sous forme de chaînes de caractères que l'on nomme chromosomes. Chaque position du chromosome est appelée un gène. Les valeurs possibles de ce gène sont appelées allèles. Un individu est généralement composé d'un chromosome ainsi que de sa valeur de performance (ou phénotype). L'ensemble des individus forme la population d'un algorithme génétique. Une génération de solutions correspond aux solutions de la population au temps t. Plus d'informations à propos de l'algorithme génétique simple sont présentées au chapitre 6.

3.4 Applications des GA en bio-informatique

La bio-informatique est un domaine complexe où la quantité de données ne cesse de grandir. Les chercheurs sont donc à la recherche d'algorithmes puissants et efficaces permettant de résoudre des problèmes dans des temps raisonnables. Il existe plusieurs applications d'algorithmes génétiques pour la bio-informatique dans la littérature. Les GA furent par exemple utilisés avec succès afin d'aligner plusieurs séquences d'ADN, d'ARN ou de protéines [76, 77, 78], d'aligner des structures secondaires d'ARN [79] ou encore, afin de prédire la structure secondaire et tertiaire de protéines [80, 81]. L'utilisation des algorithmes génétiques la plus intéressante pour ce projet de recherche est celle permettant l'étude d'associations génétiques.

Jourdan *et al.* [82, 83] construisent des ensembles de SNPs en déséquilibre de liaison (haplotypes) capables d'expliquer un phénotype particulier. Leur algorithme uti-

lise deux procédures statistiques : EH-DIALL [84] et CLUMP [85]. Leurs données sont composées de 51 SNPs pour 133 individus dont les haplotypes de longueur 3 à 6 sont connues grâce à l'énumération complète de toutes les possibilités. Dans tous les cas, l'algorithme génétique construit les bons haplotypes.

L'objectif de Braaten *et al.* [86] est d'étudier l'application d'un algorithme génétique sur des haplotypes construits à partir de 7 RFLPs (« Restriction Fragment Length Polymorphism ») afin d'étudier l'hypercholestérolémie familiale. Lors des différents tests, l'algorithme a trouvé les haplotypes associés avec un haut taux de cholestérol dans le sang.

Finalement, Pardi *et al.* [87] utilisent un GA afin de déterminer le nombre de SNPs à considérer compte tenu du nombre de sujets à génotyper (le nombre de SNPs est dépendant du nombre de sujets). À partir d'un ensemble de départ contenant 45 SNPs, leur algorithme découvre que 4 SNPs ont un intérêt particulier et que ces derniers doivent être génotypés chez environ 908 sujets. Ces résultats sont comparés à d'autres études réalisées ([10, 88]) et sont concluants selon les auteurs.

Les algorithmes génétiques présentés ci-haut présentent une alternative efficace et rapide à d'autres algorithmes déjà existants permettant la création d'haplotype à partir de SNPs en déséquilibre de liaison (Haploview [89], entre autres). Toutefois, aucun algorithme génétique permettant la création de signature d'haplotypes n'a été trouvé dans la littérature.

Les études précédentes ont utilisé un nombre restreint de SNPs. Lorsqu'une quantité beaucoup plus volumineuse de marqueurs est utilisée, le nombre d'haplotypes créés augmente aussi. Il devient donc difficile de créer une signature (sous-ensemble) d'haplotypes associée à un phénotype quelconque à l'aide des différentes méthodes existantes.

3.5 Méthodes d'analyse des haplotypes

Bien qu'il existe plusieurs programmes permettant des études d'associations à partir de marqueurs génétiques comme les SNPs (FBAT [90], merlin [91], TDT [92, 93], PDT [94], etc.), il existe peu de logiciels libres complets permettant de créer un sous-ensemble d'haplotypes lié à un phénotype particulier (une signature d'haplotypes) à partir d'un grand ensemble de départ. Les chercheurs ont généralement recours à des méthodes d'apprentissage machine ou d'heuristique comme le recuit simulé ou la recherche taboue. Malheureusement, ces algorithmes ont tendance à être moins performants lorsque la dimension des données augmente, rendant difficile la création d'une signature d'haplotypes lorsque l'ensemble de départ est volumineux. De plus, ces méthodes ont besoin d'une valeur permettant de classifier les haplotypes selon leur degré d'association. Les méthodes suivantes (entre autres) doivent donc être utilisées afin de calculer cette valeur.

Les méthodes courantes de cartographie par déséquilibre de liaison (« Linkage disequilibrium mapping ») se sont montrées efficaces afin d'étudier des maladies mendéliennes, mais leur utilité reste à prouver dans le cas des maladies multifactorielles comme l'hypertension ; plus le modèle est complexe, plus les méthodes utilisées (tel que les chaînes de Markov monte carlo ou MCMC) deviennent lentes et moins performantes [95].

Dans le cas des méthodes statistiques d'association pour les études cas et contrôles, il est simple de comparer la fréquence des différents haplotypes entre les sujets atteints et ceux non atteints, en utilisant une des nombreuses approches statistiques utilisées pour la comparaison des fréquences d'allèles. Ces méthodes sont malheureusement dépendantes du nombre d'haplotypes et de leur longueur ; plus l'haplotype est long, plus il risque d'être composé d'allèles aléatoires. De plus, il est difficile d'ajuster les résultats en présence de covariables environnementales [95].

Les modèles de régression pour les haplotypes sont les méthodes les plus avantageuses dans ce genre d'étude, car il est possible d'ajuster les résultats pour des covariables non génétiques. De plus, l'interaction entre les haplotypes et l'environnement peut être modelée. Finalement, le diagnostic de régression est bien développé. Le désavantage de ces méthodes est que leurs performances ainsi que leur temps d'exécution sont dépendants du nombre d'haplotypes étudiés [95]. Le module « HaploStats » [96, 97] est un exemple de ces modèles de régression.

Selon Schaid [95], beaucoup de travail est nécessaire afin de développer des méthodes plus performantes pour détecter des associations subtiles entre les haplotypes et le phénotype à l'étude. Ces dernières sont utilisées avant les algorithmes permettant de créer les signatures d'haplotypes. L'algorithme génétique semble être un bon candidat puisqu'il permet de ne pas utiliser ces méthodes intermédiaires permettant de classifier les haplotypes selon leur degré d'association (comme il est requis par les algorithmes d'apprentissage machine). De plus, il a su se démarquer dans des espaces de recherche volumineux, compétitionnant ainsi avec les algorithmes d'apprentissage machine et les heuristiques telle la recherche taboue [71].

Deuxième partie

Méthodologie

4. Données à l'étude

Toutes les données utilisées au cours de ce projet furent compilées par le personnel médical du Centre Hospitalier de l'Université de Montréal (CHUM) et du complexe hospitalier de La Sagamie à Chicoutimi. Elles proviennent de nombreuses familles canadiennes-françaises natives de la région du Saguenay-Lac-Saint-Jean de la province de Québec. L'avantage de l'utilisation de cette population pour notre étude est la présence de nombreux documents généalogiques datant de l'époque des fondateurs venus de France au 17^{e} siècle. De plus, un important effet fondateur est produit grâce à un patrimoine génétique provenant d'un nombre restreint d'ancêtres communs [98, 99]. On peut définir l'effet fondateur comme étant le phénomène par lequel un groupe migre d'une population initiale vers une autre région où une nouvelle population sera fondée. Ceci a pour conséquence une homogénéité génétique dans cette nouvelle population pour certains loci. Cette homogénéité apporte une prévalence exceptionnellement élevée de certaines maladies spécifiques à cette région du SLSJ, comme la tyrosinémie et l'ataxie spastique [100, 101].

Les critères de sélection des familles sont la présence d'au moins deux enfants de mêmes parents atteints de l'hypertension essentielle[1], de dyslipidémie[2], âgés de 18 à 55 ans et présentant un BMI inférieur à 35 kg/m^2 [14].

Le prélèvement des différents phénotypes inclus dans la banque de données du

[1]ayant une SBP $>$ 140 mm Hg et/ou un DBP $>$ 90 mm Hg à deux occasions ou utilisant de la médication anti-hypertenseur

[2]présentant un niveau de cholestérol sanguin $\geq 5,2$ mmol/litre et/ou un niveau de HDL $\leq 0,9$ mmol/litre ou utilisant de la médication diminuant le taux de lipides

centre de recherche du CHUM est expliqué en détail par Kotchen *et al.* [102]. Cette même banque de données contient différents marqueurs génétiques, dont les SNPs. Ces SNPs ont été recherchés à l'aide de puces d'ADN *geneChips 50K Xba* de la compagnie *Affymetrix* [103]. Un total de 58 000 SNPs couvrant tous le génome furent identifiés chez 468 sujets. La distance moyenne entre ces marqueurs est de 49,8 kb.

5. Problématique et Hypothèse

Afin d'étudier la susceptibilité à l'hypertension de la population à l'étude, il faut trouver un ensemble de marqueurs génétiques (les SNPs) qui sont présents seulement chez les individus atteints d'hypertension. Pour se faire, la méthode naïve serait de considérer toutes les combinaisons possibles de SNPs. Ceci est infaisable à cause du grand nombre de SNPs par individu (concept de dimension). En effet, plus la dimension d'un problème est grande, plus il est difficile de le résoudre dans un temps adéquat. Afin de diminuer la dimensionalité des données, plusieurs approches existent. Hamet, Merlo *et al.* ont utilisé la construction de blocs d'haplotypes. Ces derniers, en plus de capturer l'information sur le déséquilibre de liaison, sont généralement plus informatifs que les SNPs individuels compte tenu de leur nature multiallélique et de leur plus grande hétérozygosité [2, 104].

Comme il est expliqué à la section 1.2.2, les blocs d'haplotypes sont constitués d'un ensemble de SNPs qui sont en déséquilibre de liaison. Ils permettent donc une diminution importante dans le nombre de dimensions à considérer dans la présente étude. En effet, le nombre de dimensions subit une réduction d'environ 85,3 %, passant de 58 000 SNPs par individu à 8 539 blocs d'haplotypes par individu. Malgré cette énorme réduction, le nombre de dimensions reste encore très élevé. En effet, si nous considérons la méthode naïve (qui considère toutes les combinaisons de blocs d'haplotypes possibles, soit environ $3,1 \times 10^{2570}$ combinaisons), il faudrait un temps exorbitant afin de résoudre le problème. Par exemple, si nous fixons une seconde par analyse de combinaisons (ce qui est minime comparé au temps réel), il faudrait environ $9,9 \times 10^{2562}$

années de calculs. Il est certain que cette estimation de temps est faite en considérant une méthode naïve de résolution du problème, mais elle donne une bonne idée de la quantité de travail à fournir afin de trouver des résultats concluants.

Notre hypothèse de recherche est que l'utilisation d'un algorithme génétique est appropriée afin de trouver une signature d'haplotype liée à l'hypertension artérielle. Alors que cet algorithme apporte un moyen efficace (temps de calcul et qualité de solution) pour résoudre un problème où la dimensionnalité fait échouer les approches traditionnelles, la bio-informatique permet de combiner les connaissances courantes du problème en question (hypertension, génétique humaine, tests statistiques, etc.) avec les connaissances algorithmiques et informatiques nécessaires au projet de recherche. De plus, à l'aide de ces heuristiques, il est possible d'ajouter de l'information telle que des données environnementales aux données génétiques déjà existantes. Notre objectif est donc de créer un algorithme génétique spécifique à l'étude de la susceptibilité à l'hypertension de la population du Saguenay-Lac-Saint-Jean tout en le validant statistiquement.

6. Algorithme génétique simple

Avant de pouvoir imaginer un algorithme génétique spécifique à la résolution d'un problème quelconque, il faut absolument comprendre le fonctionnement d'un GA simple. Son mécanisme est relativement facile à implanter [71]. Il suffit en fait de modifier (faire évoluer) des chaînes de caractères représentant des solutions à un problème. Un algorithme génétique comporte donc plusieurs opérateurs sur des chaînes de caractères, des techniques de sélection de chaînes selon leurs performances et autres.

Le fonctionnement d'un algorithme génétique s'effectue selon différentes étapes. Après la création d'une population initiale (génération 0), les individus (« parents ») sont sélectionnés selon leur performance et les opérateurs de croisement et de mutation sont appliqués à leur chromosome afin de créer de nouveaux individus (nommés « enfants »). Par la suite, un remplacement de génération est effectué pour remplacer les individus parents par les nouveaux individus (génération 1). Le processus est répété jusqu'à ce qu'un critère d'arrêt soit vérifié. La Figure 6.1 représente schématiquement le déroulement d'un algorithme génétique.

6.1 Représentation

Dans un algorithme génétique simple, chaque solution est encodée sous forme de chaîne de bits (vecteur). Par exemple, pour l'optimisation d'une fonction simple $f(x)$, l'encodage sous forme de chaînes de bits de la variable x est trivial. La longueur l des chromosomes dépendra de la précision de x que nous voulons. Par la suite, il suffit de convertir le chromosome de la base deux à la base dix afin d'obtenir sa perfor-

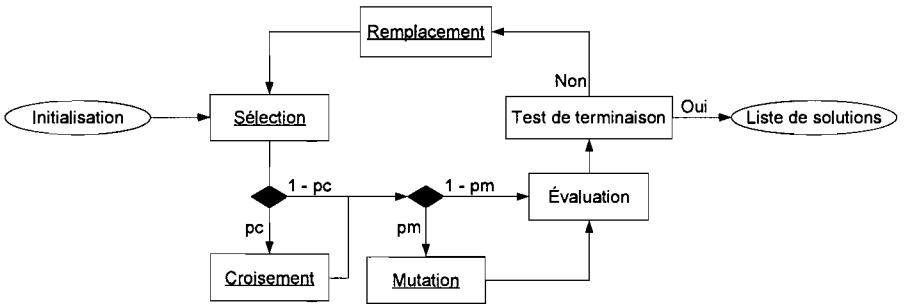


Figure 6.1: Représentation schématique d'un algorithme génétique. Représentation graphique des différentes étapes d'un algorithme génétique simple. Les losanges noirs correspondent à une variable aléatoire uniforme permettant d'omettre les opérateurs de croisement et de mutation avec une probabilité de $1 - pc$ et $1 - pm$ respectivement.

mance [105]. Beaucoup d'autres problèmes d'optimisation peuvent être encodés sous forme binaire tel le problème du sac à dos [70].

Certains problèmes plus complexes ne peuvent être représentés sous forme de chaînes de bits. En effet, l'encodage sous forme de chaînes d'entiers peut devenir nécessaire (dans le problème du TSP[1], par exemple, il existe trois représentations différentes). Bref, l'encodage va dépendre du type de problème d'optimisation à résoudre et il est impossible de les décrire en entier.

6.2 Population initiale

La création d'une population initiale est très simple. Il suffit de générer des chromosomes aléatoirement et de calculer leur performance respective selon la fonction objectif choisie pour optimiser le problème.

Dans la plupart des cas, selon Grefenstette [106] et Schaffer *et al.* [107], une population initiale de trente individus est largement suffisante. Par contre, Goldberg *et al.* [108] suggère que la taille de celle-ci doit augmenter de façon linéairement dépendante à la taille des chromosomes.

[1] « Traveling Salesman Problem »

6.3 Opérateur de sélection

La sélection des individus doit considérer la performance de ceux-ci afin de simuler la théorie de la sélection naturelle. Plusieurs méthodes de sélection différentes ont été créées, chacune ayant ses avantages et ses inconvénients. Elles sont de beaucoup différentes à la simple sélection aléatoire, décrite sommairement par Congdon [109].

6.3.1 Sélection proportionnelle

L'implémentation initiale de la sélection dans l'algorithme génétique imaginé par Holland est la sélection proportionnelle, ou « Roulette Wheel Selection » (RWS) [70]. Cette méthode utilise une distribution probabilistique où la probabilité de sélection d'un individu est directement proportionnelle à sa performance. La Figure 6.2 représente graphiquement ce type de sélection. Chaque individu est assigné à une portion du disque où l'angle est de $2\pi f_i/\bar{f}$ où f_i est la performance de cet individu, et $\bar{f}$ est la moyenne des performances de tous les individus [75].

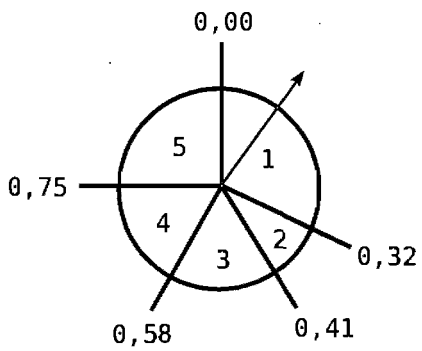


Figure 6.2: Sélection proportionnelle. Représentation de la sélection proportionnelle pour cinq individus avec 0,32, 0,09, 0,17, 0,17 et 0,25 comme valeur de performance respective [70]. Cet exemple représente la sélection de l'individu 1.

Lors de l'utilisation du RWS, certains problèmes majeurs peuvent apparaître. Un cas à problème survient lorsqu'il y a présence d'individus dont la performance est relativement haute comparée à la moyenne des performances de la population. Ces

individus seront appelés « super individus » puisque la probabilité de sélectionner ceux-ci augmente beaucoup. Ceci impliquera donc une convergence prématurée des solutions vers un optimum local, puisque l'opérateur de sélection ne permettra pas aux autres individus de performance plus faible d'être sélectionnés pour créer une nouvelle génération de solution. À l'inverse, si la variance des performances est faible (tous les individus ont une performance relativement semblable à $\bar{f}$), une sélection quasi aléatoire des individus aura lieu. La force de l'algorithme génétique à converger vers l'optimum global se retrouve donc réduite de beaucoup et elle se comparera à une « marche aléatoire » sur l'espace de recherche.

6.3.2 Sélection universelle stochastique

Un mauvais générateur de nombres aléatoires peut gêner le bon fonctionnement d'un algorithme génétique lorsque la RWS est utilisée comme opérateur de sélection. En effet, il s'agit d'une approche ayant une variabilité stochastique assez grande. Le nombre actuel de fois qu'un individu C est sélectionné (N_C) à une génération de solutions quelconque peut donc différer de beaucoup de l'espérance $E[N_C]$ attendue. La sélection universelle stochastiquee, ou « Stochastic Universal Selection », de Baker [110] (SUS) est une façon assez efficace de régler ce problème. Au lieu de choisir un seul individu à chaque étape de la RWS, le nombre N_C est produit pour chaque individu dans la population simultanément (voir Figure 6.3 page 35). Des études expérimentales produites par Hancock [111, 112] démontrent la supériorité de cette méthode.

6.3.3 Ajustement linéaire

Afin de contrecarrer le problème dû aux « super individus », la notion d'ajustement linéaire, ou « linear scaling », peut être ajoutée à la RWS. Une transformation linéaire pour convertir la fonction objectif g en f est souvent utilisée, c'est-à-dire

$$f = ag + b$$

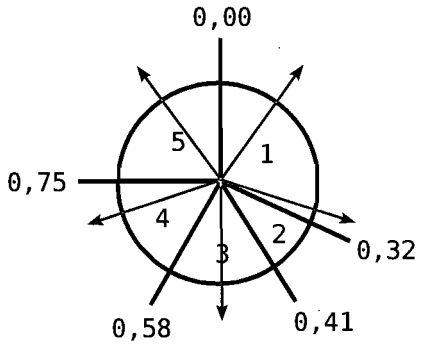


Figure 6.3: Sélection proportionnelle avec la sélection universelle stochastique. Représentation de la sélection proportionnelle pour cinq individus avec 0, 32, 0, 09, 0, 17, 0, 17 et 0, 25 comme valeur de performance respective. Utilisation de la sélection universelle stochastique [70]. Les individus 1 (deux fois), 3, 4 et 5 sont sélectionnés simultanément.

où les paramètres a et b sont obtenus à l'aide de deux conditions. Premièrement, les moyennes des deux fonctions doivent être les mêmes.

$$\bar{f} = \bar{g}$$

Deuxièmement, la performance maximale doit être égale à une constante multipliée par sa moyenne.

$$f_{\max} = \phi \bar{f}$$

La méthode d'ajustement linéaire permet de diminuer la performance des « super individus » tout en augmentant celle des individus dont la performance est de beaucoup inférieure à la moyenne (voir Figure 6.4 (a)). Cette méthode de sélection est plutôt encombrante, car les valeurs de performance maximales évoluent tout au long de la recherche et il faut sans cesse calculer de nouvelles valeurs des variables a et b [70]. Un autre problème survient lorsque des individus de très faibles performances (comparé à la moyenne) apparaissent. Ces derniers auront une nouvelle valeur de performance négative et empêcheront le bon déroulement de la RWS (voir Figure 6.4 (b)) [71].

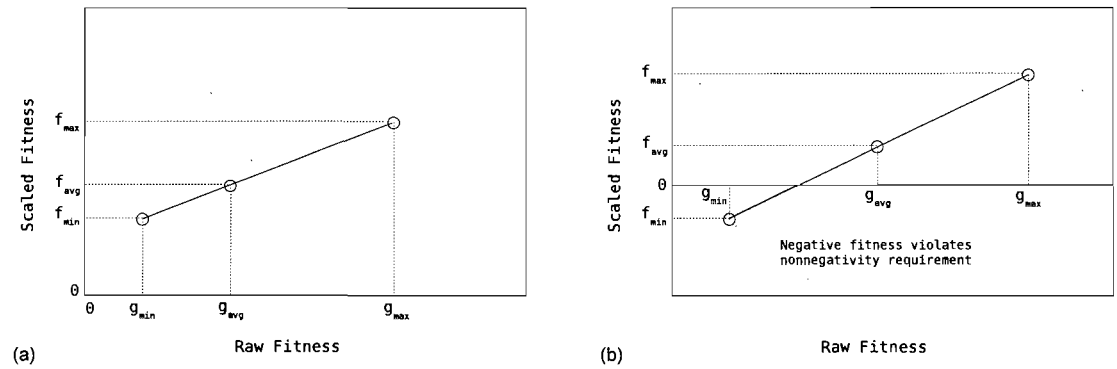


Figure 6.4: Ajustement linéaire. (a) Représentation de l'effet de l'ajustement linéaire. (b) Représentation d'un problème de l'ajustement linéaire. Il s'agit de l'apparition de valeurs négatives après ajustement [71].

6.3.4 Sélection par tournoi

Une méthode simple et efficace pour la sélection des individus est la sélection par tournoi (« tournament selection »). Un sous-ensemble de τ individus est choisi aléatoirement. Leur performance est ensuite comparée. Le gagnant, c'est-à-dire l'individu le plus performant, est ensuite sélectionné pour la reproduction. Lors d'une itération de l'algorithme (génération de n nouveaux individus), chaque individu sera comparé τ fois en moyenne [70].

Il existe deux formes de tournoi : le tournoi strict (« strict tournament ») et le tournoi doux (« soft tournament »). La première forme correspond à une probabilité $p = 1$ de choisir l'individu le plus performant à chaque tournoi. La deuxième forme correspond à une probabilité $0,5 < p \leq 1$ de choisir l'individu le plus performant à chaque tournoi. Ceci permet donc de modifier la pression sélective lors de l'évolution des solutions de l'algorithme génétique [70]. Une forte pression de sélection permettra la convergence rapide vers une solution optimale, parfois locale. Une faible pression de sélection aura pour conséquence de visiter une plus grande partie de l'espace de recherche, mais de façon moins dirigée vers l'optimum global.

6.3.5 Méthode des rangs et SUS

Une autre technique de sélection d'individus grandement utilisée est la méthode des rangs conjointement à la méthode SUS. En premier lieu, les individus doivent être classés en ordre de performance f (ordre non croissant ou non décroissant). Selon leur ordre, chaque individu recevra un rang i. Par la suite, une nouvelle mesure de performance f' est calculée :

$$f'_i = \max - \frac{(\max - \min)(i-1)}{n-1}.$$

Finalement, il suffit de générer un nombre aléatoire unique et d'ajouter 1 à ce nombre afin de sélectionner tous les individus. La Figure 6.5 représente graphiquement la méthode des rangs. L'inconvénient de cette méthode est qu'elle est insensible à l'amplitude de la différence entre les performances de deux individus.

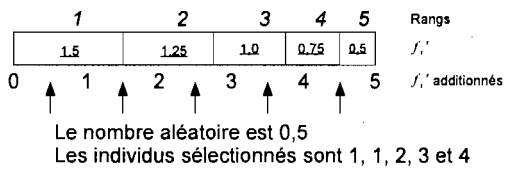


Figure 6.5: Méthode des rangs. Figure représentant la méthode de sélection par rangs utilisée conjointement avec la SUS.

6.4 Opérateur de croisement

Une fois l'opérateur de sélection appliqué à la population de la génération au temps t, une génération intermédiaire est créée. L'opérateur de croisement sélectionnera donc deux individus aléatoirement et formera deux descendants possédant des caractéristiques issues des deux parents. L'opérateur de croisement va donc favoriser l'exploration de tout l'espace de recherche, mais dirigé vers des points précis par les caractéristiques de ses deux parents.

Il existe plusieurs opérateurs de croisement différents, mais tous ont été créés par analogie au phénomène de croisement dans les molécules d'ADN, lors de la division cellulaire pour la formation des gamètes.

6.4.1 Croisement à n-point

L'idée de base du croisement à n-point peut se résumer en trois étapes. Pour commencer, les deux individus parents, choisis aléatoirement, sont alignés. Soit les parents p^0 et p^1 de taille dix (dix gènes).

$$p^0 = \quad p_0^0 \quad p_1^0 \quad p_2^0 \quad p_3^0 \quad p_4^0 \quad p_5^0 \quad p_6^0 \quad p_7^0 \quad p_8^0 \quad p_9^0$$

$$p^1 = \quad p_0^1 \quad p_1^1 \quad p_2^1 \quad p_3^1 \quad p_4^1 \quad p_5^1 \quad p_6^1 \quad p_7^1 \quad p_8^1 \quad p_9^1$$

Ensuite, il faut choisir aléatoirement n points de brisure afin de fragmenter chaque parent en $n+1$ fragments. Si $n = 3$,

$$p^0 = \quad p_0^0 \quad p_1^0 \;|\; p_2^0 \quad p_3^0 \quad p_4^0 \quad p_5^0 \;|\; p_6^0 \;|\; p_7^0 \quad p_8^0 \quad p_9^0$$

$$p^1 = \quad p_0^1 \quad p_1^1 \;|\; p_2^1 \quad p_3^1 \quad p_4^1 \quad p_5^1 \;|\; p_6^1 \;|\; p_7^1 \quad p_8^1 \quad p_9^1$$

La dernière étape consiste à créer les deux enfants c^0 et c^1 à l'aide des fragments des parents créés précédemment. Le premier enfant, c^0, est créé en copiant le premier fragment du premier parent (p^0), le deuxième fragment du deuxième parent (p^1), le troisième fragment du premier parent et ainsi de suite. Le deuxième enfant s^1 est créé de la même manière, mais en inversant l'ordre des parents (en commençant par le premier fragment du deuxième parent p^1).

$$\begin{array}{ll} p^0 = & p_0^0 \quad p_1^0 \;|\; p_2^0 \quad p_3^0 \quad p_4^0 \quad p_5^0 \;|\; p_6^0 \;|\; p_7^0 \quad p_8^0 \quad p_9^0 \\ p^1 = & p_0^1 \quad p_1^1 \;|\; p_2^1 \quad p_3^1 \quad p_4^1 \quad p_5^1 \;|\; p_6^1 \;|\; p_7^1 \quad p_8^1 \quad p_9^1 \\ \hline c^0 = & p_0^0 \quad p_1^0 \;|\; p_2^1 \quad p_3^1 \quad p_4^1 \quad p_5^1 \;|\; p_6^0 \;|\; p_7^1 \quad p_8^1 \quad p_9^1 \\ c^1 = & p_0^1 \quad p_1^1 \;|\; p_2^0 \quad p_3^0 \quad p_4^0 \quad p_5^0 \;|\; p_6^1 \;|\; p_7^0 \quad p_8^0 \quad p_9^0 \end{array}$$

Le nombre de points de croisement devrait théoriquement être choisi aléatoirement, mais de meilleurs résultats sont trouvés à l'aide de 3 points ou moins. Un des avantages/inconvénients de l'opérateur de croisement à n-point est que la probabilité que deux gènes « voyagent » ensemble est affectée par la distance entre ces deux gènes [113].

6.4.2 Croisement à n-point avec mélange

Le croisement à n-point avec mélange ($sh - n$) est très semblable au croisement à n-point. La procédure est la même, sauf que l'ordonnancement des gènes est mélangé avant de créer les $n + 1$ fragments. À la fin du croisement, les gènes retrouvent leur ordonnancement initial.

Soit les parents p^0 et p^1 de l'exemple précédent ; il suffit de les mélanger de façon à conserver les « colonnes » de gènes. Ainsi, un gène à la position i dans le parent p^0 doit se retrouver à la position i du parent p^1.

$$p^0 = \begin{array}{cccccccccc} p_2^0 & p_5^0 & p_6^0 & p_9^0 & p_0^0 & p_7^0 & p_8^0 & p_4^0 & p_3^0 & p_1^0 \end{array}$$

$$p^1 = \begin{array}{cccccccccc} p_2^1 & p_5^1 & p_6^1 & p_9^1 & p_0^1 & p_7^1 & p_8^1 & p_4^1 & p_3^1 & p_1^1 \end{array}$$

Les deux prochaines étapes sont les mêmes que le croisement à n-point, soit la séparation des parents en $n + 1$ fragments et la création des deux enfants c^0 et c^1.

$$\begin{array}{ll} p^0 = & \begin{array}{cc|cccccc|cccc} p_2^0 & p_5^0 & p_6^0 & p_9^0 & p_0^0 & p_7^0 & p_8^0 & p_4^0 & p_3^0 & p_1^0 \end{array} \\ p^1 = & \begin{array}{cc|cccccc|cccc} p_2^1 & p_5^1 & p_6^1 & p_9^1 & p_0^1 & p_7^1 & p_8^1 & p_4^1 & p_3^1 & p_1^1 \end{array} \\ \hline c^0 = & \begin{array}{cc|cccccc|cccc} p_2^0 & p_5^0 & p_6^1 & p_9^1 & p_0^1 & p_7^1 & p_8^0 & p_4^1 & p_3^1 & p_1^1 \end{array} \\ c^1 = & \begin{array}{cc|cccccc|cccc} p_2^1 & p_5^1 & p_6^0 & p_9^0 & p_0^0 & p_7^0 & p_8^1 & p_4^0 & p_3^0 & p_1^0 \end{array} \end{array}$$

Finalement, il s'agit de replacer les chromosomes selon l'ordre des parents au départ.

$$\begin{array}{ll} p^0 = & \begin{array}{cccccccccc} p_0^0 & p_1^0 & p_2^0 & p_3^0 & p_4^0 & p_5^0 & p_6^0 & p_7^0 & p_8^0 & p_9^0 \end{array} \\ p^1 = & \begin{array}{cccccccccc} p_0^1 & p_1^1 & p_2^1 & p_3^1 & p_4^1 & p_5^1 & p_6^1 & p_7^1 & p_8^1 & p_9^1 \end{array} \\ \hline c^0 = & \begin{array}{cccccccccc} p_0^1 & p_1^1 & p_2^0 & p_3^1 & p_4^1 & p_5^0 & p_6^1 & p_7^1 & p_8^0 & p_9^1 \end{array} \\ c^1 = & \begin{array}{cccccccccc} p_0^0 & p_1^0 & p_2^1 & p_3^0 & p_4^0 & p_5^1 & p_6^0 & p_7^0 & p_8^1 & p_9^0 \end{array} \end{array}$$

Bien que les gènes seront différents du résultat du croisement à n-points, le nombre de gènes copiés d'un parent sera le même avec l'opérateur *sh-n*. La différence majeure est que la probabilité que deux gènes « voyagent » ensemble n'est plus affectée par la distance séparant ces deux gènes, puisque les gènes sont ordonnancés aléatoirement [113].

6.4.3 Croisement uniforme

La description complète du croisement uniforme se retrouve dans l'ouvrage de Syswerda [114]. Soit deux parents de longueur l, chaque parent copie $l/2$ gènes à chaque enfant, où la sélection des gènes est déterminée de façon indépendante à l'aide d'une variable aléatoire suivant une loi normale. De façon mathématique, le croisement uniforme peut se représenter ainsi :

$$enfant_1 = a_1, a_2, \ldots, a_n \,|\, a_i \in parent_1 \text{ si } \tau < 0.5,\ a_i \in parent_2 \text{ sinon}$$

où $\tau \sim Uni(0,1)$, une variable aléatoire uniforme entre 0 et 1. Si $enfant_1$ obtient l'allèle a_i du $parent_1$, l'$enfant_2$ obtient l'allèle a_i du $parent_2$, et vice-versa.

Il est plutôt difficile de choisir un opérateur de croisement en particulier pour un problème donné. Selon Pawlowsky [113], le choix de l'opérateur adéquat est dépendant du problème à résoudre ainsi que des différents paramètres évolutifs utilisés. Il est donc important, lors de la création d'un algorithme génétique, d'essayer différents opérateurs de croisement. Tous les opérateurs de croisement présentés donnent de bons résultats lors de tests empiriques.

Il existe plusieurs autres opérateurs de croisement dans la littérature. Par contre, ceux-ci ne sont pas reproductibles, car ils nécessitent un encodage des solutions en chromosomes différents de celui décrit à la section 6.1. Ils seront donc utilisables seulement dans une classe très restreinte de problèmes d'optimisation. Par exemple, dans le problème du TSP, il existe environ six opérateurs de croisement différents [115, 116, 117, 118, 119].

6.5 Opérateur de mutation

Une fois les individus enfants créés suite à l'application de l'opérateur de croisement, l'opérateur de mutation est appliqué sur son résultat. Cet opérateur permet d'ajouter du bruit et empêche ainsi l'homogénéité des individus dans la population à une génération quelconque. Il empêche donc l'évolution des individus de figer dans un optimum local.

Lors de l'utilisation d'un encodage binaire, l'opérateur de mutation est très simple. Un gène est choisi aléatoirement et la valeur de l'allèle de ce gène est changée. Un masque est généré en utilisant une distribution de Bernoulli permettant de modifier l'allèle d'un gène avec une faible probabilité. Il suffit de faire l'addition, sans retenue, du masque et du chromosome de l'individu, comme l'illustre la Figure 6.6. Dans tous les autres types d'encodages, l'opérateur de mutation fonctionne de la même manière : il modifie la valeur à une position donnée, et ce, de façon aléatoire.

```
Mutation Mask   0 1 0 0 0 1 0
        Child   1 1 0 1 1 0 1  ⊕
                ---------------
Mutated Child   1 0 0 1 1 1 1
```

Figure 6.6: Opérateur de mutation. Illustration de l'opérateur de mutation à l'aide d'un masque [70].

6.6 Opérateur de remplacement

Une fois tous les enfants créés suite à l'opérateur de sélection, de croisement et de mutation, il faut remplacer certains, voir tous les individus de la population de la génération au temps t afin de créer une nouvelle génération (au temps $t + 1$). Il existe trois façons d'y arriver [109].

6.6.1 « Generation gap »

Au lieu de remplacer toute la population de la génération au temps t, certains de ces individus peuvent être conservés pour la génération suivante. Normalement, ces individus sont sélectionnés aléatoirement afin de maintenir une certaine diversité et d'éviter une convergence trop rapide de l'algorithme vers un optimum local. Cette stratégie implique que le nombre d'enfants nouvellement créés soit inférieur au nombre total d'individus dans la population [120].

6.6.2 Élitisme

Cette stratégie est similaire à la précédente. Par contre, les individus amenés à rester d'une génération t à une autre $(t+1)$ sont les individus les plus performants. Ceci implique donc que le meilleur individu rencontré au cours de la recherche se retrouvera dans la population finale (lorsque le critère d'arrêt sera atteint) [121].

6.6.3 « Steady-State »

Afin d'utiliser cette stratégie de remplacement générationnel, il faut créer un nombre d'enfants égal au nombre total d'individus dans la population. Ensuite, les deux enfants créés remplaceront les deux parents qui les ont créés, ou deux individus choisis aléatoirement dans la population au temps t.

6.7 Immigrant aléatoire

Afin d'éviter la convergence prématurée de l'algorithme génétique ainsi que l'homogénéisation de la population, la stratégie d'immigrant aléatoire a souvent été utilisée. Elle consiste à remplacer les individus les moins performants de la population par des individus nouvellement créés de la même façon aléatoire que la création de la population initiale. L'opérateur d'immigrant aléatoire est généralement appliqué lorsque le chromosome le plus performant rencontré au cours de l'exécution de l'algorithme

génétique n'a pas changé depuis un certain nombre d'itérations.

6.8 Fondements mathématiques

Une fois les rouages de l'algorithme génétique connus, il est important de se pencher sur ses fondements mathématiques. À ce moment, il n'existe aucune preuve de convergence asymptotique des algorithmes génétiques vers un optimum global. Il existe toutefois des résultats théoriques indiquant que la recherche est « sensée ». Les fondements mathématiques reposent sur la théorie des schémas [105], du problème du bandit armé à k bras (« k-armed bandit problem ») [70], ainsi que sur l'hypothèse des blocs de construction (« building blocks hypothesis ») [71].

6.8.1 Théorie des schémas

Un schéma est construit en introduisant un symbole « don't care » ('$*$') dans l'alphabet des gènes. Dans le cas d'un problème binaire, l'alphabet d'un schéma sera composé de $\{0, 1, *\}$. Un schéma représente toutes les chaînes de caractères qui satisfont toutes les positions autres que '$*$'. En d'autres mots, il représente un sous-ensemble de l'espace de recherche. Par exemple, si nous considérons les chaînes de caractères de longueur quatre, le schéma $(*\,1\,1\,0)$ représentera les deux chaînes suivantes :

$$\{(1\,1\,1\,0), (0\,1\,1\,0)\}$$

Bien entendu, un schéma ne comportant pas de caractère « don't care » représente une unique chaîne de caractères. À l'inverse, un schéma ayant à chaque position un caractère '$*$' représente toutes les chaînes possibles de même longueur. Le nombre de chaînes de caractères qu'un schéma peut représenter correspond à 2^r, où r est le nombre de symboles '$*$'. De plus, chaque chaîne de caractères de longueur m peut être représentée par 2^m schémas différents. Si nous considérons des chaînes de longueur m, il y a un total de 3^m schémas possibles. Dans une population de taille n, entre 2^m et $n \cdot 2^m$

différents schémas peuvent être représentés.

Différents schémas peuvent avoir différentes caractéristiques. L'ordre d'un schéma S, que l'on note $o(S)$, est le nombre de positions fixes (le nombre de positions ayant comme caractère '0' ou '1'). En d'autres mots, il s'agit du nombre de caractères du schéma moins le nombre de caractères '$*$'. L'ordre définit la spécialité d'un schéma. Par exemple, les trois schémas suivants, chacun de taille dix,

$$S_1 = (* \ * \ * \ 0 \ 0 \ 1 \ * \ 1 \ 1 \ 0),$$
$$S_2 = (* \ * \ * \ * \ 0 \ 0 \ * \ * \ 0 *),$$
$$S_3 = (1 \ 1 \ 1 \ 0 \ 1 \ * \ * \ 0 \ 0 \ 1),$$

ont comme ordre :

$$o(S_1) = 6,\ o(S_2) = 3, \text{ et } o(S_3) = 8,$$

et le schéma S_3 est le plus spécifique. La notion d'ordre est importante pour le calcul de la probabilité de survie d'un schéma au niveau de l'opérateur de mutation.

La longueur effective d'un schéma S (« defining length »), que l'on note $\delta(S)$, est la distance entre la première et la dernière position fixe. Cette notion définit la compactibilité de l'information contenue dans un schéma et permet de calculer la probabilité de survie d'un schéma à l'opérateur de croisement. Il est à noter qu'un schéma ne contenant qu'une position fixe a une longueur effective de 0. Par exemple, la longueur effective des schémas de l'exemple précédent est :

$$\delta(S_1) = 6,\ \delta(S_2) = 4, \text{ et } \delta(S_3) = 9.$$

L'effet de la sélection, du croisement et de la mutation sur les différents schémas peut

se résumer ainsi :

$$\xi(S, t+1) \geq \xi(S, t) \cdot \frac{eval(S, t)}{\overline{F(t)}} \left[1 - p_c \cdot \frac{\delta(S)}{m-1} - o(S) \cdot p_m\right]$$

où $\xi(S, t)$ représente le nombre de chromosomes appartenant au schéma S dans la population au temps t, $eval(S, t)$ représente la moyenne des performances des chromosomes appartenant au schéma S au temps t, $\overline{F(t)}$ représente la performance moyenne de tous les chromosomes de la population au temps t, p_c et p_m est la probabilité que les opérateurs de croisement et de mutation soient effectués, respectivement. L'évolution du nombre de représentants du schéma S dépend maintenant de trois facteurs : (1) la performance moyenne des représentants du schéma S par rapport à la performance moyenne de l'ensemble de la population, (2) la longueur effective du schéma S et (3) l'ordre du schéma S. Il est clair que les schémas ayant une performance au-dessus de la moyenne avec une longueur effective courte et un ordre faible croîtront plus rapidement que les autres [105]. Ceci amène le théorème suivant :

Schema Theorem *Short, low-order, above-average schemata receive exponentially increasing trials in subsequent generations of a genetic algorithm.* [105]

Cette théorie suppose que tous les schémas utiles sont dans la population initiale et qu'une solution optimale peut être obtenue en redistribuant ces schémas entre les chromosomes au cours de l'évolution. De plus, cette théorie oublie un aspect fort utile des opérateurs de croisement et de mutation : l'introduction de nouveaux schémas au sein de la population qu'on ne retrouve pas au sein de la population initiale.

6.8.2 Blocs de construction

La théorie des schémas apporte une hypothèse plutôt simple à concevoir, l'hypothèse des blocs de construction.

Building Block Hypothesis *A genetic algorithm seeks near-optimal performance through the juxtaposition of short, low-order, high-performance schemata, called the building blocks.* [105]

En d'autres mots, des schémas qui ont une performance moyenne supérieure à la moyenne de la population, qui sont courts et qui comptent peu de positions fixées sont sélectionnés, recombinés et réarrangés de façon à former des chaînes avec une performance supérieure potentielle. Au lieu de construire des chromosomes hautement performants en essayant toutes les combinaisons possibles, des chaînes plus performantes sont créées à l'aide de solutions partielles des étapes précédentes (schémas).

6.8.3 Bandit armé à k bras

Le problème stochastique du bandit armé à 2 bras est un problème bien connu : soit deux leviers (bras), qui, une fois abaissés, donnent une « récompense » selon deux distributions probabilistiques différentes : l'un des deux bras a une récompense moyenne plus grande que l'autre. Le problème est de maximiser le profit futur en considérant les résultats précédents. Holland [74] pense qu'un algorithme génétique approximise une stratégie optimale, qui alloue un nombre d'essais augmentant de façon exponentielle sur le meilleur des deux leviers. Le problème peut être généralisé au problème du bandit armé à k bras. Supposons qu'un ensemble de schémas soit en compétition lors de la reproduction. À noter que deux schémas S_1 et S_2 sont en compétition si pour toute position i, $i = 1, \ldots, m$, on a $(S_{1_i} = S_{2_i} = *)$ ou $(S_{1_i} \neq *, S_{2_i} \neq *)$ et $S_{1_i} \neq S_{2_i}$ pour au moins une position i. Par exemple :

$$
\begin{array}{cccccc}
* & 0 & 0 & * & 0 & * \\
* & 0 & 0 & * & 1 & * \\
* & 0 & 1 & * & 0 & * \\
* & 0 & 1 & * & 1 & * \\
* & 1 & 0 & * & 0 & * \\
* & 1 & 0 & * & 1 & * \\
* & 1 & 1 & * & 0 & * \\
* & 1 & 1 & * & 1 & *
\end{array}
$$

nous avons un bandit armé à 8 bras. Il est important de noter que dans un algorithme génétique, plusieurs problèmes de bandit armé à k bras sont résolus en parallèle. Par exemple, pour tout $i, i = 1, \ldots, m$, la longueur d'un chromosome, nous avons $m!/(i!(m-i)!)$ problèmes de bandit armé à 2^i bras qui se résolvent simultanément.

7. Algorithme génétique spécialisé

Par ses caractéristiques telles que la représentation des solutions et l'opérateur de mutation (voir chapitre 6), l'algorithme génétique simple ne peut être appliqué à nos données de génotypes. Bien qu'il soit possible d'exprimer un ensemble de blocs d'haplotypes par une représentation binaire, celle-ci est beaucoup trop coûteuse en terme de mémoire (il y a environ 75 500 valeurs différentes à représenter). De plus, cet espace mémoire se trouve à être gaspillé puisque la majorité des valeurs dans la chaîne binaire sera des '0'. Puisque la complexité au niveau de l'espace mémoire est importante (beaucoup de données à entreposer), une autre représentation des solutions est utilisée dans le cadre de ce projet. Une importante modification aux éléments de l'algorithme génétique simple doit donc être appliquée.

7.1 Représentation

Les solutions de l'algorithme génétique sont encodées de façon intuitive. Il s'agit d'une liste des blocs/valeurs d'allèles considérés lors du calcul de la fonction objectif. Chaque position de la chaîne représentant le chromosome de l'algorithme génétique est composée de trois valeurs : le numéro d'identification du bloc d'haplotypes ainsi que les numéros d'identification des deux allèles sélectionnés pour ce bloc.

$Bloc_1$:	$allèle_1$	$allèle_2$	$Bloc_2$:	$allèle_1$	$allèle_2$	$\cdots$	$Bloc_n$: $allèle_1$ $allèle_2$

De façon mathématique, cette chaîne binaire principale peut être représentée ainsi :

$$chr\_géno = \{c_1, \ldots, c_n \mid c_i = (bloc_j, a_1, a_2), a_1 \wedge a_2 \in bloc_j\_allèle, \forall i \in [1, \ldots, n]\} \quad (7.1)$$

où $bloc_j\_allèle$ est un allèle d'haplotype, *i.e.* la valeur que peut prendre un certain bloc d'haplotypes j :

$$bloc_j\_allèle = \{s_1, \ldots, s_n \in bloc_j \mid s_i = v_i, v_i \in \{1, 2\}, \forall i \in [1, \ldots, n]\}$$

Finalement, *bloc* représente le bloc d'haplotypes, *i.e.* un ensemble de SNPs en déséquilibre de liaison.

$$bloc = \{s_1, \ldots, s_n \in \{SNPs\} \mid |LD(s_i, s_j)| > X, \forall i \neq j \in [1, \ldots, n]\}$$

où $\{SNPs\}$ est l'ensemble de tous les SNPs disponibles et $LD(s_i, s_j)$ est la valeur de déséquilibre de liaison entre le SNP i et le SNP j. Pour qu'une SNP i soit en déséquilibre de liaison, il faut que $LD(s_i, s_j) > |X|$, où X est le seuil. De plus, $LD(s_i, s_j)$ doit être significatif après un test de χ^2. Il est à noter que ces blocs ont été créés par le groupe de recherche du Docteur Pavel Hamet en deux étapes : le calcul du déséquilibre de liaison à l'aide de la méthode d'Haploview [89], et l'estimation de la phase de liaison aux chromosomes parentaux à l'aide de PHASE [122, 123, 124, 125].

Une chaîne binaire secondaire (optionnelle) est ajoutée aux chromosomes afin de pouvoir représenter le sexe et l'âge des sujets à considérer dans le calcul de la fonction objectif. Cette chaîne peut aussi être utilisée afin de représenter d'autres caractéristiques des sujets (facteurs environnementaux ou autres, tels le BMI, l'âge, le ratio sodium/potassium excrété, la consommation d'alcool, etc.). Cette chaîne secondaire facilite le calcul de la fonction objectif tout en séparant les éléments génétiques des éléments environnementaux. Dans le cas de l'algorithme génétique développé dans ce projet, les deux premiers caractères de cette chaîne binaire représentent le sexe des sujets. Les positions suivantes servent à représenter les différentes tranches d'âge (de

15 à 95 ans, par intervalle de 5 ans).

$$\underbrace{p_0 \quad p_1}_{\text{sexe}} \quad \underbrace{p_2 \quad p_3 \quad \cdots \quad p_{17}}_{\text{âge}}$$

Il est entendu que cette chaîne binaire optionnelle de notre algorithme génétique spécifique augmente le nombre de dimensions à considérer. Ceci ne cause pas de problème, puisque le nombre de dimensions de départ est tellement grand que cet ajout n'est pas significatif. De plus, les algorithmes génétiques sont assez efficaces pour permettre ce type d'ajout.

Les deux algorithmes seront utilisés au cours de cette étude (soit l'algorithme qui n'utilise pas la chaîne binaire optionnelle, et celui qui l'utilise). L'algorithme génétique spécifique simple fait référence à l'algorithme n'utilisant pas la chaîne binaire optionnelle.

7.2 Population initiale

La population initiale de l'algorithme génétique spécifique n'est pas entièrement créée de façon aléatoire comme dans l'algorithme génétique simple (voir section 6.2). Puisqu'il y a un grand nombre de blocs d'haplotypes possible (environ 8 400 blocs) ayant chacun une moyenne de 4 allèles de blocs différents, il est impossible de créer aléatoirement une solution apte à représenter un nombre maximal de sujets présentant le phénotype choisi. Afin de contourner ce problème, l'opérateur de création de population initiale choisit un sujet de la population du SLSJ aléatoirement (présentant le phénotype ou non). Ensuite, l'opérateur choisit entre 4 et 10 blocs d'haplotypes différents (avec la valeur de l'allèle de bloc correspondant au sujet préalablement choisi). Il est donc certain que cette solution créée de façon quasi aléatoire représentera au moins un sujet de notre population à l'étude. De façon mathématique,

$$chr\_géno_{\text{init}} = chr\_géno \,|\, n = \{4, \ldots, 10\}, c_i \in génotype\_sujet_k \, \forall i \in [1, \ldots, n]$$

où k est un identificateur de sujet aléatoirement choisi et i est un tuple $(bloc_j, a_1, a_2)$ inclus dans $génotyope\_sujet_k$, le génotype complet du sujet k. À noter que *chr_géno* provient de l'équation 7.1. Puisque les sujets présents dans la banque de données du SLSJ sont reliés entre eux (famille), une solution initiale créée de cette façon représente entre un et vingt sujets différents, en moyenne.

7.3 Opérateurs de sélection

Un total de deux opérateurs de sélection ont été implémentés dans le cadre de ce projet, soit la sélection par tournoi ainsi que la méthode des rangs + SUS (section 6.3), chacun d'eux présentant des avantages et des inconvénients différents. La sélection par rang a été choisie pour sa capacité à sélectionner correctement des individus performants (selon la fonction objectif) malgré l'utilisation d'un générateur de nombres aléatoires biaisé. La sélection par tournoi a été utilisée pour sa simplicité et la possibilité d'ajouter un paramètre permettant de contrôler la rapidité de la convergence de l'algorithme vers un optimum de la fonction : la pression de sélection (voir section 6.3).

7.4 Opérateurs de croisement

Trois opérateurs de croisement ont été implémentés et peuvent donc être utilisés lors de différentes exécutions de l'algorithme : le croisement uniforme, le croisement à n-point et le croisement à n-point avec mélange. Les caractéristiques de ces opérateurs sont présentées à la section 6.4. Aucune modification n'est nécessaire, puisque ces opérateurs travaillent sur des chaînes de caractères, que ce soit des chiffres binaires ou des ensembles d'entiers.

7.5 Opérateurs de mutation

L'opérateur de mutation classique (GA simple) est utilisé pour la chaîne optionnelle du chromosome (la chaîne binaire représentant l'âge et le sexe, ainsi que d'autres caractéristiques optionnelles).

$$chr_{\text{muté}} = a_1, a_2, \ldots, a_n \,|\, a_i = a_{i_{\text{non-muté}}} \oplus 1 \text{ si } \tau < 0.001 \; \forall i = 1, \ldots, n$$

où τ est une variable aléatoire uniforme bornée entre 0 et 1 et $a_{i_{\text{non-muté}}}$ provient du chromosome avant sa mutation.

L'opérateur de mutation doit être modifié considérablement afin de permettre son utilisation sur la chaîne principale de l'algorithme génétique spécifique (soit la chaîne représentant le génotype des sujets). Cet opérateur est divisé en quatre opérations différentes :

1. mutation d'ajout d'un bloc ;
2. mutation de suppression d'un bloc ;
3. mutation d'un bloc ;
4. mutation d'un allèle d'un bloc.

7.5.1 Mutation d'ajout

La mutation d'ajout consiste à l'ajout d'un bloc dans la chaîne principale du chromosome. Il faut s'assurer que le bloc ajouté lors de la mutation n'appartient pas déjà à l'ensemble de blocs avant la mutation.

$$chr_{\text{muté}} = \{bloc_{\text{nouveau}}\} \cup chr_{\text{non-muté}} \,|\, bloc_{\text{nouveau}} \notin chr_{\text{non-muté}}$$

Le bloc choisi pour l'ajout ($bloc_{\text{nouveau}}$) est sélectionné aléatoirement de la même façon que pour la création de la population initiale (voir section 7.2).

7.5.2 Mutation de suppression

La mutation de suppression consiste en la réduction du nombre de blocs d'haplotypes dans la chaîne principale du chromosome. Il suffit de choisir un bloc au hasard et de l'enlever de la chaîne.

$$chr_{\text{muté}} = chr_{\text{non-muté}} - \{bloc_i\} \mid bloc_i \in chr_{\text{non-muté}}$$

où $bloc_i$ est choisi aléatoirement dans l'ensemble de blocs $chr_{\text{non-muté}}$ à l'aide d'une variable aléatoire uniforme. Ainsi, chacun des blocs présents dans l'ensemble $chr_{\text{non-muté}}$ a la même probabilité d'être « supprimé ».

7.5.3 Mutation d'un bloc

La mutation d'un bloc consiste à enlever un bloc choisi de manière aléatoire (voir la mutation de suppression à la section 7.5.2) et de le remplacer par un autre bloc, choisi lui aussi de manière aléatoire (voir la mutation d'ajout à la section 7.5.1).

$$chr_{\text{muté}} = \Big(chr_{\text{non-muté}} - \{bloc_i\}\Big) \cup \{bloc_{\text{nouveau}}\} \mid bloc_{\text{nouveau}} \notin chr_{\text{non-muté}} \ni bloc_i$$

où $bloc_i$ est choisi aléatoirement dans l'ensemble de blocs $chr_{\text{non-muté}}$ de la même façon que pour la mutation à la section précédente.

7.5.4 Mutation d'un allèle de bloc

La mutation d'un allèle de bloc consiste à remplacer la valeur (allèles) d'un bloc sélectionné aléatoirement par une autre valeur choisie, elle aussi, de manière aléatoire. Le bloc ajouté représente donc le même ensemble de SNPs que le bloc supprimé.

$$chr_{\text{muté}} = \Big(chr_{\text{non-muté}} - \{bloc_i\}\Big) \cup \{bloc_{i\,\text{nouveau}}\} \mid bloc_{i\,\text{nouveau}} \notin chr_{\text{non-muté}} \ni bloc_i$$
$$\text{et } (a_j, a_k) \in bloc_{i\,\text{nouveau}} \neq (a_l, a_m) \in bloc_i$$

où $bloc_i$ est choisi aléatoirement dans l'ensemble de blocs $chr_{\text{non-muté}}$.

7.6 Opérateurs de remplacement

Les trois opérateurs de remplacement présentés dans la section 6.6 ont été implémentés (élitisme, « steady-state » et « generation gap »). Ces opérateurs sont le remplacement par « generation gap », le remplacement élitisme ainsi que le remplacement « steady state ». Puisque ces opérateurs de remplacement agissent sur des ensembles de chromosomes (solutions de l'algorithme génétique) indépendamment de l'encodage choisi, ils n'ont pas à être modifiés afin de bien fonctionner avec l'algorithme génétique spécifique.

7.7 Immigrant aléatoire

L'immigrant aléatoire, expliqué à la section 6.7, est utilisé afin d'éviter une homogénéité prématurée dans la population de l'algorithme génétique. De façon mathématique, l'immigrant aléatoire peut se représenter ainsi :

$$nouvelle\_pop_t = \left(pop_t - min\_inds\right) \cup RI\_inds$$

où min_inds est l'ensemble des n pires individus, pop_t est la population créée après une itération de l'algorithme génétique (au temps t) et RI_inds est l'ensemble des individus créés aléatoirement par l'opérateur d'immigrant aléatoire, de la même façon que la population initiale. Il est à noter que l'opérateur d'immigrant aléatoire sera utilisé lorsque la meilleure solution de l'algorithme génétique n'aura pas changée après k itérations.

7.8 Fonction objectif

Il y a plusieurs fonctions objectifs possibles permettant de créer un ensemble de blocs d'haplotypes représentant bien les personnes hypertendues de la population du SLSJ. Il faut donc développer une fonction mathématique représentant ce que nous voulons trouver. Plusieurs de ces fonctions peuvent être équivalentes, mais certaines d'entre elles peuvent être beaucoup plus efficaces. Trouver une fonction objectif adéquate nécessite beaucoup de temps et d'effort, quelle que soit la solution que nous voulons trouver. Une fonction naïve serait de maximiser le nombre de sujets hypertendus représentés par l'ensemble de blocs d'haplotypes.

$$f_{\text{temp\_1}} = nbHypertendus$$

Une fois cette fonction analysée, il est facile de conclure que la solution optimale est l'ensemble de tous les blocs d'haplotypes possibles. Cette solution représentera donc toute la population de la banque de données (les hypertendus et les normotendus), ce qui n'est pas ce que nous cherchons. Il faut donc ajouter de l'information à la fonction objectif naïve. Une idée serait d'utiliser un ratio, soit le quotient du nombre d'hypertendus représentés à l'aide de l'ensemble de blocs d'haplotypes par le nombre total d'individus représentés.

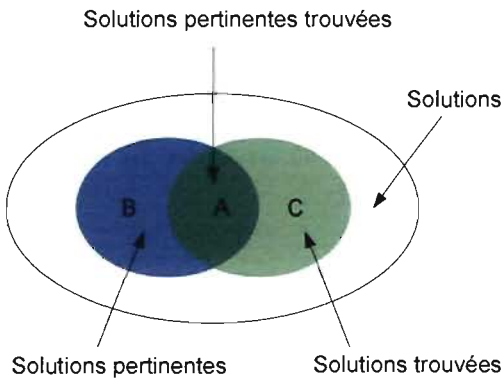
$$f_{\text{temp\_2}} = \frac{nbHypertendus}{nbTotal}$$

Encore une fois, cette fonction objectif ne réalise pas ce que nous voulons, puisque la solution optimale (un quotient de 1 qui est équivalent à 100 %) s'obtient facilement lorsque nous considérons l'ensemble de tous les blocs d'haplotypes d'un individu hypertendu. Cet ensemble ne représente que l'individu en question et va donc donner un ratio de 1, le maximum. L'idée est d'utiliser une fonction objectif qui va permettre de maximiser le nombre de personnes représentées tout en maximisant le nombre d'hypertendus représentés (ou en minimisant le nombre de normotendus représentés).

Afin de réaliser ceci, la fonction objectif utilisée pour notre algorithme génétique spécifique utilise les concepts de précision et de rappel [126, 127].

$$\text{précision} = \frac{|\{\text{solutions pertinentes}\} \cap \{\text{solutions trouvées}\}|}{|\{\text{solutions trouvées}\}|}$$
$$\text{rappel} = \frac{|\{\text{solutions pertinentes}\} \cap \{\text{solutions trouvées}\}|}{|\{\text{solutions pertinentes}\}|}$$

En d'autres mots, la précision (de l'anglais « precision ») représente le pourcentage de solutions pertinentes par rapport aux solutions trouvées lors de la recherche (vrais positifs). La précision va donc minimiser le nombre d'individus normotendus représenté par la solution optimale. Le rappel (de l'anglais « recall »), aussi nommé sensitivité, représente le pourcentage de solutions pertinentes trouvées après la recherche par rapport au nombre total de solutions pertinentes (vrais positifs et faux négatifs). Le rappel va donc maximiser le nombre de personnes hypertendues représenté par la solution optimale. Ces deux concepts permettent donc de nous renseigner sur le nombre de vrais positifs et de faux positifs trouvés lors de la recherche, ainsi que le nombre de faux positifs (voir Figure 7.1). Ils sont équivalents à la sensibilité et à la spécificité qui sont généralement utilisées en médecine [128].



$$\text{précision} = \frac{A}{A+C}$$
$$\text{rappel} = \frac{A}{A+B}$$

Fig. 7.1: Précision et Rappel. Représentation graphique de la précision et du rappel, deux statistiques couramment utilisées dans la théorie de l'information.

À partir de ces concepts, il est possible de créer une fonction mathématique servant à maximiser la précision et le rappel simultanément. Il est à noter qu'à mesure où la précision augmente, le rappel diminuera et vice-versa puisque ceux-ci évoluent de façon inversement proportionnelle. Il est donc important de pouvoir mettre en relation ces deux valeurs.

L'algorithme génétique utilise donc une fonction objectif qui fixera une des deux valeurs (dans ce cas, le rappel) et optimisera l'autre (la précision). Cette fonction objectif permettra donc de créer une courbe montrant la relation entre la précision et le rappel, tout en maximisant le nombre de vrais positifs et en limitant le nombre de faux négatifs. La fonction est la suivante :

$$f = \text{précision} - dist_{\text{rappel}}$$

où $dist_{\text{rappel}}$ est la « distance » entre $curr_{\text{rappel}}$, le rappel courant, et un certain seuil.

$$dist_{\text{rappel}} = \begin{cases} 0 & \text{si } curr_{\text{rappel}} \geq seuil \\ \dfrac{seuil - curr_{\text{rappel}}}{seuil} & \text{sinon} \end{cases} \tag{7.2}$$

Dans le deuxième cas (soit lorsque $curr_{\text{rappel}} < seuil$), il est important de diviser $seuil - curr_{\text{rappel}}$ par le seuil afin de rapporter cette valeur entre 0 et 1, permettant la comparaison directe entre cette valeur et la précision, un ratio aussi borné par 0 et 1. Cette distance devient donc aussi importante dans le calcul de la fonction objectif que la précision. La raison pour laquelle $dist_{\text{rappel}}$ est équivalent à 0 lorsque $curr_{\text{rappel}}$ est plus grand ou égal au seuil est que nous ne voulons pas pénaliser la performance des solutions qui ont un meilleur rappel que le seuil. Ce cas est plutôt rare, puisque le rappel a tendance à diminuer lorsque la précision augmente. En utilisant différentes valeurs de seuil, il est possible de créer un graphique de la précision en fonction du rappel afin de représenter les limites de notre algorithme génétique spécifique.

Finalement, la Figure 7.2 représente de façon schématique le fonctionnement de l'algorithme génétique spécialisé pour l'analyse de la susceptibilité à l'hypertension chez la population du SLSJ.

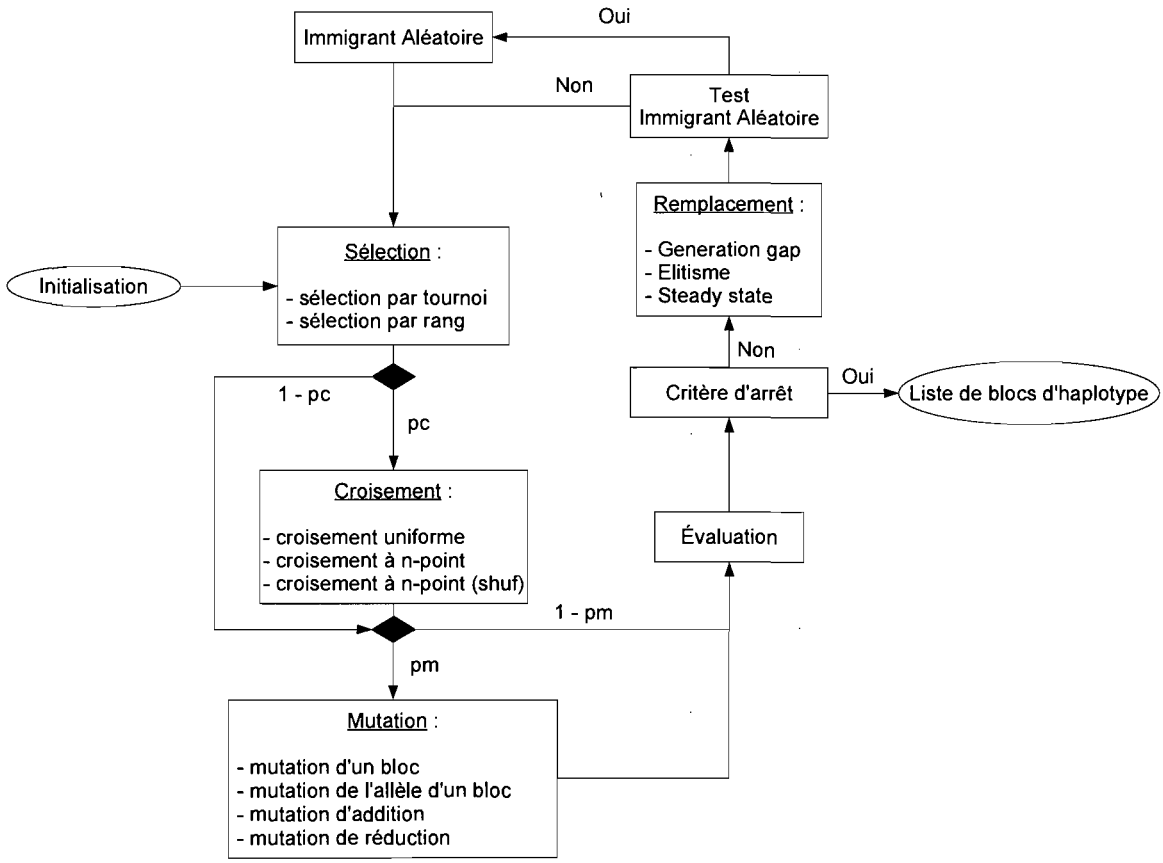


Figure 7.2: Plan d'adaptation de l'algorithme génétique spécifique. Schéma du plan d'adaptation de l'algorithme génétique spécialisé pour l'analyse de la susceptibilité à l'hypertension chez la population du SLSJ. Les losanges noirs correspondent à une variable aléatoire uniforme permettant d'omettre les opérateurs de croisement et de mutation avec une probabilité de $1 - pc$ et $1 - pm$ respectivement.

8. Expérimentation

8.1 Phénotypes considérés

Les expérimentations ont été effectuées sur deux phénotypes. Le premier correspond au sujet principal de l'étude, soit l'hypertension. Les sujets sont considérés atteints d'hypertension lorsque leur SBP $\geq$ 140 mm Hg et/ou leur DBP $\geq$ 90 mm Hg ou lorsqu'une médication antihypertenseur est utilisée [14]. Le deuxième phénotype analysé est l'indice de masse corporelle (le BMI). Ce phénotype fut utilisé pour le projet de recherche puisqu'il s'agit d'un bon prédicteur de la pression artérielle (voir section 2.3.3). Deux valeurs de BMI seront utilisées : $>$ 27 kg/m^2 (représentant les individus présentant un surpoids) et $>$ 30 kg/m^2 (représentant les individus obèses) [129], que nous nommerons BMI27 et BMI30, respectivement. Un individu possède un BMI élevé si sa valeur dépasse le premier ou le deuxième seuil, dépendamment de l'analyse réalisée.

8.2 Opérateurs utilisés

Toutes les combinaisons possibles des opérateurs de sélection, de croisement et de remplacement sont utilisées pour un même seuil de rappel, et ce, pour les trois phénotypes analysés. Ceci permet donc de comparer les différentes variantes d'opérateurs entre elles (voir Tableau VIII.I).

8.3 Validation statistique des résultats

Il est difficile de valider statistiquement les algorithmes génétiques pour les tests multiples. La méthode utilisée dans le cadre de ce projet de recherche est basée sur le rééchantillonnage présenté par Westfall et Young [130]. Cette méthode vise à conserver autant que possible les caractéristiques des données réelles. Il faut pour cela créer des données de simulation à partir des données réelles à l'aide de permutations (sans remplacement).

L'algorithme génétique est, en premier lieu, utilisé sur les données génétiques réelles afin de calculer une valeur de précision maximale pour un seuil de rappel donné à l'aide de la fonction objectif (voir équation 7.2). Par la suite, des permutations sur les phénotypes sont réalisées, afin de créer des données aléatoires, tout en conservant la proportion de sujets présentant le phénotype étudié. L'algorithme génétique est finalement utilisé sur les données créées à l'aide des permutations afin de calculer une nouvelle valeur de précision pour un même seuil de rappel. La p-valeur empirique est calculée en divisant le nombre d'échantillons aléatoires qui ont produit une précision supérieure ou égale à la précision du test original par le nombre d'échantillons (voir équation 8.1).

$$p\text{-valeur empirique} = \frac{\text{nb de précision} \geq \text{précision réelle}}{\text{nb d'échantillons}} \tag{8.1}$$

Des p-valeurs empiriques furent calculées pour des seuils de rappel de 0,3, 0,6 et 0,9 pour l'algorithme génétique spécifique considérant l'âge et le sexe des individus. Le nombre d'échantillon réalisé diffère d'un seuil de rappel à l'autre, car le temps nécessaire pour réaliser les simulations est grand. Pour cette même raison, les simulations n'ont pas été réalisées pour les phénotypes BMI27 et BMI30.

8.4 Outils informatiques

Toutes les expérimentations ont été faites sous le système d'exploitation *Linux Fedora 7* [131] à l'aide d'un *Intel®Core™Duo mobile processor* (technologie *Centrino® Duo*) cadencé à 2,16 GHz avec 1024 Mo de mémoire RAM. Il est à noter que pour réduire le temps de calcul total, deux instances de l'algorithme génétique spécifique sont calculées simultanément de façon indépendante, prenant ainsi avantage du processeur à deux cœurs d'Intel®. L'algorithme génétique est implémenté à l'aide de *Python™* version 2.5.1 [132]. Les graphiques sont réalisés à l'aide de *R* version 2.6.1 [133].

Tableau VIII.I: Combinaisons des opérateurs de l'algorithme génétique. Tableau représentant les différents opérateurs utilisés lors des différentes valeurs de seuil pour le rappel. Ainsi, à chaque valeur constante de rappel, toutes les combinaisons des différents opérateurs implémentés seront utilisées.

Sélection	Remplacement	Croisement
Tournoi	Steady-State	Uniforme
		n-point
		n-point avec mélange
	Generation Gap	Uniforme
		n-point
		n-point avec mélange
	Élitisme	Uniforme
		n-point
		n-point avec mélange
Rang + SUS	Steady-State	Uniforme
		n-point
		n-point avec mélange
	Generation Gap	Uniforme
		n-point
		n-point avec mélange
	Élitisme	Uniforme
		n-point
		n-point avec mélange

Troisième partie

Résultats

9. Résultats des analyses

Les résultats de nos différentes analyses peuvent être classés en deux catégories. La première catégorie inclut les résultats démontrant la performance de notre algorithme génétique spécifique en fonction des différents opérateurs utilisés en ce qui a trait à la recherche d'un optimum global et au temps d'exécution Ceux-ci permettent d'analyser la convergence des solutions vers une solution optimale lorsque des opérateurs distincts de sélection, de croisement et de remplacement de population sont utilisés. Des graphiques montrant le temps d'exécution (en secondes) d'une instance de l'algorithme génétique en fonction du nombre de solutions dans la population (le nombre de chromosomes) et en fonction du nombre d'itérations de l'algorithme seront présentés.

La deuxième catégorie, quant à elle, permet de visualiser les changements dans la précision en fonction du rappel des blocs d'haplotypes trouvés. Ces derniers résultats donnent donc une idée de la distribution de ces deux valeurs et présentent les limites de la résolution de notre problème à l'aide d'un algorithme génétique.

9.1 Étude de la convergence de l'algorithme

Tel qu'il est mentionné dans le chapitre 7, plusieurs opérateurs essentiels au bon fonctionnement de l'algorithme génétique furent implémentés (sélection, croisement, mutation, remplacement). Chacune des combinaisons possibles de ces opérateurs aura une répercussion différente sur la convergence de la population de solutions vers la solution optimale. Le degré de cette convergence peut nous informer quant à la performance de l'algorithme de recherche pour la détection d'un optimum global.

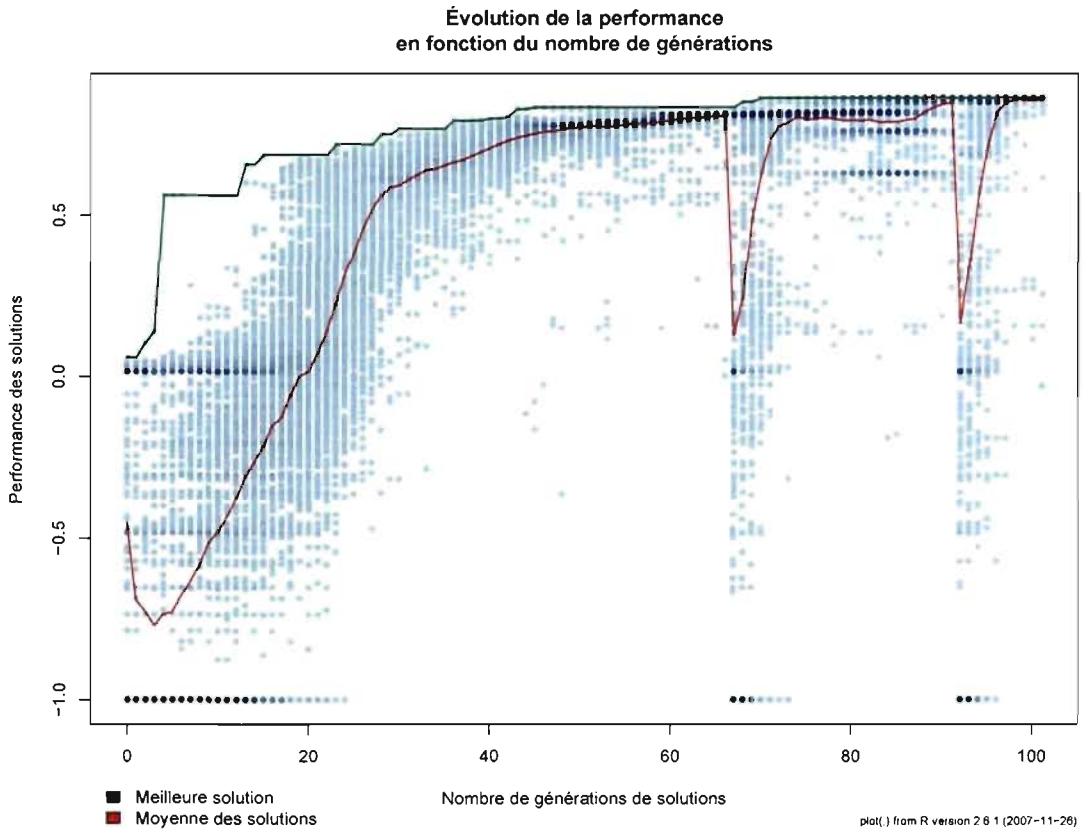


Figure 9.1: Évolution des solutions en fonction du nombre d'itérations de l'algorithme génétique (opérateurs tournoi, « steady-state » et « n-point »). Représentant de la convergence de l'algorithme vers une solution optimale. Les opérateurs utilisés pour la sélection, le remplacement et le croisement sont respectivement le tournoi, le « steady-state » et le « n-point ». Le seuil de rappel est de 0,3 et la solution optimale a une performance de 0,855.

Afin de bien représenter cette évolution en fonction du nombre d'itération de l'algorithme génétique, des graphiques représentant la relation entre la mesure de performance (valeur de la fonction objectif) et le nombre d'itérations sont ici représentés. Il est à noter qu'une itération d'un algorithme génétique représente la complétion de la série d'opérateurs suivante : sélection → croisement → mutation → remplacement. On peut aussi parler d'itérations en terme du nombre de générations de solutions. En effet, une itération au temps t correspond à la création d'une nouvelle génération, aussi au temps t.

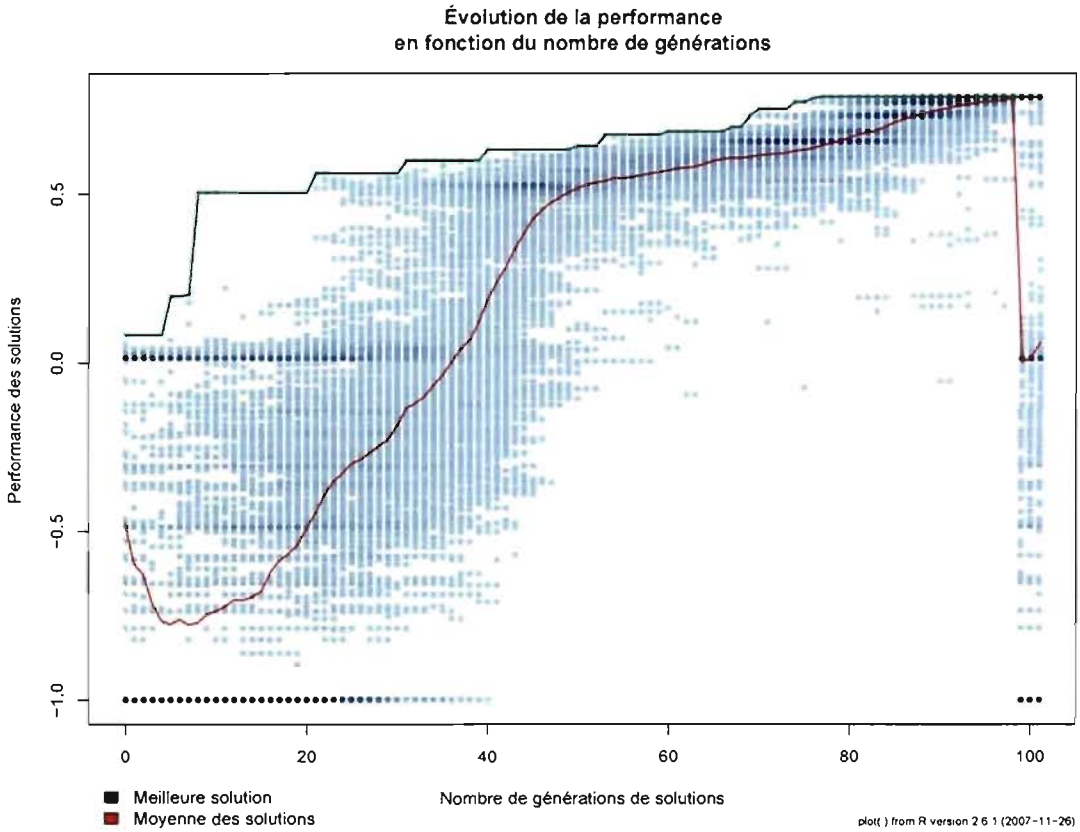


Figure 9.2: Évolution des solutions en fonction du nombre d'itérations de l'algorithme génétique (opérateurs tournoi, « generation gap » et « n-point »). Graphique représentant la convergence de l'algorithme vers une solution optimale. Les opérateurs utilisés pour la sélection, le remplacement et le croisement sont respectivement le tournoi, le « generation gap » et le « n-point ». Le seuil de rappel est de 0,3 et la solution optimale a une performance de 0,783.

Les différences les plus notables dans l'évolution de la solution optimale se retrouvent lorsqu'une comparaison est faite entre les différents opérateurs de remplacement « Steady-State », « Generation Gap » et Élitisme (voir Section 6.6 pour de plus amples informations). En effet, la simple utilisation d'un opérateur de croisement ou de sélection différent ne présente pas de changement significatif dans la plupart des cas.

Les trois graphiques des Figures 9.1, 9.2 et 9.3 représentent les résultats obtenus lors de trois analyses différentes. Les points bleus représentent la mesure de performance de chaque solution retrouvée dans la population de solutions à la génération au temps t donné. Plus le point est foncé, plus il y a de solutions ayant la même valeur de per-

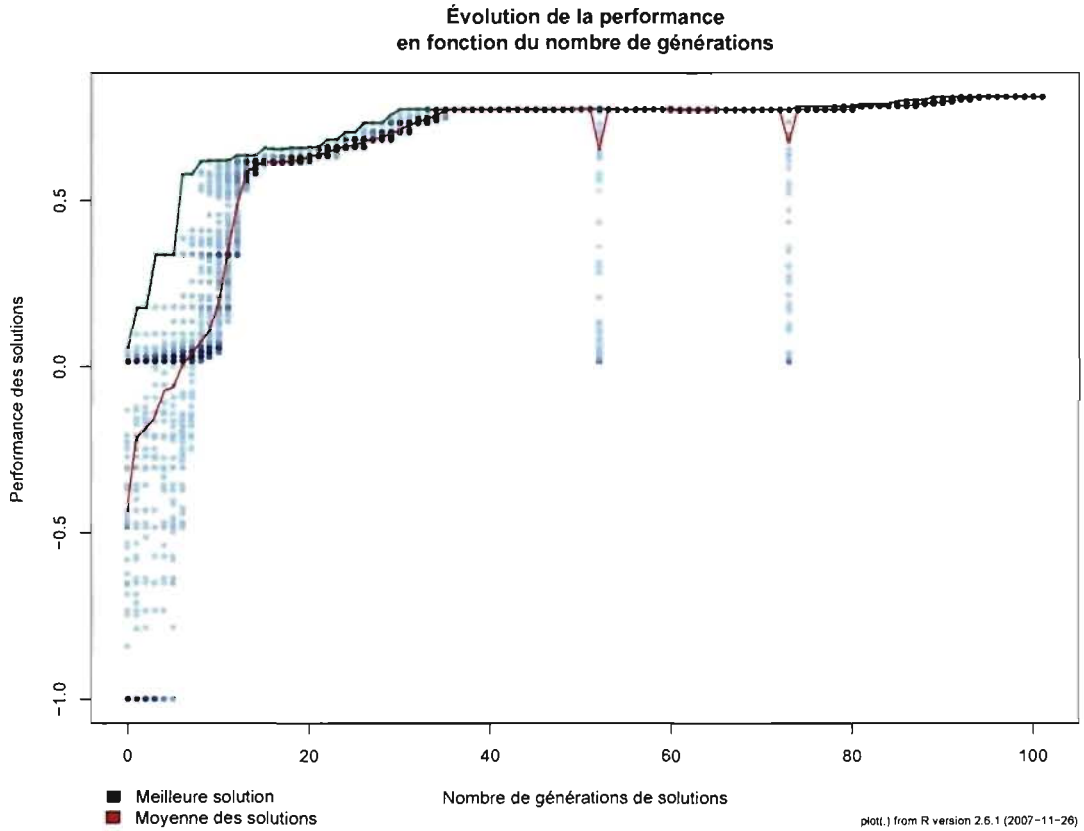


Figure 9.3: Évolution des solutions en fonction du nombre d'itérations de l'algorithme génétique (opérateurs tournoi, élitisme et « n-point »). Graphique représentant la convergence de l'algorithme vers une solution optimale. Les opérateurs utilisés pour la sélection, le remplacement et le croisement sont respectivement le tournoi, l'élitisme et le « n-point ». Le seuil de rappel est de 0,3 et la solution optiomale a une performance de 0,809.

formance. La ligne verte suit l'évolution de la meilleure solution (la solution présentant la plus haute valeur de performance) trouvée au temps t. La ligne rouge, quant à elle, représente la moyenne des valeurs de performance de toutes les solutions au temps t.

La Figure 9.1, représentant l'évolution des solutions suite à l'utilisation de l'opérateur de remplacement « steady state », montre un total de 100 itérations (générations de solutions). L'opérateur d'immigrant aléatoire (RI) est appelé deux fois, soit dans les environs des itérations 65 et 90 (on peut y remarquer une baisse de la moyenne des performaces des solutions lorsque l'opérateur RI est appelé). On peut aussi remarquer que le premier appel de l'opérateur RI amène la création d'une solution plus

performante que les autres trouvées, démontrant l'utilité de cet opérateur. Au début, la moyenne des performances des solutions diminue pour finalement augmenter après cinq itérations environ. Cette dernière se rapproche de la meilleure solution trouvée après une quarantaine d'itérations.

La Figure 9.2, représentant l'évolution des solutions suite à l'utilisation de l'opérateur de remplacement de génération « Generation gap », montre aussi un total de 100 itérations. Par contre, l'opérateur RI n'est appelé qu'une fois, vers la centième itération. La moyenne des performances des solutions prend plus de temps pour se rapprocher de la meilleure solution (soit après environ 90 itérations).

Finalement, la Figure 9.3, représentant l'évolution des solutions suite à l'utilisation de l'opérateur de remplacement de génération « élitisme », montre des courbes très différentes des deux autres figures. Premièrement, on remarque que la moyenne des performances des solutions est toujours très près de la meilleure solution trouvée, et ce, à chaque itération de l'algorithme. L'opérateur RI est appelé deux fois, mais a peu d'effet sur la moyenne des performances. On peut remarque que ce dernier améliore la meilleure solution trouvée lors de sa deuxième exécution.

La Figure 9.4 représente de façon graphique le temps d'exécution d'une instance de l'algorithme génétique en fonction du nombre de chromosomes utilisés (le nombre de solutions dans la population). En (a), il s'agit du temps d'exécution de l'algorithme génétique spécifique simple et en (b), de l'algorithme génétique prenant en compte l'âge et le sexe des individus. Dans les deux cas, différents nombres d'itérations de l'algorithme sont représentés, soit 100, 150 et 200 itérations (présentés respectivement en rouge, vert et bleu). L'augmentation du nombre d'itérations pour une instance de l'algorithme est variable, puisque le temps requis pour ce dernier afin de converger vers une solution optimale est lui aussi variable. Chaque point des graphiques représente la

moyenne du temps d'exécution en fonction du nombre de chromosomes et a un nombre d'itérations fixé. L'écart type σ à la moyenne est représenté à l'aide des indicateurs d'erreurs sur chaque point. Sa formule est la suivante :

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

où N est le nombre de valeurs à considérer dans le calcul, x_i est la valeur courante et $\bar{x}$ est la moyenne des valeurs. La droite de régression linéaire $\hat{y} = a + bx$ a été calculée à l'aide de la formule suivante :

$$b = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, \quad a = \bar{y} - b\bar{x}$$

9.2 Étude de la précision et du rappel

Comme il est mentionné à la section 7.8, la fonction objectif (équation 7.2) utilisée dans le cadre de cet algorithme génétique spécifique tente de maximiser conjointement la précision ainsi que le rappel des solutions optimales trouvées. De plus, grâce à l'utilisation d'un seuil pour le rappel, elle permet de mettre en relation ces deux valeurs, puisque pour une valeur de rappel fixe, nous obtenons la précision maximale. Il est donc normal de présenter ces résultats sous forme de graphique représentant la précision des solutions optimales en fonction de leur rappel respectif. Il ne faut pas oublier qu'en général, lorsque la précision augmente, le rappel a tendance à diminuer et vice versa.

La première figure (Figure 9.5) représente cette relation lorsque le phénotype de l'hypertension est analysé. Le graphique en (a) correspond aux analyses réalisées à l'aide de l'algorithme génétique spécifique simple, c'est-à-dire la version de l'algorithme qui n'utilise pas la chaîne binaire secondaire optionnelle. L'âge et le sexe ne sont donc

pas considérés lors du calcul de la fonction objectif. Le graphique en (b), par contre, est le résultat des analyses effectuées à l'aide de l'algorithme génétique avancé, tenant compte de l'âge et du sexe des personnes concernées. Les mêmes paramètres s'appliquent aux Figures 9.6 et 9.7. La seule différence est que la Figure 9.6 représente les analyses faites à l'aide du phénotype BMI27 (soit les individus présentant un BMI supérieur ou égal à 27 kg/m^2) alors que la Figure 9.7 représente le phénotype BMI30 (soit les individus qui présentent un BMI supérieur ou égal à 30 kg/m^2). Dans tous les graphiques, chaque point correspond à une solution optimale trouvée suite à une recherche à l'aide de l'algorithme génétique (chaque solution a une précision et un rappel). À remarquer que chacune des courbes est inclinée vers la droite et présente une courbe semblable (donc une même distribution de précision et de rappel).

Il est à noter que pour tous ces graphiques, la courbe bleue n'a pas été calculée à l'aide d'une régression. Il s'agit en fait d'une spline, ce qui facilite et accélère les calculs. Une spline consiste en une fonction définie par morceaux à l'aide de polynômes. Soit un ensemble de points, nommés nœuds. Une spline tentera de s'approcher le plus près possible de tous les nœuds selon un poids prédéterminé. Il est donc possible de réduire le poids des nœuds représentant des valeurs extrêmes afin d'obtenir une spline plus représentative des données en général. Dans notre cas, tous les nœuds sont considérés également dans le calcul. La fonction « `smooth_spline` » du projet de statistique *R* version 2.5.1 fut utilisée [134, 135, 136].

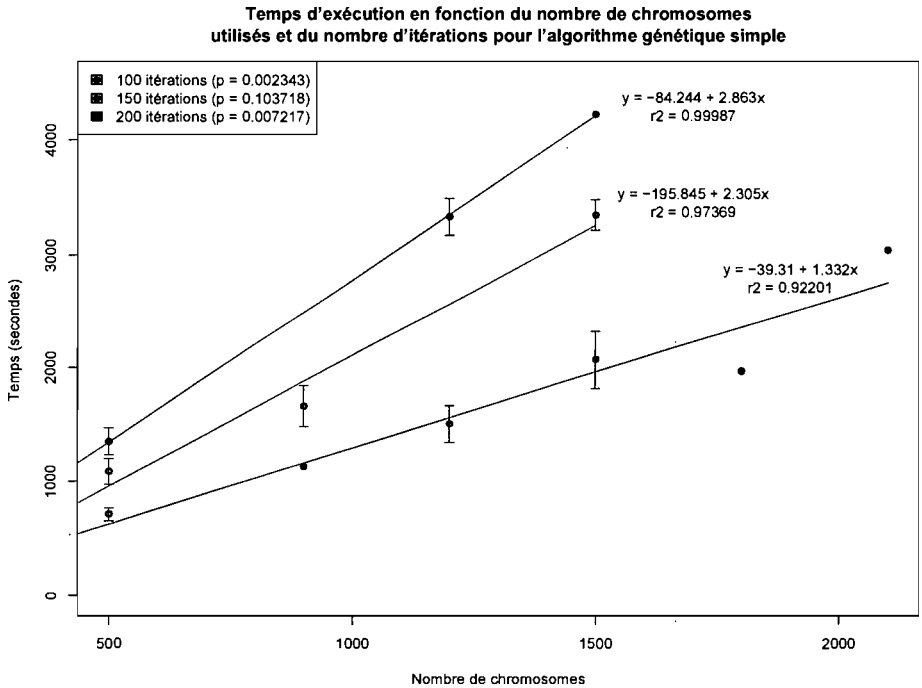
Les données brutes des graphiques mentionnés précédemment sont représentées dans les Tableaux IX.II et IX.III sous forme de moyenne et d'écart type. On y retrouve les différentes valeurs de moyennes et d'écarts types selon un seuil de rappel donné. Pour chaque seuil de rappel, il y a dix-huit instances de l'algorithme (un par combinaison d'opérateurs) représentées au Tableau VIII.I. Chacune des lignes de ce tableau représente donc une instance de l'algorithme génétique.

9.3 Validation statistique par simulation

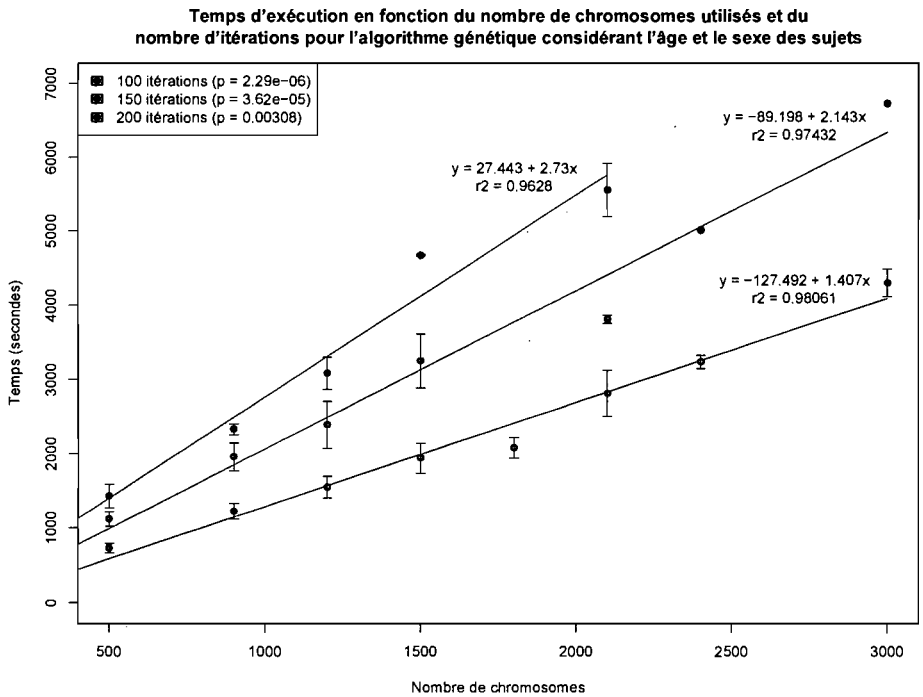
Puisque la validation statistique à l'aide de simulation est longue dans le cas des algorithmes génétiques, uniquement trois seuils de rappel furent vérifiés pour l'algorithme considérant l'âge et le sexe des individus pour la création de ce mémoire. Le Tableau IX.I montre les p-valeurs empiriques calculées suite à un nombre n de simulations. À remarquer qu'une simulation correspond à un échantillon aléatoire, donc à une permutation (sans remplacement) des phénotypes des individus (atteints d'hypertension ou non). Les différentes p-valeurs empiriques sont plus petites à une certaine valeur (symbole $<$) puisque qu'aucune simulation n'a donné de meilleure précision que la précision réelle.

Tableau IX.I: Les différentes p-valeurs empiriques obtenues suite à des simulations de l'algorithme génétique pour le phénotype de l'hypertension. Tableau représentant les différentes p-valeurs empiriques calculées à la suite des différentes simulations pour trois seuils de rappel distincts (0,3, 0,6 et 0,9). L'intervalle de confiance a été calculé à l'aide de la fonction « *binom.confint()* » de *R*.

Seuil de rappel	p-valeur empirique	Intervalle de confiance	Nombre de simulations
0,3	$< 1{,}28\times10^{-3}$	$[0{,}0\text{-}5{,}91\times10^{-3}]$	780
0,6	$< 2{,}04\times10^{-3}$	$[0{,}0\text{-}9{,}38\times10^{-3}]$	490
0,9	$< 3{,}13\times10^{-3}$	$[0{,}0\text{-}1{,}43\times10^{-2}]$	240

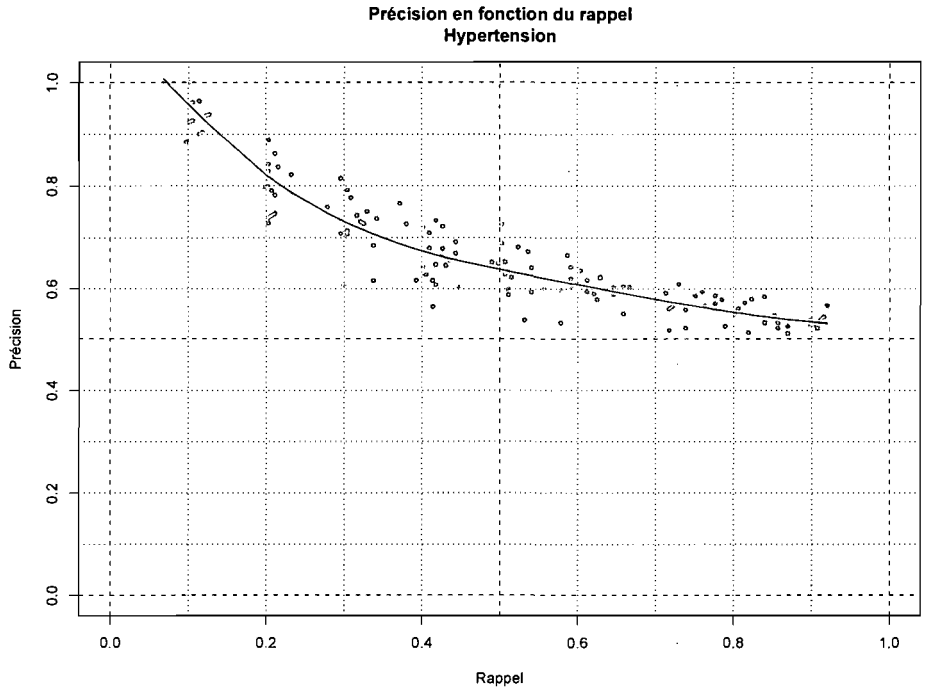


(a)

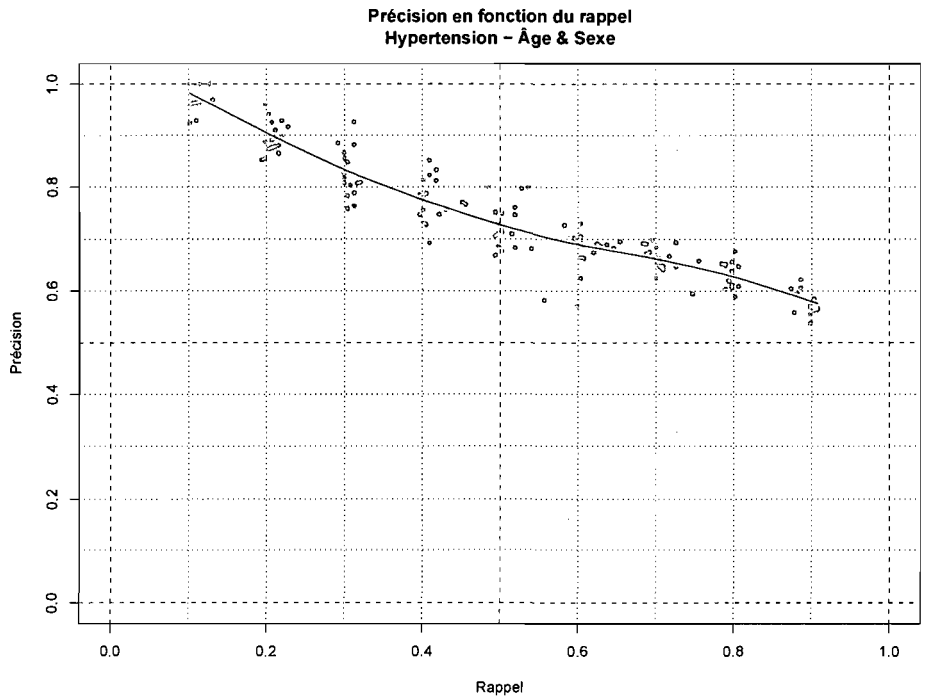


(b)

Figure 9.4: Temps moyen d'exécution de l'algorithme génétique spécifique en fonction du nombre de chromosomes utilisés et du nombre d'itérations. Graphique représentant le temps d'exécution, en secondes, de l'algorithme génétique spécifique simple (en (a)) et de l'algorithme génétique spécifique prenant en compte l'âge et le sexe des individus (en (b)). L'équation représente la droite de régression linéaire simple. Les barres d'erreur représentent les écarts types.

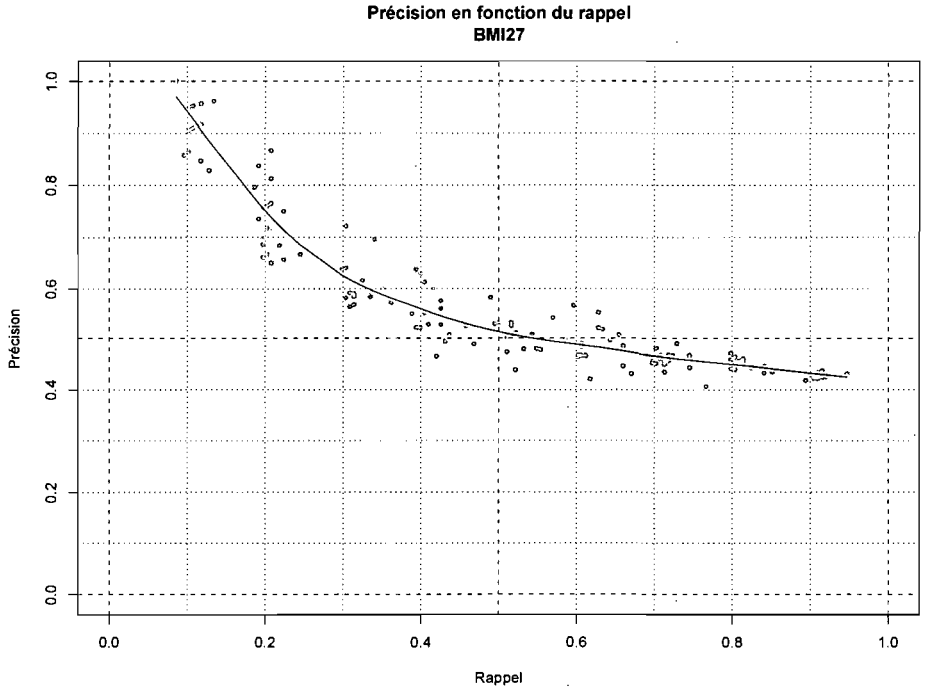


(a)

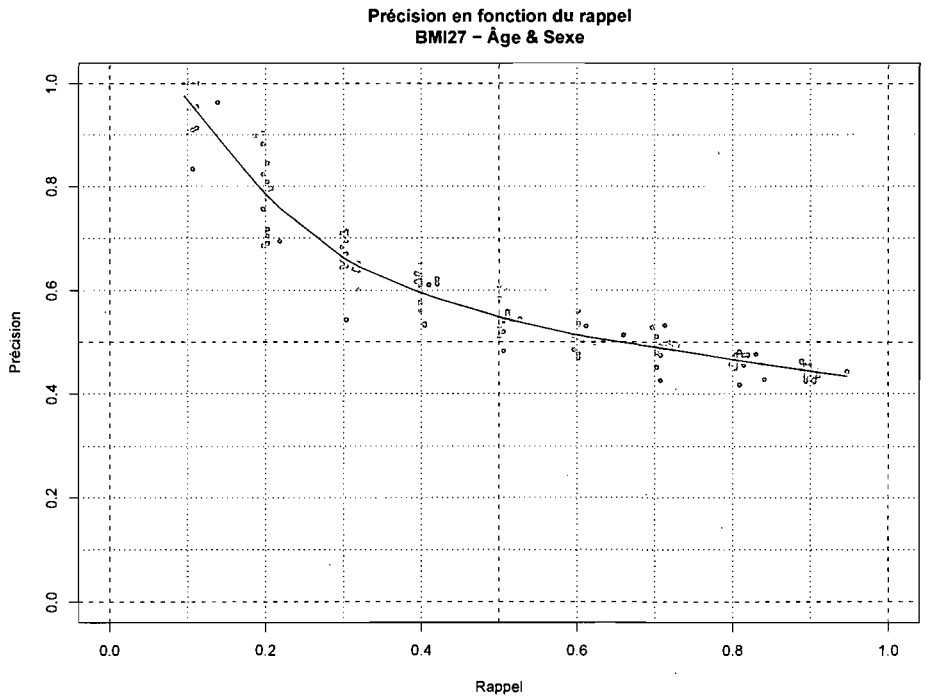


(b)

Figure 9.5: Précision de la recherche en fonction du rappel pour l'hypertension (hyp). Graphique représentant la précision de la recherche à l'aide de l'algorithme génétique spécifique en fonction du rappel pour le phénotype de l'hypertension pour (a), l'algorithme génétique spécifique simple et (b), l'algorithme génétique spécifique considérant l'âge et le sexe des individus.

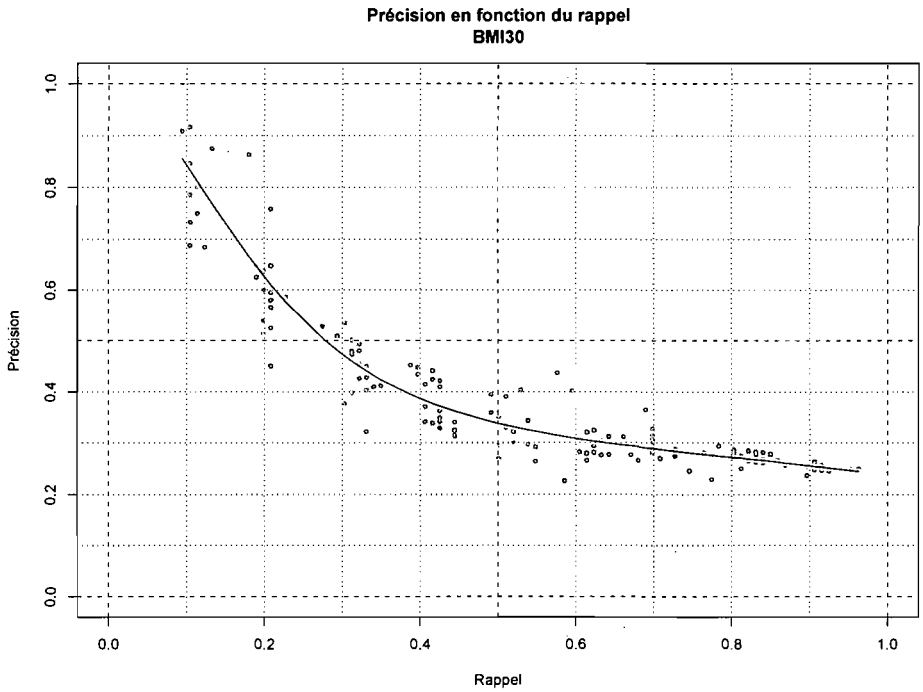


(a)

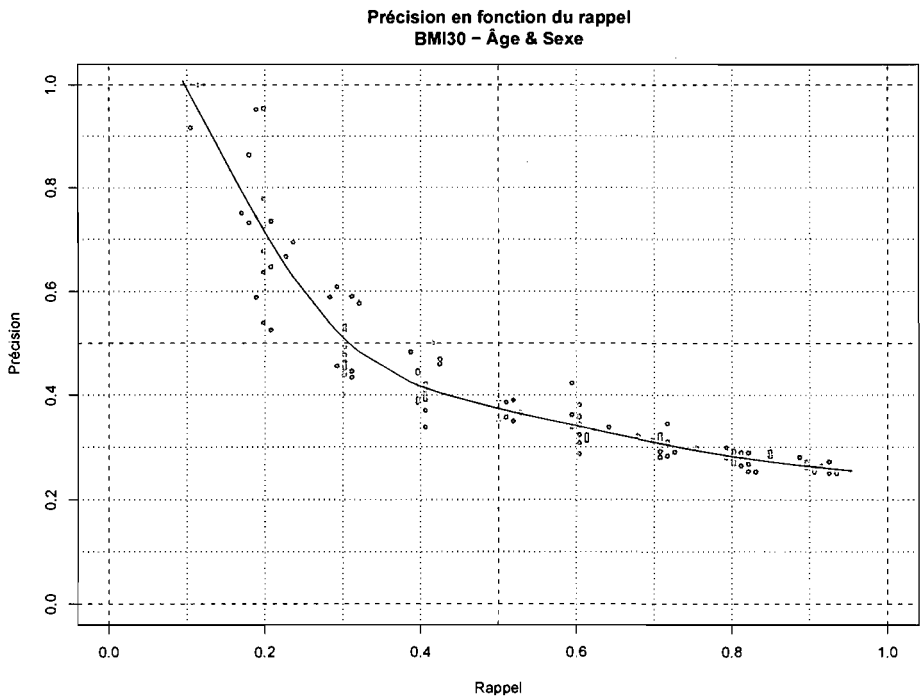


(b)

Figure 9.6: Précision de la recherche en fonction du rappel pour le BMI (BMI27). Graphique représentant la précision de la recherche à l'aide de l'algorithme génétique spécifique en fonction du rappel pour le phénotype BMI ≥ 27 pour (a), l'algorithme génétique spécifique simple et (b), l'algorithme génétique spécifique considérant l'âge et le sexe des individus.



(a)



(b)

Figure 9.7: Précision de la recherche en fonction du rappel pour le BMI (BMI30). Graphique représentant la précision de la recherche à l'aide de l'algorithme génétique spécifique en fonction du rappel pour le phénotype BMI ≥ 30 pour (a), l'algorithme génétique spécifique simple et (b), l'algorithme génétique spécifique considérant l'âge et le sexe des individus.

Tableau IX.II: Moyennes et écarts types de la précision et du rappel selon le seuil de rappel utilisé et le phénotype pour l'algorithme génétique spécifique simple. Tableau représentant les moyennes ainsi que les écarts types de la précision et du rappel selon un certain seuil de rappel et un phénotype donné. Ce tableau fait référence aux Figures 9.5 (a), 9.6 (a) et 9.7 (a) des pages 72, 73 et 74, respectivement.

Phénotype	Rappel			Précision	
	Seuil	Moyenne	Écart type	Moyenne	Écart type
Hypertension[1]	0,1	0,109	0,009	0,942	0,035
	0,2	0,213	0,018	0,799	0,044
	0,3	0,327	0,035	0,722	0,053
	0,4	0,429	0,030	0,660	0,044
	0,5	0,522	0,029	0,641	0,036
	0,6	0,631	0,044	0,602	0,029
	0,7	0,585	0,253	0,640	0,164
	0,8	0,796	0,179	0,574	0,106
	0,9	0,627	0,350	0,668	0,207
BMI27[2]	0,1	0,109	0,012	0,929	0,054
	0,2	0,206	0,014	0,731	0,064
	0,3	0,323	0,025	0,604	0,046
	0,4	0,421	0,024	0,550	0,052
	0,5	0,532	0,034	0,502	0,034
	0,6	0,651	0,058	0,485	0,036
	0,7	0,724	0,046	0,456	0,015
	0,8	0,775	0,170	0,477	0,128
	0,9	0,912	0,011	0,429	0,005
BMI30[3]	0,1	0,106	0,009	0,837	0,115
	0,2	0,205	0,020	0,607	0,091
	0,3	0,324	0,022	0,443	0,051
	0,4	0,425	0,027	0,373	0,048
	0,5	0,517	0,024	0,330	0,049
	0,6	0,625	0,027	0,312	0,045
	0,7	0,729	0,047	0,282	0,021
	0,8	0,819	0,022	0,271	0,014
	0,9	0,915	0,018	0,251	0,006

[1]SBP $\geq$ 140 mm Hg et/ou DBP $\geq$ 90 mm Hg ou prise d'anti-hypertenseur
[2]BMI $\geq$ 27 kg/m^2
[3]BMI $\geq$ 30 kg/m^2

Tableau IX.III: Moyennes et écarts types de la précision et du rappel selon le seuil de rappel utilisé et le phénotype pour l'algorithme génétique spécifique considérant l'âge et le sexe des individus. Tableau représentant les moyennes ainsi que les écarts types de la précision et du rappel selon un certain seuil de rappel et un phénotype donné. Ce tableau fait référence aux Figures 9.5 (b), 9.6 (b) et 9.7 (b) des pages 72, 73 et 74, respectivement.

Phénotype	Rappel			Précision	
	Seuil	Moyenne	Écart type	Moyenne	Écart type
Hypertension[1]	0,1	0,109	0,009	0,979	0,025
	0,2	0,208	0,009	0,898	0,031
	0,3	0,305	0,007	0,825	0,043
	0,4	0,413	0,017	0,773	0,040
	0,5	0,508	0,015	0,727	0,041
	0,6	0,612	0,018	0,682	0,035
	0,7	0,708	0,019	0,662	0,026
	0,8	0,783	0,055	0,633	0,029
	0,9	0,894	0,012	0,578	0,021
BMI27[2]	0,1	0,106	0,009	0,963	0,046
	0,2	0,201	0,006	0,774	0,073
	0,3	0,305	0,007	0,658	0,039
	0,4	0,402	0,007	0,598	0,032
	0,5	0,506	0,007	0,546	0,031
	0,6	0,608	0,015	0,509	0,021
	0,7	0,710	0,009	0,490	0,024
	0,8	0,816	0,022	0,460	0,019
	0,9	0,901	0,013	0,445	0,011
BMI30[3]	0,1	0,102	0,007	0,995	0,019
	0,2	0,199	0,016	0,714	0,117
	0,3	0,302	0,008	0,497	0,060
	0,4	0,404	0,010	0,421	0,041
	0,5	0,506	0,010	0,371	0,018
	0,6	0,611	0,020	0,339	0,030
	0,7	0,712	0,012	0,305	0,061
	0,8	0,810	0,014	0,278	0,013
	0,9	0,908	0,021	0,263	0,009

[1]SBP $\geq$ 140 mm Hg et/ou DBP $\geq$ 90 mm Hg ou prise d'anti-hypertenseur
[2]BMI $\geq$ 27 kg/m^2
[3]BMI $\geq$ 30 kg/m^2

Quatrième partie

Discussion

10. Discussion

À toute évidence, pris individuellement, un SNP ayant un impact fonctionnel ne peut mener directement à une maladie complexe comme l'hypertension. Il s'agirait plutôt d'un ensemble de SNPs (ou un ensemble de blocs d'haplotypes), pouvant être distribué à travers le génome et possédant chacun un effet modeste sur la fonction. Il est donc important d'étudier les différentes interactions entre les blocs d'haplotypes. C'est pourquoi notre algorithme génétique spécifique à l'analyse de la susceptibilité génétique à l'hypertension de la population canadienne-française du SLSJ utilise ces blocs. L'avantage de notre méthode sur les approches conventionnelles (apprentissage machine, par exemple) est de pouvoir sélectionner des blocs d'haplotypes liés génétiquement à une maladie multifactorielle aussi complexe que l'hypertension, malgré un nombre considérable de ces blocs.

Afin de déterminer si notre algorithme génétique est efficace pour l'optimisation de notre fonction objectif, il faut analyser sa convergence vers une solution optimale. En d'autres mots, il faut déterminer si les solutions incluses dans la population tendent, après un certain temps t, à s'approcher d'un optimum de la fonction objectif implantée dans l'algorithme. Il faut aussi vérifier la solution obtenue à la suite d'une exécution de l'algorithme. Il est évident que, pour des problèmes où l'on ne connait pas la solution exacte, il est plutôt difficile d'évaluer la qualité d'une solution quelconque. L'utilisation de deux paramètres, soit la précision et le rappel, longuement utilisés en recherche d'informations et lors de classifications statistiques, devient donc très utile.

10.1 Convergence de l'algorithme

Une convergence adéquate (ni trop lente, ni trop rapide) de l'algorithme génétique est l'un des éléments les plus importants puisque c'est elle qui va dicter l'évolution des solutions vers une solution optimale. Une convergence trop rapide génère habituellement des optimums locaux, qui peuvent parfois se situer relativement loin des optimums globaux. Une convergence trop lente affecte le temps d'exécution de l'algorithme lors de la recherche d'une solution optimale. La vitesse d'une convergence adéquate est relative et dépendra du problème à résoudre et de l'algorithme en soi, puisque la vitesse de convergence est dépendante des des différents opérateurs utilisés.

10.1.1 Convergence rapide

Puisque toutes les solutions convergent rapidement vers une solution unique, la population devient rapidement homogène. L'opérateur de croisement n'a donc plus d'effet sur les solutions et l'opérateur de mutation devient l'unique moyen de continuer à faire progresser l'algorithme. La recherche s'apparente alors davantage à une recherche aléatoire sur l'espace de recherche et la probabilité de « sortir » de l'optimum local devient relativement faible. C'est ce qui est illustré à la Figure 9.3.

Bien que ce graphique représente une instance de l'algorithme génétique ayant obtenu une bonne solution (lorsqu'on la compare aux solutions d'autres instances), on voit bien que la convergence des solutions se fait très rapidement (en environ 35 générations de solutions). Ceci est dû à l'utilisation de l'élitisme comme opérateur de remplacement. Comme il est expliqué à la section 6.6, l'opérateur de remplacement élitisme conserve les meilleures solutions de la population au temps t tout en remplaçant les moins performantes par les meilleures créées par croisement/mutation afin de former la nouvelle génération au temps $t+1$. En d'autres mots, la nouvelle génération sera constituée des meilleures solutions de la génération au temps t et des meilleures solutions « enfants » nouvellement créées. Comme les solutions nouvellement créées le

sont à l'aide des meilleures solutions au temps t, elles ressemblent énormément à ces dernières (homogénéité). La vitesse de convergence s'en trouve donc augmentée.

En général, lors de nos analyses, pour des valeurs de seuils de rappel élevées (*i.e.* > 0,7), lorsque l'opérateur de remplacement élitisme est utilisé, l'algorithme génétique arrête parfois sa recherche dans un optimum local ayant comme valeurs de rappel et de précision 0,1 et 1, respectivement (représentés par certains des points situés dans le coin supérieur gauche des graphiques présentés aux figures 9.5, 9.6 et 9.7). Puisqu'à chaque fois que le rappel augmente, la précision diminue, il est très difficile d'augmenter les valeurs de rappel pour atteindre le seuil. En effet, la fonction objectif, lorsque la précision est de 1 et le rappel, de 0,9, est de 0,11. Si l'on augmente le rappel de 0,1, la précision augmentera à environ 0,8 (selon les données présentées dans le Tableau IX.III). La mesure de performance diminue donc à 0,022 (diminution de 0,089). L'algorithme génétique préférera la première solution à la deuxième, puisque sa performance est meilleure. Il faut donc un changement drastique de rappel et de performance pour permettre d'atteindre le seuil de rappel afin de créer la courbe mettant en relation la précision et le rappel.

10.1.2 Convergence lente

Alors qu'une convergence trop rapide a pour conséquence l'homogénéisation de la population de solutions, une convergence trop lente, quant à elle, a pour effet de conserver une population hétérogène beaucoup plus longtemps, maintenant ainsi un grand nombre de solutions différentes. Même si l'algorithme réussit à obtenir une solution optimale, il le fera en un laps de temps beaucoup plus grand. La Figure 9.2 montre une convergence plus lente.

Ce graphique présente aussi une solution optimale adéquate, comparée aux autres solutions. La convergence, par contre, se fait beaucoup plus lentement (en environ 75

générations de solution, soit 40 de plus que l'exemple précédent). L'opérateur de remplacement utilisé dans ce cas-ci est le « generation gap » (voir section 6.6 pour plus d'informations). En résumé, les solutions qui vont remplacer celles au temps t (afin de former la génération au temps $t + 1$) sont choisies aléatoirement. Il y a donc une grande probabilité que la nouvelle génération ne soit constituée que des pires solutions de la génération au temps t et des pires « enfants » nouvellement créés. Ceci a pour conséquence de ralentir beaucoup la convergence de l'algorithme génétique.

L'inconvénient majeur d'une convergence trop lente n'est pas de rester bloqué dans un optimum local, mais bien d'obtenir des temps d'exécution beaucoup trop lents comparativement à une convergence « normale » de l'algorithme (compromis entre les convergences rapides et lentes).

10.1.3 Convergence « normale »

Tout algorithme génétique tente d'obtenir un juste milieu entre une convergence trop rapide et trop lente des solutions vers l'optimum. Une convergence dite « normale » représente le compromis entre la qualité des solutions et le temps d'exécution. La Figure 9.1 représente une convergence « normale ». En effet, dans cet exemple, la convergence se fait en environ 40 générations de solutions (comparé à 75 générations de solutions [convergence lente] et à 35 générations de solutions [convergence rapide]). La solution optimale trouvée lors de cette recherche a une performance de 0,855 alors que les solutions optimales avec les convergences lente et rapide ont une performance de 0,783 et 0,809 respectivement.

Nous pouvons donc conclure que la vitesse de convergence vers une solution optimale est dépendante des différents opérateurs utilisés ainsi que des différents paramètres de l'algorithme génétique. De plus, l'obtention d'une solution optimale est reliée à la vitesse de convergence ; plus la convergence est rapide, plus nous risquons d'accéder

à un optimum local. À l'inverse, plus la convergence est lente, plus le temps d'exécutions devient grand, et l'algorithme pourrait s'arrêter avant l'obtention d'une solution optimale. Nous pouvons aussi voir une différence dans le bruit de fond (solutions non optimales restant tout de même dans la population de l'algorithme génétique). Plus la convergence est rapide, moins il y a présence de bruit de fond. Malgré différentes vitesses de convergence de notre algorithme (conséquence de l'utilisation d'opérateurs distincts), ce dernier fournit tout de même des résultats similaires en des temps raisonnables, ce qui est un avantage sur les méthodes d'apprentissage machine.

10.2 Effet de l'immigrant aléatoire

Nous voyons bien l'effet de l'opérateur d'immigrant aléatoire sur la population de solutions dans les trois Figures 9.1, 9.2 et 9.3. Le moment où l'algorithme exécute l'opérateur dépend de la vitesse de convergence. Plus celle-ci se fera précocement, plus l'effet de l'opérateur apparaîtra rapidement. Ceci s'explique par le fait que plus la population de solutions est homogène, moins la meilleure solution trouvée aura tendance à changer (à devenir plus performante). Comme il est expliqué à la section 6.7, l'opérateur d'immigrant aléatoire s'applique lorsque la solution optimale ne s'est pas améliorée après k itérations (lors de nos analyses, k se situe entre 15 et 30 itérations). À chaque fois que cet opérateur est utilisé, il remplace entre 60 % et 90 % de la population, choisissant les solutions les moins performantes. Son rôle est d'ajouter de la diversité dans la population de solutions tout en ajoutant des solutions qui, dans certains cas, n'ont jamais été considérées. Ceci a pour conséquence de ralentir la convergence de l'algorithme vers une solution optimale.

À la Figure 9.3 (représentant une convergence plus rapide), l'immigrant aléatoire opère à deux reprises, soit à environ 50 et 75 générations de solutions. Nous pouvons remarquer qu'à la première exécution de l'opérateur, il n'y a aucun effet sur la meilleure solution trouvée jusqu'à présent. Lors de la deuxième exécution, par contre, on peut

voir une augmentation de la performance de la meilleure solution trouvée (valeur de la fonction objectif). Dans les deux cas, nous pouvons remarquer que la moyenne des performances des solutions diminue à chaque exécution de l'immigrant aléatoire. Ceci est dû au fait que l'opérateur crée ses solutions de façon aléatoire. Leurs performances ont donc une probabilité accrue d'être plus faibles que la moyenne des performances courantes, car ces dernières ont eu le temps « d'évoluer ». Puisque dans ce cas l'opérateur de remplacement élitisme est utilisé, nous pouvons apercevoir que l'effet de l'immigrant aléatoire est court (soit une seule génération de solutions) étant donné que les solutions nouvellement créées seront enlevées très rapidement par l'opérateur de sélection.

À la Figure 9.2 (représentant une convergence plus lente), l'immigrant aléatoire opère une seule fois, à l'itération 95 environ, et de façon beaucoup plus tardive. Malheureusement, l'effet de l'opérateur sur la meilleure solution n'est pas visible, puisque l'algorithme arrête son évolution à la l'itération 100. Par contre, en général, lors de l'utilisation du « generation gap » comme opérateur de remplacement, nous n'apercevons que très rarement l'effet de l'immigrant aléatoire sur la meilleure solution ; la population de solution étant déjà hétérogène, l'ajout de nouvelle solution n'a pas toujours d'effet. La fonction de l'opérateur change donc pour ajouter un peu de bruit dans les solutions. Par contre, lorsque nous regardons la moyenne des performances des solutions, nous pouvons remarquer que celle-ci diminue. Encore une fois, l'effet à long terme sur la moyenne n'est pas visible, mais généralement celui-ci est beaucoup plus long que lors d'une convergence plus rapide (comme à l'exemple précédent). L'effet peut aller jusqu'à une vingtaine de générations de solutions en moyenne, lors de nos expériences.

À la Figure 9.1 (représentant une convergence « normale »), l'immigrant aléatoire opère approximativement aux itérations 65 et 95. Le moment de l'exécution de l'opérateur lors d'une convergence « normale » se situe entre ceux des convergences rapide

et lente. Nous pouvons apercevoir une amélioration de la meilleure solution lors de la première exécution de l'opérateur. L'effet sur la moyenne des performances est d'environ 5 générations de solutions.

Dans tous les cas, l'opérateur d'immigrant aléatoire est synonyme d'hétérogénéité de la population de l'algorithme génétique. Il est essentiel d'ajouter du bruit de fond dans les solutions afin de permettre une exploration plus grande de l'espace de recherche. Ceci est d'autant plus vrai lors d'une convergence rapide de l'algorithme. En effet, dans la majorité des cas, l'opérateur d'immigrant aléatoire a pour effet la création de nouvelles solutions permettant une amélioration de la meilleure solution trouvée. L'ajout de l'opérateur d'immigrant aléatoire permet de faire un saut dans l'espace de recherche et de trouver d'autres solutions possibles lors d'une même recherche. Ceci ne peut pas être fait facilement avec certaines méthodes heuristiques utilisées comme la recherche taboue (utilisant un voisinage de solutions) ; la recherche doit être recommencée afin de modifier les solutions initiales et le voisinage de solutions.

L'opérateur d'immigrant aléatoire permet aussi d'investiguer des haplotypes apparaissent peu dans la population (à faible fréquence). En effet, il est normal (mais non souhaitable) que l'algorithme tende à analyser les haplotypes apparaissant en grand nombre dans la population. Puisque l'opérateur d'immigrant aléatoire ajoute des solutions de façon aléatoire à l'ensemble de solutions courant, la probabilité d'ajouter un haplotype ayant une faible fréquence est la même que celle d'ajouter un haplotype se retrouvant en grande quantité chez les sujets étudiés.

10.3 Temps d'exécution

Le temps d'exécution d'une instance de l'algorithme génétique spécifique est important. Un algorithme permettant de trouver rapidement de très bonnes solutions à un problème est plus intéressant qu'un autre algorithme donnant les mêmes solutions

mais de manière beaucoup plus lente. Les deux graphiques de la Figure 9.4 représentent le temps d'exécution, en secondes, en fonction du nombre de solutions (chromosomes de l'algorithme génétique) utilisées. On peut remarquer que dans les deux versions de l'algorithme (soit l'algorithme génétique simple, en (a), et l'algorithme génétique considérant l'âge et le sexe des sujets, en (b)), le temps évolue de façon linéaire en fonction du nombre de solutions utilisées dans la population et ce, dans l'intervalle étudié. Dans tous les cas, sauf un, la p-valeur de la régression linéaire est significative (soit $< 0{,}05$). Les différentes valeurs de r^2 sont très bonnes dans tous les cas (elles sont près de 1).

Lorsque nous comparons les temps d'exécution pour les deux versions de l'algorithme génétique pour un nombre d'itération donné, on peut remarquer que ceux-ci varient peu, et ce, même si la version évoluée incorpore de l'information supplémentaire sur l'âge et le sexe des sujets à l'étude. En effet, toutes les analyses faites à l'aide d'un test de Student ont donné des p-valeurs non-significatives. Il n'y a donc pas de différence en temps d'exécution entre les deux versions de l'algorithme génétique spécifique.

Le temps d'exécution est malheureusement grand lorsque nous considérons un nombre élevé de solutions. En effet, il peut atteindre une heure cinquante minutes lorsque nous considérons 150 itérations avec 3 000 solutions dans la population. Cependant, il est à noter que ces extrêmes ne se produisent que très rarement (soit lorsque le seuil de rappel est élevé). Dans la majorité des cas, l'algorithme a convergé vers une solution optimale après 100 itérations et avec une population de 500 solutions (chromosomes). Si nous utilisons la formule représentant la droite de régression linéaire simple correspondante (soit $y = -127,492 + 1,407x$), nous obtenons une moyenne de 576 secondes d'exécutions (soit 9 minutes, 36 secondes), compétitionnant avec les méthodes d'apprentissage machine qui peuvent prendre environ trois heures avec les mêmes données. Ce temps moyen d'exécution est bon, si l'on considère que l'algorithme n'a pas encore été optimisé à cet effet.

Il existe plusieurs façons d'optimiser un algorithme génétique en ce qui a trait au temps d'exécution. Puisqu'en général l'évaluation de la fonction objectif nécessaire au calcul de la performance des solutions est l'opération qui demande le plus de temps, il est nécessaire de la perfectionner. Pour ce faire, la littérature propose l'utilisation d'algorithmes génétiques parallèles. Cette approche n'a pas été utilisée dans le cadre de ce projet de recherche puisqu'elle nécessite l'utilisation d'une grappe de processeurs. Plus de détails concernant les algorithmes génétiques parallèles sont présentés à l'annexe II.

10.4 Précision et rappel

Afin d'analyser la qualité des solutions optimales trouvées suite à une recherche à l'aide de l'algorithme génétique spécifique, nous employons deux paramètres couramment utilisés : la précision et le rappel [126, 127]. Tel que mentionné précédemment, la précision nous informe sur le nombre de vrais positifs et de faux négatifs, alors que le rappel nous informe sur le nombre de faux positifs. Dans notre recherche, le nombre de vrais positifs représente le nombre de sujets présentant le phénotype d'intérêt, possédant tous les blocs d'haplotypes trouvés à l'aide de l'algorithme génétique (solution optimale). Le nombre de faux négatifs, quant à lui, représente le nombre de sujets présentant le phénotype d'intérêt, mais ne possédant pas tous les blocs d'haplotypes de la solution optimale. Finalement, les faux positifs représentent le nombre de sujets ne présentant pas le phénotype d'intérêt et possédant tous les blocs d'haplotypes de la solution de l'algorithme.

Les graphiques des Figures 9.5, 9.6 et 9.7 représentent la qualité des solutions. Pour chaque valeur de seuil de rappel, l'algorithme optimise la précision. Nous voyons donc apparaître une distribution de ces deux paramètres (soit la précision en fonction du rappel). Puisque la précision varie de façon inversement proportionnelle à la variation du rappel et vice versa, il est ardu de décider quel paramètre nous allons « sacrifier »

afin d'optimiser l'autre, puisque ces deux valeurs ont chacune leur importance.

Une distribution de ces deux paramètres nous permet de visualiser conjointement les deux valeurs de précision et de rappel, l'une en fonction de l'autre. Nous pouvons donc décider par la suite si nous privilégions la précision ou le rappel. Par exemple, si nous voulons obtenir un ensemble de blocs d'haplotypes représentant uniquement des sujets hypertendus (une précision de 100 %) suite à une analyse avec l'algorithme génétique spécifique simple, nous savons que, à l'aide de la Figure 9.5 (a), nous représenterons environ 10 % de tous les sujets hypertendus (rappel). À l'inverse, si nous voulons obtenir des blocs d'haplotypes représentant tous les sujets présentant un BMI supérieur ou égal à 27 kg/m^2 (un rappel de 100 %), suite à une analyse avec l'algorithme génétique considérant l'âge et le sexe des sujets, la Figure 9.6 (b) nous permet de constater que 45 % des sujets représentés par l'ensemble de blocs d'haplotypes présenteront le phénotype, alors que 55 % ne le présenteront pas (faible précision).

Un autre avantage de ces graphiques est de permettre la comparaison rapide entre les différentes versions de l'algorithme génétique spécifique. Si nous analysons les distributions présentées à la Figure 9.5 (a) et (b), nous voyons une légère augmentation des valeurs de précision et de rappel lorsque l'algorithme considère l'âge et le sexe des sujets de l'étude sur l'hypertension. Le même phénomène se produit lors de l'étude des deux autres phénotypes, soit le BMI27 et le BMI30 (Figure 9.6 (a) et (b) et Figure 9.7 (a) et (b)).

Cette différence est résumée lors de la comparaison des Tableaux IX.II et IX.III. Le Tableau X.I représente la comparaison directe des différentes valeurs de précision pour différents seuils de rappel fixés, classées selon le phénotype étudié. On peut remarquer que dans tous les cas (sauf deux), la moyenne de précision augmente de 1 % à 16 % lorsque nous ajoutons de l'information sur l'âge et le sexe des patients (validée statisti-

quement par un test de Student). Nous pouvons donc affirmer que le fait d'ajouter ces renseignements supplémentaires aide à la classification des différents blocs d'haplotypes lors d'une étude de liaison génétique, et ce, même si cela implique un accroissement du nombre de dimensions des données. Il est probable que l'augmentation d'informations sur l'environnement direct des sujets à l'étude, tels l'apport en sel et la consommation de tabac, augmenterait la précision moyenne pour des seuils de rappel donnés. Le fait d'ajouter de l'information à la recherche permet tout de même à l'algorithme génétique de trouver une signature d'haplotypes valable (voir section 10.5 pour la validation statistique). Il est probable que cet ajout rendrait difficile l'utilisation des approches traditionnelles comme l'apprentissage machine, par exemple, puisque cette dernière à déjà de la difficulté à trouver des solutions dans de gros jeux de données (selon des expérimentations faites et non présentées dans le texte).

10.5 Validation statistique par simulation

Le Tableau IX.I de la page 70 présente les différentes p-valeurs empiriques calculées à l'aide de simulations pour certains seuils de rappel. On peut y remarquer que celles-ci sont toutes significatives ($< 0{,}05$) et qu'elles sont inclusent dans l'intervalle de confiance calculé. Nous pouvons donc rejeter l'hypothèse nulle ; les résultats obtenus ne le sont pas par hasard et il existe une association réelle entre les signatures d'haplotypes trouvées et le phénotype étudié. À titre de comparaison, lors d'un test fait à l'aide d'un algorithme d'apprentissage machine sur le même jeu de données, les résultats obtenus ne furent pas statistiquement significatifs (test fait par un membre du laboratoire, non présenté dans ce document).

Tous les seuils de rappel n'ont pas été vérifiés puisque le temps pour accomplir cette tâche est considérable. En prenant des seuils de rappel situés stratégiquement (à gauche, au centre et à droite tels que 0,3, 0,6 et 0,9), on peut donner une bonne idée de la significativité des résultats pour toutes les valeurs de précision et de rappel

présentées par les courbes des Figures 9.5, 9.6 et 9.7 des pages 72, 73 et 74. Pour la même raison que mentionnée ci-haut, uniquement les résultats portant sur le phénotype de l'hypertension furent validés statistiquement.

10.6 Autres utilisations de l'algorithme

Il est connu que les SNPs, la fréquence des allèles, le déséquilibre de liaison (et donc la longueur des blocs d'haplotypes) diffèrent beaucoup d'une population à une autre [2], en raison de l'historique des populations (âge, nombre de fondateurs, évolution, migration) [137, 138, 139]. Il est donc normal que la signature d'haplotypes trouvée à l'aide de notre algorithme ne puisse être utilisée que sur la population du Saguenay-Lac-Saint-Jean. L'avantage de notre méthode est l'utilisation du concept de numéro d'identification unique pour les blocs et les allèles (voir chapitre 7). Ce concept n'utilise pas la composition exacte du bloc d'haplotypes en question (énumération des SNPs). Il est donc possible d'utiliser notre GA spécifique sur d'autres populations comportant de grandes différences en ce qui a trait à la composition des haplotypes sans connaître en détail la structure exacte de ces derniers, ce qui est un avantage marqué.

L'algorithme peut aussi être utilisé sur d'autres phénotypes complètement différents de l'hypertension, pourvu qu'ils soient binaires ; un individu présente ou ne présente pas le phénotype. Par exemple, au cours de ce projet de recherche, l'algorithme fut utilisé pour étudier l'indice de masse corporelle sans modifications majeures de l'algorithme. Il est aussi possible d'ajouter de l'information sur l'environnement des sujets à l'étude (habitudes alimentaires, exercice physique, etc.). Malheureusement, l'utilisation de phénotype discret n'est pas supportée par l'algorithme implémenté.

Tableau X.I: Différence de précision entre les deux versions de l'algorithme génétique spécifique. Tableau représentant la différence de précision entre l'algorithme génétique spécifique simple et celui considérant l'âge et le sexe des individus. Les chiffres rouges représentent une diminution de la précision d'une version de l'algorithme à l'autre. Les chiffres bleus représentent une augmentation de la précision supérieure à 10 %. Les chiffres verts représentent les p-valeurs non significatives (soit $\geq$ 0,05) suite à un test de Student. Les données sont tirées des Tableaux IX.II et IX.III des pages 75 et 76, respectivement.

Phénotypes	**Seuils de Rappel**	**Précision**		**Différence**	**Student**
		GA_{sp_simple} (A)	$GA_{sp\_A\&S}$ (B)	$B - A$	p-**valeur**
Hypertension	0,1	0,942	0,979	0,038	$1{,}19\times10^{-3}$
	0,2	0,799	0,898	0,099	$1{,}25\times10^{-8}$
	0,3	0,722	0,825	0,103	$4{,}75\times10^{-7}$
	0,4	0,660	0,773	0,113	$4{,}62\times10^{-9}$
	0,5	0,641	0,727	0,085	$2{,}33\times10^{-7}$
	0,6	0,602	0,682	0,080	$2{,}26\times10^{-8}$
	0,7	0,640	0,662	0,022	0,600
	0,8	0,574	0,633	0,059	0,0389
	0,9	0,668	0,578	-0,090	0,0924
BMI27	0,1	0,929	0,963	0,034	0,0568
	0,2	0,731	0,774	0,043	0,0771
	0,3	0,604	0,658	0,054	$8{,}79\times10^{-4}$
	0,4	0,550	0,598	0,048	$3{,}09\times10^{-3}$
	0,5	0,502	0,546	0,043	$3{,}91\times10^{-4}$
	0,6	0,485	0,509	0,024	0,0276
	0,7	0,456	0,490	0,035	$2{,}10\times10^{-5}$
	0,8	0,477	0,460	-0,017	0,584
	0,9	0,429	0,445	0,017	$5{,}65\times10^{-6}$
BMI30	0,1	0,837	0,995	0,159	$2{,}50\times10^{-5}$
	0,2	0,607	0,714	0,106	$5{,}78\times10^{-3}$
	0,3	0,443	0,497	0,054	$8{,}19\times10^{-3}$
	0,4	0,373	0,421	0,048	$3{,}74\times10^{-3}$
	0,5	0,330	0,371	0,041	$3{,}62\times10^{-3}$
	0,6	0,312	0,339	0,027	0,0438
	0,7	0,282	0,305	0,024	$8{,}52\times10^{-4}$
	0,8	0,271	0,278	0,007	0,145
	0,9	0,251	0,263	0,012	$6{,}09\times10^{-5}$

Cinquième partie

Conclusion

11. Conclusions & perspectives

11.1 Conclusions

Nous avons développé une approche permettant d'étudier la susceptibilité génétique à l'hypertension de la population du Saguenay-Lac-Saint-Jean en sélectionnant des traits génétiques (blocs d'haplotypes) pouvant être associés au phénotype d'intérêt (l'hypertension) ; une signature d'haplotypes. Cette approche consiste en l'utilisation d'une métaheuristique nommée algorithme génétique. Deux versions de l'algorithme furent développées : l'algorithme génétique spécifique simple et l'algorithme génétique spécifique considérant l'âge et le sexe des sujets à l'étude. Ce dernier permet l'ajout de plusieurs autres informations sur l'environnement direct des sujets, permettant ainsi d'affiner la recherche de blocs d'haplotypes représentant un phénotype donné à l'aide de covariables telles que la consommation de tabac ou d'alcool, par exemple. Dans le cadre de ce projet, uniquement l'âge et le sexe des individus furent considérés comme covariables.

Les deux algorithmes génétiques ont démontré une convergence vers une solution optimale respectable. Souvent, à l'aide d'une combinaison des différents opérateurs essentiels au bon fonctionnement de l'algorithme, il est possible de modifier cette convergence pour qu'elle se fasse plus ou moins rapidement. Par exemple, l'utilisation de l'opérateur de remplacement de génération élitisme engendrera une convergence beaucoup plus rapide que l'utilisation de l'opérateur « generation gap ». L'utilisation d'opérateurs différents, ayant pour conséquence de modifier la vitesse de convergence de

l'algorithme, permet une meilleure couverture de l'espace de recherche, rendant possible la détection de divers optimums locaux et globaux.

Les différents temps d'exécution des deux versions de l'algorithme génétique sont relativement courts, compte tenu du nombre de calculs nécessaires afin de permettre l'obtention de solutions détenant une certaine qualité (selon le seuil de rappel). De plus, suite à un test de Student, les différents temps de calcul ne changent pas de façon significative entre les deux versions de l'algorithme, même si l'une d'elle considère certains éléments de l'environnement des sujets. Une optimisation peut par contre être faite lorsque l'on considère les avantages des algorithmes génétiques parallèles. Cependant, afin de permettre ce type d'optimisation, il faut envisager un autre langage de programmation, puisque Python ne s'adaptait pas pour l'utilisation de multiprocesseurs lors de l'implantation de l'algorithme génétique.

Nous avons aussi représenté les différentes distributions des valeurs de précision et de rappel pour les trois phénotypes analysés. Ces distributions permettent de montrer la relation directe du rappel sur la précision de notre algorithme de recherche portant sur notre ensemble de données du Saguenay-Lac-Saint-Jean. Par exemple, notre algorithme est capable d'obtenir des solutions ayant des précisions moyennes de 98 %, de 73 % et de 63 % pour des rappels de 10 %, de 50 % et de 80 % respectivement. Nous avons aussi remarqué une augmentation de la précision de l'algorithme de recherche lorsque nous ajoutons de l'information supplémentaire à l'information génétique déjà disponible, tels l'âge et le sexe des sujets à l'étude. Cet ajout d'information a permis une meilleure sélection des blocs d'haplotypes (représentée par une augmentation des valeurs de précision pouvant aller jusqu'à 10 % pour un seuil de rappel donné). Cette augmentation a été validée statistiquement à l'aide d'un test de Student.

En plus des valeurs de précision et de rappel nous permettant d'étudier la qualité

des solutions obtenues, nous avons validé statistiquement, à l'aide de la méthode de rééchantillonnage, nos résultats. Les p-valeurs empiriques calculées suite à cette analyse sont toutes statistiquement significatives puisqu'elles sont inférieures à 0,05 (95 % de confiance) et qu'elles se retrouvent dans leurs intervalles de confiance respectifs. Les p-valeurs empiriques varient entre $1{,}28\times10^{-3}$ et $3{,}13\times10^{-3}$. Uniquement certains seuils de rappel furent étudiés pour le phénotype de l'hypertension, puisque le temps nécessaire afin de valider tous les seuils pour les trois phénotypes (hypertension, BMI27 et BMI30) est très grand.

Notre algorithme génétique spécifique montre des avantages marqués lorsqu'il est comparé aux méthodes courantes de création de signature d'haplotypes. En effet, il permet la création de signatures statistiquement valides avec un temps d'analyse relativement court (lorsqu'il est comparé aux méthodes d'apprentissage machine testées sur le même jeu de données). De plus, il permet facilement l'ajout de covariables binaires sans allonger significativement le temps d'exécution.

11.2 Perspectives

La prochaine étape du projet de recherche serait de rechercher une fonction objectif améliorée permettant l'obtention de meilleures solutions optimales plus rapidement, ce qui ferait augmenter les valeurs de précision et de rappel tout en diminuant le temps d'exécution de l'algorithme.

Une sélection des haplotypes étudiés permettrait une meilleure réduction dans la dimensionnalité des données et pourrait engendrer des valeurs de rappel et de précision plus élevées. De plus, une analyse plus approfondie des résultats obtenus à l'aide de notre algorithme pourrait être réalisée dans le cadre d'un autre projet de recherche. Pour ce faire, il faudrait scruter chaque ensemble de blocs d'haplotypes sélectionnés par l'algorithme de recherche et regarder dans la littérature leur emplacement respectif

dans le génome humain, afin de trouver des gènes situés à proximité. Cette approche permettrait de renforcer la validation de la liaison génétique de ces marqueurs avec le phénotype d'intérêt. Ces analyses ne furent pas effectuées, puisqu'elles dépassaient le but de ce projet de recherche, qui consistait en la création d'un algorithme permettant la sélection de marqueurs génétiques (blocs d'haplotypes) pouvant représenter un phénotype d'intérêt. Les valeurs de précision et de rappel nous suffisaient pour étudier la qualité de la sélection des différents blocs d'haplotypes.

Bibliographie

[1] The Wellcome Trust Sanger Institute / European Bioinformatics Institute. Ensembl Genome Browser. `http://www.ensembl.org/index.html`, Dec 2007.

[2] J. C. Stephens, J. A. Schneider, D. A. Tanguay, J. Choi, T. Acharya, S. E. Stanley, R. Jiang, C. J. Messer, A. Chew, J. H. Han, J. Duan, J. L. Carr, M. S. Lee, B. Koshy, A. M. Kumar, G. Zhang, W. R. Newell, A Windemuth, C. Xu, T. S. Kalbfleisch, S. L. Shaner, K. Arnold, V. Schulz, C.M. Drysdale, K. Nandabalan, R. S. Judson, G. Ruano, and G. F. Vovis. Haplotype variation and linkage disequilibrium in 313 human genes. *Science*, 293(5529) :489–493, July 2001.

[3] D. L. Rimoin, J. M. Connor, R. E. Pyeritz, and B. R. Korf, editors. *Emery and Rimoin's Principles and Practice of Medical Genetics.* Churchill Livingstone Elsevier, United States of America, fifth edition, 2007.

[4] R. L. Nussbaum, R. R. McInnes, and H. R. Willard. *Thompson & Thompson Genetics in Medicine.* Saunders, United States of America, sixth edition, 2004.

[5] A. Ciechanowicz, Z. Dolezel, G. Placha, J. Starha, J. Góra, Z. Gaciong, A. Brodkiewicz, and G. Adler. Liddle syndrome caused by P616R mutation of the epithelial sodium channel β subunit. *Pediatric nephrology*, 20 :837–838, 2005.

[6] F. S. Collins, L. D. Brooks, and A. Chakravarti. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Research*, 8(12) :1229–1231, December 1998.

[7] A. Chakravarti. Population genetics—making sense out of sequence. *Nature Genetics*, 21 :56–60, January 1999.

[8] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273 :1516–1517, 1996.

[9] D. Gordon, S. C. Heath, X. Liu, and J. Ott. A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. *American Journal of Human Genetics*, 69 :371–380, 2001.

[10] M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander. High-resolution haplotype structure in the human genome. *Nature Genetics*, 29 :229–232, October 2001.

[11] S. B. Gabriel, S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly, and D. Altshuler. The structure of haplotype blocks in the human genome. *Science*, 296 :2225–2229, June 2002.

[12] P. Y. Liu, Y. Y. Zhang, Y. Lu, J. R. Long, H. Shen, L. J. Zhao, F. H. Xu, P. Xiao, D. H. Xiong, Y. J. Liu, R. R. Recker, and H. W. Deng. A survey of haplotype variants at several disease candidate genes : the importance of rare variants for complex diseases. *Journal of medical genetics*, 42(3) :221–227, March 2005.

[13] A. V. Chobanian, G. L. Bakris, H. R. Black, W. C. Cushman, L. A. Green, J. L. Izzo, D. W. Jones, B. J. Materson, S. Oparil, J. T. Jr. Wright, E. J. Roccella, and The National High Blood Pressure Education Program Coordinating Committee. Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure. *Hypertension*, 42(6) :1206–1252, December 2003.

[14] P. Hamet, E. Merlo, O. Šeda, U. Broeckel, J. Tremblay, M. Kaldunski, D. Gaudet, G. Bouchard, B. Deslauriers, F. Gagnon, G. Antoniol, Z. Pausová, M. Labuda, M. Jomphe, F. Gossard, G. Tremblay, R. Kirova, P. Tonellato, S. N. Orlov, J Pintos, J. Platko, T. J. Hudson, J. D. Rioux, T. A. Kotchen, and A. W. Cowley Jr. Quantitative founder-effect analysis of french canadian families identifies specific loci contributing to metabolic phenotypes of hypertension. *American Journal of Human Genetics*, 76 :815–832, 2005.

[15] P. M. Kearney, M. Whelton, K. Reynolds, P. Muntner, P. K. Whelton, and J. He. Global burden of hypertension : analysis of worldwide data. *Lancet*, 365 :217–223, 2005.

[16] S. S. Franklin, M. J. Jacobs, N. D. Wong, G. J. L'Italien, and P. Lapuerta. Predominance of isolated systolic hypertension among middle-aged and elderly US hypertensives : analysis based on National Health and Nutrition Examination survey (NHANES) III. *Hypertension*, 37 :869–874, 2001.

[17] M. R. Joffres, P. Hamet, D. R. MacLean, G. J. L'italien, and G. Fodor. Distribution of blood pressure and hypertension in Canada and the United States. *American Journal of Hypertension*, 14 :1099–1105, 2001.

[18] P. Hamet. The burden of blood pressure : Where are we and where should we go? *Canadian Journal of Cardiology*, 16(12) :1483–1487, 2000.

[19] V. M. Hawthorne, D. A. Greaves, and D. G. Beevers. Blood pressure in a scottish town. *British Medical Journal*, 3 :600–603, September 1974.

[20] A. W. Cowley Jr. The genetic dissection of essential hypertension. *Nature Reviews. Genetics*, 7 :829–840, November 2006.

[21] J. Kanellis, T. Nakagawa, J. Herrera-Acosta, G. F. Schreiner, B. Rodríguez-Iturbe, and R. J. Johnson. A single pathway for the development of essential hypertension. *Cardiology in Review*, 11(4) :180–196, 2003.

[22] A. C. Guyton, T. G. Coleman, and A. V. Cowley Jr. *et al.* Arterial pressure regulation. overriding dominance of the kidneys in long-term regulation and in hypertension. *American Journal of Medicine*, 52 :584–594, 1972.

[23] P. Hamet, Z. Pausova, V. Adarichev, K. Adaricheva, and J. Tremblay. Hypertension : genes and environment. *Journal of Hypertension*, 16 :397–418, 1998.

[24] B. Lippe. Turner syndrome. *Endocrinology and Metabolism Clinics of North America*, 20 :121–152, 1991.

[25] J. Nielsen and M. Wohlert. Chromosome abnormalities found among 34,910 newborn children : results from a 13-year incidence study in Arhus, Denmark. *Human Genetics*, 87 :81–83, 1991.

[26] M. Elsheikh, B. Casadei, G. S. Conway, and J. A. H. Wass. Hypertension is a major risk factor for aortic root dilatation in women with turner's syndrome. *Clinical Endocrinology*, 54(1) :69–73, January 2001.

[27] W. H. Price, J. F. Clayton, S. Collyer, R. De Mey, and J. Wilson. Mortality ratios, life expectancy, and causes of death in patients with Turner's syndrome. *Journal of Epidemiology and Community Health*, 40 :97–102, 1986.

[28] R. W. Naeraa, C. H. Gravholt, J. Hansen, J. Nielsen, and S. Juul. Mortality in Turner sundrome. In K. Albertsson-Wikland and M. B. Ranke, editors, *Turner Syndrome in a Lifespan Perspective : Research and Clinical Aspects*, page 323. Elsevier, Amsterdam, 1995.

[29] D. W. Hall. Resistant hypertension, secondary hypertension, and hypertensive crises. *Cardiology Clinics*, 20 :281–289, 2002.

[30] B. E. Akpunonu, P. J. Mulrow, and E. A. Hoffman. Secondary hypertension : Evaluation and treatment. *Disease-a-Month*, 42(10) :612–722, October 1996.

[31] T. M. Weinberger. Systemic hypertension. In W. N. Kelly, editor, *Textbook of internal medicine*, pages 268–283. Lippincott, Philadelphia, 1989.

[32] E. Onusko. Diagnosing secondary hypertension. *American Family Physician*, 67(1) :67–74, January 2003.

[33] P. Hamet and J. Tremblay. Genetic determinants of the stress response in cardiovascular disease. *Metabolism*, 51(6) :15–24, June 2002.

[34] V. Ruppert and B. Maisch. Genetics of human hypertension. *Herz*, 28(8) :655–662, 2003.

[35] H. Smulyan, R. G. Asmar, A. Rudnicki, G. M. London, and M. E. Safar. Comparative effects of aging in men and women on the properties of the arterial tree. *Journal of the American College of Cardiology*, 37(5) :1374–1380, 2001.

[36] K. F. Gangar, S. Vyas, M. Whitehead, D. Crook, H. Meire, and Campbell S. Pulsatolity index in internal carotid artery in relation to transdermal œstradiol and time since menopause. *Lancet*, 338 :839–842, 1991.

[37] W. M. Gilbert. Anatomy and physiology of the placenta, fetal membranes and amniotic fluid. In T. R. Moore, R. C. Reiter, Rebar R. W., and V. V. Baker, editors, *Gynecology and Obstetrics*, pages 209–222. Churchill Livingstone, New York, NY, 1993.

[38] M. L. S. Gass. Physiology and pathophysiology at the postmenopausal years. In T. R. Moore, R. C. Reiter, R. W. Rebar, and V. V. Baker, editors, *Gynecology and Obstetrics*, pages 883–898. Churchill Livingstone, New York, NY, 1993.

[39] M. R. Joffres, P. Ghadirian, J. G. Fodor, A. Petrasovits, A. Chockalingam, and P. Hamet. Awareness, Treatment, and Control of Hypertension in Canada. *American Journal of Hypertension*, 10(10) :1097–1101, 1997.

[40] T. Gordon and D. Shurtleff. Section 29 : Means at each examination and interexamination variation of specified characteristics : Framingham Study, exam 1 to exam 10. In W. B. Kannel and T. Gordon, editors, *The Framingham Study : An epidemiological investigation of cardiovascular disease*, pages 74–478. US DHEW (National Institutes of Health), Washington DC, 1977.

[41] P. S. Vokonas, W. B. Kannel, and L. A. Cupples. Epidemiology and risk of hypertension in the elderly : the Framingham Study. *Journal of Hypertension*, 6 :S3–S9, 1988.

[42] B. Chamontin, L. Poggi, T. Lang, J. Ménard, H. Chevalier, H. Gallois, and O. Crémier. Prevalence, treatment, and control of hypertension in the french population - data from a survey on high blood pressure in general pratice, 1994. *American Journal of Hypertension*, 11 :759–762, 1998.

[43] W. Januszewicz, J. Chodakowska, and G. Styczyński. Secondary hypertension in the elderly. *Jornal of Human Hypertension*, 12 :603–606, 1998.

[44] R. S. Vasan, A. Beiser, S. Seshadri, M. G. Larson, W. B. Kannel, R. B. D'Agostino, and D. Levy. Residual lifetime risk for developing hypertension in middle-aged women and men. *JAMA*, 287(8) :1003–1010, February 2002.

[45] J. Staessen, R. Fagard, and A. Amery. The relationship between body weight and blood pressure. *Journal of Human Hypertension*, 2 :207–217, 1988.

[46] H. M. Whyte. Blood pressure and obesity. *Circulation*, 19 :511–516, 1959.

[47] M. Gong and N. Hubner. Molecular genetics of human hypertension. *Clinical Science*, 110 :315–326, 2006.

[48] D. Spiegelman, R. G. Israel, C. Bouchard, and W. C. Willett. Absolute fat mass, percent body fat, and body-fat distribution : which is the real determinant of blood pressure and serum glucose? *The American Journal of Clinical Nutrition*, 55(6) :1033–1044, June 1992.

[49] P. Hamet, E. Mongeau, J. Lambert, F. Bellavance, M. Daignault-Gélinas, M. Ledoux, and L. Whissel-Cambiotti. Interactions among calcium, sodium, and alcohol intake as determinants of blood pressure. *Hypertension*, 17[suppl I] :I–150–I–154, 1991.

[50] L. Dahl. Possible role of salt intake in the development of hypertension. In P. Cottier and K. D. Bock, editors, *Essential Hypertension : An International Symposium*, pages 53–65. Springer-Verlag, Berlin, 1960.

[51] A. Lev-Ran and M. Porta. Salt and hypertension : a phylogenetic perspective. *Diabetes/Metabolism resarch and reviews*, 21 :118–131, 2005.

[52] H. G. Tian, Y. Nan, and R. C. Shao *et al.* Association between blood pressure and dietary intake and urinary excretion of electrolytes in a chinese population. *Journal of Hypertension*, 13 :49–56, 1995.

[53] D. Simmons. Blood pressure, ethnic group and salt intake in belize. *Journal of Epidemiology and Community Health*, 37 :38–42, 1983.

[54] M. J. Klag, J. He, and J Coresh *et al.* The contribution of urinary cations to the blood pressure diffrences associated with migration. *American Journal of Epidemiology*, 142 :295–303, 1995.

[55] M'Buyamba-Kabangu Jr, R. Fagard, and P. Lijnen *et al.* Blood pressure and urinary cations in urban bantu of zaire. *American Journal of Epidemiology*, 124 :957–968, 1986.

[56] K. M. O'Shaughnessy and F. E. Karet. Salt handling and hypertension. *Annual Review of Nutrition*, 26 :343–365, 2006.

[57] F. J. Haddy. Role of dietary salt in hypertension. *Life Sciences*, 79 :1585–1592, 2006.

[58] B. Rodriguez-Iturbe and N. D. Vaziri. Salt-sensitive hypertension—update on novel findings. *Nephrology Dialysis Transplantation*, 22 :992–995, 2007.

[59] F. M. Sacks, L. P. Svetkey, W. M. Vollmer, L. J. Appel, G. A. Bray, D. Harsha, E. Obarzanek, P. R. Conlin, E. R. Miller III, D. G. Simons-Morton, N. Karanja, and P.-H. Lin. Effects on blood pressure of reduced dietary sodium and the dietary approaches to stop hypertension (DASH) diet. *The New England Journal of Medicine*, 344(1) :3–10, January 2001.

[60] F. D. Fuchs, L. E. Chambless, P. K. Whelton, F. J. Nieto, and G. Heiss. Alcohol consumption and the incidence of hypertension : the atherosclerosis risk in communities study. *Hypertension*, 37 :1242–1250, 2001.

[61] A. G. Shaper, G. Wannamethee, and P. Wincup. Alcohol and blood pressure in middle-aged British men. *Journal of Human Hypertension*, 2 :71–78, 1988.

[62] L. J. Beilin and I. B. Puddey. Alcohol and hypertension. *Hypertension*, 47 :1035–1038, 2006.

[63] J. Rehm, R. Room, M. Monteiro, G Gmel, K. Graham, N. Rehn, C. T. Sempos, and D. Jernigan. Alcohol as a risk factor for global burden of disease. *European Addiction Research*, 9 :157–164, 2003.

[64] P. Biron, J. G. Mongeau, and D. Bertrand. Familial aggregation of blood pressure in 558 adopted children. *Canadian Medical Association Journal*, 115(8) :773–774, October 1976.

[65] N. Kato. Genetic analysis in human hypertension. *Hypertension research*, 25 :319–327, 2002.

[66] P. Hamet and O. Šeda. Current status of genome-wide scanning for hypertension. *Current Opinion in Cardiology*, 22 :292–297, 2007.

[67] R. L. Hanson, M. G. Ehm, D. J. Pettitt, M. POrochazka, D. B. Thompson, D. Timberlake, T. Foroud, S. Kobes, L. Baier, D. K. Burns, L. Almasy, J. Blangero, W. T. Garvey, P. H. Bennettl, and W. C. Knowler. Autosomal genomic scan for loci linked to type II diabetes mellitus and body-mass index in pima indians. *American Journal of Human Genetics*, 63 :1130–1138, 1998.

[68] T. A. Kotchen, U. Broeckel, C. E. Grim, P. Hamet, H. Jacob, M. L. Kaldunski, J. M. Kotchen, N. J. Schork, P. J. Tonellato, and A. W. Cowley Jr. Identification of hypertension-related QTLs in African American sib pairs. *Hypertension*, 40 :634–639, 2002.

[69] C. D. Langefeld, L. E. Wagenknecht, J. I. Rotter, A. H. Williams, J. E. Hokanson, M. F. Saad, D. W. Bowden, S. Haffner, J. M. Norris, S. S. Rich, and B. D. Mitchell. Linkage of the metabolic syndrome to 1q23-q31 in Hispanic families : the Insulin Resistance Atherosclerosis Family Study. *Diabetes*, 53 :1170–1174, 2004.

[70] Colin R. Reeves and Jonathan E. Rowe. *Genetic Algorithms - Principles and Perspectives, A Guide to GA Theory.* Kluwer Academic Publishers, 2003.

[71] D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley Publishing Company, inc., 1989.

[72] K. Popper. *Toute vie est résolution de problèmes.* Actes Sud, 1997.

[73] J. L. Ribeiro and P. C. Treleaven. Genetic-algorithm programming environments. *Computer*, 27(6) :28–43, June 1994.

[74] J. H. Holland. *Adaptation in natural and artificial systems : an introductory analysis with applications to biology, control, and artificial intelligence.* Ann Arbor : University of Michigan Press, 1975.

[75] M. Srinivas and L. M. Patnaik. Genetic algorithms : a survey. *Computer*, 27(6) :17–26, June 1994.

[76] C. Notredame and D. G. Higgins. SAGA : sequence alignement by genetic algorithm. *Nucleic Acids Research*, 24(8) :1515–1524, 1996.

[77] C. Notredame, D. G. Higgins, and J. Heringa. T-Coffee : A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302 :205–217, 2000.

[78] C. Zhang and A. K.C. Wong. A genetic algorithm for multiple molecular sequence alignment. *CABIOS*, 13(6) :565–581, 1997.

[79] C. Notredame, E. A. O'Brien, and D. G. Higgins. RAGA : RNA sequence alignment by genetic algorithm. *Nucleic Acids Research*, 25(22) :4570–4580, 1997.

[80] J. T. Pedersen and J. Moult. Genetic algorithms for protein structure prediction. *Current Opinion in Structural Biology*, 6(2) :227–231, April 1996.

[81] S. Schulze-Kremer. Genetic algorithms for protein tertiary structure prediction. In *Applications of Genetic Slgorithms, IEEE Colloqium on*, pages 1–5, March 1994.

[82] L. Jourdan, C. Dhaenens, and E.-G. Talbi. Discovering haplotypes in linkage disequilibrium mapping with an adaptive genetic algorithm. *Lecture Notes in Computer Science*, 2611 :66–75, January 2003.

[83] L. Vermeulen-Jourdan, C. Dhaenens, and E.-G. Talbi. Linkage disequilibrium study with a parallel adaptive GA. *International Journal of Foundations of Computer Science*, 16(2) :241–260, 2005.

[84] J. D. Terwilliger and J. Ott. *Handbook of human genetic linkage.* Johns Hopkins University Press, Baltimore, June 1994.

[85] P. C. Sham and D. Curtis. Monte carlo tests for associations between disease and alleles at highly polymorphic loci. *Annal Human Genetic*, pages 97–105, 1995.

[86] Ø. Braaten, O. K. Rødningen, I. Nordal, and T. P. Leren. The genetic algorithm applied to haplotype data at the LDL receptor locus. *Computer Methods and Programs in Biomedicine*, 61 :1–9, 2000.

[87] F. Pardi, Lewis C. M., and J. C. Whittaker. SNP selection for association studies : maximizing power across SNP choice and study size. *Annals of Human Genetics*, 69 :733–746, 2005.

[88] M. C. Byng, J. C. Whittaker, A. R. Cuthbert, C. G. Mathew, and C. M. Lewis. Snp subset selection for genetic association studies. *Ann Hum Genet*, 67 :543–556, 2003.

[89] J. C. Barrett, B. Fry, J. Maller, and M. J. Daly. Haploview : analysis and visualization of ld and haplotype maps. *Bioinformatics*, 21(2) :263–265, January 2005.

[90] N. Laird, S. Horvath, and X. Xu. Implementing a unified approach to family based tests of association. *Genet Epidemiol*, 19(Suppl 1) :S36–S42, 2000.

[91] G. R. Abecasis, S. S. Cherny, W. O. Cookson, and L. R. Cardon. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet*, 30 :97–101, 2002.

[92] R. S. Spielman, R. E. McGinnis, and W. J. Ewens. Transmission test for linkage disequilibrium : the insulin gene region and insulin-dependent diabetes mellitus (iddm). *Am J Hum Genet*, 52(3) :506–516, Mar 1993.

[93] R. S. Spielman and W. J. Ewens. The TDT and other family-based tests for linkage disequilibrium and association. *American Journal of Human Genetics*, 59 :983–989, 1996.

[94] E. R. Martin, S. A. Monks, L. L. Warren, and N. L. Kaplan. A test for linkage and association in general pedigrees : the pedigree disequilibrium test. *Am J Hum Genet*, 67(1) :146–154, Jul 2000.

[95] D. J. Schaid. Evaluating associations of haplotypes with traits. *Genetic Epidemiology*, 27 :348–364, 2004.

[96] D. Schaid, C. Rowland, R. Jacobson, and G. Poland. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *American Journal of Human Genetics*, 70 :425–434, 2002.

[97] S. Lake, H. Lyon, K. Tantisira, E. Silverman, S. Weiss, N. Laird, and D. Schaid. Estimation and tests of haplotype-environmental interaction when linkage phase is ambiguous. *Human Heredity*, 55 :56–65, 2003.

[98] M. Labuda, D. Labuda, M. Korab-Laskowska, D. Cole, E. Zietkiewicz, J. Weissenbach, E. Popowska, E. Pronicka, A. Root, and F. Glorieux. Linkage disequilibrium analysis in young populations : Pseudo-vitamin d-deficiency rickets and the founder effect in french canadians. *American Journal of Human Genetics*, 59 :633–643, 1996.

[99] D. Labuda, E. Zietkiewicz, and M. Labuda. The genetic clock and the age of the founder effect in growing populations : A lesson from French Canadians and Ashkenazim. *American Journal of Human Genetics*, 61 :768–771, 1997.

[100] M. De Braekeleer, V. Lamarre, C. R. Scriver, J. Larochelle, and G. Bouchard. Fertility in couples heterozygous for the tyrosinemia gene in saguenay lac-st-jean. *Genet Couns*, 1 :259–264, January 1990.

[101] Scriver C. R. Human genetics : lessons from Quebec populations. *Annual review of genomics and human genetics*, 2 :69–101, January 2001.

[102] T. A. Kotchen, J. M. Kotchen, C. E. Grim, V. George, M. L. Kaldunski, A. W. Cowley, P. Hamet, and T. H. Chelius. Genetic determinants of hypertension : Identification of candidatre phenotypes. *Hypertension*, 36(1) :7–13, July 2000.

[103] Affymetrix. Data Sheet – GeneChip® Human Mapping 100K Set. `http://www.affymetrix.com/support/technical/datasheets/100k_datasheet.pdf`, Mar 2008.

[104] H. Zhao, R. Pfeiffer, and M. H. Gail. Haplotype analysis in population genetics and association studies. *Pharmacogenomics*, 4(2) :171–178, March 2003.

[105] Z. Michalewicz. *Genetic Algorithms + Data Structures = Evolution Programs.* Springer, third, revised and extended edition, 1996.

[106] J. J. Grefenstette. Optimization of control parameters for genetic algorithms. *IEEE-SMC*, 16 :122–128, 1986.

[107] J. D. Schaffer, R. A. Caruana, L. J. Eshelman, and R. Das. A study of control parameters affecting online performance of genetic algorithms for function optimization. In J.D. Schaffer, editor, *Proceedings of 3rd International Conference on Genetic Algorithms*, pages 51–60, San Mateo, CA, 1989. Morgan Kaufmann.

[108] D. E. Goldberg, K. Deb, and J. H. Clark. Genetic algorithm, noise, and the sizing of populations. *Complex Systems*, 6 :333–362, 1992.

[109] C. B. Congdon. *A comparison of genetic algorithms and other machine learning systems on a complex classification task from common disease research.* PhD thesis, University of Michigan, 1995.

[110] J. E. Baker. Reducing bias and inefficiency in the selection algorithm. In J. J. Grefenstette, editor, *Proceedings of the 2nd International Conference on Genetic Algorithms*, pages 14–21, Hillsdale, New Jersey, 1987. Lawrence Erlbaum Associates.

[111] P. J. B. Hancock. An empirical comparison of selection methods in evolutionary algorithms. In T. C. Fogarty, editor, *Evolutionary Computing : AISB Workshop, Leeds, UK*, pages 80–94, Berlin, April 1994. Springer-Verlag.

[112] P. J. B. Hancock. Selection methods for evolutionary algorithms. In L. Chambers, editor, *Practical Handbook of Genetic Algorithms : New Frontiers*, volume II, pages 67–92. CRC Press, Boca Raton, FL, 1996.

[113] M. A. Pawlowsky. Crossover operators. In L. Chambers, editor, *Practical Handbook of Genetic Algorithms : Applications*, volume I, chapter 4, pages 101–114. CRC Press, Boca Raton, FL, 1995.

[114] G. Syswerda. Uniform crossover in genetic algorithms. In *Proceedings of the Third International Conference on Genetic Algorithms*, pages 2–9, San Francisco, CA, USA, 1989. Morgan Kaufmann Publishers Inc.

[115] J.-Y. Potvin. Genetic algorithms for the traveling salesman problem. *Annals of Operations Research*, 63 :339–370, 1996.

[116] J. J. Grefenstette, R. Gopal, B. Rosmaita, and D. Van Gucht. Genetic algorithm for the tsp. In J. J. Grefenstette, editor, *Proceedings of the First Internation Conference on Genetic Algorithms*, pages 160–168, Hillsdale, NJ, 1985. Lawrence Erlbaum Associates.

[117] D. E. Goldberg and R. Lingle. Alleles, Loci, and the TSP. In J.J. Grefenstette, editor, *Proceedings of the First Internation Conference on Genetic Algorithms*, pages 154–159, Hillsdale, NJ, 1985. Lawrence Erlbaum Associates.

[118] L. Davis. Applying adaptive algorithms to epistatic domains. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 162–164, 1985.

[119] I. M. Oliver, D. J. Smith, and J. R. C. Holland. A study of permutation crossover operators on the traveling salesman problem. In J. J. Grefenstette, editor,

Proceedings of the Second International Conference on Genetic Algorithms, pages 224–230, Hillsdale, NJ, 1987. Lawrence Erlbaum Associates.

[120] L. Davis. *Handbook of Genetic Algorithms.* an Nostrand Reinhold, New York, 1991.

[121] K. A. De Jong. *An analysis of the behavior of a class of genetic adaptive systems.* PhD thesis, University of Michigan, 1975.

[122] Matthew S. Software for association mapping, genotype imputation, haplotype estimation etc. `http://stephenslab.uchicago.edu/software.html`, Mar 2008.

[123] M. Stephens and P Donnelly. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics*, 73 :1162–1169, 2003.

[124] M. Stephens, N. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, 68 :978–989, 2001.

[125] M. Stephens and P. Scheet. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American Journal of Human Genetics*, 76 :449–462, 2005.

[126] A. Singhal. Modern information retrieval : A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4) :35–43, 2001.

[127] R. R. Korfhage. *Information Storage and Retrieval.* Wiley, 1997.

[128] D. G. Altman and J. M. Bland. Diagnostic tests. 1 : Sensitivity and specificity. *British Medical Journal*, 308(6943) :1552, 1994.

[129] W. Nystad, H. E. Meyer, P. Nafstad, A. Tverdal, and A. Engeland. Body mass index in relation to adult asthma among 135,000 Norwegian men and women. *American Journal of Epidemiology*, 160(10) :969–976, 2004.

[130] P. H. Westfall and S. S. Young. *Resampling-based multiple testing : examples and methods for p-value adjustment.* Jonh Wiley & Sons, New York, New York, 1993.

[131] Red Hat Inc. Fedora Project. `http://fedoraproject.org/`, Dec 2007.

[132] Python Software Fondation. Python Programming Language – Official Website. `http://www.python.org/`, Dec 2007.

[133] The R Project for Statistical Computing. `http://www.r-project.org/`, Dec 2007.

[134] J. M. Chambers and T. J. Hastie. *Statistical Models in S.* Wadsworth & Brooks/Cole, 1992.

[135] P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalized Linear Models : A Roughness Penalty Approach.* Chapman and Hall, 1994.

[136] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models.* Chapman and Hall, 1990.

[137] M. C. Ng, Y. Wang, W. Y. So, S. Cheng, S. Visvikis, R. Y. Zee, A. Fernandez-Cruz, K. Lindpaintner, and J. C. Chan. Ethnic differences in the linkage disequilibrium and distribution of single-nucleotide polymorphisms in 35 candidate genes for cardiovascular diseases. *Genomics*, 83(4) :559–565, April 2004.

[138] X. Zhu, D. Yan, R. S. Cooper, A. Luke, M. A. Ikeda, Y. P. Chang, A. Weder, and A. Chakravarti. Linkage disequilibrium and haplotype diversity in the genes of the renin-angiotensin system : Findings from the family blood pressure program. *Genome Res*, 13(2) :173–181, February 2003.

[139] D. E. Reich, M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, and E. S. Lander. Linkage disequilibrium in the human genome. *Nature*, 411(6834) :199–204, May 2001.

[140] E. Cantú-Paz. A survey of parallel genetic algorithms. *Calculateurs parallèles, Réseaux et Systèmes répartis*, 10(2) :141–171, 1998.

[141] P. Adamidis. Review of parallel genetic algorithms bibliography. Technical report, Aristote University of Thessaloniki, Thessaloniki, Greece, 1994.

[142] V. S. Gordon and D. Whitley. Serial and parallele genetic algorithms as function optimizers. In S. Forrest, editor, *Proceedings of the Fifth International Conference on Genetic Algorithms*, pages 117–183, San Mateo, CA, 1993. Morgan Kaufmann.

[143] S.-C. Lin, W. Punch, and E. Goodman. Coarse-grain parallel genetic algorithms : Categorization and new approach. In *Sixth IEEE Symposium on Parallel and Distributed Processing*, Los Alamitos, CA, October 1994. IEEE Computer Society Press.

[144] D. Abramson and J. Abela. A parallel genetic algorithm for solving the school timetabling problem. In *Proceedings of the Tifteenth Australian Computer Science Conference*, volume 14, pages 1–11, 1992.

[145] Branke J., H. C. Andersen, and H. Schmeck. Parallelising global selection in evolutionary algorithms. Submitted to journal of Parallel and Distributed Computing, January 1997.

[146] T. C. Fogarty and R. Huang. Implementing the genetic algorithm on transputer based parallel processing systems. *Oarakkek Oribken Sikvubg from Nature*, pages 145–149, 1991.

[147] R. Hauser and R. Männer. Implementation of standard genetic algorithm on MIMD machines. In Y. Davidor, H.-P. Chwefel, and R. Männer, editors, *Parallel Problem Solving from Nature, PPSN III*, pages 504–513, Berlin, 1994. Springer-Verlag.

[148] B. Shapiro and J. Navetta. A massivelu parallel genetic algorithm for RNA secondary structure prediction. *The Journal of Supercomputing*, 8 :195–207, 1994.

[149] H. C. Braun. On solving travelling salesman problems by genetic algorithms. In H.-P. Schwefel and R. Männer, editors, *Parallel Problem Solving from Nature*, pages 129–133. Springer-Verlag, Berlin, 1990.

[150] E. Cantú-Paz and D. E. Goldberg. Modeling idealized bounding cases of parallel genetic algorithms. In J. Koza, K. Deb, M. Dorigo, D. Fogel, M. Garzon, H. Iba, and R. Riolo, editors, *Genetic Programming 1997 : Proceedings of the Second Annual Conference*, San Francisco, CA, 1997. Morgan Kaufmann.

[151] M. Munetomo, Y. Takai, and Y. Sato. An efficient migration scheme for subpopulation-based asynchronously parallel genetic algorithms. In S. Forrest, editor, *Proceedings of the Fifth International Conference on Genetic Algorithms*, page 649, San Mateo, CA, 1993. Morgan-Kaufmann.

[152] F. Gruau. *Neural network synthesis using cellular encoding and the genetic algorithm.* PhD thesis, Université Claude-Bernard-Lyon I, 1994.

[153] R. Bianchini and C. M. Brown. Parallel genetic algorithms on distributed-memory architectures. In S. Atkins and A.S. Wagner, editors, *Transputer Research and Applications*, pages 67–82. IOS Press, Amsterdam, 1993.

Annexe I
Hypertension

Tableau XII.1: Symptômes suggérant une hypertension secondaire. Tableau représentant les différentes causes de l'hypertension secondaire (tiré de [32]).

Findings	*Disorder suspected*	*Further diagnostic studies*
Snoring, daytime somnolence, obesity	Obstructive sleep apnea	Sleep study
Hypernatremia, hypokalemia	Aldosteronism	Ratio of plasma aldosterone to plasma renin activity, CT scan of adrenal glands
Renal insufficiency, atherosclerotic cardiovascular disease, edema, elevated blood urea nitrogen and creatinine levels, proteinuria	Renal parenchymal disease	Creatinine clearance, renal ultrasonography
Systolic/diastolic abdominal bruit	Renovascular disease	Magnetic resonance angiography, captopril (Capoten)-augmented radioisotopic renography, renal arteriography
Use of sympathomimetics, perioperative setting, acute stress, tachycardia	Excess catecholamines	Confirm patient is normotensive in absence of high catecholamines.
Decreased or delayed femoral pulses; abnormal chest radiograph	Coarctation of aorta	Doppler or CT imaging of aorta
Weight gain, fatigue, weakness, hirsutism, amenorrhea, moon facies, dorsal hump, purple striae, truncal obesity, hypokalemia	Cushing's syndrome	Dexamethasone-suppression test
Use of drug in table 2	Drug side effect	Trial off drug, if possible
High salt intake, excessive alcohol intake, obesity	Diet side effects	Trial of dietary modification
Erythropoietin use in renal disease, polycythemia in COPD	Erythropoietin side effect	Trial off drug, if possible
Paroxysmal hypertension, headaches, diaphoresis, palpitations, tachycardia	Pheochromocytoma	Urinary catecholamine metabolites (vanillylmandelic acid, metanephrines, normetanephrines) Plasma free metanephrines
Fatigue, weight loss, hair loss, diastolic hypertension, muscle weakness	Hypothyroidism	TSH levels
Heat intolerance, weight loss, palpitations, systolic hypertension, exophthalmos tremor, tachycardia	Hyperthyroidism	TSH levels
Kidney stones, osteoporosis, depression, lethargy, muscle weakness	Hyperparathyroidism	Serum calcium, parathyroid hormone levels
Headaches, fatigue, visual problems, enlargement of hands, feet, tongue	Acromegaly	Growth hormone level

CT = computed tomography; COPD = chronic obstructive pulmonary disease; TSH = thyroid-stimulating hormone.

Tableau XII.II: Formes d'hypertension mendélienne. Tableau représentant les différentes causes de l'hypertension mendélienne. Adapté d'Hamet *et al.*[23].

Syndrome	Inheritance	Chromosome	Gene
Defects in steroid metabolism			
Glucocorticoid-remediable hyperaldosteronism	ADI	8q22	CYP11B1/CYP11B2
Male pseudo-hermaphroditism	ARI	10q24.3	CYP17
Female pseudo-hermaphroditism	ARI	8q22	CYP11B1
Apparent mineralcorticoid excess	ARI	16q22	HSD11B2
Defects in ion transport			
Liddle's syndrome	ADI	16p12-p13	SCNN1B
			SCNN1G
Gordon's syndrome	ADI	1q31-q42	PHAIIA (locus)
		12p13.3	WNK1
		17q21	WNK4
Overproduction of catecholamines (pheochromocytoma)			
Isolated pheochromocytoma	ADI	1p	?
Multiple endocrine neoplasia, type II A	ADI	10q11.2	RET proto-oncogene
Multiple endocrine neoplasia, type II B	ADI	10q11.2	RET proto-oncogene
von Huppel-Lindau syndrome	ADI	3p26-25	VHL tumour suppressor gene
Neurofibromatosis, type I	ADI	17q11.2	NF1 gene
Other			
Hypertension and brachydactyly	ADI	12p12.2-p11.2	HTNB

ADI, autosomal dominant mode of inheritance ; ARI, autosomal recessive mode of inheritance.

Tableau XII.III: Risque résiduel de l'hypertension en fonction de l'âge*. Tableau représentant le risque d'être hypertendu selon l'âge et le sexe (Figure tirée de [44]).

	Risk for Hypertension, % (95% Confidence Interval)			
	Women, Age, y		Men, Age, y	
Time, y	55 (n = 709)	65 (n = 549)	55 (n = 589)	65 (n = 438)
10	52 (46-58)	64 (60-69)	56 (49-63)	72 (67-78)
15	72 (68-76)	81 (77-84)	78 (74-82)	85 (81-89)
20	83 (80-86)	89 (86-92)	88 (85-91)	90 (87-93)
25	91 (89-93)	...	93 (91-95)	...

*For 55-year-old subjects, the risk for developing hypertension over 25 years represents their lifetime risk. For 65-year-old subjects, the risk for developing hypertension over 20 years indicates their lifetime risk. Ellipses indicate not applicable.

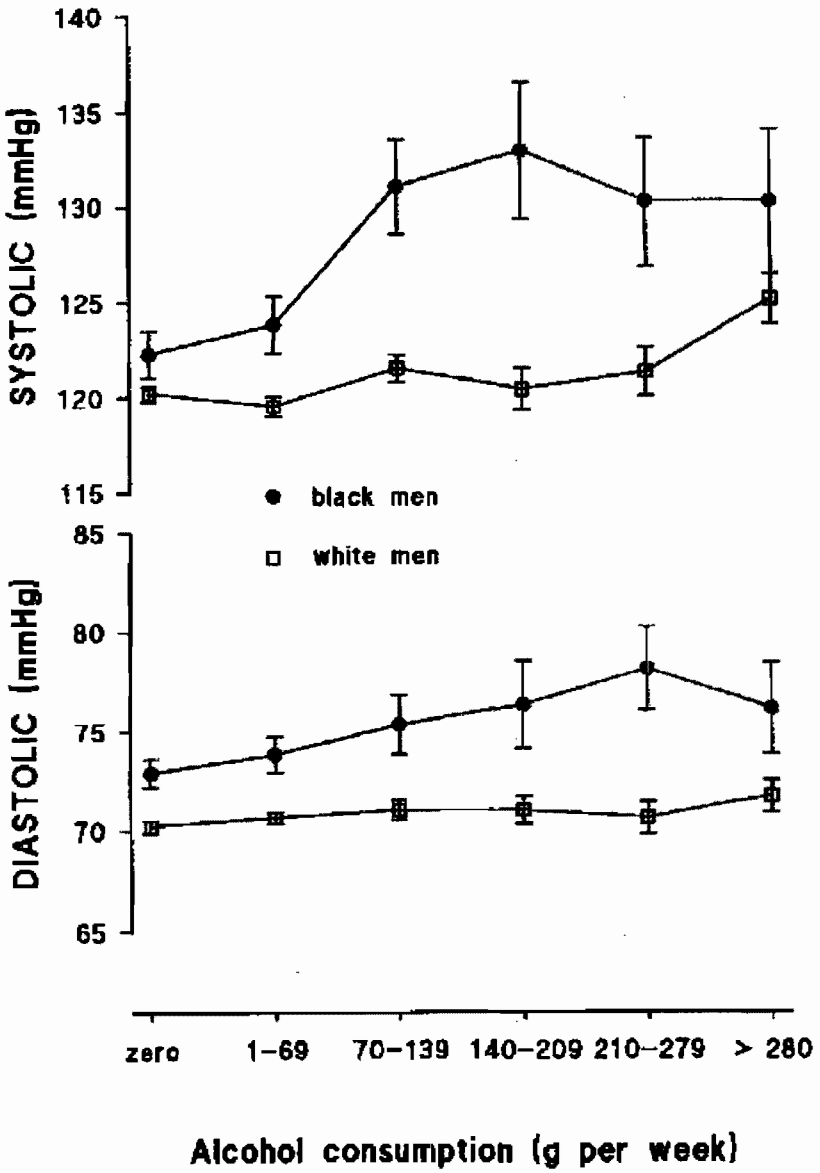


Figure 12.1: Variation de la pression artérielle en fonction de la consommation d'alcool. Moyenne de la SBP et DBP selon la consommation d'alcool (ajustée pour l'âge, le BMI, l'éducation, l'activité physique et le diabète). Figure tirée de [60]

Annexe II
Algorithmes génétiques parallèles

Optimisation des algorithmes génétiques

Les algorithmes génétiques sont aptes à découvrir des solutions aux problèmes avec des temps raisonnables (comparativement aux autres techniques de recherche). Par contre, plus les problèmes deviennent complexes, plus le temps d'exécution augmente radicalement. Plusieurs recherches ont été faites afin de diminuer le temps d'exécution de ces heuristiques, dont la parallélisation [140, 141, 142, 143].

Il y a quatre grandes classes d'algorithmes génétiques parallèles : les AGs globaux à population unique utilisant la parallélisation maître-esclave, les AGs à population unique « fine-grained », les AGs multidème (multiple population) et les AGs parallèles hiérarchiques. Ces derniers mélangent les caractéristiques des trois premières classes mentionnées. Dans ce présent ouvrage, les quatre différentes classes d'algorithmes génétiques parallèles sont expliquées sommairement. Plus de détails sur l'implantation de ces algorithmes sont présentés dans [140].

Parallélisation maître-esclave

Dans les AGs utilisant la parallélisation maître-esclave, il existe une population unique identique à la population d'un algorithme génétique simple. La parallélisation se fait au niveau du calcul de la mesure de performance des individus qui se retrouve distribuée sur plusieurs processeurs (voir Figure 12.2 (A) de la page xix). Puisque dans ce type d'AG parallèle la sélection et le croisement considèrent la totalité de la population, il se nomme aussi algorithme génétique global [140].

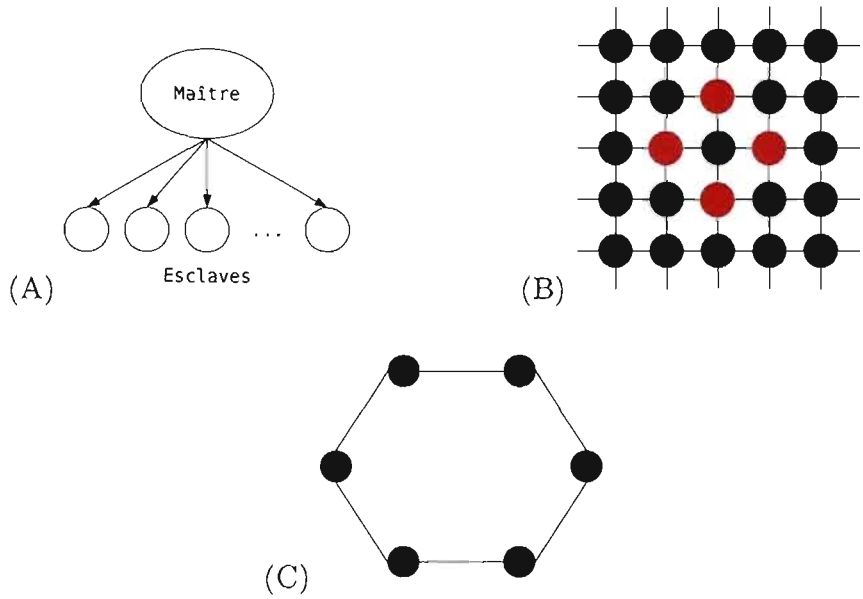


Figure 12.2: Trois types de parallélisation. (A) La parallélisation maître-escalve. Schéma représentant le fonctionnement de l'algorithme génétique utilisant une parallélisation maître-esclave. Le maître enregistre la population et exécute les différents opérateurs d'un AG. Il distribue ensuite les individus aux esclaves afin que ceux-ci calculent la performance de chacun [140]. (B) La parallélisation « Fine-Grained ». Chaque cercle représente un processeur contenant un seul individu. Le voisinage dans lequel l'individu bleu peut se reproduire et se comparer est représenté par les cercles rouges [140]. (C) La parallélisation multidème. Schéma représentant le fonctionnement de l'algorithme génétique utilisant une parallélisation multidème. Chaque cercle représente un processeur. Chaque processeur est responsable d'une sous-population. L'opérateur de migration (représenté par une ligne) est responsable d'échanger des individus d'une sous-population à une autre [140].

Comme mentionné précédemment, la parallélisation se fait au niveau du calcul de la performance des individus. Ceci facilite l'implantation, puisque la performance est indépendante du reste de la population et aucune communication n'est nécessaire entre les processeurs. En fait, la seule communication se fait au début de la phase d'évaluation, lorsque l'esclave reçoit un sous-ensemble de la population à calculer et à la fin de cette même phase lorsque les esclaves retournent la performance de chaque individu dans la sous-population. Plusieurs utilisations de ce type d'algorithme génétique sur des problèmes différents ont été réalisées avec succès, dont [144, 145, 146, 147].

Parallélisation « Fine-Grained »

Un algorithme génétique utilisant une parallélisation « fine-grained » contient une seule population. Celle-ci est structurée de façon à ce qu'il n'y ait qu'un individu par processeur. Ainsi, la sélection et le croisement sont restreints à un petit voisinage d'individus (voir Figure 12.2 (B) de la page xix). En d'autres mots, un individu ne peut se reproduire et compétitionner qu'avec ses voisins. Puisque les voisinages s'entrecoupent, une bonne solution peut tout de même se propager, quoique lentement, dans la population en son entier. Ceci empêche donc la convergence prématurée des solutions vers un optimum local. Une application intéressante d'un algorithme génétique utilisant la parallélisation « fine-grained » est présentée par Shapiro *et al.* [148] où le problème était de prédire la structure secondaire d'ARN.

Parallélisation multidème

Les algorithmes génétiques multidèmes (ou multiples populations) sont beaucoup plus sophistiqués puisqu'ils doivent gérer plusieurs sous-populations où certains individus doivent être échangés occasionnellement (voir Figure 12.2 (C) de la page xix). Cet échange d'individus est nommé « migration » et doit être contrôlé par plusieurs paramètres [140]. Ce type d'algorithme apporte plusieurs changements au niveau des opérateurs de l'algorithme génétique simple.

Dans la majorité des algorithmes multidèmes, la migration est synchrone, ce qui signifie que la migration a lieu dans toutes les sous-populations dans un laps de temps prédéterminé. Dans le cas d'une migration asynchrone, chaque dème propagera son meilleur individu seulement après qu'un évènement précis ne soit apparu. Plusieurs auteurs ont présenté un algorithme où la migration se fait uniquement lorsque toutes les sous-populations ont convergé complètement [149, 150, 151]. Le but de ce type de migration est de restaurer la diversité dans les sous-populations.

Algorithmes génétiques hiérarchiques

Cette catégorie d'algorithmes génétiques parallèles unit le multidème à la parallélisation maître-esclave ou « fine-grained ». Selon [140], ce type d'algorithme combine tous les bénéfices de la parallélisation utilisée et assure de meilleures performances. Lorsque deux méthodes de parallélisation sont combinées, elles forment une hiérarchie. Au plus haut niveau de l'algorithme se retrouve la multipopulation. Certains hybrides adoptent une parallélisation « fine-grained » à un niveau plus bas (voir Figure 12.3 (A) de la page xxi) [152]. D'autres, comme Bianchini *et al.* [153], utilisent une parallélisation maître-esclave sur chaque dème (voir Figure 12.3 (B) de la page xxi). Finalement, un autre modèle présente un multidème aux deux niveaux (voir Figure 12.3 (C) de la page xxi).

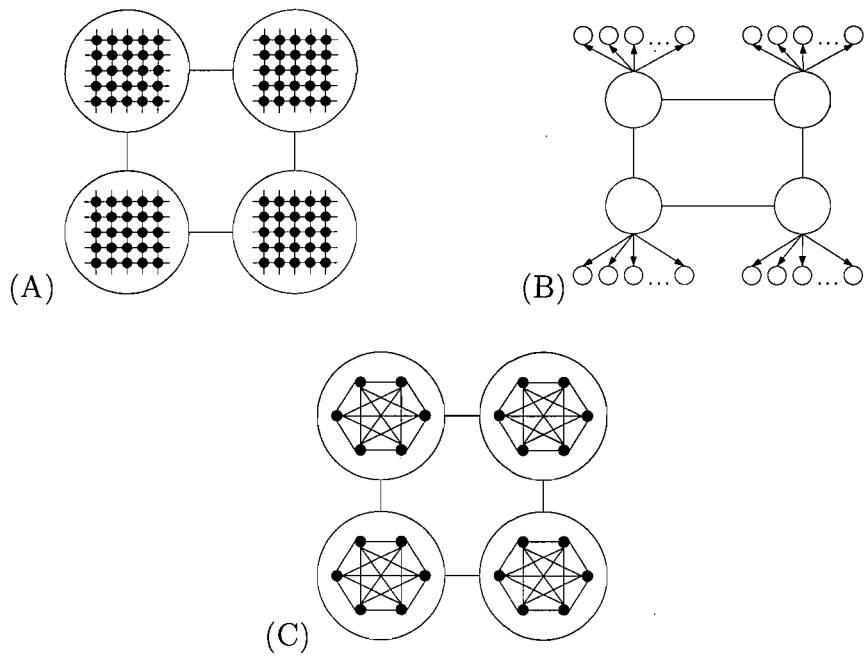


Figure 12.3: Parallélisation hiérarchique. (A) Parallélisation hiérarchique combinant le multidème (niveau le plus haut) et le « fine-grained » à un niveau plus bas [140]. (B) Parallélisation hiérarchique combinant le multidème (niveau le plus haut) et le maître-esclave à un niveau plus bas [140]. (C) Parallélisation hiérarchique combinant deux multidèmes (niveau le plus haut et niveau le plus bas) [140].