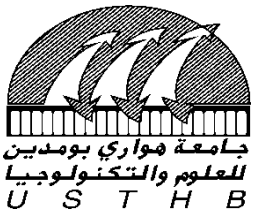
N° d'ordre : 40/2016-C /IN

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université des Sciences et de la Technologie Houari Boumediène

Faculté d'Electronique et d'Informatique



THESE

Présentée pour l'obtention du **diplôme** de **DOCTORAT 3eme Cycle**

En : INFORMATIQUE

Spécialité : Intelligence Artificielle

Par : **Aicha BOUTORH**

Sujet

Génération par techniques hybrides de data mining de relations génomiques fonctionnelles pour le diagnostic de maladies.

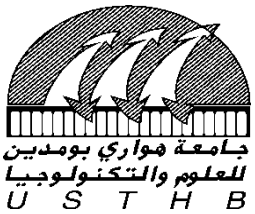
Soutenue publiquement, le **Lundi 11 Juillet 2016,** devant le jury composé de :

Mr. Ahmed Riadh BABA ALI	Professeur	à l'USTHB/FEI	Président
Mr. Ahmed GUESSOUM	Professeur	à l'USTHB/FEI	Directeur de thèse
Mr. Abdelouahab MOUSSAOUI	Professeur	à l'UNIV SETIF	Examinateur
Mr. Mourad DAOUDI	MC/A	à l'USTHB/FEI	Examinateur
Mr. Mohamed Chaouki BATOUCHE	Professeur	à l'UNIV CONST	Examinateur
Mme. Chahrazed IGHILAZA	MA/A	à l'USTHB/FEI	Invitée

Order: 40/2016-C/IN

DEMOCRATIC AND POPULAR REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
University of Science and Technology Houari Boumediene

Faculty of Electronics and Informatics
Department of Informatics



THESIS

Submitted in partial fulfillment for the degree of **Doctor of Philosophy**

In: Computer Science

Specialty: Artificial Intelligence

By: **Aicha BOUTORH**

Subject

Hybrid Data Mining Techniques for The Generation of Functional Genomics Relationships for Disease Diagnosis

Publicly defended, on **Monday July 11th, 2016,** before the jury composed of:

Mr. Ahmed Riadh BABA ALI	Professor	USTHB/FEI	President
Mr. Ahmed GUESSOUM	Professor	USTHB/FEI	Supervisor
Mr. Abdelouahab MOUSSAOUI	Professor	UNIV SETIF	Examiner
Mr. Mourad DAOUDI	MC/A	USTHB/FEI	Examiner
Mr. Mohamed Chaouki BATOUCHE	Professor	UNIV CONST	Examiner
Mrs. Chahrazed IGHILAZA	MA/A	USTHB/FEI	Invited

DEDICATION

To My Dear Beloved Parents ***AbdAllah & Malika*** …

The reason of what I become today.

Who instilled in me the virtues of perseverance and commitment
and relentlessly encouraged me to strive for excellence.

Who dedicated themselves over the years for my education and intellectual development, as well as for my sister *Asma*, and my two brothers *Abderrahim* and *Abdelwadoud,* whom I wish the brightest future.

This thesis is dedicated to you, my father and mother, for all that you are.

Thank you for being such great parents.
Thank you for your endless support and encouragement.
Thank you for being always there for me, for listening to me, for being patient with me the best you can.

I am grateful for all the sacrifice, the care and the unconditional love,
I am grateful for being the source of my motivation and strength during moments of despair and discouragement,
I am grateful for your advice and all your prayers.

Without you this would never have happened.
All I have and I will accomplish are possible only thanks to you.

I will be grateful all the days of my life for having had the honor to be your daughter.

Dad, Mom … You always picked me up on time,
and encouraged me to go forward every adventure.

I love you in totality for eternity, May Allah bless you
and give me the strength to make you always proud of me

Your Eldest Daughter: ***AICHA***

Abstract

The world of biological research is experiencing unprecedented enthusiasm for Bioinformatics; a domain that uses computer technology, including intelligent techniques, for the analysis of large volumes of data available on the human genome. The completed human genome sequence, that encodes the genetic instructions for human physiology, was characterized as a *"tremendous foundation on which to build the science and medicine of the 21st century"* (*Dr.Bruce Alberts*).

Common diseases are caused by a combination of multiple genetic and environmental factors. Understanding the relationship between individual DNA variation and susceptibility to disease is a major objective of human genetics and genetic epidemiology that lead to better prevention, diagnosis and treatment of diseases. Functional Genomics aim to understand the functional aspect of a biological system by understanding the role of genes and the complex relationship between Genotype and Phenotype with Genetic Association Studies that can point scientists in the direction of a treatment.

On the other hand, the development of a new drug is a costly and time-consuming process. One strategy that aims to address this issue is Drug Repositioning, which is known as the use of already known drug for a new application,that can help to reduce the time and cost of drug discovery. One of the most followed direction by these computational techniques is the identification of connections between drug and disease based on the connections between biological entities.

The rapid development of the Biology and Genomic area have generated a surprising amount of data which lead to great challenges in Computational Biology and made an accurate analysis of existing methods of calculation impossible. The development of new techniques specifically for biological data is essential to realize the potential of functional genomics, hence, the interest in Intelligent techniques based on machine learning and data mining. The results of data mining can be referred to as knowledge in the form of rules. Association Rule Mining (ARM) is one of the popular data mining methods widely used in different areas since rules provide important, concise and easily understood information, and solve problems such as classification and association. Recently,

hybrid intelligent systems are becoming popular due to their ability to handle many real world complex problems. The awareness in the academic communities that the hardest problem in Artificial Intelligence can be solved and treated by combining and integrating different approaches, led to a wide investment in the Hybrid Intelligent Systems.

This thesis aims to develop an Hybrid Intelligent paradigm on Genomic data for finding interesting relationships that are able to explain the function of genes to address problems related to Genetic Diseases and Drug Discovery. Several contributions have been made within this thesis in the field of Bioinformatics by proposing rule mining methods using Grammatical Evolution, called GEARM, and creating hybrid models in order to solve two important issues in Computational Biology that are Genetic Diseases relationships and Drug Repositioning.

Firstly, hybrid technique is based on combined GEARM with ANN, one of the powerful machine learning techniques, and GA in the same model. GA-NN-GEARM is capable to deal with the complex genetic databases and find interesting genetic interaction based on SNPs that contribute to the development of the diseases. The proposed model deals with the "curse of dimensionality problem" by providing a specific way for feature selection which takes into consideration the relation among the variables rather than dealing with each of them separately. GA-NN-GEARM had a high performance in dealing with Genetic Diseases relationships.

Secondly, as another key contribution of the thesis, we aim to find hidden associations between Drug and Disease based on their biological entities. GEARM is used for Drug Repositioning (DR). A set of ARs is extracted between Genes and Pathways targets Drug and Disease in order to find new pairs of drug-disease. Two kinds of ARs between Drug and Disease are considered and differ in the antecedent and the consequent of the rule. GEARM was able to discover several pairs of (Drug, Disease), where some are reported in the literature and some others are new, unknown pairs which can be a new indication for drugs.

In every proposed paradigms, innovative approaches have been developed to perform a specific task in the biological contexts. The intelligent computational approaches presented here, that are based on Association Rule Mining, Artificial Neural Network and Evolutionary Algorithms, are either novel in the conception of the algorithm, or are innovatively tailored and combined to be used for specific biomedical and genomic problem areas, and their performance was compared to several techniques that have been used in the literature.

ملخص

يعرف عالم الأبحاث البيولوجية حماسًا غير مسبق للمعلوماتية الحيوية، المجال الذي يستخدم تكنولوجيا الكمبيوتر بما في ذلك التقنيات الذكية، لتحليل الكميات الكبيرة من البيانات المتاحة عن الجينوم البشري. اكتمال تسلسل جينوم الإنسان ، الذي يشفر التعليمات الجينية لعلم وظائف الأعضاء البشرية، وُصفت بأنها *"مؤسسة هائلة لبناء العلم والطب في القرن ال*21

الأمراض الشائعة تحدث بسبب مزيج من عدد من العوامل الوراثية والبيئية . هدف مهم ورئيسي لعلم الوراثة البشرية وعلم الأوبئة الوراثية هو فهم العلاقة الانتقالية بين وجود فروق فردية في الحمض النووي والاختلاف في الحساسية للمرض، والتي تتيح تحسين الوقاية والتشخيص والعلاج من الأمراض. علم الجينوم الوظيفي يهدف لفهم الجانب الوظيفي للنظام البيولوجي من خلال فهم دور الجينات والعلاقة المعقدة بين النمط الجيني و النمط الظاهري الذي يمكن أن يوجه العلماء في اتجاه العلاج.

من ناحية أخرى، تطوير دواء جديد هو عملية مكلفة وتستغرق قتا طويلا. احدى الاستراتيجيات التي تهدف إلى معالجة هذه المسألة تُعرف بإعادة تموضع الدواء، هذه الاستراتيجية تتمثل في استخدام الأدوية المعروفة الفعل لتطبيق جديد، والتي يمكن أن تساعد على تقليل الوقت وتكلفة اكتشاف أدوية جديدة. احد أهم الاتجاهات المتبعة من طرف التقنيات الحاسوبية هي تحديد روابط جديدة بين المرض والدواء اعتمادا على روابط بين العناصر البيولوجية.

التطور السريع للبيولوجيا وعالم الجينوم أنتج كميات هائلة من البيانات أدّت إلى تحديات كبيرة في علم الأحياء، وجعلت التحليل الدقيق للطرق القائمة على الحساب مستحيل. تطوير تقنيات جديدة لتحليل البيانات الضخمة أصبح ضروري لتحقيق فعاليات الجينوم الوظيفي، ومنه أهمية التقنيات الذكية على أساس التعلم الآلي والتنقيب على البيانات. نتائج التنقيب يمكن أن تكون معلومات في شكل قواعد. استخراج قواعد الروابط بين البيانات، هي تقنية من أهم تقنيات التنقيب في البيانات، تستعمل على نطاق واسع في عدة مجالات نظرا لكونها تعبر عن معلومة مهمه بطريقة مختصرة ومفهومة، ولها القدرة في ايجاد حلول لمسائل متعددة مثل التصنيف والترابط.

مؤخرا، الأنظمة الذكية الهجينة أصبحت ذو شعبية واسعة بسبب قدرتها على التعامل مع العديد من المشاكل المعقدة في العالم الحقيقي. ازدياد الوعي في الأوساط الأكاديمية أن المشاكل الأصعب في الذكاء الاصطناعي يمكن أن تحل وتعالج من خلال جمع ودمج المناهج المختلفة، أدى إلى استثمار واسع في الأنظمة الذكية الهجينة.

تهدف هذه الأطروحة إلى تطوير نموذج ذكي هجين على بيانات الجينوم البشري لإيجاد روابط مثيرة للاهتمام قادرة على شرح وظيفة الجينات لمعالجة المشاكل المتعلقة بالأمراض الوراثية واكتشاف الأدوية. عدة مساهمات أُنشئت في هذه الأطروحة في مجال المعلوماتية الحيوية عن طريق اقتراح ايجاد روابط قاعدية باستعمال تقنية " تطور القواعد" و تطوير تقنيات هجينة لحل قضيتين مهمتين في المعلوماتية الحيوية هما: العلاقات الجينية للأمراض واعادة تموضع الأدوية.

أولا، نقترح تقنية هجينة من خلال دمج تقنيتنا المقترحة " استخراج العلاقات القاعدية بتقنية تطور القواعد" مع تقنية الشبكة العصبية الاصطناعية وتقنية الخوارزمية الجينية. هذه التقنية الهجينة الجديدة قادرة على التعامل مع البيانات الجنية المعقدة وايجاد روابط جينية مهمة على أساس ''التغير النكليوتيدي البسيط'' الذي يشارك في تطوير الأمراض. النموذج المقترح يتعامل مع " مشكلة الأبعاد" من خلال توفير طريقة محددة لاختيار السّمات، هذه الطريقة تأخذ بعين الاعتبار العلاقة بين هذه السّمات بدلا من التعامل مع كل منها على حدى.

ثانيا، كمساهمة رئيسية أخرى للأطروحة، نهدف إلى إيجاد روابط خفية بين الأدوية والأمراض التي تستند إلى الكيانات البيولوجية. تقنياتنا الرئيسية استعملت لإعادة تموضع الأدوية. تم استخراج مجموعة من الروابط القاعدية بين الجينات و المسالك البيولوجية المرتبطة بالأدوية والأمراض من أجل ايجاد زوج جديد من (دواء، مرض). تقنيتنا استطاعت ايجاد أزواج كثيرة ، منها ما هو معروف مسبقا ومذكور في المراجع العلمية، ومنها ما هو جديد ومقترح.

في كل النماذج المقترحة، تم تطوير أساليب مبتكرة لتنفيذ مهمة محددة في سياقات البيولوجيا. التقنيات الحسابية الذكية المقترحة في هذه الأطروحة، والتي تقوم على الروابط القاعدية ، الشبكات العصبية الاصطناعية و الخوارزمية الجينية، إما جديدة في فكرة الخوارزمية، أو مصممة ومدمجة بشكل مبتكر لاستخدامها في مجالات المشاكل الطبية الحيوية والجينومية المحددة، ومقارنة أدائها مع العديد من التقنيات المستخدمة في البحوث السابقة.

Résumé

Le monde de la recherche biologique est une expérience enthousiaste sans précédent pour la Bioinformatique, un domaine qui utilise la technologie des ordinateurs, y compris les techniques intelligentes, pour l'analyse de grands volumes de données disponibles sur le génome humain. La séquence complète du génome humain, qui code les instructions génétiques pour la physiologie humaine, a été caractérisée comme "*fondation énorme sur laquelle sont construites la science et la médecine du 21eme siècle*" *Dr.Bruce Alberts*.

Les maladies courantes sont dues aux facteurs génétiques et environnementaux multiples qui peuvent être combinés.

La compréhension de la relation entre la variation de l'ADN de l'individu et la susceptibilité à la maladie est un objectif majeur de la génétique humaine et l'épidémiologie génétique qui conduit à une meilleure prévention, diagnostic et traitement des maladies. La Génomique fonctionnelle vise à comprendre l'aspect fonctionnel d'un système biologique par la compréhension du rôle des gènes et de la relation complexe entre génotype et phénotype à travers l'étude des associations génétiques qui peuvent faire pointer les scientifiques sur le chemin du traitement.

D'une autre part, le développement d'un nouveau médicament est un processus coûteux et consommable du temps. Une stratégie qui vise à résoudre ce problème est le repositionnement des médicaments, défini par l'utilisation d'un médicament déjà connu pour une nouvelle application, qui peut aider à réduire le temps et le coût de la découverte de nouveaux médicaments. L'un des procèdes les plus suivis par ces techniques informatiques est l'identification de liens entre le médicament et la maladie sur la base des liens entre les entités biologiques.

Le développement rapide de la biologie et de la génomique a généré une quantité surprenante de données qui conduit à de grands défis en biologie computationnelle et rend l'analyse précise par les méthodes de calcul existantes impossible. Le développement de nouvelles techniques spécifiques pour les données biologiques est essentielle pour réaliser le potentiel de la génomique fonctionnelle, d'où l'intérêt pour les techniques intelligentes basées sur l'apprentissage automatique, et la fouille de données. Les résultats de l'exploration de données peuvent être référés comme connaissances sous forme de règles. L'extraction des règles d'associations est une des méthodes populaires d'exploration de données largement utilisée dans des différents domaines vu que les règles prévoient des informations importantes, concises et faciles à comprendre. Les règles d'association ont reçu beaucoup d'attention et d'enthousiasme de la part des chercheurs de fouille de données pour leur capacités à résoudre de nombreux problèmes d'exploration de données tels que la classification et l'association.

Récemment, des systèmes intelligents hybrides deviennent populaires en raison de leur capacité à gérer de nombreux problèmes complexes du monde réel. La prise de conscience dans les milieux universitaires que les problèmes les plus difficiles en intelligence artificielle peuvent être résolus et traités en combinant des approches différentes, a conduit à un grand investissement dans les systèmes intelligents hybrides.

Cette thèse vise à développer un paradigme Intelligent hybride pour les données génomiques pour trouver des relations intéressantes capables d'expliquer le fonctionnement des gènes dans le but de traiter des problèmes liés aux maladies génétiques et la découverte de médicaments. Plusieurs contributions ont été faites dans cette thèse dans le domaine de la Bioinformatique. On a proposé des méthodes d'extraction de règles d'association par l'utilisation de la grammaire évolutionnaire, technique qu'on l'a appelé GEARM, et on a créé des modèles hybrides afin de résoudre deux problèmes importants en biologie computationnelle: la relation génotype-phénotype et le Repositionnement des médicaments.

La première technique hybride est basée sur la combinaison de GEARM avec les réseaux de neurones, l'une des techniques puissantes d'apprentissage automatique, et l'algorithme génétique dans le même modèle. GA-NN-EARM est capable d'examiner les bases de données génétiques complexes et trouver des interactions génétiques intéressantes basées sur les polymorphismes (SNP) qui contribuent au développement des maladies. Le modèle proposé a fait face au problème de la dimensionnalité en proposant une manière spécifique pour la sélection des attributs qui prend en considération la relation entre les variables plutôt que les traiter séparément. GA-NN-GEARM avait une haute performance dans le traitement des relations génétiques de maladies.

En second lieu, une autre contribution essentielle de la thèse vise à trouver des associations cachées entre des médicaments et des maladies en fonction de leurs entités biologiques. GEARM est utilisé pour le repositionnement de médicaments. Un ensemble de règles d'associations est extrait entre Gènes et Pathways cibles des médicaments et maladies dans le but de trouver de nouvelles paires de (médicament, maladie). Deux types de règles entre médicament et maladie étaient considérés, ils se diffèrent dans l'antécédente et la conséquente de la règle. GEARM a pu découvrir plusieurs paires de (médicament, maladie), où certaines sont rapportées dans la littérature et d'autres sont inconnues qui peuvent être des nouvelles indications pour certains médicaments.

Dans tous les paradigmes proposés, des approches novatrices ont été développées pour résoudre des problèmes spécifiques dans les contextes biologiques. Les approches informatiques intelligentes présentées ici, qui sont basés sur l'Extraction des Règles d'Association, les Réseaux de Neurones Artificiels et les Algorithmes Evolutionnaires, sont soit nouvelles dans la conception de l'algorithme, ou sont adaptées et combinées de manière innovante pour être utiliser aux problèmes biomédicaux et génomiques, et leur performance a été comparée à plusieurs techniques citées dans la littérature.

Acknowledgment

To finally produce and complete this work which was begun in *February 2012*, I was given all the determination and the strength against all the difficulties from the Almighty Allah, whom I can not thank enough, and I was supported, encouraged, guided and inspired by many persons who deserve my warmest thanks.

First, I would like to express my gratitude to my supervisor *Pr. Ahmed GUESOOUM*, who made the way clear and the tunnel luminous. *Mr. GESSOUM* was my teacher for two years of Master. I was impressed by his discipline, rigour and conscience at work, as well as, his high level and competences. I was impressed even more by his openness for scientific research in all fields and for all kind of fruitful collaborations. I am sincerely grateful for his encouragement, advice, ideas and availability whenever I needed his assistance. I am grateful for the wonderful atmosphere and the team spirit he created within our research group through the informal meetings. Without his guidance this challenging process could not have been completed. Thank you *Mr. GUESSOUM* for the precious work, and for helping me developing my career.

Profuse thanks go to all the members of the jury who agreed to assess my work. I thank *Pr. Ahmed Riadh BABA-ALI* for honouring my jury by presiding it, *Pr. Mourad DAOUDI*, *Pr. Abdelouahab MOUSSAOUI* and *Pr. Mohamed BATOUCHE*. I am deeply grateful for their efforts, and for taking the time to read my manuscript and participate in the defence of this thesis.

I would like to thank so deeply *Pr. Pietro Lio* for welcoming me into his group of Bioinformatics, and for giving me the opportunity to had several research discussion with interesting people in the Computer laboratory and other departments of Cambridge university in UK. I really enjoyed the few months I spent there. I enjoyed the work environment and had an amazing time, learning and having fun. A big thanks to everyone in the Bioinformatics team.

Special thanks are due to my family for believing in me and supporting me through the period of my PhD. I would like to take the chance to thank my sister *Asma*, congratulation by the way for the doctorate in pharmacy this year, for the long fruitful discussions we had about the medical and biological aspects. She reinforced my understanding to many points in the natural field.

My deepest thanks to all my friends with whom I spent wonderful time despite the hard moments. Friends, thanks for tolerating my stress throughout the PhD years, your friendship makes my life a wonderful experience.

Thanks to all my colleagues with whom I shared challenging moments during the thesis preparation. I am thankful for your support, your help in need and your kind wishes of success. I wish to all the PhD students, in my turn, all the best and great success in their career.

Thanks to all my inspiring and challenging teachers along the way of my study from the primary to the university.

I am grateful to all who wished me the best of luck and prayed for me to successfully accomplish this research.

Last and not least, I would like to thank every person who will read this thesis. Thank you for being interested in my work and I would be so happy to help in any point related to my dissertation.

This thesis is the beginning of my journey...

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done by others. I state that every single definition, sentence, paragraph, etc, taken from any other work is referenced as indicated in the research agreements text, except if the phrase is frequently used and the information is public knowledge. I further state that no part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at USTHB or any other university or similar institution except.

Aicha BOUTORH.
May 2016.

Scientific Production

- Boutorh, A., Guessoum, A., Complex diseases SNP selection and classification by hybrid Association Rule Mining and Artificial Neural Network-based Evolutionary Algorithms. *Engineering Applications of Artificial Intelligent.*(2016), pp. 58-70 (51). *http* : *//dx.doi.org/*10.1016/*j.engappai*.2016.01.004*i*

- Boutorh, A., Pratanwanich, N., Guessoum, A., and Liò, P. Drug Repurposing by Optimizing Mining of Genes Target Association. In Computational Intelligence Methods for Bioinformatics and Biostatistics (2014), pp. 209-218. Springer International Publishing

- Boutorh, A., and Guessoum, A. Classification of SNPs for breast cancer diagnosis using neural-network-based association rules. In 12th International Symposium on Programming and Systems (ISPS 2015), April, Algiers, Algeria. pp. 1-9. IEEE.

- Boutorh, A., Pratanwanich, N., Guessoum, A., and Liò, P. Rule-based Grammatical Evolution for Drug Repositioning. In 11th International Meeting, CIBB 2014, Cambridge, UK, June 26-28, 2014. pp.1-6

- Boutorh, A., and Guessoum, A. Grammatical Evolution Association Rule Mining to Detect Gene-Gene Interaction. In 5 th International Conference on Bioinformatics Models, Methods and Algorithms – BIOINFORMATICS 2014, March, Angers, Loire Valley, France. pp. 253-258

- Boutorh, A., and Guessoum, A. Rule Mining based Grammatical evolution to detect Epistasis. In LRIA Science Days 2014, May, USTHB, Algiers, Algeria.

Table of Contents

List of Figures

List of Tables

Liste of Acronyms

A Adenine. 13, 15

ACO Ant Colony Optimization. 45, 77

AI Artificial Intelligence. 47

ANN Artificial Neural Network. 35, 36, 43–45, 47, 56, 59–61, 67, 95, 98, 100, 102, 105, 111, 112, 125–129

AR Association Rule. 5, 45, 59, 69–71, 73–77, 80, 82, 87, 93–95, 98, 102, 112, 123, 127, 128

ARM Association Rule Mining. 30, 44, 45, 68, 77, 84, 93, 113, 125, 129

BAR Boolean Association Rule. 71

BN Bayesian Networks. 56

BNF Backus–Naur Form. 40, 68, 74, 76, 85

BPNN Backpropagation Neural Network. 44, 61

BSO Bees Swarm Optimization. 45, 77

C Cytosine. 13

CA Cellular Automata. 56

CB Computational Biology. 47

CFG Context Free Grammar. 67, 68

CPU Central Processing Unit. 5, 70, 78–80, 82

CTD Comparative Toxicogenomics Database. 53, 54, 114

DM Data Mining. 29, 47, 125

DNA Deoxyribonucleic Acid. 11–16, 19, 21, 27, 48, 56, 84

DR Drug Repositioning. 25–27

EA Evolutionary Algorithm. 38, 43–46, 58, 61, 100, 125

FN False Negative. 86

FP False Positive. 86

FS Feature Selection. 41, 42, 57, 59, 93, 96

FTD Focused Time Delay. 100, 108, 109, 112, 125

FTDNN Focused Time-Delay Neural Networks. 37, 38

G Guanine. 13, 16

GA Genetic Algorithm. 39, 40, 43–45, 47, 56–58, 77, 98, 100, 102, 103, 112, 125, 127, 128

GA-NN-GEARM Genetic Algorithm- Neural Network- Grammatical Evolution Association Rule Mining. 127

GE Grammatical Evolution. 39, 40, 44, 45, 47, 62, 67–69, 71, 74, 82, 85, 93, 113, 125, 128

GEARM Grammatical Evolution Association Rule Mining. 5, 67–69, 71, 74–80, 82–91, 93–96, 98, 102, 104, 110–112, 123, 126–128

GEARM-OE Grammatical Evolution Association Rule Mining for Overall Extraction. 127

GEARM-PE Grammatical Evolution Association Rule Mining for Parallel Extraction. 127

GEDT Grammatical Evolution Decision Tree. 50, 62, 84, 89, 90, 126

GENN Grammatical Evolution Neural Network. 44, 62, 84

GEO Gene Expression Omnibus. 52, 91, 92

GMM Gaussian Mixture Model. 35

GP Genetic Programming. 39, 40, 43, 45, 57, 61

GPNN GP-optimized Neural Network. 43, 44, 61, 62

GPU Graphics Processing Unit. 128

GWAS Genome Wide Association Studies. 10, 11, 14, 55, 56, 126

He Heritability. 51, 52, 84, 88–90

KEGG Kyoto Encyclopedia of Genes and Genomes. 54, 114

KNN K-Nearest Neighbors. 35, 61

MAF Minor Allele Frequencies. 52, 84, 88, 90

MDR Multifactor Dimension Reduction. 56, 62

ML Machine Learning. 47, 125

MLP Multi Layer Perceptron. 36, 37, 61, 100, 108, 109, 112, 125

NCBI National Centre for Biotechnology Information. 52, 91, 92, 103

NN-GEARM Neural Network-Grammatical Evolution Association Rule Mining. 127

OMIM Online Mendelian Inheritance in Man. 53, 114

QAR Quantitative Association Rule. 71, 72

RBF Radial Basis Function. 37, 100, 108, 109, 112, 125

RF Random Forest. 56

RNA Ribonucleic Acid. 14, 48

SE Sensitivity. 86

SNP Single Nucleotide Polymorphism. 16, 23, 35, 42, 50, 55–61, 83–86, 88, 90–96, 98, 100, 102–105, 107–112, 125–127, 129

SP Specificity. 86

SVM Support Vector Machines. 35, 59, 61

T Thymine. 13, 15, 16

TN True Negative. 86

TP True Positive. 86

TWIST Training with Input Selection and Testing. 60

XOR Exclusive OR. 22, 50, 84, 90

Part I

General Introduction

Problem Statement

The field of functional genomics tries to find answers to questions related to the DNA function at the level of genes. A key characteristic to answer these questions by functional genomics studies, is their Genome-Wide approach. The aim of Functional Genomics is thus to understand the functional aspect of biological systems by understanding the role of genes and the complex relationship between Genotypes and Phenotypes using Genome-Wide approaches. The later generally involve high-throughput methods rather than a more traditional "gene-by-gene" approach, in order to reach a better prevention, diagnosis and treatment of diseases.

Does a disease always develop due to external factors or does arise from the individual's own genes?

The intuitive answer to this question is to look outside our bodies, to the environment and to the life style as they are visible factors which play a role in the development of several diseases. The environmental factors have widely been studied by researchers due to their visibility. But what about the internal factors? What about the genetic risks ?

For centuries scientists have worked hard to untangle the genetic relationships and causes of complex diseases. The heritability of traits is a remarkable feature of human beings; one can easily see the resemblance between parents and their children. This obvious fact has eventually led to an important scientific fact: just as the physical traits are inherited in a complex way, transmitted from generation to another as genetic information, so can diseases be inherited. Researchers have indeed discovered that almost all diseases are influenced by many genetic and environmental factors in a complex way.

An important goal of human genetics and genetic epidemiology is to understand the mapping relationship between inter individual variations in DNA sequences, variations in environmental factors and variations in disease susceptibility. Stated differently, the question is to know how one or more changes in an individual's DNA sequence increase or decrease his/her risk of developing a disease through complex networks of biomolecules that are hierarchically organized, highly interactive and dependent on environmental exposures. Thus, understanding of the role of genomic variation in disease susceptibility is likely to improve prevention diagnosis and treatment.

On the other hand, the discovery of a new drug is a costly and time consuming process. Drug Repositioning or Drug Repurposing, a new application of a known drug, is considered as a viable strategy that can reduce the cycle time of drug discovery since their side effects in a clinical environment have already been studied. The currently available computational power has been exploited to improve the effectiveness and efficiency of drug

discovery. One of the common themes shared by most of these computational approaches is to identify new links between entities and complete the drug-disease connection using domain knowledge.

Success in this important public health endeavour "Disease Genetics and Drug Repositioning" will depend critically on the amount of non-linearity of the Genotype-Phenotype-Drug relationships and our ability to address it.

Motivation

Disease-Gene prediction, one of the most significant problems in biomedical research, is the task that consists of finding and identifying the most plausible candidate Disease-Gene associations.

With high-throughput technologies, data on the interaction between genes has grown quickly and currently covers almost the entire genome. As such network-based methods to tackle the problem are becoming prominent, and a variety of approaches have been proposed.

In the last few years, prediction of novel Drug-Disease pairs by Drug Repositioning has been growing in importance. Several Computational approaches have been developed to improve the efficiency and success of drug repositioning.

As the repositioned drug has already passed a significant number of toxicity and other tests, it is known safe with a reduced risk of failure, which is considered as major advantage of the strategy of Drug Repositioning and lead to gain in time and cost.

Machine learning and data mining techniques has been successfully applied to solve various important biomedical problems such as genome annotation, pattern recognition, classification of microarray data, prediction of drug-target and discovery of gene-gene interaction in disease data. In particular, they have been applied to identifying disease associated genes and new pairs of drug-disease.

The problem of detecting the relationship between the genotype and phenotype and drug can be approached using data mining and machine learning techniques.

Genes-Disease association can be formulated as a classification task. To date, a number of supervised learning techniques, the binary-class classification, and various types of gene annotation data have largely been used to solve the disease gene classification problem.

Disease-Gene-Drug can be formulated as association task, where the task is to find hidden relations between drug targets and disease-related genes/pathways (or other data) so as to find new hypotheses of new drug-disease pairs. Most of the previous methods exploit the similarities between drugs on one side and between diseases on the other, independently, and assume at the mean time that similar drugs are likely to combat similar diseases and vice versa. Under this assumption, new pairs of drugs and diseases can be linked.

However, to the best of our knowledge, there has been no study that works perfectly, it is generally true that no classifier is better than others for all classification problems, no models is better for all association problems. Beside that a single intelligent approach may perform well but while combined to other intelligent techniques it can give better results.

Recently, hybrid intelligent systems are becoming popular due to their ability to handle many real world complex problems. The awareness in the academic communities that the hardest problem in Artificial Intelligence can be solved and treated by combining and integrating different approaches, led to a wide investment in the Hybrid Intelligent Systems.

Motivated by the great importance of the generation of genetic disease relation and the discovery of new pairs of Drug-Disease by drug repositioning, and by the success of hybrid intelligent systems in solving different real complex problems, in this thesis it is presented several novel frameworks and Hybrid Intelligent models seeking to mine hidden knowledge embedded in large-scale biological/biomedical datasets.

In every proposed paradigms, innovative approaches have been developed to perform a specific task in the biological contexts. The intelligent computational approaches presented here, that are based on Association Rule Mining, Artificial Neural Network and Evolutionary Algorithms, are either novel in the conception of the algorithm, or are innovatively tailored and combined to be used for specific biomedical and genomic problem areas, and their performance was compared to several techniques that have been used in the literature.

Contributions

In this thesis we have studied a small aspect of the big role played by functional genomics in identifying genetic factors that are responsible for complex diseases, and contributing in the discovery of proper treatment. We focus mainly on Disease-Gene relationships based on interactions and Drug Repositioning based on targets Genes and Pathways.

To this end, we have developed a new intelligent approach based on Association Rule Mining (ARM), well known data mining technique which is widely used to discover hidden relationships in large sets of data. The process of extracting Association Rules was optimised using Grammatical Evolution (GE), an Evolutionary Algorithm, that lead to the technique we have named **Grammatical Evolution Association Rule Mining (GEARM)**. We have used **GEARM** with other intelligent techniques to design a new hybrid intelligent paradigm used in additional contributions to solve problems related to functional genomics.

First, the performance of GEARM was evaluated, where it was tested on different databases, with different transactions and item sizes, frequently used in data mining, and we compared our results to well-known exact and optimized techniques used for the extraction of Association Rules (ARs) as reported in the literature according to the fitness function and the Central Processing Unit (CPU) time.

Then, **GERAM** was used to detect *Gene-Gene interactions*, where it served for classifying individuals in case and control samples with the detection of the interaction between SNPs and the identification of the functional SNPs (those related to the disease). The technique was applied on a simulated SNP dataset which represents different Epistasis models. The results obtained with **GERAM** have been compared to GEDT (Grammatical Evolution Decision Tree) and have shown the high performance reached by our proposed technique.

Then, **GEARM** was combined with Artificial Neural Networks (ANN) giving the **NN-GEARM** model which was applied to a real SNP dataset for *breast cancer diagnosis*. In this contribution, **GEARM** served as a feature selection technique. The main idea here was to extract ARs between SNPs in order to find dependencies between them in such a way that the SNPs appearing in the consequent part of the rules depend on those appearing in the antecedence part. In other words, the SNPs of the antecedent part are considered as dominant ones and will be selected as best features that lead to the highest classification accuracy, where the classification was performed using *ANNs*. We have proposed and compared two different ways of rule extraction. First, we extracted the ARs using **GEARM** on the whole dataset which contains case and control samples; we called it **GEARM-OE** for Overall Extraction. Second, we extracted the ARs by **GEARM** from two separated sets of data, one set of case samples and another set of control samples, and called this **GEARM-PE** for Parallel Extraction. The two sets of features that were selected from the two rule extraction techniques (**GEARM-OE** and **GEARM-PE**) were used as input to the *NN* and compared on classification accuracy. The results indicated that **GEARM-PE** performed better than **GERAM-OE**, while both gave promising results.

To improve the **NN-GEARM** model and apply it for other diseases, we have optimised it with a Genetic Algorithm (GA) to create a new hybrid intelligent model **GA-NN-GEARM** which was applied on four different SNP datasets of *Autism*, *Mental Retardation*, *Colon Cancer* and *Breast cancer* for the *diagnosis of complex diseases*. **GEARM** was used for dimensionality reduction, which is the key for high classification accuracy, **PE** being the version of **GEARM** that we used for the selection of the best SNPs. The set of best features that were selected was used as input to the *NN* to perform the classification for diseases diagnosis. GA was used to optimise the parameters of the model, for which it finds the best architecture for the NN (hidden neurons, number of iterations, ...) and the number of rules to be extracted by GEARM. We have used three types of *NNs*: Multi Layer Perceptrons (MLP), Radial Basis Function (RBF) and Focused Time Delay (FTD) a special type of *Dynamic NNs*. The results proved the high performance of our proposed approach for the diagnosis of complex diseases based on SNP interactions compared to other feature selection and classification techniques.

GEARM was also used for the discovery of drugs in the context of *Drug Repositioning (DR)* which is defined as finding new uses for a known drug. We have extracted a set of ARs between Genes and Pathways target Drugs; and Genes and Pathways related Disease in order to find new pairs of Drug-Disease. Two kinds of rules were considered: (1) rules where Genes/Pathways target Drugs represent the antecedent part and the Disease related Genes/Pathways represent the consequent part, and (2) rules where the Disease-related Genes/Pathways represent the antecedent part and the Genes/Pathways target Drugs represent the consequent part. A set of *(Drug-Disease)* pairs was successfully found for each kind of extracted rules, where some of them are already known in the literature and some others are considered as new discovered pairs.

In the sequel, each contribution is presented carefully. We explain the detailed process of each proposed model and present its strengths by performing series of tests as well as comparisons to other successful techniques in the domain.

Organization of this Thesis

This thesis is organized in four parts. The first part is the general ***Introduction*** which presents the problem addressed in this thesis, the motivation for this work, the contributions, and this section of thesis organization.

The second part presents the ***Background*** which is needed for the rest of the thesis and is divided into *three chapters*.

Chapter 1 gives an overview of the Biological world defines the basic concepts of biology that are needed for our work, genetic variations, genotype-phenotype relationship, genetic susceptibility to diseases, where four diseases are presented, genetic interaction, biological pathways, as well as drug repositioning.

Chapter 2 gives an overview of the computational methods we needed for this work, including Data Mining and Machine Learning Techniques, Feature Selection Techniques and Hybrid Techniques.

Chapter 3 presents several biological datasets and a survey of the use of Data Mining and Machine Learning techniques in Biology for both studies of our interest, Gene-Disease relationships and Drug Repositioning. This mating between the two domains is known as Bioinformatics or Computational Biology.

The Third part is the part where we present our ***Contribution***. This part is also divided into three different chapters.

Chapter 4 presents in detail the main approach we propose and which we call *GEARM*. The chapter gives a description of the general form of this technique and evaluate its performance on several transaction datasets comparing to different optimized and exact techniques to extract ARs, as well as it gives the different possible uses of *GEARM*. The following chapters in this part present the application of GEARM as a basic algorithm combined with other techniques of data mining and machine learning to solve different problems related to functional genomics.

In chapter 5 we focus mainly on finding genomic relationships for disease diagnosis. Two major sections are included. Section 5.2 presents the use of *GEARM* on simulated data to detect gene-gene interaction and to identify the functional SNPs related to a specific disease. The process of *GEARM* applied to this problem is presented in detail by specifying the form of the used grammar and the evaluation function and showing the results that are reached. Section 5.3 presents the hybridization of *GEARM* with a *NN* and *GA*. First *GEARM* is hybridised with *NN* and a new technique is created which we named *NN-GEARM*. The proposed *NN-GEARM* approach is applied on real breast cancer SNP data. Then, *NN-GEARM* is optimised and improved by using a *GA*, and the new model *GA-NN-GEARM* is used for four different SNP data bases. The process of *GA-NN-GEARM* is described in detail: the grammar and the evaluation function used for this problem are specified, and the role of each combined method in this contribution is presented. Moreover, a series of tests and a comparison are presented to show the high performance of the hybrid intelligent approach we propose.

In Chapter 6 we focus on finding genomic relationships for the drug repositioning problem. We present all the detailed steps of *GEARM* applied on Genes/Pathways datasets target Drugs and related to diseases in order to discover new pairs of Drug-Disease. The results we have obtained are provided and discussed.

The fourth and the last part of this thesis is about the ***Conclusion and Future Work***. In this part, we sum up the work we have accomplished for this thesis. We evaluate the results we have reached with respect to the different problems we have addressed under the main general problem. We also present some work that can further be done and problems that can be tackled in the area of intelligent Bioinformatics.

Part II

Background

Chapter 1

Overview of Biology Concepts

1.1 Introduction

The Functional Genomics is the field that study the biological function of the genes and their products. It was properly defined as " approaches under development to ascertain the biochemical, cellular, and/or physiological properties of each and every gene product" (Gibson and Muse, 2009). A key feature of the studies of Functional Genomics is the Genome wide approaches. The goal of Genome Wide Association Studies (GWAS) is to identify genetic risk factors for common traits. These studies are useful to find genetic variations that contribute to complex diseases by examining genetic variants and comparing different individuals for a specific disease or trait. GWAS can reveal the reasons of having a healthy life and of being predisposed towards diseases. The identified genetic associations can be used by researchers to design and develop new powerful techniques and strategies to detect, treat and prevent diseases. With the birth and the development of Genomics, the understanding of the hidden aspects of Biology, especially those related to human genetics, was improved. Many questions in genetics arose looking for explanations. For instance, what makes a person different from another? How are the genetic variations associated with diseases? How can genomics be used for treatment and drug discovery? Diving into the world of Biology, these questions and many others can find their answers. In this chapter, we present an overview of different Biological concepts to help the reader better grasp the problem addressed in the thesis.

1.2 Basic Concepts

Numerous are the phenomena that can be understood if we consider the Genome as the book of life (Bodmer and McKie, 1997). Every person's book would contain the same

chapters and sections written using the same letters and arranged in the same order. It is more understandable to describe the genome book as a set of *23* chapters called *Chromosomes*, devised into sections called *Genes*, formed by a collection of words called Deoxyribonucleic Acid (DNA) and written with only four letters *A*, *T*, *G* and *C* (Singh, 2012). Figure 1.1 illustrates the whole genomic information of the Human Genome, and each of them is described below.

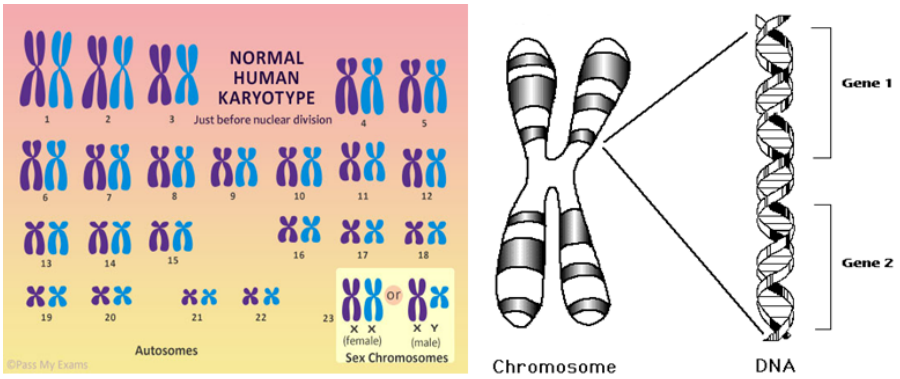


FIGURE 1.1: Human Genome
(Biology Notes for IGCSE, 2014), (Pass My Exams Biology, 2016)

1.2.1 Genome

The term Genome was coined in 1920 by professor of botany *Hans Winkler* (Winkler, 1920) at the University of Hamburg, Germany. It is known as a set of inherited instructions that constitute the genetic material of all the Organisms. The Genome is responsible on passing life from one generation to another by building, running, and maintaining the organism (DeWeerdt et al., 2003). Every cell in the organism contains the whole copy of the genome that guides its function and carries the organism's whole genetic information (Ridley, 2013).

Each species of the earth has its own particular genome, including plants like the wheat genome, animals like the horse genome, bacterium like Escherichia coli and virus genome (DeWeerdt et al., 2003). The human genome is a massive text, where if its three billion letters were printed out on standard paper and heaped up, they would require a pile of books almost as high as the Washington monument (Collins and Mansoura, 2001).

All human genomes are quite similar, yet every human being has a slightly different version (DeWeerdt et al., 2003). The difference between individuals can be studied by GWAS (Guttmacher and Collins, 2003). Sequencing the genome will be such a great help for the scientists to find genes much more easily. Simply, genome sequence is constructed in a mysterious language as a very long string of letters (DeWeerdt et al., 2003).

1.2.2 Chromosome

The genome of every organism is divided into Chromosomes, which are found in pairs in the human species. A chromosome is a package that contains some of the organism's genes. It is made of *DNA* tightly coiled many times around proteins forming a long chain of nucleotides. The nucleotides are serried and arranged in the form of chromatin which allows huge DNA molecules to fit into eukaryotic cells. Chromosomes aid a cell preserve a large amount of genetic information neat, organized, and compact (DeWeerdt et al., 2003).

Chromosomes have in general different sizes and shapes. Each eukaryotic organism has a very specific number of chromosomes per cell and most of them have a linear shape just as for a human. The chromosomes of a human are diploid, which means we have two copies of each chromosome. Humans have *46* chromosomes arranged in *23* pairs which consist of *22* pairs of autosomes and one pair of sex chromosomes (Figure 1.1). Females have two copies of the X chromosome, while males have one X and one Y. Children get half of their chromosomes from their mother and half from their father (Lea, 2009).

Each chromosome is divided into two section or *"arms"* by a constriction point called the centromere. The *"p arm"* is the short arm of the chromosome and the *"q arm"* is the long arm of the chromosome. The characteristic shape of each chromosome is given by the location of the centromere, which can be used to help describe the location of specific genes. The particular region of the chromosome where a specific gene is located is called *locus* (Parakhia, 2009) (Gillham, 2011).

1.2.3 Gene

A gene is the basic physical and functional unit of heredity passed from parents to their children. Genes, which are made up of DNA, act as instructions to make molecules named proteins which are further responsible for specific traits and functions in the organisms (Slack, 2014). The Human Genome Project has estimated that in humans there are between 25,000 and 30,000 genes, and the size of each gene varies from a few hundred DNA bases to more than 2 million bases. (DeWeerdt et al., 2003).

Every person inherits one copy of gene from each parent, having at the end two copies of each gene. Genes are the unit that is responsible for the differences between the species in general and between individuals in particular. Moreover, genes are responsible for the similarities between human beings. For example, humans are similar in having "eyes color gene" which codes for eyes color but they are different in which color each one has, in people we can find different color of the eyes, black, brown, grey, blue, green etc.

Mainly, all human beings have similar genes (around 99%) for every and each trait, only a small number of genes are slightly different between individuals; these are *alleles*, the single variants of genes. The variations in the physical appearance, phenotype, of people are due to these small differences (less than 1%) which contribute to each person's unique physical features. The more the functionalities of the genes can be comprehensible, the more doors will be opened for a better understanding of rare, complex and common diseases, and the more researchers will have opportunities for drug development (Feero et al., 2010a).

1.2.4 DNA

The Deoxyribonucleic Acid (DNA) is the hereditary material in all the organisms including humans (Genetics Home Reference, 2016b). It is the basic biochemical entity of the gene and genome inside the nucleus of a cell that carries the genetic instructions for making a living organism (Figure 1.1) (Avery et al., 1944) (Hershey and Chase, 1952).

Almost, all the cells in a person's body have the same DNA. The information in the DNA is stored as a code written in a language of four bases: *Adenine (A)*, *Thymine (T)*, *Cytosine (C)* and *Guanine (G)*. Like the letters of the alphabet create specific words when used in a particular order, the arranging of the four bases that form the DNA in a particular order define the genetic information responsible for building and maintaining an organism (Genetics Home Reference, 2016b). These bases formed of sugar and the phosphate group make up the nucleotides which are the chemical units of the DNA (Singh, 2012). Nucleotides that form a spiral shape in two long strands called a double helix (Genetics Home Reference, 2016b). The sequence of these nucleotides determines the biological instructions on genes (National Human Genome Reseach Institute, 2011)

The DNA gets transmitted across generations and its coded language guides every cell in its function and organization (Hershey and Chase, 1952).

The human DNA is composed of about *3 billion* bases, nucleotides, and more than *99 %* of those nucleotides are identical in all individuals (Genetics Home Reference, 2016b).

The double helical structure of the DNA was discovered in 1953 by *James D. Watson* and *Francis Crick* (Watson and Crick, 1993), who worked together in the Cavendish laboratory in Cambridge, England. They began their work in the early 1950s (DeWeerdt et al., 2003). In their study, they explained the probable pairing of the *Adenine* (*A*)- *Thymine* (*T*) and *Cytosine* (*C*)- *Guanine* (*G*) (Watson and Crick, 1953).

The DNA has an important property is to make a copy of itself. In another word, the DNA has the ability to replicate, where each strand plays the role of a pattern for

duplicating the sequence of bases. In the division of the cells, each new cell will have an exact copy of the DNA which was present in the old cell (Genetics Home Reference, 2016b).

1.3 Genetic Variations

Genome books of human beings are unique, about 99.6% of base pairs are identical from person to person, which makes people unique for the most part (Feero et al., 2010a). But, in spite of that, every individual's genome is slightly different from another's.

Given the variety of the human species, all the human genome are mutant and no genomic sequence is normal. Differences between individuals are in general referred to *variations* and are found in specific location in the human genome (Feero et al., 2010a). The observation of the phenotypic variation can lead to the identification of the genetic variation, and the inheritance of variations in the genome leads to differences in phenotypes, which can increase the risk of diseases.

The variation in the order of bases in nucleotides causes genetic variations. A genetic variation is located on a specific position of a genomic sequence (DNA). When it affects a transcribed region, the change is propagated to the transcribed sequence (Ribonucleic Acid (RNA)), and if it affects a coding region, the change is also spread in the amino acid sequence (protein).

Today scientists are able to identify more genetic variations of common diseases such as heart disease, diabetes, asthma, and common cancers by sequencing the DNA, thanks to new technologies and tools that researchers now have. One such research approach is GWAS. To find genetic variations associated with a specific disease, GWAS look at the complete sets of variable DNA markers (up to about a million!) in many individuals. Through a successful identification of these genetic associations, this information can be used by researchers to improve the methods of detecting, treating, and preventing the disease. GWAS are setting the foundations of personalized medicine (Lea, 2009).

Two popular types of genetic variations (Feero et al., 2010b):

1. "Mutations": which are random changes in one or more base pairs known to be pathologic, and
2. "*Alleles*": in which the variants are are called *Polymorphisms* if the frequency of the minor allele is greater than 1%, and defined as a single base variations or differences present in each individual DNA but these are not mutations.

1.3.1 Allele

An alternative form of gene is known as Allele, which is situated at a particular position on a particular chromosome. Different traits that are passed on from parents to offspring are defined by such DNA codings (Angelini, 2009). The word *" allele "* is a short form for *allelomorph*. Allele was used for characterizing variant gene forms detected as various phenotypes in the early days of genetics (Craft, 2013).

An individual inherits two copies of each gene, explained in Figure 1.2, one from each parent, which may have distinct phenotypic effects, and that is referred to as Allele. The Alleles can be the same, as they may be different. If we look at the situation where and individual has a base "A" in one position, and at the same position another individual has "T" , this case is called as "*Heterozygous*", because the two alleles are different. This situation of two different nucleotides represents two alleles of the same gene. The other possible situation, is where the two alleles are the same, and this is called as "*Homozygous*" (Singh, 2012).

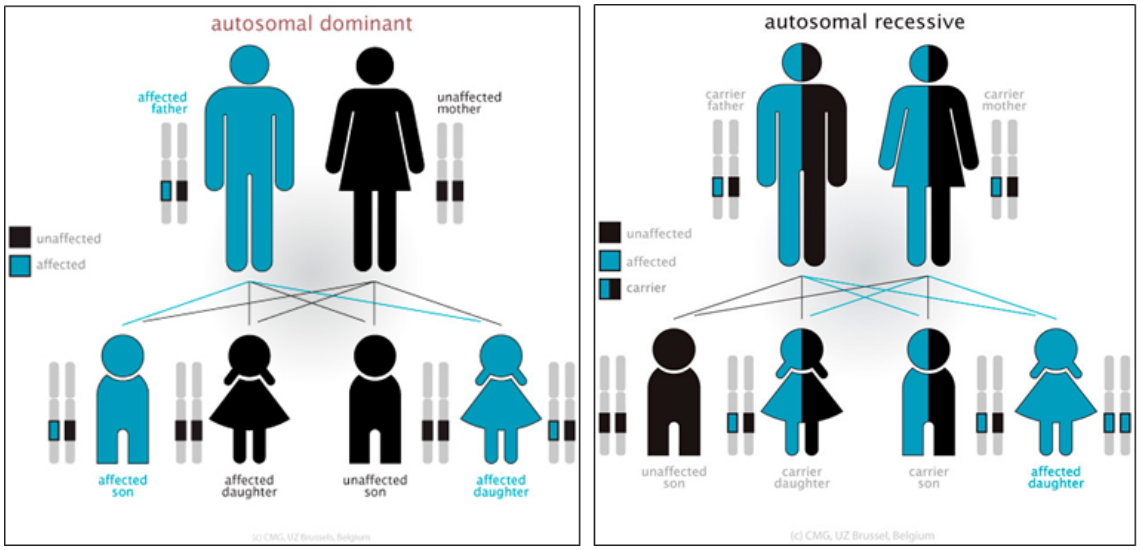


FIGURE 1.2: The Inheritance Process (Genetics course at UW-Madison, 2014)

The fraction of all the chromosomes in the population that carry a particular Allele is known as *Allele frequency*, and represents the relative frequency of an allele at a specific locus in a population. *Allele frequencies* show the reflection of genetic diversity in population genetics and qualify the quantity of variation at a specific locus or through multiple loci (Gillespie, 2010).

1.3.2 Single Nucleotide Polymorphism (SNP)

Genetic alterations that occur in more than 1 percent in a DNA nucleotide at specific positions among individuals are called Single Nucleotide Polymorphisms (SNPs) (Genetics Home Reference, b). We can explain it as one individual possibly having *"T"* at a specific position whereas another individual has *"G"*. At this specific base position there is a SNP, and the *T* or *G* are said to be alleles for this base position (Figure 1.3).

Polymorphisms are very common DNA variations that lead to many normal differences between individuals traits like eye colour and shape, hair colour and type, skin colour and type, blood type and many others. Despite the fact that many SNPs have no negative impact on individual's health, there are some polymorphisms that can contribute to the development of particular diseases (Genetics Home Reference, b). The severity of the disease and how our bodies respond to treatments are also genetic variations demonstrations events. To determine if a particular genetic variation is associated with a trait or a disease, Association Studies are developed (Zhang et al., 2004).

SNPs are the most common type of sequence variation, contribute about 90% of the total sequence variation (Collins et al., 1998). SNPs occur on average nearly every 100 to 300 bases (Ke et al., 2008). They can be present in coding and non-coding DNA. The SNPs that have a high chance to produce functional differences are those existing in the coding regions of genes (cSNPs). Despite that most SNPs have no effect on the function of a gene, a big number of them can be used as markers to find SNPs that do affect gene function (Collins et al., 1998).

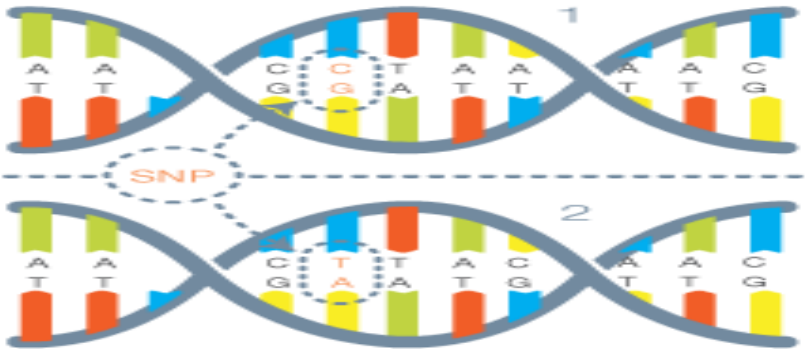


FIGURE 1.3: Single Nucleotide Polymorphism (SNP) (ADN – Evolutions, 2015)

1.4 Genotype-Phenotype Relationship

The distinction between *Genotype* and *Phenotype* was first proposed by the scientist *Wilhelm Johannsen* in Denmark, and introduced in his textbook on heredity research, titled " *Elemente der exakten Ereblichkeitslehre* " (The Elements of an Exact Theory

of Heredity) in 1909. The two concepts were then developed more in 1911 by the same researcher (Peirson, 2013). The point of the distinction between the Genotype and the Phenotype is situated in the distinction between the inherited genetic information of organisms and the ways these information express themselves to physical characteristics (Peirson, 2013). According to *Johannsen*, in the process of development the phenotype of an organism is defined by its genotype which may be influenced by environmental factors (Peirson, 2013). Improving the understanding of the connection between genotype and phenotype is an ultimate goal of genetic research.

1.4.1 Genotype

The genotype is the genetic make-up of an individual organism which is the set of data carried by the genome of this individual, in other words all of its genetic material which is encoded as DNA. The genotype is the data bases used by cells to define the characters of an individual. The functions of the genotype represented the instructions for the growth and development of the body. The word "genotype" is commonly used to describe the genetics of a specific traits (like eye, hair, skin colour) (Science Learning, 2011).

1.4.2 Phenotype

The phenotype is the set of observable physical or biochemical traits and characteristics of an individual. It results from the expression of its genotype (Science Learning, 2011). An important thing to consider is that the expression of the genotype, and thus the resulting phenotype, are sensitive to the influence of multiple factors : the time of life, the environment, nutrition, stress, disease or medication.

1.4.3 Genetic Susceptibility to Diseases

Genetic susceptibility to a disease, in medicine, is assigned to genetic predisposition to a health issue, which is an increased and an inherited risk to develop a disease based on the genetic makeup of an individual (Eldridge, 2015). Such genetic change plays a role in the development of disease but it does not produce it in a direct way. Having a genetic predisposition for a disease does not mean that you will certainly get that disease. Some people with a genetic predisposition to a specific disease may live a healthy life and never get this disease, where as some others will have it, even if they belong to the same family (Genetics Home Reference, 2016a). The development of the disease is not assured; its appearance is influenced by other conditions like lifestyle, nutrition and environmental

factors. Diseases that are produced by a combination of different factors are described as multifactorial (Genetics Home Reference, 2016a).

Certain mutations in certain genes will have an increased risk of developing a related disease, while some other mutations will have small effects (Genetics Home Reference, 2016a).

Success in understanding the role of genomic variation and environmental context in disease susceptibility will help the improvement of the diagnosis, prevention and treatment. This depends mainly on the ability to address the mapping of genotype to phenotype and the sum of non-linearity which is defined as an outcome that cannot be easily predicted by the amount of the individual genetic markers, and can arise from phenomena such as (Moore et al., 2010):

- Locus heterogeneity: the same phenotype can be determined by different DNA sequence variations.
- Phenocopy: phenotypes determined by environmental factors and do not have a genetic basis.
- The dependence of the effects produced by the genotype on the environmental factors: gene-environmental interactions.
- Genotypes at other loci: gene-gene interactions or epistasis.

Among the complex diseases developed by genetic variations, we give here a brief presentation of four different diseases that are considered in our contribution. For each disease, we give some of its characteristics summarized as follow.

1.4.3.1 Autism

Autism is a neurodevelopmental disorder that has as characteristics weakness in social interaction like impairments in social reciprocity, challenges in verbal and non-verbal communication and the presence of restricted and repetitive behaviors. It is incomprehensible how this disorder happens, but it is known that an alteration of the organization and the connection of the nerve cells and their synapses is due because of the affectation of the information processing by the Autism disease (Susan E et al., 2009).

In general, the symptoms on the children are noticed by their parents from the early childhood, roughly in the first two years of their child's life, (Myers and Johnson, 2007) (Stefanatos, 2008) (Association, 2013), and affect the daily functions. Children with

Autism suffer from many difficulties, they often lack the capability of engaging with other children to play and they tend to stay isolated, they avoid eye contact with other people and they may fail to respond to many intuitive questions like their names. In addition, people affected by Autism are unable to express themselves and talk about their feelings, also it is hard for them to understand other people's feelings.

Despite the fact that Autism is highly heritable, environmental factors beside the genetic ones may be significant causes together (Chaste and Leboyer, 2012).

Autism is known to have a strong complex genetic associations, outcomes from either rare mutations with major effects or rare multigene interactions of common genetic variants (Abrahams and Geschwind, 2008) (Buxbaum, 2009). Recently, other genetic mutations in children affected by Autism have been discovered.

The complexity of this disorder occurs due to environmental factors, epigenetic factors and multiple genes interactions (Rapin and Tuchman, 2008). Scientists believe that the genetic factors and the environmental factors have an important role in developing and increasing Autism risk, but, yet no specific environmental reasons have been characterized, and most of mutations have not been identified. Characteristically, Autism can not be associated to a single gene mutation (Abrahams and Geschwind, 2008).

The availability of the actual rapid, precise gene-sequencing tools and the accessibility to the large numbers of DNA samples, have lead to a major advance in identifying genetic factors associated to Autism disorder. These developed tools have lead, as well, to a considerable advance in specifying a number of environmental factors to be consider for future studies, as it is quite complicated where researchers need to identify how the environment can interact with the genetic information (American Speech-Language-Hearing Association, 2016).

Despite that there is no known cure (Myers and Johnson, 2007), early speech or behavioural interventions can be a great assistance to children with Autism to develop their skills of communication and gain self-care. It should be mentioned that, not all children with Autism will have an isolated life by reaching their adulthood, some become successful (Howlin et al., 2004) and there have been some cases who recovered (Helt et al., 2008).

Recently, an autistic culture has known a considerable development, some people are looking for a cure, while others believe that Autism should be accepted as a difference and not treated as a disorder (Silverman, 2008).

1.4.3.2 Mental Retardation

Mental retardation disease (Islam and Islam, 2015) is known also as intellectual disability (ID), intellectual development disorder (IDD) or general learning disability (Wilmshurst, 2012).

It is defined as a generalized neurodevelopmental disorder with an IQ (Intelligence Quotient) score below 70 that lead to significant limitation of personal skills including difficulties in adaptive behaviours of daily life and impaired intellectual adaptive functioning. (Schalock et al., 2010)

Researchers have found many causes of Mental Retardation including genetic disorder (Daily et al., 2000). Occasionally, disability is caused by abnormal genes inherited from parents, errors in the combination of genes, or other reasons (Disability, 2011).

Four possible levels of intellectual disability: mild, moderate, severe, and profound. Serious cases of Mental Retardation are diagnosed at birth. However, for the other cases parents can notice some failures of their child to reach common daily life goals, until the age of 18 years where the Intellectual Disability is diagnosed (Schalock et al., 2010).

1.4.3.3 Colon Cancer

Cancer begins when cells of the body start to grow abnormally and out of control. Colorectal cancer is the case where the cancer starts in either the colon or the rectum and spread to other parts of the body. It is named also as colon cancer or rectal cancer according to where it starts. It can begins as a polyp, a small growth, that starts in the colon or rectum and rises to the center, and it affects mostly people older than 50. Colorectal cancer is the third leading cause of cancer death in both men and women, an estimated 49,190 deaths are expected to occur in 2016 (Society, 2008).

Screening test can be helpful to detect cancer in its early stage in order to start the treatment which work best when the cancer is found early before spreading in the body (Health-Conditions, 2013).

Signs and symptoms may include rectal bleeding, pain in the belly, blood in the stool, a change in the bowel habits, the feeling that the bowel is not completely empty, constant tiredness, in rare cases, unexplained weight loss (Society, 2008), (Health-Conditions, 2013).

Although the causes of colon cancer are not well known, there are some common risk factors that increase the chance of developing this diseases such as lifestyle, older age,

and inherited genetic disorders (Stewart and Wild, 2015). Epigenetic factors, such as abnormal DNA methylation of tumor suppressor promoters play a role in the development of colorectal cancer (Schuebel et al., 2007).

Other risk factors include diet, smoking, alcohol, lack of physical activity, family history of colon cancer and colon polyps, presence of colon polyps, race, exposure to radiation, and even other diseases. Globally, colorectal cancer is the third most common type of cancer making up about 10% of all cases (Schuebel et al., 2007).

1.4.3.4 Breast Cancer

The cancer that is developed from breast tissue is known as breast cancer, and has different signs like change the shape of the breast, dimpling or red scaly patch of skin, a lump in the breast, a fluid coming from the nipple. After the spread of the diseases, other signs may appear like bone pain, swollen lymph nodes, shortness of breath, yellow skin (Stages et al., 2016).

Genetics susceptibility play a significant role in developing breast cancer, where about 5–10% of all cases get the disease by inheriting genes from their parents including BRCA1 and BRCA2 (Pasche, 2010).

Women whose mother was diagnosed before 50 have an increased risk of 1.7 and those whose mother was diagnosed at age 50 or after has an increased risk of 1.4 (Gage et al., 2012) (Colditz et al., 2012).

The BRCA1 and BRCA2 gene mutation are up to 90% of the total genetic influence (Pasche, 2010). Genetics play a greater role in less than 5% of cases in causing a hereditary breast–ovarian cancer syndrome for those who carry these mutations (Hendrick, 2010).

Breast cancer is the most common invasive cancer in women. It affects approximately 12% of women worldwide. Based on U.S. statistics in 2015 there were 2.8 million women affected by breast cancer (World Health Organization, 2012).

1.5 Epistasis or Gene-Gene Interaction in Complex Diseases

In the disorders research area, researchers are learning that almost diseases have a genetic factor. Some of them are due to a mutation in a single gene, while some others are much more complex. The complex diseases are probably due to the effects of genetic factors,

TABLE 1.1: Exclusive OR (XOR) interaction model for two binary variables.

X	Y	Z
0	0	0
0	1	1
1	0	1
1	1	0

life style and environment factors, thus they are known as *complex* or multifactorial disorders. Some of these diseases are: Autism, Mental Retardation, Diabetes, Alzheimer, Cancer, Heart disease and many others.

Complex diseases are characterized by their difficulties to study and treat due to the combination of different factors. Researchers are still looking for responsible factors that have not been identified yet by developing improved strategies as the standard single-locus tests might not be able to identify such effects (Genetics Home Reference, a). The detection of interactions between loci will allow to illustrate the biological and biochemical pathways that promote disease. In this section the interactions between genetic loci that contribute to human genetic complex disease, which is known as *Gene-Gene interaction* or *Epistasis*, is presented.

It is known now that genes can mask and alter the effects of other genes. In fact, understanding these interactions that exist between genes or Epistasis maybe a key to better understand and master complex diseases.

Gene-Gene interaction or Epistasis, is a phenomena of collective actions of multiple genes. It is a basic concept in genetic that describes any interaction between two or more loci. Epistasis has been in use for almost a century (Musani et al., 2007). Gene-Gene interaction consists of the dependence of one genes on other genes, it can be one or more. In another words, the expression of one gene being affected by the expression of one or more other genes which affects in its turn a specific phenotypic trait (Carlborg and Haley, 2004). Epistasis is recognized broadly and acknowledged as an important contributor to genetic variation in complex diseases (Musani et al., 2007).

Attribute or variable interactions, i.e. statistical Epistasis, can be illustrated by an example of non linear attribute interaction the *Exclusive OR (XOR)* model, as shown in Table 1.1, and that was presented and described in (McKinney et al., 2006) as follow:

Let X and Y be independent variables and Z be a class variable. If the relationship between X and Z depends on Y, it is said that X and Y interact.

Let *SNP1* with Alleles *'A'* and *'a'* and *SNP2* with Alleles *'B'* and *'b'* be two Single Nucleotide Polymorphisms (*SNPs*). By ignoring the Allele order, each *SNP* has three possible states known as genotypes.

- For *SNP1*, the genotypes are AA, Aa and aa,
- For *SNP2*, the genotypes are BB, Bb and bb,

The penetrance values given in the Table 1.2 present the probability that an individual will have the disease for each of the nine possible genotype combinations. High disease risk is dependent on inheriting a heterozygous genotype (Aa) from one *SNP* or a heterozygous genotype from a second *SNP* (Bb), but not both.

The biological allele frequency in this example is of $p = q = 0.5$ with genotype frequency (McKinney et al., 2006):

- p^2 for AA and BB
- $2pq$ for Aa and Bb
- q^2 for aa and bb

The marginal penetrance for a specific loci is the product of its peneterance values (vector)and the genotype frequencies of the other genetic variables. It represent the probability of this specific genotype to be associated to disease risk independently. For example, the marginal penetrance of Aa is $(1, 0, 1)(0.25, 0.50, 0.25)^T = 0.5$.

All the marginal penetrance values of this example represented in Table 1.2 are equal. The presented model is a purely epistasis without main effect. In another words, the disease risk is affected only in the presence of SNP interactions, and is not affected by an independent genetic variables although the inequality of penetrance values. This example presents a challenging genetic model and one possible worst-case scenario in genetic analysis (McKinney et al., 2006).

1.6 Biological Pathways

The chains of chemical reactions that involve proteins, genes and metabolites in the cellular level of the human body are called biological pathways, and are often visualised in graphical diagrams.

TABLE 1.2: Penetrance values for combinations of two SNPs genotypes in the absence of effects (McKinney et al., 2006).

Genotype	Genotype			Marginal Penetrance
	AA (0.25)	Aa (0.5)	aa (0.25)	
BB (0.25)	0	1	0	0.5
Bb (0.50)	1	0	1	0.5
bb (0.25)	0	1	0	0.5
Marginal Penetrance	0.5	0.5	0.5	

A biological pathway is a series of actions in a cell among molecules that leads to a certain change in a cellular state or process, their most common types are Metabolic pathway, Genetic pathway and Signal transduction pathway. Figure 1.4 show an example of biological pathways that play key role in advanced studies of Genomics (National Human Genome Research Institute, 2015).

Among the important actions of Gene-regulation pathways is turning genes on and off. This task is essential because the proteins are produced by the recipe provided by genes. Proteins are known to be the key components necessary to achieve almost every task in our body (National Human Genome Research Institute, 2015).

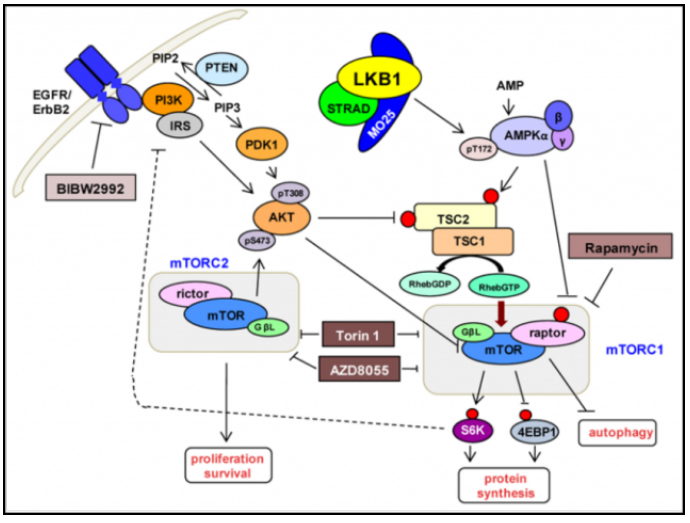


FIGURE 1.4: A typical signalling pathway (Andrade-Vieira et al., 2013).

Biological pathways have the ability to act over short or long distances. Among their roles is controlling person's response to the world like some pathways affect how the body processes drugs, others maintain balance while a person is walking and so on. Biological Pathways are known to work together to perform tasks. A biological network is formed when multiple biological pathways interact between them. Many important biological

pathways have been discovered. Like any component in the organism, Pathways are vulnerable to malfunction. A specific disease can be developed if the pathway doesn't work properly or something goes wrong in it (National Human Genome Research Institute, 2015).

Information about biological pathways are used to develop new and more efficient drugs as well.

More research on biological pathways and the genetic profiles of particular tumors, might allow to developers of drugs to concentrate only one some pathways, where patients could after receive the drugs most likely to repair these pathways affected. . Pathways offer helpful resources of exploring the effect of drugs on the human body (National Human Genome Research Institute, 2015).

1.7 Drug Repositioning

Drug development is time-consuming requires at least 10 years, and very expensive requires at least billion dollars. Any business model based on products made within 10 to 15 years and costs about $1.3 billion per successful product needs to be reconsidered and constantly evaluated (Persidis, 2011). Drug discovery and development has been attracting increasing attention from both academia and industry.

Drug Repositioning (DR) also known as *Re-tasking*, *Re-profiling*, *Drug Repurposing* or *Therapeutic Switching* has long been an efficient strategy of drug development as it can renew a failed drug or expand the number of indications for a successful one. One thing beneficial to the patient while finding new indications for a specific drug, is to have potential new therapy earlier (Persidis, 2011).

Drug Repositioning is defined as the application of known drugs and compounds to new indications i.e., new diseases, in another words, it is finding new indications for approved drugs (Sleigh and Barton, 2010). Its major advantage compared to other strategies to discoverer and develop new drugs, is that the risk of failure is reduced and it is safe, because the positioned drug is known, it has already passed several clinical tests (Ashburn and Thor, 2004).

Drug Repositioning strategies were proposed in response to the failure risk associated with novel drug discovery, where more than 90% fail (DiMasi et al., 1991). The start of DR from approved molecule, increase its ability to rapidly enter the clinical phase. Figure 1.5 presents a drug repositioning strategy.

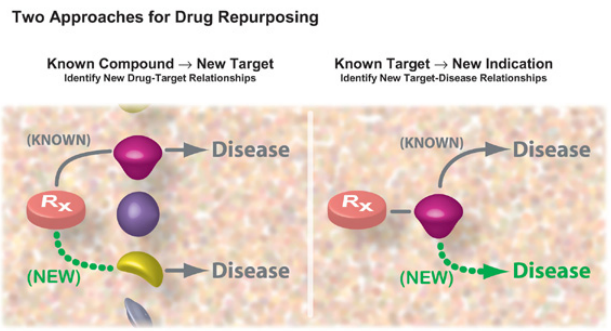


FIGURE 1.5: Drug Repositioning Strategy (Biomedecine, 2012)

The number of repositioning success stories is increasing. One of the most famous drug repositioning examples is the successful repurposing of *Sildenafil*, designed to treat *Pulmonary Arterial Hypertension* originally, to treat *Erectile Dysfunction* (Boolell et al., 1996).

Making them in points; the arguments of DR are simple and have been summarized in (Persidis, 2011) as follow:

- **Safety**: The known drugs that are approved or have been proven safe in late-stage trials, their inherently reduced development risk can be leveraged into potentially new indications. A significant development advantage of repositioned drug is that about 30% of drug failures in clinical trials (Persidis, 2011).
- **Save money**: Developing new drug seems to cost 160 million times more than successfully bringing a repositioned drug to market (Persidis, 2011).
- **Market potential**: Two very good examples are *Celgene's Thalomid*, which is repositioned *thalidomide*, and its derivative *Revlimid* (*lenalidomide*), that represent a revenue of more than $2.8 billion. However, it should not be assumed that such success for a particular drug repositioning signifies the same success for all repositioned drugs. Many factors are necessary for potential for market success including market need, competition, excellent product, a successful strategy and so on (Persidis, 2011).
- **The return on investment potential**: The repositioned drugs will always represent a better return on investment (Persidis, 2011).

Despite of the success reached by the strategy of drug repositioning, it is still facing some challenging problems from the clinical and the commercial point of views.

Finding new applications for an approved drug had a great impact both on commercial, where it has the potential for significant returns, and on research, where true innovation

were generated in the understanding of the basic biology of disease (Persidis, 2011). As such, Drug Repositioning (DR) future is assured and bright.

1.8 Conclusion

In this chapter Biological overview has been presented. It focused mainly on defining the basic concepts of biology (DNA, Genes, Chromosomes and Genome) and their interactions and relationships with phenotypes; more precisely complex diseases. Genetic codes profoundly influence our bodies, our behaviours, our minds and everything from eye colors and height to aging and disease.

A part of treatment has been covered in this chapter as well, by giving the definition and the description of the strategy of Drug Repositioning (DR) and its relationships with genetics targets.

With the completion of the human genome sequence, the necessity for powerful tools to deal with this huge amount of biological information in the medical sciences become urgent. The different complex issues related to the human genome, including the analysis of genetic variation, access to genetic information and so on, gave a birth to new discipline known as Bioinformatics.

During the past few years, Bioinformatics, defined as using informatics techniques to solve biological problems, has become one of the most highly visible fields of modern science. Life scientists working in mainstream research is strongly affected by the explosion of information and data. Researchers had no choice but to be attached to Bioinformatics techniques in order that a considerable impact in research will be always preserved.

Chapter 2

Overview of Intelligent Computational Methods

2.1 Introduction

The amount of data stored in files, databases, and other repositories gives a great motivation to develop powerful analysis and interpretation tools. Previously, decisions were only made based on decision makers' intuitions and instincts, because of the lack of powerful techniques to extract the valuable information and knowledge hidden in the huge volume of data. The answer to this issue is " Data mining", which has emerged from a number of different disciplines, including statistics, machine learning, artificial intelligence, information retrieval, and so on. The reason to be attracted to such strategies is due to the understanding that "we are data-rich but information poor". There is huge volume of data but they are hardly turned into useful knowledge for decision making. It turns out that a serious obstacle facing the productive use of the large amount of available data is what is named as the "curse of dimensionality". This problem refers to the huge number of variables (characteristics) of a given problem and, hence the insufficient data for its analysis. To this end, various statistical and intelligent techniques have been proposed to address this issue but, the remaining difficult question is : Which features should be used to create a powerful predictive model? . Several techniques of feature selection, machine learning and data mining have been combined together in order to increase their power. Intelligent techniques may work rather well separately, but when combined with each other, they tend to lead to a better performance. This combination can be with the aim of reducing the dimensionality or optimizing a costly process in time or otherwise solving any other problems related to a specific technique. The central goal

of Artificial Intelligence and machine learning is getting computers to solve problems automatically, encompassed by what Turing called " Machine Intelligence " (Turing 1948, 1950). In the 1950s, the terms " Machine Intelligence ", " Artificial Intelligence ", and " Machine Learning " all referred to the goal of getting " machines to exhibit behavior, which if done by humans, would be assumed to involve the use of intelligence" (to quote Arthur Samuel, 1983).

In the current chapter, we provide a brief introduction to data mining and machine learning, focusing on some widely used and successful techniques in different fields. We present the dimensionality reduction problem and the different techniques used to deal with the kind of data that has a high number of features. We also present the advantages of combining several techniques in the same model and how the resulting performance can be improved.

2.2 Data Mining and Machine Learning Techniques

During the past two decades, Various data mining and machine learning techniques and tools have been proposed, designed, and implemented to tackle the challenging problems by finding interesting patterns from data with the objective of providing useful assistance in different domain (Han et al., 2011).

We present in the sequel the Data Mining and Machine Learning Techniques that will be used in our solution in Part III of this thesis.

Data Mining

Data Mining (DM) is a new approach to data analysis and knowledge discovery that was emerged in the mid 1990's. It is a relatively new concept. The first ACM Conference on Knowledge Discovery and Data Mining (a.k.a. SIGKDD) was held in the USA in 1995 (Yoo et al., 2012). In 2001, MIT's Technology Review identified data mining as one of the ten emerging technologies that would change the world (Yoo et al., 2012).

There are different definitions of data mining from different perspectives (Larose, 2014). Here are some of these definitions (Delpisheh, 2010):

- "Data mining is a process of automatically analyzing data in large data repositories to extract interesting patterns, such as relationships, from data" (Jiawei and Kamber, 2001).

- "Data mining is a process of selection, exploration, and modeling of large-volume data to discover relationships that are previously unknown, aiming at obtaining clear and useful results for the owner of the data" (Paolo, 2003).
- "Data mining is a process of data-driven extraction of implicit, previously unknown, and potentially useful (or actionable) information from large databases" (Sumathi and Sivanandam, 2006).
- "Data mining is a process of discovering various models, summaries, and derived values from a given collection of data" (Kantardzic, 2011).

In the work of this thesis, we refer to data mining as an automatic process to analyze and extract hidden relationships from a large volume of data, in order to reach higher performance in decision making.

Machine Learning

Machine learning refers to mathematical models and algorithms that allow computers to learn from experience. There are two types of learning process: Supervised and Unsupervised (John Lu, 2010).

- In supervised machine learning, where the learning processes is guided, the goal is to predict the output variable given a set of input variables. The measure of performance of the machine represents how well was the learning process from the training data by evaluating its ability to predict outcomes from independent test data.
- In unsupervised learning, the output variables are unavailable and the goal of machine learning is to describe hidden structures from unlabeled data.

In the literature, there are major data mining and machine learning tasks, and, in this section, we briefly discuss some of these that are successfully and widely used in different fields.

2.2.1 Association Rule Mining (ARM)

Association Rule Mining (ARM) is one of the important unsupervised techniques and the most studied data mining paradigm (Ian and Eibe, 2005). It is defined as the task of discovering and describing relationships in the form of *If-Then* rules among

different items in large databases (Agrawal et al., 1993). The associations produced by this technique represent interesting relations between variables and are in general easy to understand and interpret. In association rule mining, interesting patterns are represented as association rules. For example, in the basket analysis, $milk, eggs \longrightarrow bread$ is an association rule, which says that if *milk* and *eggs* are bought together by a customer, then *bread* is likely to be bought as well.

Formally, the problem of mining ARs is formulated as follows: let T be a set of transactions $T = \{t_1, t_2, t_3, ..., t_n\}$, and let I be a set of items $I = \{i_1, i_2, i_3, ..., i_k\}$, an AR is represented as an implication of the form $X \longrightarrow Y$ where $X \subset I$, $Y \subset I$ and $X \cap Y = \oslash$. X is an itemset called the antecedent of the rule and Y another itemset called its consequent.

There are different measures to evaluate the association rules obtained by an ARM process. The two commonly used measures are the *Support* and the *Confidence*.

Let $P(X)$ be the probability of appearance of itemset X in T and let $P(Y|X)$ be the conditional probability of appearance of itemset Y given that itemset X appears. The *Support* $Sup(X)$ of an itemset $X \subseteq I$ is defined as the ratio of transactions $t_i \in T$ such that $X \subseteq t_i$. Namely $Sup(X) = P(X)$, while, the *Support* of the Rule $Sup(X \longrightarrow Y) = P(X \cup Y)$. The *confidence* of the rule $Conf(X \longrightarrow Y)$ is defined as $P(Y|X) = P(X \cup Y)/P(X) = Sup(X \cup Y)/Sup(X)$. The standard problem of mining association rules is to find all rules whose *Support* and *Confidence* are equal to or greater than the *minimum Support* s, and *minimum Confidence* c thresholds, respectively (Jiawei and Kamber, 2001).

One measure of independence between X and Y is *Lift* measure, that is defined $Lift(X \longrightarrow Y) = Conf(Y \longrightarrow X) = Conf(X \cup Y)/Sup(Y)$. Values close to 1 indicate that X and Y are independent and the rule is not interesting.

Another measure to evaluate the interestingness of the rule is the *Conviction* $Conv(X \longrightarrow Y)$ (Brin et al., 1997), which compares the probability that X appears without Y. The *Conviction* overcomes the weakness of confidence and lift. It attempts to measure the degree of implication of a rule, and is defined as $Conv(X \longrightarrow Y) = (1 - (Sup(Y)))/(1 - Conf(X \longrightarrow Y))$

The *Conviction* values range from $0.5, to + \infty$. Its value is 1 in case of independence and is infinite for logical implications (confidence 1). Unlike *Lift*, *Conviction* is sensitive to rule direction ($Conv(X \longrightarrow Y) \neq Conv(Y \longrightarrow X)$, and unlike confidence, the support of both antecedent and consequent are considered. Conviction values that are far from 1 indicate interesting rules.

In the past, many algorithms were developed by researchers for Boolean and Fuzzy association rule mining such as Apriori (Agrawal and Srikant, 1994), FP-Growth (Han et al., 2004), etc. The approach ARM is mainly based on the Apriori algorithm suggested by Agrawal et al. This algorithm works in two phases: the first step is to find the frequent itemsets, whose supports are more than a user-specified *minimum support*; the second step is to generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset

Apriori Algorithm:

1. **Begin**
2. Input: Transactional Database (DB)
3. Output: Frequent Itemset
4. Let k = 1
5. $L_1 \longleftarrow$ { Generate frequent itemsets of length 1 }
6. Repeat until no new frequent itemsets are identified
   (a) Generate length (k+1) candidate itemsets from length k frequent itemsets
   (b) Prune candidate itemsets containing subsets of length k that are infrequent
   (c) Count the support of each candidate by scanning the DB
   (d) $L_k \longleftarrow$ { frequent items of size k } Eliminate candidates that are infrequent
7. Return $\cup L_k$
8. **End**

A typical example of association rule mining, as previously mentioned, is the basket analysis problem. It analyzes customers' shopping habits by finding associations among the different items that customers place in their shopping baskets.

The Apriori algorithm is explained by applying it to the transactional database example shown in Table 2.1. The minimum support for finding frequent itemsets is set to 33%. This is equivalent to requiring that a frequent itemset must appear in at least [33% * 9] =3 transactions, where 9 is the total number of transactions in Table 2.1.

First, the set of frequent 1-itemsets is found as *C1* by scanning the database to accumulate the count for each item, and collecting those items that satisfy the minimum support criterion. The resulting set is denoted as *L1*.

TABLE 2.1: Transactional Database for Basket Analysis.

	Bread	Milk	Sugar	Eggs	Butter
T1	1	1	0	0	1
T2	0	1	0	1	0
T3	0	1	1	0	0
T4	1	1	0	1	0
T5	1	0	1	0	0
T6	0	1	1	0	0
T7	1	1	1	0	0
T8	1	1	1	0	1
T9	1	1	1	0	0

C1= { *Sup(Bread)= 66%, Sup(Milk)= 89%, Sup(Sugar)= 66%, Sup(Eggs)= 22%, Sup(Butter)= 22%* }
L1= {{*Bread*}, {*Milk*}, {*Sugar* }}

Next, *L1* is joined to itself to generate the set of candidate 2-itemsets, denoted as *C2*. Given the Apriori property, we only select those frequent 2-itemsets, denoted together as *L2*, from C2 for further consideration.
C2={*Sup(Bread,Milk)= 55%, Sup(Bread, Sugar)= 44%, Sup(Milk, Sugar)= 55%* }
L2= {{*Bread, Milk*}, {*Bread,Sugar*}, {*Milk, Sugar*}}

Then, using the join operation again, we generate *C3*, from which we select the frequent 3-itemsets, denoted as *L3*.
C3= { *Sup(Bread, Milk, Sugar)= 33%* }
L3= { {*Bread, Milk, Sugar*}}

This process is repeated each time generating larger frequent itemsets, until no more frequent itemsets can be found.

The Apriori algorithm terminates, after having found all of the frequent itemsets in L= $L1 \cup L2 \cup L3$

For each frequent k-itemset l, every nonempty proper subset $f \subset l$ is checked whether the rule $R : f \longrightarrow l - f$ satisfies $Conf(R) \geqslant Min\_Conf$ where *Min_Conf* is the minimum confidence threshold. The resulting association rules are shown below, each with its confidence:

- R1: Bread $\longrightarrow$ Milk 5/6= 83%
- R2: Milk $\longrightarrow$ Bread 5/8= 62.5%

- R3: Bread ⟶Sugar 4/6= 66%
- R4: Sugar ⟶ Bread 4/6= 66%
- R5: Milk ⟶ Sugar 5/8= 62.5%
- R6: Sugar ⟶ Milk 5/6= 83%
- R7: Bread ⟶ Milk, Sugar 3/6= 50%
- R8: Milk, Sugar ⟶ Bread 3/5= 60%
- R9: Milk ⟶ Bread, Sugar 3/8= 32.5%
- R10: Bread, Sugar ⟶ Milk 3/4= 75%
- R11: Sugar ⟶ Bread, Milk 3/6= 50%
- R12: Bread, Milk ⟶ Sugar 3/5= 60%

If the minimum confidence is taken to be 60%, then only the association rules R1, R2, R3,R4, R5, R6, R8, R10 and R12 are returned.

Association rules were widely used in various areas such as telecommunication networks, market analysis, risk management, inventory control, web usage mining, intrusion detection, medical diagnosis, Bioinformatics, etc.

2.2.2 Classification

The Classification is one of the most common data mining and machine learning tasks. It is defined as assigning a class label to an unknown new sample. Each sample is represents by a number of attributes and a class label. The classification process is performed on two steps (Delpisheh, 2010):

- The first step is the learning step or the training step. This step consists of determining the sets of data classes. Primarily, it finds a function, $y = f\ (x)$, that is able to predict the class label y of a new unknown sample x based on the attributes values of x.
- The second step is the test step. It consists of applying the classifier on new set of data called test set to estimate its prediction accuracy. If the accuracy is acceptable, the classifier will be put into use to classify future data samples with unknown class labels.

The algorithms that are widely used for classification include Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Gaussian Mixture Model (GMM), Decision Trees (), Artificial Neural Networks (Artificial Neural Network (ANN)) etc. We present in this section Artificial Neural Network.

2.2.2.1 Artificial Neural Networks (ANNs)

Machine learning is an approach for prediction and classification able to deal with the dimensionality problem in a smooth manner. One of the very promising machine learning techniques is Artificial Neural Networks (Grossi and Buscema, 2007).

Computation using ANNs is a popular machine-learning model inspired by the biological neural network, the brain. One of the reasons of their success is their ability to solve both supervised and unsupervised problems. This mathematical model is characterised by: the ability to learn complex, nonlinear input-output relationships; the use of sequential training procedures; and the adaptation to the data (Sharma and Kaur, 2013).

The architecture of a Supervised ANN is a mathematical model which consists of several layers of inter-connected small processing units called neurons. The first layer is given by the inputs (e.g., an appropriate representation of an SNP genotype ((Motsinger-Reif et al., 2008a)), each of the middle layer(s) (or hidden layer(s)) contain(s) a number of neurons that perform some mathematical operations which allow to process the input they receive and forward their results to the next hidden layer neurons or to the output layer neurons, in a pipelined fashion. Each connection between two neurons has a weight **w** that represents the strength of the signal exchanged between these two neurons. This is referred to as the Feed-Forward processing in ANNs. A graphical representation of an ANN is given in Figure 2.1

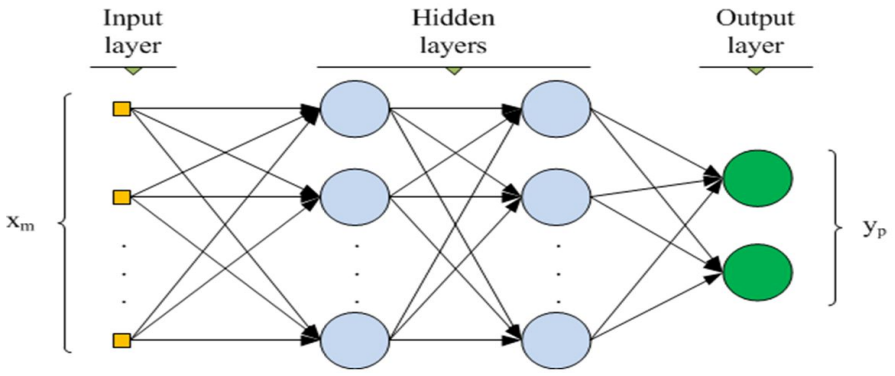


FIGURE 2.1: Artificial Neural Network Architecture (Jing et al., 2012)

Classification is one of the fundamental types of problems in multivariate analysis and consists basically in finding mathematical models that can recognize the membership of samples to their proper class (Lavine and Rayens, 2014). The process of building a Supervised ANN involves a training phase during which the network takes the inputs and their corresponding target data repeatedly and produces some output. This (actual) output is compared to the desired output (target) and an error vector is computed. In order to decrease the computed error, the strengths of the connections between the network neurons are adjusted according to a learning algorithm during each iteration.

One of the problems associated with the use of supervised ANNs that can lead to "under-fitting" or "over-fitting" and affects (positively or negatively) the prediction performance of the model, is the selection of the network architecture It is well known for instance that neural networks with an overly complex structure are more likely to over fit the training data than equivalent ones with a simpler structure (Myung, 2000), (Dreyfus, 2005).

The Over-fitting appears when a model starts to memorize training data instead of learns to generalize trends from it.

To avoid the possible cases of over-fitting problems, it is important to well perform the initialization and the configuration of the ANN parameters. One situation where the over-fitting is likely to appear is when the database is characterized by a high number of attributes for a low number of samples.

This well documented phenomenon is frequently addressed in the literature (Baumann, 2003), (Reunanen, 2003), (Hawkins, 2004), (Cawley and Talbot, 2010), (Panchal et al., 2011). Much work has been performed in order to face the problem of over-fitting with techniques like model selection, early stopping, regularization, network pruning and weight decay (Nowlan and Hinton, 1992), (Bishop, 1995).

Neural networks can be classified in two different classes: Static and Dynamic. In static networks the output is calculated directly from the input through feed forward connections. They does not contain delays, and does not have feedback elements. Whereas, in dynamic networks, the output depends on the current or previous inputs, outputs, or states of the network (Demuth et al., 1992).

§a Multi Layer Perceptron (MLP) : The most common neural network statistical model is the Multi Layer Perceptron (MLP), which has been a popular machine learning paradigm since the 1980s. This type of ANN is known as a supervised network; it is a feedforward artificial neural network model that has as goal to map sets of input data onto a set of appropriate outputs. using "historical data" so that the model can

then be used to produce the output when the desired output is unknown (Rosenblatt, 1961).

An MLP consists of multiple layers of nodes in a directed graph. The input layer receives vectors of patterns to process. Each layer is connected to the next one (Skapura, 1996). MLP utilizes a supervised learning technique called backpropagation for training the network (Rosenblatt, 1961). MLP can distinguish data that is not linearly separable (Cybenko, 1989).

Multilayer perceptrons are fast and reliable networks suited for simple pattern recognition problems. For any supervised learning process, they use the standard backpropagation algorithm. MLPs have been used in diverse fields such as speech recognition, image recognition, machine translation software, and so on (Wasserman and Schwartz, 1988).

§b Radial Basis Functions (RBF) : Another type of neural networks is the Radial Basis Function (RBF) which are powerful tools for classification that use Radial Basis Functions. Radial functions are functions whose performance decreases (or increases) monotonically with distance from a central point. The centre, the disctance scale, and the precise shape of the radial function are parameters of the model that represents the transfer function used on the network layers. Radial Basis Networks can require more neurons than standard feedforward backpropagation networks. They work best when many training vectors are available (Chen et al., 1991), (Lowe, 1995).

§c Focused Time Delay (FTD) : Dynamic networks can be divided into two categories: networks that have only feedforward connections, and networks that have feedback, or recurrent, connections. The recurrent-dynamic networks typically have a longer response than the feedforward-dynamic networks (Demuth et al., 2008).

Dynamic networks have been applied in several applications such as speech recognition, prediction in financial markets, prediction of protein structure, and many others. The training of Dynamic networks is a little difficult, however they are in general more powerful than static networks. They can be trained to learn sequential or time-varying patterns using their memory. The training of Dynamic networks can be performed using the same gradient-based algorithms that are used for static networks (Demuth et al., 2008).

Focused Time-Delay Neural Networks (FTDNN) are the most straightforward dynamic networks; they consist of a feedforward network with a tapped delay line at the input (Demuth et al., 2008).

FTDNN is part of a general class of dynamic networks, called focused networks in which the dynamics appear only at the input layer of a static multilayer feedforward network (Demuth et al., 1992).

2.2.3 Evolutionary Algorithms (EA)

Evolutionary Algorithms (EAs) are stochastic and adaptive population-based search methods based on the principles of natural evolution. They have number of variants that share the same objective which is trying to find the optimal solution using the operations of reproduction, mutation, recombination, and natural selection on a population of candidate solutions. Figure 2.2 presents the general scheme of the evolutionary process (Yu and Gen, 2010): it starts by generating an initial population of individuals represented as a chromosome, each of which is a potential solution to the problem. A fitness function is generally defined by taking into account the domain knowledge and an objective function which reflects the quality of solutions to the problem that is to be solved. Each individual is evaluated according to this fitness function and individuals with the highest scores are considered as better solutions. A subset of those individual solutions will be selected to create a new population of the same size as that of the initial population. This is done by applying the crossover and mutation operations from one generation to the next, until a predefined termination criterion is met. The process stops either when reaching a fixed number of maximum generations or an optimal fitness value is found. In this section, we focus mainly on presenting three important variants of EAs: Genetic Algorithms and Grammatical Evolution.

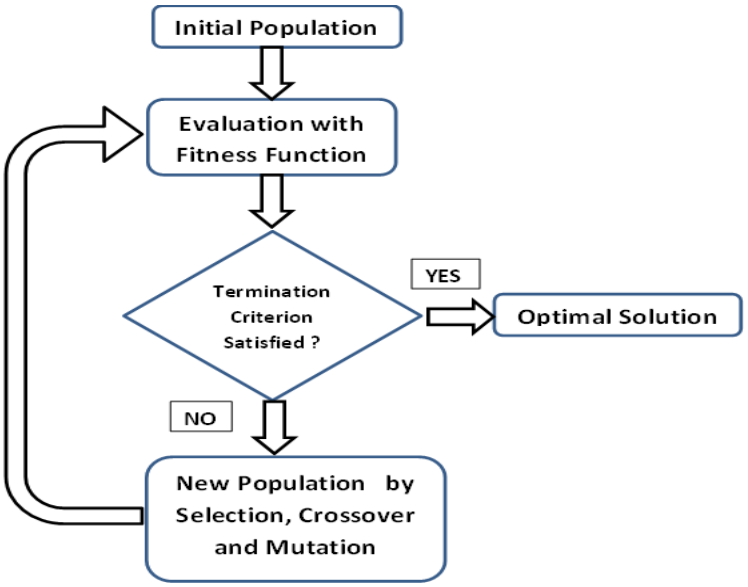


FIGURE 2.2: Evolutionary Algorithm Process

2.2.3.1 Genetic Algorithms (GA)

Genetic Algorithms (GAs) were invented and developed in 1960s by John Holland at the University of Michigan (Mitchell, 1998). It is an optimization techniques inspired by natural genetics which has been successfully applied in many machine learning and optimization problems, to generate useful solutions. GA is a stochastic general search method, capable of effectively exploring large search spaces. It requires five essential constituents:

- An appropriate way of encoding solutions to the problem, which can be handled by the Algorithm.
- A way of initializing the population of chromosomes.
- An appropriate fitness function.
- The operators that can be applied to parents when they reproduce to change their genetic information.
- Parameter settings for the algorithm, operators, etc.

2.2.3.2 Grammatical Evolution (GE)

Grammatical Evolution (GE) is one of the most frequently used evolutionary algorithms, in addition to GA and Genetic Programming (GP). It is a form of evolutionary computation inspired by the biological process of generating a protein (phenotype) from the genetic material (DNA genotype). It uses a specific grammar to translate populations made of linear genomes into a computer program (O'Neill and Ryan, 2003) in a deterministic process of genotype-phenotype mapping, where each genotype is always mapped to the same phenotype.

The grammar consists of production rules which are defined in terms of non-terminals and terminals. Only non-terminals can appear on the left-hand side of the production rule, whereas the right-hand side may consist of any combination of terminals and/or non-terminals. Formally, the grammar is defined by four sets $< S, N, T, P >$, where, S is the start symbol, N is the set of Non-terminal symbols that be substituted by terminal symbols, T is the set of Terminal symbols that appear in the language by applying corresponding production rules and P is the set of production rules.

Each generated individual will be evaluated using a fitness function, and evolutionary operators will be applied at the chromosomal level (strings) to create subsequent generations. Before the evaluation of each individual, the following steps take place in Grammatical Evolution (O'Neill and Ryan, 2003):

- The genotype (integer string) is used to map the start symbol and the non-terminal symbols of the Backus–Naur Form (BNF) grammar definition into terminals. The grammar is used to specify the desired phenotypes.
- The integer values (from the integer string) are then transformed into an appropriate production rule through the mapping process by using the grammar.
- The production rule is selected using the MOD operation presented in the formula of the Equation 2.1, where P-rule is the index of the selected production rule and nb-al is the number of alternatives (i.e. rules) defining the current non-terminal.

$$P - rule = (Value) MOD (nb - al) : \tag{2.1}$$

  The non-terminals are replaced by the alternative which gets selected through the MOD operation and this process continues until only terminal elements remain.

The wrapping process is one of the main characteristics of GE: if the end of the chromosome is reached and the program still has non-terminal elements, the algorithm returns to the start of the chromosome to obtain the next integer T number of times (O'Neill and Ryan, 2003).

GE has some advantages comparing to GA. Among the drawbacks of GA is the fixed length of the chromosome, whereas GE uses variable lengths for the chromosomes (Sharma and Tivari, 2012). Despite the success of GPs, they still have limitations. GE differs from GP in several ways.

First, GE uses linear genome like GA rather than tree structures. Second, GE manipulates a string of integers to perform genetic operations unlike GP in which the genetic operations are performed on trees which is costly both in execution time and complexity. Furthermore, by using a grammar, substantial changes can be made through simple manipulations of the specified grammar. These features are important improvements of GE over other techniques (O'Neill and Ryan, 2001).

2.3 Feature Selection Techniques

The modern world has known a fast growing of data from different domain. The high dimensional need powerful techniques to deal with its volume. Techniques of Machine Learning were developed to construct effective classifiers, and they are facing the challenge of the hight dimensionality. The important question was how to deal with all this amount of features? The evident solution is to try to reduce the number of variable in the data. For this reason, several feature selection approaches were proposed (Guyon and Elisseeff, 2003).

Feature Selection (FS) techniques played a key role in different research areas like machine learning, pattern recognition and data mining. If a traditional classifier is used to classify a sample based on all the variables of the data, a low accuracy is expected. FS techniques aim to remove all the redundant and irrelevant features in the data that cause the reduction of the accuracy of the classifier model (Shardlow, 2009).

The high number of features and the relatively small number of observations (samples), as microarray data for example, is a common phenomenon known in machine learning as the "curse of dimensionality" problem. The objectives of FS thus involve both minimizing the number of features that get selected and maximizing the classification performance. Reducing the attributes will reduce the complexity of the model and make it simple to understand and explain.

FS operation for dimensionality reduction is defined as the process of removing unneeded and irrelevant attributes to generate a reduced dataset with all the important information of the initial data, so that the learning algorithm focuses only on the most informative features that were selected for the training data, and that are beneficial for the analysis and predictions (Chen et al., 2006), (Christin et al., 2013), (Guyon and Elisseeff, 2003).

In other words, given an $\boldsymbol{n}$ dimensional dataset, we try to find a subset of $\boldsymbol{k}$ dimensions, where $\boldsymbol{k} < \boldsymbol{n}$, which preserves all the information of the original data (Fodor, 2002). The goals of FS are mentioned in three points is (Guyon and Elisseeff, 2003) as follow: 1) increasing the prediction performance of the model, 2) creating a cost-effective model with a fast process, and 3) a better understanding of data generation process.

Finding a reduced set od data with the best features is in general intractable. This problem of FS have been proved to be NP-hard (Kohavi and John, 1997). FS methods may be divided into univariate techniques and multivariate techniques. The former are fast and scalable but ignore feature dependencies (Bolón-Canedo et al., 2014), (Saeys et al., 2007); the latter overcome the drawback of the univariate type and incorporate

feature dependencies but at the cost of being slower and less scalable (Bolón-Canedo et al., 2014).

There are usually three types of feature selection methods: filters, wrappers and embedded methods (Mohamad et al., 2009), (Seo and Oh, 2012), and many others techniques have been proposed to select the most informative features.

2.3.1 Filter Methods

Filtering approaches evaluate the attributes based on general characteristics of the training data to select features that are independent of any predictor (i.e. statistical measures). These techniques use some measures that indicate the usefulness of the feature for the classifier by ranking these later and selecting the best ones (Shardlow, 2009). They use an "easy-to-calculate metric" for a faster ranking of the features and selection of the top-ranking ones. Classical filtering methods were used on microarray data, like Correlation Feature Selection (CFS), Fast Correlation-Based Filters (FCBF), and ReliefF (Saeys et al., 2007). The main advantage of filter methods is that they are much faster than wrapper methods, whereas their disadvantage is that they lack the interaction with the classifier which lead to a lower performance results.

2.3.2 Wrapper Methods

Wrapper methods use a machine learning algorithm, as part of the selection process, to check the effect of various subsets of features (Ian and Eibe, 2005), (Ruiz et al., 2006). In general, the principal advantage of wrappers is that they lead to a better classification accuracy (Kohavi and John, 1997). However, their common disadvantage is that they need to build the classifier model many times in order to evaluate each selected set of features. The space of the features is very large, which makes the search of every possible combination a complex process and computationally expensive, especially for SNP datasets. This means that new FS techniques and some heuristic search methods are needed to be developed in order to reduce the complexity of the process of finding optimum sets of features and increase the accuracy of the classifier.

2.3.3 Embedded Methods

Embedded approaches are as an intermediate solution to overcome the drawback of filters and wrappers. They generally use machine learning models for classification which

eliminate features as part of the training process. Embedded approaches can be more effective with their strategy of incorporating the feature selection in the process of training. This efficiency can be mainly resulting from the better use of the data by not needing to split it into training and validation sets, and the avoidance of the retraining the model from scratch for every selected variable subset (Guyon and Elisseeff, 2003). Thus, the advantage of the Embedded methods is that they implicate the interaction with the classifier, without being computationally expensive like wrapper methods (Saeys et al., 2007).

2.4 Hybrid Techniques

Several techniques have been combined together in order to obtain better performance. Hybridizing techniques is defined as a combination of two or more techniques in order to benefit from the power of each techniques together rather than separately. Artificial Intelligence techniques work good separately, but while combined in a smart way they may work better.

2.4.1 Artificial Neural Network-based Evolutionary Algorithm

Although ANN architecture is still considered an unsolved task, Xu and Chen (2008) overview several approaches to the selection of appropriate architectures. One approach is the use of Evolutionary Algorithms (Floreano et al., 2008). No guarantee is assured by these algorithms to find the best architecture of the ANN. However, they are likely able to find, in a reasonable time, a good solution automatically (Manning et al., 2013).

GA, GP and GE are the most frequent EAs used to optimise the architectures of ANNs (Rivero et al., 2010), (Soltanian et al., 2013),(Ahmadizar et al., 2015); they have recently given a good performance on Bioinformatics data (Motsinger-Reif et al., 2008a), (Turner et al., 2010), (Ahmad et al., 2012), (Koo et al., 2013), (Li et al., 2014a), (Moore and Hill, 2015).

In literature, combining neural network to genetic programming improved its power (Ritchie et al., 2003). The authors proposed a new technique that uses GP as evolutionary algorithm to optimize the inputs and the architecture of ANNs. Each GP binary expression tree represents an ANN. This kind of method is named GP-optimized Neural Network (GPNN), and aims to generate the appropriate architecture of neural network for the provided dataset.

Another type of machine learning method, GE, was combined to ANNs to optimize their inputs, weights, and architecture. Like GPNN, Grammatical Evolution Neural Network (GENN) was used to improve the performance of ANN by finding an optimal architecture (Motsinger et al., 2006) (Motsinger-Reif et al., 2008a). GENN optimizes the inputs from a large variables, the weights, the number of hidden layers and the number of nodes in the hidden layer. Thus, an optimal neural network architecture for a given dataset is automatically generated.

In (Motsinger et al., 2006) GENN outperformed traditional Backpropagation Neural Network (BPNN), a random search algorithm, and GPNN in larger datasets. This is due to the flexibility of the model, where GE was able to optimize the ANN more efficiently and with less computational cycles than GPNN. The use of the grammar gave more flexibility in changing the way ANN is built by a simple modification of grammar file, the led to a decrease in development time and an increase in flexibility (Motsinger and Ritchie, 2008).

2.4.2 Association Rule Mining-based Evolutionary Algorithm

In General, the extraction of association rules is based on the Apriori algorithm suggested by Agrawal et al. (Agrawal and Srikant, 1994). The algorithm works in two phases: the algorithm starts by generating all the frequent items whose supports are greater than or equal to the specified minimum support, then it extracts all the valid rules that satisfy the minimum confidence constraint (see Section 2.2.1).

For large datasets, calculating all possible association rules i.e. all possible itemstets is computationally inefficient. In order to deal with the problem of extracting all the rules from large databases, different approaches have been explored. Various new proposed techniques were presented in several work as promising methods that outperformed previously existing algorithms.

EAs were used to overcome this drawback and to find an optimal or near-optimal solution to the problem when many solutions exist (Aguiar et al., 2010), (Nunkesser et al., 2007). Using EAs and Heuristics is one of the solutions to optimize the process of rules extraction. GA is one of the most used EA with ARM (Mata et al., 2001) (Mata et al., 2002). An improved algorithm based on GA was presented in (Yan et al., 2009) called ARMGA. An interesting work that provided the performance analysis of ARM based on GA is described in (Indira and Kanmani, 2012).

Genetic Programming was used as well to extract AR. In (Olmo et al., 2011) the authors used the grammar guided genetic programming (G3P) to avoid invalid individuals found by GP process.

Studies based on Ant Colony Optimization (ACO) for serial ARM were presented in (Kuo and Shih, 2007) (Moslehi et al., 2011). In (Djenouri et al., 2014) an ARM algorithm based on an improved version of Bees Swarm Optimization (BSO) BSO-ARM was described. the authors proposed three different heuristics for exploring the search area. Some of the proposed optimization algorithms in the literature suffer from some important limitations such as generating false rules.

The proposed techniques of Association Rule extraction using GA have been used in several real problems like Biological problems. Despite the fact that GP (Espejo et al., 2010) has been successfully used to generate ARs in different data sets, there are still limitations to evolving ARs using this type of machine learning algorithms. New techniques based on Evolutionary Algorithms are needed to deal with the challenge of optimizing the extraction of Association Rules.

2.5 Conclusion

In this chapter an overview of Intelligent Computational methods was presented. Strategies of Data Mining played a key role to solve the problem of "we are data rich but information poor". We have given a brief definitions of data mining and machine learning, beside introducing some interesting techniques widely used in different areas.

In the last years, ARM has been a key technique in many research areas with the different proposed optimized methods that improve the running time with a good quality of the generated rules. ANNs are one of the very promising machine learning techniques for prediction and classification capable of dealing with the dimensionality problem in a smooth manner. EAs are stochastic and adaptive population-based search methods. They are based on the principles of natural evolution widely used mainly for optimization and variable identification. They have various variants that share the same objective which is trying to find the optimal solution using the operations of reproduction, mutation. Among the most used variants, we presented GA and GE, that have been successfully used with intelligent techniques like ARM and ANN with very promising results.

The combination of EA to association studies and classification task, has proven its efficiency in different studies from the literature. Thus, they are a promising strategies that need to be more developed to generate new sophisticated models with higher performance.

Chapter 3

Bioinformatics and the use of Intelligent Computational Models in Biology

3.1 Introduction

The rapid development of the Biological and Computational techniques have generated surprising amounts of data which created great challenges for Computational Biology (CB).

Artificial Intelligence (Artificial Intelligence (AI)), the science refers to enabling computers to do things that require human intelligence has had long and complex interrelationships with Biology throughout its history. Biology represents an area of inspiration for many Artificial Intelligence processes. We can mention as examples Artificial Neural Networks (ANN) that have been designed as a mimic of the neuronal behaviour of the human brain, Genetic Algorithms (GA) that mimic the mutation and crossover operations in human genes, Grammatical Evolution (GE) that has tried to copy the process of making proteins, and so on.

On the other hand, advanced technologies in AI in their turn provided biology with effective solutions to its problems, especially the complex ones. A variety of Machine Learning (Machine Learning (ML)) and Data Mining (DM) techniques have been used to reveal and explain comprehensive and complex biological mechanisms. For AI, storing, organizing and mining the streams of biomedical and biological data is a fundamental challenge, where the need for new advanced Computational Biology techniques is evident.

In this Chapter we highlight one side of the efforts made in Intelligent Computational Technologies to address specific Biological problems related to Functional Genomics, for Genetic Diseases and Drug Repositioning. To do so we give a brief survey of the Hybrid Data Mining and Machine Learning techniques applied on Biomedical and Genomic data.

3.2 Computational Biology

Before the advent of computational biology, biologists have been unable to access large amounts of data. Computational Biology is the science of using developing algorithms and relations among different biological systems by using biological data. Bioinformatics is an interdisciplinary field that combines different specialities: computer science, statistics, mathematics, and engineering to develop models and software tools in order to understand, analyze and interpret Biological data (Hogeweg, 2011).In the beginning of the 1970s the term of Bioinformatics started to be used and developed. This field was defined as "the study of informatic processes in biotic systems". Artificial Intelligence research was exploring new techniques of knowledge representation in that same period inspired by biological systems (Hogeweg, 2011).

Computational Biology, this new discipline, has attracted much attention since 1990s, it has been become very important domain of development of emerging technologies dedicated to the biology field (Moody, 2004) and its areas, we cite (Elumalai and Eswaraiah, 2013):

- Experimental molecular biology area: from huge amounts of data, Bioinformatics techniques aim to find beneficial results.
- Genetics and genomics area: Bioinformatics techniques help in detecting genetic mutations and in sequencing and annotating genomes. These techniques were very efficient in understanding different aspects of molecular biology, as well as they are a key for the analysis of gene and protein expression and regulation, the text mining of biological literature and ontologies development, the analyze and catalogue of the biological pathways and networks, and so on.
- Structural biology area: Bioinformatics techniques assist the simulation and modeling of DNA, RNA, molecular interactions and protein structures.

Increasing the understanding of the biological process is the elementary objective of Bioinformatics, by the development and the application of computationally intensive

techniques. Main research efforts in the field of Bioinformatics and Computational Biology include, but are nor limited to, gene finding, genome assembly, sequence alignment, drug design, drug discovery, protein structure alignment, protein structure prediction, prediction of gene expression and protein–protein interactions, genome-wide association studies, the modelling of evolution and cell division/mitosis (Elumalai and Eswaraiah, 2013). Among the used techniques, we can cite : pattern recognition, visualization, data mining and machine learning algorithms for sequence alignment, genome wide association studies, gene expression, drug design and drug discovery.

3.3 Biological Databases

Biological databases are large collections of structured, organized, indexed and update biological data. This data come from different research areas like genomics, proteomics, metabolomics, microarray gene expression, and phylogenetics (Altman, 2004). In other words, biological databases are libraries of life science information, and, like some other databases, are usually associated with computerized software designed to update, query, and retrieve components of the data stored within the system. The data is represented in different formats, and include gene function, structure, localization, mutation, textual descriptions, attributes and ontology classifications, citations, and many other biological information, mainly collected from scientific experiments, published literature, high-throughput experiment technology, and computational analysis (Attwood et al., 2011).

Biological databases have two main functions (Per, 2001):

1. **The availability of the biological data to the scientific community:**

   A specific type of biological information should be available in the same locality as possible to make the access of the scientists easier. Collecting published data from literature is very time-consuming, and find and access to it may be difficult. Beside that, not all data is published in articles.

2. **The computer-readable form of the biological data:** Having the data in computer-readable form is very important and useful, which help in analysing the biological data that always involves computers.

Most Biological databases are available through web sites and are public for download. They are organized, as well, in ways that allow users to access and browse it online. The biological data can be in different formats such as text, sequence data, protein structure and links. They can be obtained from certain sources, some of them are described in the following sub-sections.

3.3.1 Simulated SNPs

The simulated SNPs data [1] described in the study *Grammatical Evolution Decision Tree (Grammatical Evolution Decision Tree (GEDT))* (Motsinger-Reif et al., 2010) is presented here where epistatic genetic models were generated with varying effect sizes (Cordell, 2002).

To represent epistatic genetic models the Authors used penetrance functions (defined in Section 1.5 "probability of disease given a particular genotype combination"). Single-nucleotide polymorphisms (SNPs) are the used genetic variations modelled. The researchers generated 100 datasets for each combination of effect size and genetic model (Figure 3.1).

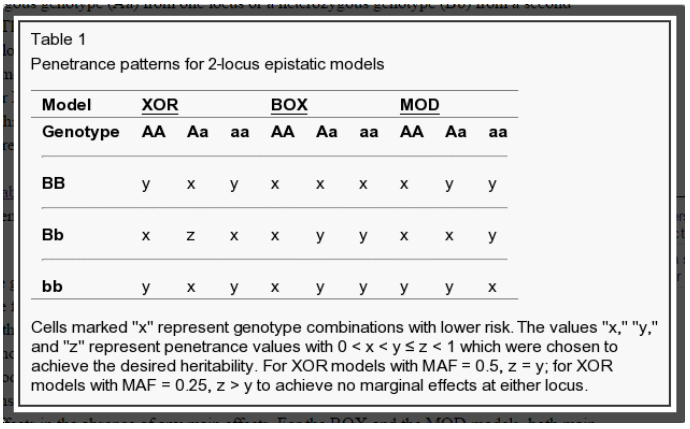


Table 1
Penetrance patterns for 2-locus epistatic models

Model	XOR			BOX			MOD		
Genotype	AA	Aa	aa	AA	Aa	aa	AA	Aa	aa
BB	y	x	y	x	x	x	x	y	y
Bb	x	z	x	x	y	y	x	x	y
bb	y	x	y	x	y	y	y	y	x

Cells marked "x" represent genotype combinations with lower risk. The values "x," "y," and "z" represent penetrance values with 0 < x < y ≤ z < 1 which were chosen to achieve the desired heritability. For XOR models with MAF = 0.5, z = y; for XOR models with MAF = 0.25, z > y to achieve no marginal effects at either locus.

FIGURE 3.1: Penetrance patterns for 2-locus epistatic models ((Motsinger-Reif et al., 2010))

Two interaction models with main effects and a model with absence of main effects were used. Models without main effects is a challenge for the method to find interactions in a complex dataset. Figure 3.1 presents the used general penetrance functions which are described as follow:

- The first model is the *XOR* model (described with an example in Section 1.5), for this model the authors modified the model initially described by Li and Reich (Li and Reich, 2000). The low risk of disease for this model correspond to inheriting a *heterozygous* genotype (Aa) from one locus or a *heterozygous* genotype (Bb) from a second locus, but not both.
- The second model is the *BOX* model, it is a symmetric two-locus interaction with main effects at both loci (described by Neuman and Rice (Neuman et al., 1992))Low risk of disease in this model correspond to inheriting two low-risk alleles at either one or both loci (AA and/or BB).

[1] This data was used for the Contribution in **??**.

Model Number	Heritability (%)	Minor Allele Frequency	Genetic Model
1	1	0.25	XOR
2	1	0.5	XOR
3	2.5	0.25	XOR
4	2.5	0.5	XOR
5	5	0.25	XOR
6	5	0.5	XOR
7	7.5	0.25	XOR
8	7.5	0.5	XOR
9	10	0.25	XOR
10	10	0.5	XOR
11	1	0.25	Box
12	1	0.5	Box
13	2.5	0.25	Box
14	2.5	0.5	Box
15	5	0.25	Box
16	5	0.5	Box
17	7.5	0.25	Box
18	7.5	0.5	Box
19	10	0.25	Box
20	10	0.5	Box
21	1	0.25	Mod
22	1	0.5	Mod
23	2.5	0.25	Mod
24	2.5	0.5	Mod
25	5	0.25	Mod
26	5	0.5	Mod
27	7.5	0.25	Mod
28	7.5	0.5	Mod
29	10	0.25	Mod
30	10	0.5	Mod

FIGURE 3.2: Simulated Models: Summary characteristics for the simulated models are listed, including the minor allele frequency, the heritability of the model, and the genetic model used (Motsinger-Reif et al., 2010).

- The third model is the *MOD* model, it has an asymmetric risk pattern that represents a modifying model on an *exclusive OR* function (Li and Reich, 2000).

Five different effect sizes (Heritability (He)) and two different minor allele frequencies (MAF) were used for each genetic model (*XOR*, *BOX*, and *MOD*). Genotypes were

generated according to Hardy-Weinberg proportions at two different allele frequencies, *0.25* and *0.5*. A summary of the characteristics of all the simulated models are presented in the Figure 3.2. GenomeSim software (Dudek et al., 2005) was used to simulate the data.

- The Minor Allele Frequencies (MAF) is the frequency of the least common allele in given population. In another word, it is the number of occurrence of the least common allele in a population of *"N"* individuals, divided by the total number of alleles in this population which equals to *"2*N" alleles* (each individual has two alleles). Allele Frequency is defined in Section 1.3.1
- The Heritability (He) is a statistical estimation that measures the proportion of variation in a phenotypic trait that can be due to genetic factors, i.e. genetic variation, in a population (Wray and Visscher, 2008). If the phenotypic variance is zero, all measurements are identical.

3.3.2 Gene Expression Omnibus (GEO)

The Gene Expression Omnibus (GEO) is an international public repository (Barrett et al., 2013). It contains microarray, next-generation sequencing (NGS) and other forms of high-throughput functional genomic data freely distributed. The database is built and maintained by the National Centre for Biotechnology Information (NCBI), a division of the National Library of Medicine, located on the campus of the National Institutes of Health in Bethesda, MD, USA. Data in GEO is original scientific publicly available for download in different formats. GEO has supporting data and links to almost 20 000 published manuscripts (Barrett et al., 2013). GEO provides also beside the public archive, tools to help users identify, analyse and visualize data relevant to their specific interests. These tools include a powerful search engine, sample comparison applications and gene expression profile charts. Many different attributes can be used to search in GEO, like keywords, organism, DataSet type and authors. The GEO database is growing up and is being actively developed to facilitate data mining and scientific discovery (Barrett et al., 2013).

3.3.3 DrugBank

In History, drug information were only found in books, journals and expensive commercial databases. Recently, the scientific community has known a radical change, where most drugs and their information are freely available over the internet (Knox et al., 2011).

DrugBank is described as "a richly annotated database of drug, drug target information and drug action information developed, maintained and enhanced by extensive literature surveys performed by domain-specific experts and skilled biocurators". (Law et al., 2014). DrugBank first released in 2006 (Wishart et al., 2006), it has data on nomenclature, ontology, chemistry, structure, function, action, pharmacology, metabolism and information on the target diseases, proteins, genes and organisms on which drugs act (Knox et al., 2011). For 2014 (Law et al., 2014) DrugBank has been improved to support the increasing quantities of Drug knowledge. All the information has been expanded and updated including drug structures and targets and others. Beside that, several new drugs have been added and new tools have been developed to facilitate more the process of finding information. The latest version 4.0 contains 7677 drug entries. Additionally, 4270 non-redundant protein (i.e. drug target/enzyme/transporter/carrier) sequences are linked to these drug entries. The quality of DrugBank data made it the referential drug data source and so popular in the rsearch scientific community including pharmaceutical researchers, medicinal chemists, clinicians, educators and the general public (Law et al., 2014).

3.3.4 Online Mendelian Inheritance in Man (OMIM)

Online Mendelian Inheritance in Man (OMIM) is a continuation of Dr Victor A. McKusick's Mendelian Inheritance in Man (MIM) (Amberger et al., 2009). OMIM is considered the elementary repository of genetic-phenotypes relationships and all their information (Amberger et al., 2015). OMIM has been online since 1987 after publishing around 12 editions between 1966 and 1998 (Amberger et al., 2015). This database has an important role in the classification of genetic phenotypes because of the rapid increase in the reports of gene-phenotype relationships (Amberger et al., 2015). OMIM.org provides a user easily searchable portal of literature to aid in clinical and molecular genetic research (Amberger et al., 2015). October 2014, OMIM is comprised of over 22,634 entries describing 14,831 genes and 7,894 phenotypes (Amberger et al., 2015), and this data is available for FTP download (Hamosh et al., 2002).

3.3.5 The Comparative Toxicogenomics Database (CTD)

Ten years ago, the Comparative Toxicogenomics Database (CTD) was developed in order to provide formalize, harmonize and centralize several genes and proteins information in response to environmental toxic agents through various species (Davis et al., 2014). The initial approach of CTD was to make the comparison of nucleotide and protein sequences of toxicologically significant genes easier using electronic annotation of these

entities with chemical terms from their associated references (Davis et al., 2014). After that, CTD has known a large extend in order to represent a triad of chemical–gene, chemical–disease and gene–disease interactions. Today, CTD includes 24 million toxicogenomic connections relating chemicals/drugs, genes/proteins, diseases, taxa, phenotypes, Gene Ontology annotations, pathways and interaction modules (Davis et al., 2015).

3.3.6 Kyoto Encyclopedia of Genes and Genomes (KEGG)

Kyoto Encyclopedia of Genes and Genomes (KEGG) is a collection of databases used for research in bioinformatics. It supports various biological entities such as genomes, biological pathways, diseases, medications and others. Among the researches carried out by bioinformatics are cited data analysis in genomics, modeling and simulation in systems biology, and translational research in drug development. The project of KEGG database was initiated in 1995 by Minoru Kanehisa, Professor at the Institute for Chemical Research, Kyoto University (Kanehisa, 1997) (Kanehisa and Goto, 2000). A knowledge base has been manually created in the forms of molecular networks called KEGG pathway maps, BRITE functional hierarchies and KEGG modules (Kanehisa et al., 2011). KEGG for consequence is considered as a reference knowledge base data widely used, as it integrate several biological information generated by genome sequencing and other high-throughput experimental technologies. Besid that, KEGG is growing to reach more practical applications that integrate human diseases, drugs and other health-related substances (Kanehisa et al., 2014).

3.4 The Need for Data Mining and Machine Learning Methods

Success in understanding the role of genomic variation and the role of environmental factors in disease susceptibility will help the improvement of the diagnosis, prevention and treatment. This depends mainly on the ability to address the mapping of genotype to phenotype and the sum of non-linearity which can arise from different phenomena (see Section 1.4.3).

One of the greatest challenges in the field of human genetics is the identification of genetic and environmental factors which cause susceptibility to common, complex diseases (Moore et al., 2010). Epistasis, or gene-gene interaction, is a well-known challenge that has given rise to the development of different linear model and parametric statistical approaches techniques (Moore and Williams, 2005).

The biggest disadvantage of these techniques is that, due to the complexity of the problem, they are not well suited to detect gene-gene interaction. A key reason for this decrease in performance of the statistical techniques in solving this problem is the high dimensionality of the data. This is due to either the large number of SNPs that get generated for these problems or the interactions that occur between more than two polymorphisms. To overcome the limitations of traditional approaches, data mining and machine learning techniques have widely been explored (McKinney et al., 2006) (Koo et al., 2013).

To successfully identify genetic variation associated with disease by genome wide approach, three challenges must be overcome which were reviewed in (Moore and Ritchie, 2004) :

1. Powerful data mining and machine learning methods will be needed to model the relationship between SNPs and environmental factors with disease susceptibility in a computational way due to the limitation of parametric methods. Traditional parametric statistical approaches lack the ability of modelling high-order non-linear interactions that are likely important in the aetiology of complex diseases (Moore and Williams, 2002).

2. The selection of the most relevant SNPs from thousands or millions that should be included in the analysis. Feature selection algorithms will show a great importance in GWAS because it is exhaustive to evaluate all the huge combinations of SNPs using modern computational techniques.

3. The biological interpretation of non-linear genetic models. Even when the computational models succeed to identify the SNPs related to a specific disease, this can not be translated to treatment without the interpretation of the results in the world of Biology. This translating process may be the most difficult challenge (Moore and Ritchie, 2004).

3.5 Intelligent Models in Genetic Association Studies

During the past decade, the knowledge and the understanding of disease genetics have been improved. Thanks to GWAS, thousands of SNPs have been associated with diseases and other complex traits.

In order to find associations between SNPs and a specific phenotype, Linear parametric statistical approaches rely, in general, on single-locus test. Yet, this is considered as

a simple approach unable to tackle the complexity of the underlying biological mechanisms. Due to the complexity of the non-linearity of genotype-phenotype relationships, Machine Learning and Data Mining approaches become strongly needed to overcome this shortcoming (Upstill-Goddard et al., 2013) (Niel et al., 2015). In recent years, several intelligent techniques have been developed and used on disease genetic data, facing some interesting challenges with significant success.

During the past decade, large amounts of biological data have been generated. A growing set of evidences have shown that complex interactions among genes have effective effects in human disease aetiology and may exert the predominant effect. The high dimensionality of the data and the complex interactions between two or more polymorphisms made traditional statistical methods inappropriate to detect gene-gene interactions. The consideration of all possible genotype combinations effect on the disease risk is a combinatorial challenge du to the large number of loci. Thus, Intelligent approaches are more sophisticated and flexible techniques that represent alternatives to statistical methods for detecting combinations of variants that are associated with a phenotype (Niel et al., 2015) (McKinney et al., 2006). Several intelligent methods have been used for gene-gene interaction detection, like GAs, ANNs, , Bayesian Networkss (BNs), Random Forests (RFs), Cellular Automata (CA), Multifactor Dimension Reduction (MDR), etc. Detecting and characterizing attribute interactions, as calling statistical Epistasis either, and identifying the classification variables that indicate the outcome prediction of a disease are well known addressed challenges by Data Mining and Machine Learning techniques (McKinney et al., 2006) (Moore et al., 2010) (Upstill-Goddard et al., 2013) (Niel et al., 2015).

We focus in this section on some hybrid models that have been used to detect and characterize the combination of genes through the SNP datasets. Mainly we interest in recent and current approaches that used Association Rule Mining, Artificial Neural Network and Evolutionary Computation techniques.

3.5.1 Dimensionality Reduction of SNP Data

Microrarry datasets have motivated a new research direction in Bioinformatics. Microarray data sets represent a great challenge for computational techniques, due to their large dimensionality and small sample sizes (Somorjai et al., 2003). Major efforts have been made to study the functional and structural of SNPs computationally (Mooney, 2005). This DNA sequence variation resulting from an alteration of a single nucleotide in the genome, is an important source of the human genome variability (Collins et al., 1998) (see Section 1.3.2). Numerous studies, like the field of GWAS, have shown that

SNPs may have important biological effects and have been implicated in and characterize several human diseases (Sachidanandam et al., 2001). In this context, machine learning and data mining techniques have been widely used to analyse SNP data (Chen et al., 2008), (He et al., 2010), (Schwender and Ickstadt, 2008).

SNP data is currently used in the development of effective algorithms for the classification of complex diseases. However, SNP datasets are characterised by their high dimensionality. They all tend to contain high levels of noise and be small in size. These factors make it difficult to develop an efficient classifier (Knudsen, 2011).

Despite the success reached by standard artificial intelligence techniques in genetic area as analysing gene expression data, analysing large-scale data with only single standard intelligent approaches become inefficient for the classification problem. As many pattern recognition techniques are not able to perform a good classification on a large amount of features, it is necessary to combine them to FS techniques. The most important objectives of feature selection for this area are (Saeys et al., 2007):

- Improving the performance of the model and avoid over-fitting,
- Providing faster and profitable models.
- Gaining a deeper insight into the generated data processes.

3.5.2 Combined Approaches for Genetic Diseases

Recently, due to their efficiency, hybrid intelligent approaches are becoming more popular in various fields.
EAs have been widely used in Bioinformatics for different problems such as Feature Selection and parameter estimation (Li et al., 2014b).

Recently, researchers have been inspired by the natural processes while designing novel optimization techniques used to select most informative features and tune parameters (McKinney et al., 2006). In keeping with the biological and genetic theme, approaches that simulate mutant, hereditary, hybrid and evolution in nature have been successfully applied in gene-gene interaction detection and research of gene regulatory networks. Evolutionary algorithms, such as GA and GP, are based on the mechanics of Darwinian evolution by natural selection. These algorithms have been applied in various bioinformatics issues, with the aim to find answers to some still unanswered questions. This can be possible by developing efficient models able to select, assemble, analyse, and interpret Microarray data.

The analysis of SNP data is a key to disease-gene association studies. Recently, due to their efficiency, hybrid intelligent approaches based on evolutionary algorithms are becoming more and more popular in various fields. Evolutionary learning methods have already been successfully used in different Microarray studies (Deutsch, 2003), (Jirapech-Umpai and Aitken, 2005).

One of the uses of GAs in bioinformatics is biomarker identification which is equivalent to feature selection. In general GA works as a wrapper method with a classifier. The selected features by the GA in each iteration are evaluated by the classifier to estimate the performance of the model (Li et al., 2014b). For example, Authors in (Hong and Cho, 2006) used a Simple Genetic Algorithm (SGA) and a Neural Network. According their experimentation results, they show that SGANN, the combined approach, were able to reduce the dimensionality more efficiency than SGA alone.

The hybrid models based EAs have been used with success to identify SNPs associated with diseases, beside the gene selection (Carlson et al., 2004), (Mahdevar et al., 2010). In (Anekboon et al., 2014) a practical scoring classification machine learning technique for feature selection based on a genetic algorithm was proposed. It was designed for case/control samples classification. A Multi-Layer Perceptron was used to construct a prediction model on the selection of predictive SNPs from a Crohn's disease data set. The authors showed that the proposed SNP prediction framework outperformed previously proposed techniques in terms of accuracy. Indeed, Selecting the best features that increase the performance of the classifier is also a challenge as this performance depends of detecting the relevant features and discarding irrelevant ones (Kesharaju and Nagarajah, 2015).

It is commonly accepted that many complex diseases such as cancer arise from complex interactions among multiple SNPs (Zhang et al., 2008). This is known as multi-locus interactions (Cordell, 2002). Different ensemble methods have been proposed to identify SNP-SNP interaction which may increase the classification performance of the disease of interest (Upstill-Goddard et al., 2013). Feature extraction is the key to good pattern recognition since even the best classifier will perform poorly if the features are not well selected.

Association rules are widely used in various areas and domains. More particularly, these techniques have proven their value in biological data analysis (Park and Kim, 2012). (Naulaerts et al., 2015) gives an overview of various algorithms of frequent itemset mining techniques , and illustrates how they can be used in several real-life Bioinformatics application domains. In (Mutalib et al., 2014), the authors have mined frequent patterns for genetic variants associated to diabetes. Based on their results

the generated patterns were informative enough to allow drawing relations between the reported risky SNPs and other unreported SNPs.

The application of frequent itemset mining have proven their value in wide range of knowledge extraction problems and was not restricted to market basket. Among these domains comes the Bioinformatics where the frequent itemset mining used to identify "biologically relevant patterns that can be interpreted in a biological context" and include (Naulaerts et al., 2015) : the interpretation of gene expression data, annotations, protein interaction networks, and biomolecular localization prediction.

Association rules can be used as classifier, such as classifying tumor and healthy samples (Cai et al., 2010) (Giugno et al., 2013) (Antonie and Bessonov, 2011), protein–protein interactions (Park et al., 2009), and many others.

Some techniques of Machine Learning function as a black box; like SVM; whereas Association rules based classifiers are considered as " whit-box" models because their prediction reasoning is transparent, and their performance is as good as SVM classifiers for Biological data (He et al., 2008) (Giugno et al., 2013) (Antonie and Bessonov, 2011) (Karabatak and Ince, 2009). The combination of association rule mining with other classification methods, such as SVM, ANN can significantly increase their accuracy (Tang et al., 2005) (Karabatak and Ince, 2009) (Naulaerts et al., 2015).

Several FS techniques were based on ARs (Chawla, 2010), (Wang and Song, 2012), (Xie et al., 2009). In (Karabatak and Ince, 2009) the authors used ARs to reduce the dimensionality of Wisconsin breast cancer data and ANN for classification. The method was applied on nine features, four of which were selected as best features by the ARs, and were passed on for classification by a ANN. The reported classification rate of the proposed system was 95.6%.

Traditional feature selection techniques tend to ignore the interaction between features (Zhao and Liu, 2009), while their combination may have a strong correlation with the target (Shen et al., 2014).

Nowadays, hybrid methods attracted more attention like combining two or more algorithms that are different in their concept to perform the feature selection and classification tasks. In (Halakou et al., 2015) the authors proposed three hybrid selection methods CNNFS, Ck-NNFS, and CRRFS. In these techniques, all Neural Network, k-Nearest Neighbours, and Ridge Regression were respectively injected in the wrapper phase as induction algorithms. The obtained results showed the performance of the proposed hybrid methods in addition to their dimensionality reduction ability in SNP selection.

Various dimensionality reduction approaches have been used to perform the selection of the most informative SNPs. In (Batnyam et al., 2013), the authors combined various existing techniques to find the most effective SNP data classification. The analysis was conducted in three stages: first, the selection of informative SNPs; second, the generation of an artificial feature from the selected SNPs; third, the classification task.

Among the most relevant problems in pattern recognition are validation protocols and input selection. The generation of the best sets of training and tests are represent the problem of interest, on the one hand, and the selection of the best features as input that can maximize the accuracy of the model in a blind test, on the other hand, are the two most important problems for machine learning. Training with Input Selection and Testing (TWIST) (Buscema et al., 2013) is an algorithm that was introduced to reduce the dimensionality of the data by selecting the most informative variables, in two generated subsets of data with a very similar probability density of distribution (see section **??**). The TWIST algorithm is designed for optimal training/test data splitting and variable selection (Buscema et al., 2013), used with neural networks with very promising results (Rotondano et al., 2011), (Coppedè et al., 2013), (Buscema et al., 2014), (Grossi et al., 2014), (Drenos et al., 2015), which prove the importance of the feature selection process for the improvement of the classification accuracy. The TWIST evolutionary system is usually composed of a population of Multilayer Perceptrons. Each ANN learns from the training set of data and is tested in a blind way on the test set of data. TWIST is an evolutionary algorithm based on Genetic Doping Systems (Buscema, 2004), already applied to medical data with very promising results (Coppedè et al., 2013), (Gironi et al., 2015), (Buscema et al., 2015).

Machine learning is an alternative approach for prediction and classification that can face the problem of dimensionality. One of the very promising machine learning techniques is Artificial Neural Networks (ANN) (Grossi and Buscema, 2007) which are widely used in many fields. The use of ANNs on genomic data has been explored in several studies.

Supervised pattern recognition methods for classification (Bishop, 1999) are currently applied across almost all research fields. In recent years, ANNs have been successfully used in medicine and in several gene-disease association studies (Sargent, 2001), (Lisboa, 2002), (Tomita et al., 2004), (Baldassarre et al., 2004), (Penco et al., 2005), (Lisboa and Taktak, 2006), (Grossi et al., 2007), (Motsinger-Reif et al., 2008a), (Buscema et al., 2010), (Rotondano et al., 2011), (Silva et al., 2013), (Upstill-Goddard et al., 2013), (Coppedè et al., 2013), (Drenos et al., 2015), and many others.

ANNs have widely been used in Bioinformatics (Manning et al., 2014). The reason for their popularity is their proven successful application to a number of challenging problems. Among the strengths of this technique in this field is its strong generalization ability, a much desired quality in many situations in Bioinformatics, domain knowledge does not need to be complete, and the robustness of their solutions and the complexity of the faced problems (Manning et al., 2014)..

The number of hidden layers of the NN and the number of neurons in each hidden layer can only be empirically defined and depend on the complexity of the problem. To date, Genome Wide Prediction (GWP) has used ANNs with a single hidden layer (González-Recio et al., 2014).

The comparison of ANNs with other techniques shows that it is not always the best for all problems. However, they are still a very powerful tool. Several papers have presented different new ANN models that avoid some drawbacks of classical ANNs like over-fitting and falling in local minima and identified where NNs outperform other machine learning techniques (Penco et al., 2005), (Buscema et al., 2006), (Grossi et al., 2007), (Buscema et al., 2014).

ANNs have also been developed and applied on genetic studies (Liu et al., 2004), and in SNP association studies (North et al., 2003), (Kooperberg and Ruczinski, 2005).

Cho and Ryu Cho and Ryu (2002) showed that the ANN Multi-Layer Perceptron performed as well as or better than the SVM approaches, while comparing it to two variations of the SVM in combination with a number of feature selection algorithms on gene expression profiles. In (Cho and Won, 2003) also, a MLP was shown to be better than a and a KNN algorithm in some cases and produced similar results in other cases for Leukemia, colon and lymphoma data sets.

In order to evaluate the ability of the combination of GP-optimized NN (GPNN) for detecting gene-gene interaction, authors in (Ritchie et al., 2003) performed simulation studies, where five different epistasis models were simulated, and GPNN was compared with the traditional backpropagation NN (BPNN). GPNN was able to model non-linear interactions as well as a traditional BPNN based on the analysis of only interacting genes, while it improved power and predictive ability compared with BPNN among noise datasets. These results indicate that this method combined the advantage of the two techniques, and the addition of EA successfully increased the power and the ability of ANN for detecting interactions and functional SNPs.

An additional research of GPNN has been processed in (Ritchie et al., 2004) and (Bush et al., 2005). The first one compared the GPNN methods associated with stepwise logistic regression and the second one compares GPNN and GP for the genetics of

complex human disease. In the two indicted studies, GPNN had higher power. This demonstrates that machine learning like NN network particularly where combined to Evolutionary Algorithm, could provide more power to detect epistasis over traditional statistical methods and could be considered as useful analytical methods for detecting genetic models.

Another promising Evolutionary Algorithm, known Grammatical Evolution (GE), was combined to machine learning and data mining techniques like Neural Networks (NNs) and Decision tree (). In (Motsinger-Reif et al., 2008b) an application of Grammatical Evolution Neural Network (GENN) to detect gene-gene interaction is presented. The author examined the power of GENN to detect interesting interactions in the presence of noise, and compared its performance to Multifactor Dimensionality Reduction (MDR). The generated results indicated that GENN is a promising method to detect gene-gene interaction, even in the presence of common types of error found in real data.

The results produced through GENN has given better results than GPNN, where GENN greatly outperforms GPNN in data sets with a large number of single nucleotide polymorphisms (Motsinger-Reif et al., 2008a). Moreover the analysis of Grammatical Evolution Decision Trees (GEDT) (Motsinger-Reif et al., 2010) has shown promising results in identifying interactions on simulated data.

3.6 Computational Models for Drug Repositioning

Despite the advances in technologies and Genomics area , the discovery of new drug has grown to be time-consuming and costly. To bring a new drug to the market it takes about 9–12 years and around billions of investment dollars (Dickson and Gagnon, 2009). Investments in pharmaceutical R&D (Research and Development) have steadily increased, while the number of new drug approvals has stagnated (Booth and Zemmel, 2004). The most important priority for pharmaceutical industry is to improving R&D productivity (Paul et al., 2010). To overcome this situation, a strategy known as Drug repositioning has emerged as an important key for new drug discovery and precision medicine paradigm (Shameer et al., 2015). Drug Repositioning could reduce the risks of development and the costs (Shaughnessy, 2011). It consist of finding and developing new clinical indications for existing drugs, or for those who are under development (Li et al., 2015).

The explosive growth of large-scale genomic and phenotypic data is allowing the development of new computational techniques for the purpose of drug repositioning (Li et al.,

2015). This strategy is an effective tool to find new uses from existing drugs. A recent survey is presented in (Li et al., 2015) and shows recent advancements of computational drug repositioning from multiple aspects.

In this section, we present some recent advancements in the critical areas of Computational Drug Repositioning based mainly on the intelligent techniques applying to biological datasets for the purpose of finding new pairs of Drug-Disease.

Several computational approaches have been developed to identify drug repositioning (Pujol et al., 2010) (Sardana et al., 2011). Some approaches are based on targets relevant to specific disease, this can represent a shared gene or feature (biological process, pathway, or phenotype) between a disease-disease, drug-drug, or a disease-drug (Wu et al., 2013) (Hurle et al., 2013)

In (Chiang and Butte, 2009) authors computed disease-disease similarity network to identify drug repositioning candidates, while (Yang and Agarwal, 2011), (Cheng et al., 2012) used drug-drug similarities, and (Fukuoka et al., 2013), (Gottlieb et al., 2011), (Keiser et al., 2009) used both disease-disease and drug-drug similarities.

In a recent review (Hurle et al., 2013) some rapidly developing computational methodologies for Drug Repositioning were covered. Some of the mentioned strategies were target-screening, genetics-based, phenotype-screening and text mining-based methods. One of the most common theme between some cited approach is to identify new link between drug and disease based on the identification of link between biological entities.

In recent years, an increasing number of machine learning methods have been proposed. A causal inference-probabilistic matrix factorization approach was used in (Yang et al., 2014) to infer drug–disease associations. The authors constructed causal networks connecting drug– target– pathway– gene– disease by integrating multilevel relations to predict novel drug–disease associations for Drug Repositioning.

PREDICT, a method that takes into account chemical, molecular and biological aspects of the drug-disease interactions, was presented in (Gottlieb et al., 2011) where authors represented separately drug-drug and disease-disease associations to train the machine learning Algorithm, they used logistic regression classifier to recognise real associations and the results were encouraging. (Napolitano et al., 2013) presented a machine learning method similar to PREDICT, the main difference is in using Support Vector Machine for this study.

Network-based analysis is one of the widely used strategies for computational drug repositioning. several studies have suggested that drug–target network, drug–drug network,

drug–disease network, protein interaction network, and others are useful in the identification of drug targets characteristics which provide new opportunities for drug discovery or repositioning (Li et al., 2015).

Each of the proposed computational drug repositioning strategies and approaches has its methodological advantages and limitations. A combination of these methods and a proposition of new other models is necessary to achieve better results. Despite several successful use cases of computational drug repositioning, challenges remain.

3.7 Conclusion

In this chapter a global overview of Bioinformatics and Computational Biology with some databases were presented, and a state of the art of the techniques used to addresses a specific Biological problems was given. The focus was specifically limited to Intelligent models developed to find gene-gene interactions for SNP datasets in complex diseases and the computational strategies for drug repositioning.

On the one hand, the knowledge about genetic variants has been extended, thanks to Genome Wide Association Studies (GWAS), influence the susceptibility to complex diseases.

Statistical techniques used to detect Epistasis suffer from a large limitation to cover the complex interactions because they test for each single-nucleotide polymorphism (SNP) separately. More powerful approaches are necessary to identify SNPs that influence disease risk jointly or in complex interactions, which were provided by techniques of Data Mining and Machine Learning. Experimental and simulated genome-wide SNP data provided by the genetic analysis afforded an opportunity to analyze the applicability and benefit of several machine learning methods. From the presented survey, it is noticed that machine learning and data mining approaches are promising techniques for genetics diseases and offer a high potential to analyse single-and multi-SNP interactions for complex human diseases. However, improved intelligent implementations and new variable selection procedures are required.

On the other hand, it is belied that computational drug repositioning research is of great significance to improve human health through discovering new uses for existing drugs. From the presented survey, we notice that a number of successful studies have widely been developed. However, a further investigation in this area will open more opportunities to accelerate drug discovery with interesting chances in several particular disease areas. Computational drug repositioning appears as a topic of growing interest

in the scientific community, and the research community should give it more attention to further develop techniques and methods toward new discoveries and breakthroughs.

The conclusions drawn from this chapter have oriented the work reported in this thesis and increased the motivation to explore functional genomics relationships for disease diagnosis and drug repositioning from the perspective of biological processes and the success of computational intelligent models. More precisely, the aim is to create intelligent models for extracting ""heading associations??"" among biological entities in order to reach a better performance for genetics relationships in complex diseases and Drug Repositioning.

Part III

Contribution

Chapter 4

Optimizing Association Rule Mining by Grammatical Evolution (GEARM)

4.1 Introduction

Several problems are encountered during the knowledge discovery process. The characteristics of Evolutionary Algorithms make them efficient to be used to solve such kind of problems. Motivated by the success of the use of GE with ANNs and , and by the fact that Association Rules (ARs) represent a promising technique for finding hidden patterns in a large data set we propose the use of GE to discover ARs. This combination yields the technique we have named GEARM for Grammatical Evolution Association Rule Mining. In this chapter we present the details of our proposed approach and we evaluate on several transactional datasets. The following chapters show the success of the use of this Algorithm to solve different problems in functional genomics.

4.2 The GEARM Technique

The GEARM algorithm is a proposal to extract association rules independently of any domain or problem. This algorithm makes use of GE to define interpretable individuals. These individuals are defined through the use of a Context Free Grammar (Context Free Grammar (CFG)). The technical details that explain the coupling of Grammatical Evolution with Association Rules using a CFG grammar are provided and the power of

the approach is evaluated by analyzing the use of the GEARM process to solve different problems in Bioinformatics.

In order to combine GE with ARM, we adapt the GE process to allow the automatic generation of valid rules. To this end, a suitable description in BNF of the CFG of the ARs must be generated. This grammar must specify the antecedents and the consequent of each rule, be consistent with the data it operates upon, and be geared towards the problem at hand.

4.3 A Grammar for Association Rule Mining

A grammar, as it was presented in Section 2.2.3.2, is defined by a set of production rules where each rule is of the form $A \Longrightarrow B$. The right-hand side (B) is a combination of terminals and/or non-terminals, whereas the left hand side contains only non-terminals. By applying the corresponding sequence of association rules, the non-terminals are eventually substituted by terminals, which are the final (atomic) elements that appear in the language.

More formally, a Context-Free Grammar is defined as a quadruple (S, N, T, P), where :

- S is the start symbol,
- N is the set of non-terminal symbols,
- T is the set of terminal symbols, and
- P is the set of production rules.

A General Grammar to extract ARs contains production rules of the form $A \Longrightarrow B$ where $A \in N$ and B $\in \{N \cup T\}$, and it is represented in the form:
$G = \{S, N, T, P\}$
$S = \{Rule\}$
$N = \{Rule, Antecedent, Consequent, VAR, VAL\}$
$T = \{VAR_1, VAR_2, ..., VAR_n, VAL_1, VAL_2, ..., VAL_m, Separator\}$
$P = \{< Rule >::=< Antecedent > Separator < Consequent >$
$< Antecedent >::=< VAR >< VAL > | < VAR >< VAL >< Antecedent >$
$< Consequent >::=< VAR >< VAL > | < VAR >< VAL >< Consequent >$
$< VAR >::= VAR_1|VAR_2|...|VAR_n$
$< VAL >::= VAL_1|VAL_2|...|VAL_m\}$

Each problem solution using GE consists of the mapping from the genotype to the phenotype. In the case of rule extraction, the two components of the solution are described as:

- a genotype, represented by a string in Grammatical Evolution,and
- a phenotype, that represents the complete rule consisting of an antecedent and a consequent.

The following rule represents the general structure of an association rule that is used in the GEARM process **If VAR1 = VAL2 and VAR4 = VAL1 then VAR3 = VAL2**.

"Separator" is a chosen character in the grammar used to separate the antecedence and the consequent of the rule.

According to the process of GE presented in section 2.2.3.2, more precisely the mapping process from *genotype* to *phenotype*, we give in follow an example to illustrate how GE is applied to generate an AR using the grammar defined in this section and the *MOD* operation presented in equation 2.1.

4.4 Illustrative Example of Rule Extraction

Let us illustrate here through an example, the mapping process from a genotype (represented as a vector of integer values) to the phenotype (association rules) using the above grammar and the equation 2.1. We assume that we have 4 variables with 3 possible values. Consider for instance the (input) vector *25,12,17,32,75,3,6,10,8*. The start symbol $< Rule >$ produces two non-terminals separated by *"SEP"* $< Antecedent > SEP < Consequent >$. The first non-terminal $< Antecedent >$ has two different alternatives, $< VAR >< VAL >$ and $< VAR >< VAL >< Antecedent >$. Using the first value of the input vector and by applying the MOD operation on the number of alternatives we obtain *25 MOD 2 = 1*. The result of the MOD operation represents the number of alternatives which will replace the current non-terminal. Since the $< VAR >< VAL >$ alternative is numbered as number 0, the non-terminal $< Antecedent >$ will be replaced by $< VAR >< VAL >< Antecedent >$ (alternative number 1 which is 25 MOD 2). The next non-terminal is $< VAR >$ (with 4 alternatives) and the next value in our vector is *12*; the process goes on until no non-terminal is left. The full example is presented in the following steps: (Antecedent= Ant, Consequent= Cons, Separator= SEP)

- $< Ant > SEP < Cons > \Longrightarrow 25,12,17,32,75,3,6,10,8 \Longrightarrow 25\ MOD\ 2 = 1$
- $< VAR >< VAL >< Ant > SEP < Cons > \Longrightarrow 12,17,32,75,3,6,10,8 \Longrightarrow 12\ MOD\ 4 = 0$
- $VAR_1 < VAL >< Ant > SEP < Cons > \Longrightarrow 17,32,75,3,6,10,8 \Longrightarrow 17\ MOD\ 3 = 2$
- $VAR_1\ VAL_3 < Ant > SEP < Cons > \Longrightarrow 32,75,3,6,10,8 \Longrightarrow 32\ MOD\ 2 = 0$
- $VAR_1\ VAL_3 < VAR >< VAL > SEP < Cons > \Longrightarrow 75,3,6,10,8 \Longrightarrow 75\ MOD\ 4 = 3$
- $VAR_1\ VAL_3\ VAR_4 < VAL > SEP < Cons > \Longrightarrow 3,6,10,8 \Longrightarrow 3\ MOD\ 3 = 0$
- $VAR_1\ VAL_3\ VAR_4\ VAL_1 SEP < Cons > \Longrightarrow 6,10,8 \Longrightarrow 6\ MOD\ 2 = 0$
- $VAR_1\ VAL_3\ VAR_4\ VAL_1\ SEP < VAR >< VAL > \Longrightarrow 10,8 \Longrightarrow 10\ MOD\ 4 = 2$
- $VAR_1\ VAL_3\ VAR_4\ VAL_1\ SEP\ VAR_3 < VAL > \Longrightarrow 8 \Longrightarrow 8$ MOD 3=2
- $VAR_1\ VAL_3\ VAR_4\ VAL_1\ SEP\ VAR_3\ VAL_3$

The generated AR is then:
If $VAR_1 = VAL_3$ **AND** $VAR_4 = VAL_1$ **THEN** $VAR_3 = VAL_3$

4.5 Performance Evaluation of GEARM in Transaction Databases

In this chapter we evaluate the performance of our proposed approach on different databases, with different transactions and item sizes, frequently used in data mining, and we compare our results to well-known exact and optimized techniques used for the extraction of ARs as reported in the literature according to the fitness function and the CPU time.

4.5.1 Grammars for Transaction Databases

Two types of problems are addressed. We aim to find ARs in two different kinds of databases: two-valued categorical data and numerical data.

The former, consists in finding associations in a table that has an attribute corresponding to each item and a record corresponding to each transaction. The value of an attribute

for a given record is "*1*" if the item corresponding to the attribute is present in the transaction corresponding to the record and " *0* " otherwise. In order to optimize the set of data, we can formulate the problem in another way, where each transaction contains only the number corresponding to the item present in it; this means that the size of each transaction is equal to the number of its items. We refer to this problem as the ***Boolean Association Rule (BAR) problem***.

The second type of problem consists in finding associations in a table of numerical attributes used in most business and scientific domains. Attributes are quantitative, i.e where each item can have one of its possible values for each transaction. We refer to this mining problem as the ***Quantitative Association Rule (QAR) problem***.

The GEARM algorithm has been designed to fit any type of dataset for any ARs problems; it can be applied for mining associations in Boolean, Quantitative, Categorical and Mixed databases, thanks to the grammar that gives this hybrid model such flexibility. To be able to generate admissible ARs, GE will facilitate the task by defining an appropriate grammar. We define for each type of data its corresponding quadruple (S; N; T; P) of the grammar as explained now.

4.5.1.1 A Boolean Association Rule (BAR) Grammar

The grammar used for this type of problems will specify the presence or the absence of each Item in the rule in a specific representation. An example of such a rule is: ***If*** I_1 ***and*** I_3 ***Then*** I_5. The grammar corresponding to the generation of BARs is defined as follows:

$S = \{Rule\}$

$N = \{Side, Item\}$

$T = \{I_1; I_2; \ldots; I_n\}$

$P = \{< Rule >::=< Side > SEP < Side >$

$< Side >::=< Item > | < Item >< Side >$

$< Item >::= I_1|I_2|\ldots|I_n\}$

4.5.1.2 A Quantitative Association Rule (QAR) Grammar

The grammar used for this type of problem will define both the antecedent and the consequent as a set of Items and their Values. For each Item the grammar will specify all the possible values it can take. This way, no item risks to get an abnormal value. An example of such a rule is: ***If*** $I_1 = V_{12}$ ***and*** $I_3 = V_{34}$ ***Then*** $I_5 = V_{51}$. The grammar

corresponding to the generation of QARs is defined as follows:
$S = \{Rule\}$
$N = \{Side, Item, Item_1; Item_2; \ldots; Item_n; Val_1; Val_2; \ldots; Val_n\}$
$T = \{I_1; I_2; \ldots; I_n; V_{11}; V_{12}; \ldots; V_{1i}; V_{21}; V_{22}; \ldots; V_{2j}; \ldots; V_{n1}; V_{n2}; \ldots; V_{nk}\}$
$P = \{< Rule >::=< Side > SEP < Side >$
$< Side >::=< Item > | < Item >< Side >$
$< Item >::=< Item_1 > | < Item_2 > |. . . .| < Item_n >$
$< Item_1 >::= I_1 < Val_1 >$
$< Val_1 >::= V_{11}|V_{12}|. . . |V_{1i}$
$< Item_2 >::= I_2 < Val_2 >$
$< Val_1 >::= V_{21}|V_{22}|. . . |V_{2j}$
. . . . . .
$< Item_n >::= I_n < Val_n >$
$< Val_n >::= V_{n1}|V_{n2}|. . . |V_{nk}\}$

4.5.2 Encoding of a Solution

The solution is encoded depending on the type of the rule. The two kinds considered here are represented as a string i.e vector of items representing the antecedent and the consequent of the rule separated by the key word '*SEP*'. The items that appear before the separator are considered the antecedent part of the rule, and the items that appear after the separator are the consequent part of the rule. The difference between the representations of the two types of rules is in the size of each type.

4.5.2.1 Boolean Association Rule (BAR) Encoding

Each solution S is represented as a vector of size "*n+1*", where "*n*" is the number of items present in the current rule, and the $(n+1)^{th}$ element is a box for the separator character.

- *S[i]= x* means the Item 'x' (I_x) appears in the rule.

Example: Let Items = {Milk, Bread, Coffee, Butter} for the market basket problem.

- *S1= [1, 3, SEP, 2, 4]* represents the rule: ***if Milk and Coffee Then Bread and Butter***.
- *S2= [1, SEP, 3]* represents the rule: ***if Milk Then Coffee***.

4.5.2.2 Quantitative Association Rule (QAR) Encoding

Each solution S is represented as a vector of size "*2n+1*" where "n" is the number of items present in the current rule and is multiplied by *2* because each item has a specific value, and the $(2n+1)^{th}$ element is a box for the separator character.

- *S[i]*= x means that Item 'x' (I_x) appears in the rule
- *S[i+1]*= V_{xy} means that Item I_x takes the value 'y'

Example: Let Items = { I_1, I_2, I_3}

- I_1 can take values in { 1, 0.5, 3 }
- I_2 can take values in { 0, 2 }
- I_3 can take values in { 0.2, 0.5, 0.01, 1.2 }

S= [1, 0.5, 2, 0, SEP, 3, 1.2] represents the rule: ***if*** I_1 ***= 0.5 and*** I_2 ***= 0 Then*** I_3 ***= 1.2***.

4.5.3 The Evolutionary Operations

The crossover and mutation are the two evolutionary operations used in this study. They are applied at the genotype level i.e at the vector of integers rather than at the phenotype level i.e at the AR. The advantage of applying these operations at the genome level is to avoid the different cases of crossover and mutation which depend on the type of the rule and may create conflicts and thus generate inadmissible rules. For example, if we use quantitative rules, lots of strategies of crossover and mutation can be considered, such as which element will be mutated, items or values.

Performing the evolutionary operations on the genotype level may solve many problems and can be applied on any type of rules, where after generating a new vector (by crossing and mutating the integers of this vector); the 'MOD' operation and the grammar ensure the generation of admissible solutions without having to consider the different possible cases of the operations that depend on the rule type.

For the crossover operation used in this work, given two individuals, we create two children using One-point Crossover with a probability "*P-Cros*" and return them. For the mutation operation, we mutate an individual with Int Flip Mutation by randomly choosing a new int with probability "*P-Mut*".

4.5.4 Fitness Function

Each generated Association Rule "R" is evaluated on the dataset and its fitness gets recorded according to Equation 4.1 , where "a" and "b" are two weight parameters to set by experimentation according to the importance we give to the confidence and support respectively. *Sup(R)* is the support of the rule R, and *Conf (R)* is the confidence of the rule R.

$$Fitness(R) = (a * Sup(R)) + (b * Conf(R)) \tag{4.1}$$

4.6 The GEARM Process and Algorithm

A detailed description of every structural block of the GE can be found in (O'Neill and Ryan, 2003), and it was briefly presented in section 2.2.3.2. GE is an evolutionary computation technique inspired by the mechanism of protein generation. It is based on a BNF grammar and a mapping process. It can be combined to different intelligent techniques in order to solve problems faced in stages of knowledge discovery process, such as extraction of AR. The different steps of the GEARM process that we introduce here are described as follows:

1. GEARM has a set of parameters that must be initialized:
    - **Population_Size**: the number of integer vectors (genotype) that will get generated randomly.
    - **Max_Generation**: the number of generations (iterations) performed by the process.
    - **Crossover_Rate**: the rate of the evolutionary operation crossover.
    - **Mutation_Rate**: the rate of the evolutionary operation mutation.
    - **Codon_Size**: the maximum values in the integers (codon) of the generated vectors.
    - **Wrap_Count**: the number of times the mapping process will wrap around the vector.
    - **Min_Chrom_Size**: the minimum size of the chromosome (the vector on the integers).
    - **Max_Chrom_Size**: the maximum size of the chromosome (the vector on the integers).

2. GEARM process begins by generating an initial population of N random individuals, where each individual is represented as a vector of integer values. The genotype-to-phenotype mapping process uses the defined grammar and always begins with the Start symbol. Each vector is mapped to an association rule by using the grammar and the *MOD (modulus)* operation which amounts to selecting the appropriate production rule p (from the set P) that will replace the current Non-Terminal.

   The production rule is selected using the formula represented in Equation 2.1 , where P-rule is the index of the selected production rule and nb-al is the number of alternatives (i.e. rules) defining the current non-terminal (see Section 2.2.3.2).

   - *P-rule = (Value) MOD (nb-al)*

   If the end of the genome is reached and the mapping process is still incomplete, then the genome is wrapped over and the integers are read again from the start of the vector. The wrapping process continues T times, where T is a predefined upper limit. If this limit is reached or if all the non-terminals are replaced, then the mapping process terminates.

3. The resulting output string then determines the set of N association rules where each individual in the initial population (genotype) is mapped onto an association rule (phenotype). Each association rule R is evaluated and its fitness gets recorded.

4. The best N-rule solutions are selected for crossover and reproduction. The crossover and mutation operations are performed at the chromosomal level (the vector of integer values), not at the level of the association rules. The new generation that gets generated, containing the best rules and equal in size to the original population, is used in the cycle time again until some criterion is met, after which GEARM stops. This criterion is either that the error is zero or a limit on the number of generations is reached.

5. The best solution is identified after each generation. At the end of the GEARM evolution, the overall best solution is selected as the optimal AR set. For some tasks (like classification), the best GEARM set is tested on a data left out to estimate the prediction error.

Since the process of finding association rules returns many rules, the definition of a good measure of fitness is necessary in order to ease this burden. There are different fundamental criteria to evaluate the quality of an extracted Association Rule, where the commonly used measures were presented in Section 2.2.1. The fitness function used in our contributions is defined by combining these evaluation measures.

The GEARM Algorithm

Algorithm 1 formalises the GEARM process described below. The *Non-terminal* symbols are substituted by *Terminal* symbols that represent the elements of the generated AR. The mapping process is ensured by the 'MOD' operation and the BNF grammar.

1. **Algorithm 1: GEARM**
2. **Input**: Transactional Database (DB)
3. **Output**: Best Set of ARs (ARs Set)
4. **Begin**
5. T= 0
6. Generate a number *Pop_Size* of vectors of integers */*Pop_Size* of individuals
7. **while** Generation <= Max_G **do** :
8. **for** each vector **do** :
   / The Mapping process using the defined Grammar*
9. i= 0
10. Codon = vector[i]
11. **while** AR is Non-terminal Symbols **and** $T \preccurlyeq$ Wrap_Count **do**:
12. P = Codon MOD Nb-Al
13. AR = replace the Non-terminal Symbol by P
14. i= i+ 1
15. Codon = vector[i]
16. **If** (end of vector)
17. Codon = vector[0]
18. T=T+1
19. **End If**
20. **end while**
21. AR = Terminal Symbols /* All the Non-terminal Symbols substituted by Terminal Symbols

22. f= Fitness (AR)
23. **end for**
24. Select best Rules for Crossover and Mutation according to their Fitness values
25. **end while**
26. ARs Set = Best N rules according to Fitness values /* $N \preccurlyeq Pop\_Size$
27. **Return**(ARs Sets)
28. **End**

4.7 Experimental Study

4.7.1 Description of The Datasets

Several scientific databases which are frequently used in data mining were consulted (Guvenir and Uysal, 2000) (Goethals and Zaki, 2003) in order to perform tests on them. Table 4.1 presents the description of the different data sets used to evaluate our proposed technique. From this table we notice that the data sets differ according to the number of transactions and the number of items. Some these datasets are very large but with a small number of items per transaction, while others are small datasets with a significant number of items per transaction.

4.7.2 Evaluation of The Results

In order to evaluate the performance of our proposed approach, we have applied GEARM to the datasets described above and compared it to several other successful approaches. The results presented are the best results obtained over several executions using different parameters. Table 4.2 summarizes the resulting fitness obtained by executing GEARM on small datasets and its comparison to *BSO-ARM* (ARM algorithm based on an improved version of BSO), *ARMGA* (an improved algorithm based on GA to identify ARs), *G3APARM* (generating ARs using grammar guided genetic programming) and *ACO* (studies based on ACO for serial ARM) from the literature (Djenouri et al., 2014). The objective is to maximize this fitness used by all the mentioned approaches (equation 4.1). We notice that *ARMGA* and *ACO* always give less performing results compared to our algorithm, whereas *GEARM* performs better than *BSO-ARM* in some datasets and gives comparable results for the others. *G3APARM* has a high performance as well but always less than *GEARM* for all the considered datasets. Our average fitness for

TABLE 4.1: Datasets Properties

Dataset	Number of Transactions	Number of Items
Bolts	40	8
Sleep	62	8
Pollution	60	16
Basket ball	96	5
Quake	2,178	4
Chess	3,196	75
Mushroom	8,124	119
Pumbs star	20,819	7,116
Korasak	80,769	7,116
Retail	88,162	16,469
Connect	100,000	999
WebDocs	1,692,082	5,267,656

this small data does not go below 0.80.

Table 4.3 summarizes the fitness obtained by *GEARM* and its comparison to *BSO-ARM* and *ARMGA* on bigger datasets from the literature. We can clearly see that *GEARM* outperforms so far the two techniques for the *Pumb* star and *Korasak* and *Connect* datasets which have a large number of transactions and items. For the *Retail* dataset we can see that *GEARM* has the lowest performance but it is almost close to the two others.

Table 4.4 shows the CPU time (in second) required by our proposed *GEARM* algorithm compared to the two popular exact algorithms *Apriori* and *FP-Growth*. The run time with the different datasets clearly varies according to the number of transactions and items.

From Table 4.4 we can see that *GEARM* by far outperforms the exact algorithms in terms of CPU time. Furthermore, the CPU runtime resulting from the execution of our approach does not exceed 1300 seconds while *Apriori* and *FPGrowth* reach 4500

TABLE 4.2: FITNESS COMPARISON ON SMALL DATASETS

Dataset	GEARM	BSO-ARM	ARMGA	G3APARM	ACO
Bolts	0.991	1.00	0.30	0.92	0.69
Sleep	0.990	1.00	0.26	0.90	0.67
Pollution	0.990	1.00	0.45	0.92	0.66
Basket ball	0.990	0.97	0.40	0.93	0.61
Quake	0.957	1.00	0.39	0.90	0.73
Chess	0.990	0.88	0.26	0.86	0.30
Mushroom	0.980	0.75	0.30	0.85	0.10

TABLE 4.3: FITNESS COMPARISON ON BIGGER DATASETS

Dataset	GEARM	BSO-ARM	ARMGA
Pumb star	0.790	0.40	0.25
Korasak	0.739	0.35	0.32
Retail	0.241	0.36	0.28
Connect	0.375	0.26	0.14
WebDocs	Blocked	Blocked	Blocked

and 3800 seconds, respectively. Based on these experiments, it can be noticed that for *GEARM* the number of transactions has more impact on the runtime than the number of items. However, for the other techniques, although the number of transactions on the *Connect* dataset is greater than that on the *Korasak* dataset, their runtime on *Connect* is less than on *Korasak* which is the opposite for *GEARM*. This can be explained that our approach uses a strategy to avoid going through all the items of the dataset, focusing only on the items present in the rule by a direct access to them in the dataset. This helps reduce the execution time in a remarkable way. An example is the Retail data which has the highest number of items but the CPU time of *GEARM* is less than for the *Connect* data because of its very large number of transactions, which represents the most influential factor for the *GEARM* technique.

Figure 4.1 shows the fitness obtained by the *GEARM* algorithm according to the number

TABLE 4.4: CPU TIME COMPARISON WITH EXACT ALGORITHMS (SEC)

Dataset	GEARM	Apriori	FP-Growth
Pumb star	250	500	600
Korasak	400	3000	2900
Retail	956	4500	3800
Connect	1276	2600	2900

of transactions. From this figure, we can clearly see that the fitness decreases as the number of transactions increases, while for the small datasets the fitness is almost 1. This shows that the number of transactions highly influences the quality of the solution.

Figure 4.2 gives the *GEARM* fitness according to the number of items. It is clear that the fitness does not much depend on the number of items, since for some large number of items the fitness is higher than for some others with a small number of items.

Figure 4.3 shows the CPU time variation with respect to the number of transactions. In this case it is clear that as the number of transactions increases, so does the execution time.

Figure 4.4 shows the variation of the CPU time with respect to the number of items. Once again, we notice that there is no constant relationship between the items size and the execution time, since in some cases the small number of items corresponds to the highest CPU time.

To test the limits of our proposed model, we have used a large benchmark *WebDocs*. On this dataset the approach is blocked as shown in Table 4.3 . This can be explained first by the huge number of transactions on the dataset (as, just explained, the number of transactions influences the fitness and the CPU, i.e the solution quality). Therefore, parallelization can be a solution to bring more efficiency to *GEARM*.

These experiments show the necessity to use the optimization methods instead of the exact ones, and the high performance of *GEARM* compared to the aforementioned optimization approaches. This has motivated us to improve *GEARM* and to see it as a promising technique for ARs extraction that can be used in different domains to solve several complex problems.

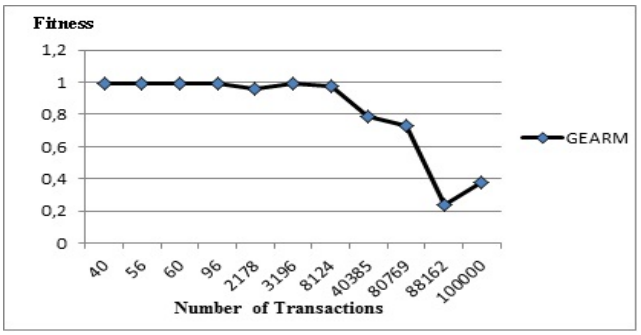


FIGURE 4.1: Fitness values according to the number of transactions in databases

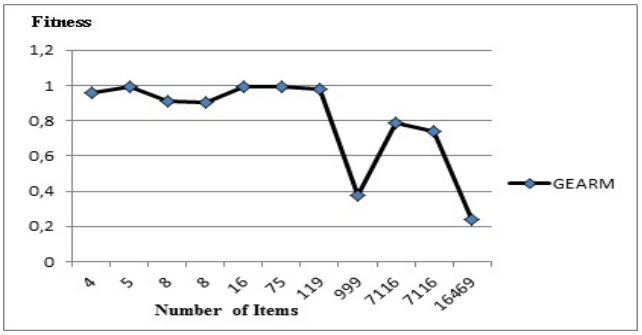


FIGURE 4.2: Fitness values according to the number of items in databases

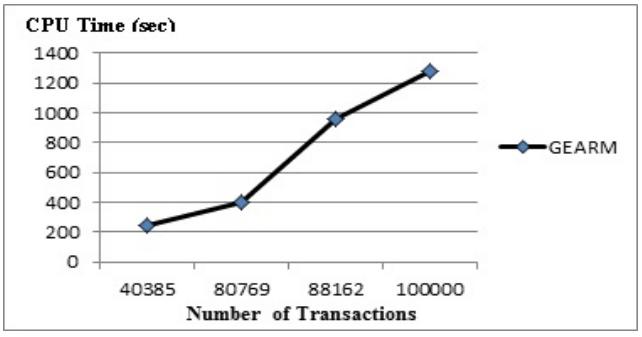


FIGURE 4.3: CPU time according to the number of transactions in databases

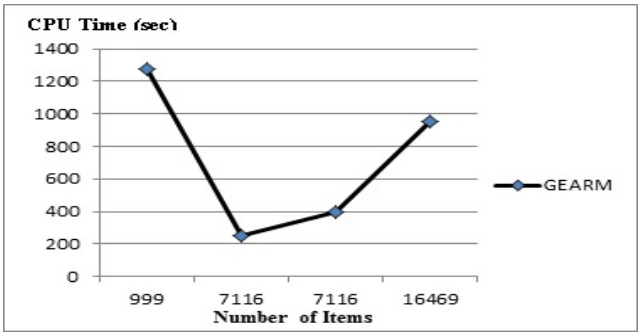


FIGURE 4.4: CPU time according to the number of items in databases

4.8 Conclusion

In this Chapter, a new hybrid Association Rules Mining algorithm based on Grammatical Evolution optimization named GEARM was introduced and evaluated. The extraction of valid rules is ensured by the grammar that guides the process that allows to move from the genotype to the phenotype. To prove the effectiveness of our suggested algorithm, it was tested on various datasets that are frequently used in data mining, and compared to different optimized and exact approaches. The results showed that our paradigm outperformed some existing optimized algorithms in terms of solution quality. Moreover, it was proved that the performance of the optimized techniques ensures a gain in CPU time compared to the exact approaches.

Our proposed Algorithm GEARM was designed to be used in different domains and to solve different problems. The use of GE evolution optimizes the extraction of ARs and enhances its flexibility; one only needs to change the grammar according to the used data to fit the problem at hand.

Within our work, we are interested in finding functional genomic relationships so as to solve problems related to Diseases Diagnosis and Drug Discovery. To this end, GEARM was combined to other powerful Data Mining and Machine Learning techniques to create strong intelligent models that have been applied on genomics data to attain the desired objective.

The following chapters present the use of hybrid versions of the basic Algorithm for applications related to genetic complex disease and drug repositioning and prove the success of our techniques through promising results.

Chapter 5

Extracting Functional Genomics Relationships for Disease Diagnosis

5.1 Introduction

In this Chapter we present hybridization variation of our GEARM algorithms, described in the previous chapter, and their applications to solve the problem of classification and dimensionality reduction using SNP datasets as biological markers for disease diagnosis.

First, GEARM is applied on simulated data for the detection of gene-gene interaction and functional SNPs related to a specific disease. The data used represent different Epistasis models that helped us show the good performance of our approach to deal with these kinds of problems.

Then, GEARM is combined to Neural Networks and is tested on real SNP data related to breast Cancer. The results presented show the high performance reached by NN-GEARM which uses GEARM for the selection of the best features and a NN for classification.

To get better results and improve our proposed model, NN-GEARM is optimised with a Genetic Algorithm which is used for setting the parameters of GEARM which is used for dimensionality reduction, and NN, responsible for pattern recognition. GA-NN-GEARM is applied on four SNP datasets related to Autism, mental Retardation, Colon Cancer and Breast Cancer. The results we have reached are compared to other techniques that had a good performance on the same datasets and clearly show that GA-NN-GEARM out performed them on the basis of to the quality of the generated solution.

In the following sections, we go through the different contributions point by point and describe the different proposed models in detail.

5.2 Detecting Gene-Gene Interaction using GEARM for Case/Control Classification in Simulated Data

The identification of the variation on DNA sequence that increase or decrease the susceptibility to particular disease is a major goal of human genetics. Complex interactions among genes and environmental factors are known to play a role in common human disease etiology. As explained in the preceding chapters, methods for ARM are highly successful; especially that they produce rules which are easily interpretable. This has made them widely used to solve various problems in genomics. GEARM, the approach we have developed and described in the previous chapter, is used on simulated data that represents Epistasis models to solve the problem of gene-gene interaction. We show that it approach was able to improve the performance of gene-gene interaction detection by reaching a high accuracy and finding the functional SNPs.

5.2.1 The Simulated SNP Dataset

To solve the problem of gene-gene interaction, we have applied GEARM on simulated data that represents Epistasis models. It was used for the GENN (Motsinger-Reif et al., 2008a) and GEDT studies (Motsinger-Reif et al., 2010), and described in Section 3.3.1.

Individual	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	.....	SNP100	Class
Indiv 1	1	1	0	2	0	2	.....	1	0
Indiv 2	2	0	1	0	1	2	.....	2	1
Indiv 3	0	2	2	1	0	2	.....	0	1
.....	.....	.....	.....	.....	.....	.....	.....	.....	....
Indiv 250	1	2	2	1	0	1	.....	0	1

FIGURE 5.1: Simulated SNP Data Set.

The data are stored in rows, where each row represents an individual and each individual is formed of 100 different SNPs and the class it belongs to. Two of the SNPs are associated with the outcome. The number of individuals is 250 (125 fo case and 125 for control) Three simulated genetic models have been used (XOR, BOX, and MOD) with different He and MAF. The XOR function exhibits interaction effects in the absence of any main effects. For the BOX and the MOD models, main and interaction effects are both observed (Motsinger-Reif et al., 2010). Figure 5.1 represents the format of the data set which is used.

5.2.2 GEARM for Epistasis

GEARM was used to process the genetic datasets to solve the problem of Epistasis detection. As explained in Chapter 4, we adapt the GE process to allow the automatic generation of valid rules. To this end, a suitable BNF description of the association rules must be generated. This grammar must specify the antecedents and the consequent of each rule so that it be consistent with the SNP data.

5.2.2.1 A Grammar for Classification

For genetic association data, the antecedents of a rule represent genotypes at specific loci, where a genotype can take one of three genotype values for a bi-allelic SNP (AA, Aa, aa), encoded as 0, 1, and 2, respectively. The set of variables and their values represent the antecedent part of the association rule. The consequent of the rule (class variable) can take one of two values, either positive '1', (for case), or negative, '0' (for control), states. Each individual is associated with case/control. All the elements that have a static form, meaning that they will not be substituted, are identified as terminals. Thus the grammar used for the genetic association data is as follows:
$G = \{S, N, T, P\}$
$S = \{Rule\}$
$N = \{Rule, Antecedent, Consequent, SNP, VAL\}$
$T = \{SNP_1, SNP_2, ..., SNP_n, 0, 1\}$
$P = \{< Rule >::=< Antecedent >< Consequent >$
$< Antecedent >::=< SNP >< VAL > | < SNP >< VAL >< Antecedent >$
$< Consequent >::= 0|1$
$< SNP >::= SNP_1|SNP_2|...|SNP_n$
$< VAL >::= 0|1|2\}$

Here is an example of rule which illustrates the general structure of an association rule that is generated by the GEARM process for the problem of gene-gene interaction:
If SNP1 = 2 and SNP4 = 0 then class = 1

5.2.2.2 Classification Rule Evaluation

The process of evaluating each individual (rule) is performed by calculating the value of the fitness function. The rule evaluation function must not only consider the instances that are correctly classified but also the ones not classified and those incorrectly classified. Thus four possible concepts are relevant (as presented in (Sousa et al., 2004)):

- **True Positive (TP)**: The number of instances covered by the rule that are correctly classified.
- **False Positive (FP)**: The number of instances covered by the rule that are wrongly classified.
- **True Negative (TN)**: The number of instances not covered by the rule, whose class differs from the training target class.
- **False Negative (FN)**: The number of instances not covered by the rule, whose class matches the training target class.

The fitness function is then a multiplication between the *Sensitivity (SE)* (the ratio of positive instance that are correctly classified) and the *Specificity (SP)* (the ratio of negative instance that are correctly classified), and it is defined in Equation 5.1 as:

$$F = \frac{TP}{TP + FN} * \frac{TN}{TN + FP} \tag{5.1}$$

The predictive accuracy of the classifier measures the proportion of correctly classified instances using the equation 5.2:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.2}$$

5.2.2.3 Classification using GEARM

GEARM process is used to classify case and control individuals by identifying gene-gene interaction according to the steps presented in Section 4.6. In this section, we briefly recall the steps of the process with particular emphasis on the features added to GEARM to be adapted to the simulated SNP data and accomplish the task of classification.

1. The first step, is the initialization of the previously defined parameters of GEARM: population size= 250 individuals (125 "case" and 125 "control"); generation size= 250; number of generated rules= 150; crossover rate= 0.9; mutation rate= 0.1; codon size= 200; wrap count= 2; minimum chromosome size= 10 and maximum chromosome size= 100.
2. We use 10-fold cross-validation for this contribution, where data is divided into 10 equal parts. 9/10 of the data is used for training, and the remaining 1/10 of the data is later used to test the model and evaluate its predictive ability.

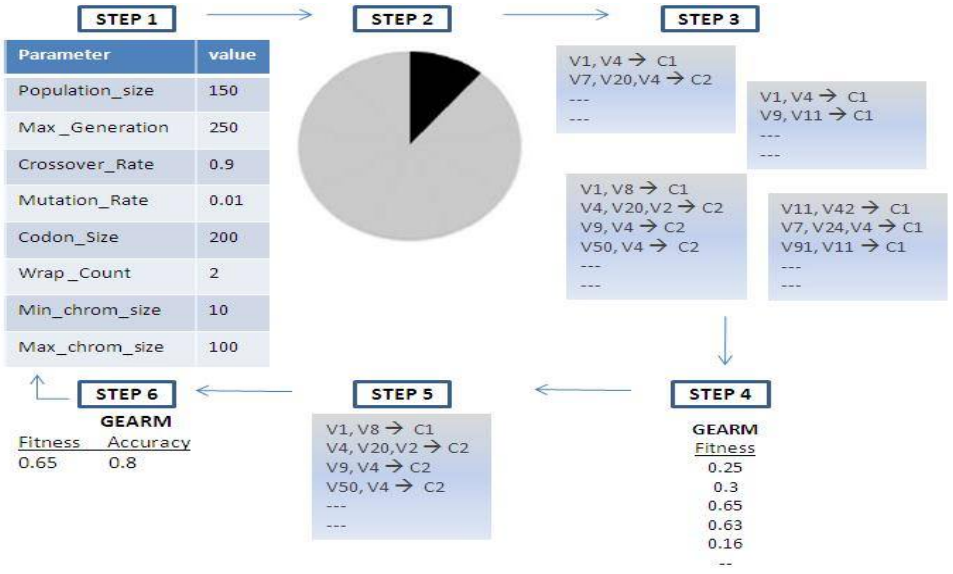


FIGURE 5.2: Different steps of the GEARM process.

3. The training step of the GEARM process begins by generating an initial population of 150 vector of integer of values from 1 to 200 according to the specified codon size. Using the above grammar and the Mod operation, we generate a set of ARs with consdering the possibility of wrapping 2 times over the vector.

4. The resulting output string then determines the set of 150 association rules where each individual in the initial population is mapped onto an AR. Each association rule R is evaluated on the training set and its fitness gets recorded using Equation 5.1.

5. The best Rules solutions are selected for crossover and mutation. The new generation that gets generated, containing the best rules and equal in size to the original population, is used for max generation.

6. The best solution is identified after each generation. At the end of the GEARM evolution, the overall best solution is selected as the optimal AR set. This best GEARM set is tested on the 1/10th of the data left out to estimate the prediction error using Equation 5.2.

   The above steps are performed 10 times using a different 9/10th of the data for training and the remaining 1/10th of the data for testing with the same parameter settings, in order to obtain the best set of association rules.

Figure 5.2 represents the different steps of GEARM for the classification task using 10-fold cross validation.

5.2.2.4 Functional SNPs Identification (FSNPs)

Each rule that yet generated by the GEARM process indicates a possible interaction among SNPs, and the final output is a list of such interactions. In order to determine the variables that have a strong influence on the epistasis, we propose two different methods to detect the functional SNPs.

§a SNPs of Equal Weights (AREW) : In this first method we count the number of times each SNP is present in the set of association rules that get generated for each 10-fold cross validation data split while giving the same weight to all the variables. The signal of each SNP ($S-SNP_x$) equals to its number of occurrence over the 10 splits of dataset (DS). The SNP that exists in the ten sets of data has a signal equal to 10. The SNP that does not exist in any set of the data has a signal of 0. Formally:

if $SNP_x \subset DS_i$ then $S_{xi} = 1$; $S-SNP_x = \sum \mathrm{S}_{xi}$; where:
$\{\ i \in [1,10]$, $S-SNP_x \in [0,10]$ with i and $S-SNP_x$ natural integers $\}$

§b SNPs of Weight of Appearance (ARWA) : In this method, we count the number of appearances of each SNP ($\#SNP_x$) in each split of data, and we calculate the weight of the SNP ($W-SNP_x$) as the number of its appearances divided by the number of SNPs in this set of rules. At the end, for each SNP we obtain 10 different values of weights for each 10-fold cross validation data split. The functional SNPs are those that have the highest sum of weights($SW-SNP_x$). Formally, the some of weights of each SNP is calculated as:

$W_i - SNP_x = \#SNP_x$ in DS_i / $\#SNP$ in DS_i; $SW-SNP_x = \sum \mathrm{W}_i - SNP_x$;
where: $\{\ i \in [1,10]$ with i natural integer $\}$

5.2.3 Evaluation of GEARM for Epistasis and Comparison to GEDT

This section provides the results obtained by applying GEARM for Epistasis models. Table 5.1 summarizes the average fitness *(Avr-F)* which is obtained on the training set of 10-fold cross-validation, and the average accuracy *(Avr-A)* obtained on the remaining test set (1/10th of the data) for each model and different Heritability (He) and Minor Allele Frequency (MAF) values. The fitness function takes into consideration all the instances that are correctly or incorrectly classified, and the ones not classified, which makes it always smaller than the accuracy that only take into account the proportion of the correctly classified rules. Through experimentation, we can confirm that the increase

in the size of a generation leads to an increase in predictive accuracy and better results in terms of quality of the generated rules.

TABLE 5.1: Evaluation results for simulated models.

G.M	He	M.A.F	Avr-F	Avr-A
XOR	2.5	0.25	0.25	0.44
XOR	2.5	0.5	0.26	0.45
XOR	7.5	0.25	0.26	0.4
XOR	7.5	0.5	0.25	0.29
XOR	10	0.5	0.24	0.4
BOX	2.5	0.25	0.29	0.38
BOX	2.5	0.5	0.3	0.44
BOX	7.5	0.25	0.3	0.55
BOX	7.5	0.5	0.32	0.6
BOX	10	0.5	0.24	0.5
MOD	2.5	0.25	0.29	0.38
MOD	2.5	0.5	0.3	0.4
MOD	7.5	0.25	0.27	0.51
MOD	7.5	0.5	0.3	0.5
MOD	10	0.5	0.27	0.42

For our power studies, we have tested our algorithm on several datasets for each genetic model and He combination. We have compared our results with those obtained by the Grammatical Evolution Decision Tree (GEDT) approach (Motsinger-Reif et al., 2010). Table 5.2 and Table 5.3 present the percentage of the power of GEARM using both "Equal Weight" (AREW) and "Weights of Appearance" (ARWA).

TABLE 5.2: Power 1 results for simulated models.

Model	He	MAF	AREW	ARWA	GEDT
XOR	2.5	0.25	1	2	0
XOR	2.5	0.5	5	4	0
XOR	7.5	0.25	7	10	3
XOR	7.5	0.5	5	5	2
XOR	10	0.5	3	5	4
BOX	2.5	0.25	20	40	13
BOX	2.5	0.5	40	30	16
BOX	7.5	0.25	70	90	72
BOX	7.5	0.5	80	70	53
BOX	10	0.5	90	80	69
MOD	2.5	0.25	30	20	7
MOD	2.5	0.5	10	15	6
MOD	7.5	0.25	30	40	79
MOD	7.5	0.5	50	50	47
MOD	10	0.5	60	80	60

"Power 1" (P1) is the number of times the algorithm correctly identified both functional loci in the data sets (Table 5.2). "Power 2" (P2) is the number of times the algorithm

TABLE 5.3: Power 2 results for simulated models.

Model	He	MAF	AREW	ARWA	GEDT
XOR	2.5	0.25	3	4	1
XOR	2.5	0.5	5	5	2
XOR	7.5	0.25	7	10	4
XOR	7.5	0.5	10	14	6
XOR	10	0.5	10	10	7
BOX	2.5	0.25	40	50	59
BOX	2.5	0.5	60	60	69
BOX	7.5	0.25	100	96	95
BOX	7.5	0.5	90	100	93
BOX	10	0.5	90	97	95
MOD	2.5	0.25	40	50	49
MOD	2.5	0.5	10	30	2
MOD	7.5	0.25	90	70	96
MOD	7.5	0.5	60	74	65
MOD	10	0.5	67	80	48

identified at least one of the two functional loci (Table 5.3). Analyzing the results, we can clearly see that (P2) is always higher than (P1). This can be explained since (P1) is considered as a subset of (P2) thus a stricter condition. We base our discussion on the power of the two methods. Tables 5.2 and 5.3, show that the powers increase as the He and the MAF increase, and this is observed for the two techniques.

The XOR model is a purely epistasis model without any main effect, whereas the BOX and MOD models are two interaction models with main effect. Analysing the results we can see the low power of the XOR model for all the combinations of He and MAF comparing to BOX and MOD models. This is due to the challenge imposed by the models that luck main effect for the computational methods to find the genetic interaction comparing to the common biological models, which have both marginal and interaction effects, where the detection of interactions are easier.

For the challenging model XOR (purely epistatic model) we can see that GEARM performs a little better compared to GEDT even if both have a weak power. This can be explained by the fact that decision trees can miss rules found by association rule mining. For example, in the case where He = 2.5, even if GEARM has shown a weak power (between 1% and 5%), GEDT could not even detect the two functional SNPs. The best results are seen for the BOX model and especially with He=7.5 for both cases where MAF equals 0.25 and 0.5. In these cases, GEARM generates the best set of rules with the highest prediction accuracy (Table 5.1).

In decision trees, the path from the root to the leaf determines all the antecedents; the consequent is determined by the leaf. Given a rule in a decision tree, it is likely that an

equivalent association rule exists. However, the opposite is not true: given an association rule, it may not be possible to find an equivalent rule in the decision tree. Furthermore, decision trees don't allow the extraction of rules based on sub-paths from the root to a leaf; indeed a rule starts from the root all the way down to leaf. This leads to longer and more complex rules, whereas association rules can find all the less complex predictive rules from a data set given a proper setting of the parameters. These results indicate that while GEARM and GEADT can both detect gene-gene interactions. GEARM can do it more efficiently and has higher power to detect two-locus interactions under either definition of power.

5.3 Hybrid GEARM with a Neural Network and a Genetic Algorithm for Dimensionality Reduction and Complex Disease SNP Classification in Real Data

Single Nucleotide Polymorphisms (SNPs) are an important source of the human genome variability and have thus been implicated in several human diseases. Using SNPs as genetic markers has widely helped researchers understand the genotype-phenotype relationship. One of the major problems related to SNP data is the high number of features which makes the task of classification complex. Our goal is to identify the complex interactions and the relationship among SNPs which may increase the performance of the classification of the disease of interest. The classification task is described as a pattern recognition problem. Feature extraction is the key to good pattern recognition since even the best classifier will perform poorly if the features are not well selected. Motivated by the success of the combination of intelligent techniques for the feature selection and classification tasks on biological data, and by the high performance of evolutionary algorithms, and knowing that traditional feature selection techniques tend to ignore the interaction between features, while their combination may have a strong correlation with the target, we propose GA-NN-GE-ARM. It is a new hybrid intelligent technique which is based on the GEARM algorithm and Neural Networks in addition to a Genetic Algorithms, used to find the best parameters of the two combined techniques. GA-NN-GEARM is applied on SNP data of different complex diseases. This data was obtained from the NCBI Gene Expression Omnibus (GEO) website, and GA-NN-GEARM reached a very high accuracy of up to 100% on this data.

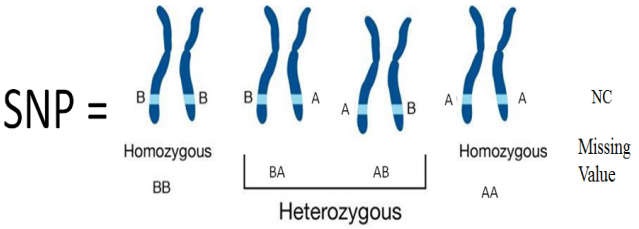


FIGURE 5.3: Homozygous, Heterozygous genotype and Missing value of SNP

5.3.1 SNP Datasets of Complex Diseases from NCBI

SNP data consists of sequences of nucleotides representing the mutations on a specific genomic region for the group of individuals under investigation.

We have used in this study four Affymetrix Mapping 250K Nsp SNP Arrays: GSE9222 (Marshall et al., 2008), GSE13117 (McMullan et al., 2009), GSE16125 (Reid et al., 2009), and GSE16619 (Kadota et al., 2009) downloaded from the NCBI Gene Expression Omnibus repository (Barrett and Edgar, 2006). The GEO project is a public repository that archives and freely distributes high-throughput functional genomic data (see Section 3.3.2).

Each studied dataset is represented by a set of individuals characterized by a set of SNPs, and two class labels (case and control). In other words, the dataset is denoted by a matrix $D_{N\times K}$, where N is the number of individuals and K is the number of studied SNPs (features). Table 5.4 summarizes the characteristics of each data set used for the current work.

Each sample in our data is represented by its genotype at specific loci. SNP values in GSE16125 are in the form of real numbers, while the other three arrays contain data in alphabetical format. As shown in Figure 5.3, according to the occurrence of mutation in the copies of genes from father and mother, a genotype can take one of four values for a bi-allelic SNP: AA, BB (representing homozygous genotype), AB (representing heterozygous genotype) or $NoCall$ (representing a missing value).

In order to apply feature selection and classification to these three datasets, the alphabetical format was converted into a numerical format. There are several ways to do so; in our case, we have used the encoding: $11, 10, 01, 00$ for AA, BB, AB, and $NoCall$, respectively.

We compare our results with results generated by (Batnyam et al., 2013) (the authors have tested several combinations of feature selection and classification techniques), who have used the same SNP datasets.

TABLE 5.4: Summary of SNP datasets

Dataset	#of SNPs	#of Samples	Description
GSE9222	250,000+	567	Autism (ASD)
GSE13117	250,000+	360	Mental Retardation
GSE16125	250,000+	42	Colon Cancer
GSE16619	500,000+	111	Breast Cancer

5.3.2 Dimensionality Reduction using GEARM

The concept of feature independence has led to the exploration of different approaches that were designed to remove a group of SNPs if they depend on others. In this contribution, we propose a new FS technique which is based on the extraction of ARs to find the hidden dependency relationships between SNPs, which in turn allows us to remove those whose effects are dominated by others, and keep only the most informative ones.

GEARM is used as a feature selection method. The difference between our proposed approach and the other feature selection methods is that the technique of ARM does not consider the features as independent; rather it takes into consideration the interaction between the different SNPs and their influence on each other. An SNP itself may have a specific correlation with the target but, when combined with other SNPs, they may together have a strong correlation with the state of the disease. Also, some SNPs are considered dominant; the presence of such SNPs can eliminate other SNPs since the function of the latter is a consequence of the former.

GEARM consists of the extraction of ARs using GE as an optimization technique to overcome the expensive computation of the classical AR mining algorithms (see Chapter 4). The output of this part is a set of optimal solutions of FS rules, the latter being taken as the input of the next part. The parameters used and their meanings are given in Table 5.5. Max Generation, crossover and mutation rates were chosen empirically; the given values are those that have yielded the best performance. The range of the codon is related to the number of SNPs used. The rest of the parameters are justified as explained below in the different steps of the technique. The grammar used here is presented as follows:

TABLE 5.5: Parameters of the GEARM Algorithm

Parameter	**Meaning**	**Value**
Pop_Size	Number of chromosomes	By GA
Max_G	Max Generations	150
Cross	Crossover rate	0.9
Mut	Mutation rate	0.01
Codon	The integers of the vector	[1-1000000]
Min_Chrom	Min Chromosome size	4
Max_Chrom	Max Chromosome size	12
Wrap	Max number of wrappings	2

$G = \{S, N, T, P\}$
$S = \{Rule\}$
$N = \{Rule, Side, SNP, VAL\}$
$T = \{Sep, SNP_1, SNP_2, ..., SNP_k, Val_1, Val_2, ..., VAL_m, Separator\}$
$P = \{< Rule >::=< Side > Sep < Side >$
$< Side >::=< SNP >< VAL > | < SNP >< VAL >< Side >$
$< SNP >::= SNP_1|SNP_2|...|SNP_k$
$< VAL >::= Val_1|Val_2|...|Val_m\}$

5.3.2.1 Association Rules Extraction Steps Between SNPs

As explained previously, the GEARM algorithm starts by generating an initial population of N vectors of integers with different sizes, where each vector represents a potential solution. The length of each solution is generated randomly with a value between *Min_Chrom* and *Max_Chrom*. *Min_Chrom* has been set to a value that ensures having at least one SNP in both the antecedent and the consequent, so that no part of the rule will be empty, and *Max_Chrom* limits the number of SNPs in the antecedent and the consequent to 3 SNPs at most so as to avoid very long and complicated rules (see Table 5.5). The extraction of the set of ARs between SNP follows the GEARM steps formalized in the *Algorithm 1*, and presented in Section 4.6.

5.3.2.2 Overall and Parallel Extraction to Select Best SNPs

For feature selection with GEARM, we propose two different techniques for AR extraction as explained below. In case of circular dependency between two SNPs, both SNPs are selected for the two techniques.

§a GEARM-OE :

The idea of GEARM-OE is to use all the samples in the data to extract the set of rules. We call it ***Overall Extraction (OE)*** in that we do not distinguish between the classes of individuals while looking for the association, and we try to find relations among SNPs based on their values for case and control samples. If a good rule is extracted with high fitness, the SNP in the antecedent part are considered as the dominant and the more informative elements. For example, the rule " if SNP_1= AA and SNP_7= AB then SNP_3= BB " indicates that SNP_1 and SNP_7 together influence SNP_3 i.e. SNP_3 depends on the other two SNPs. Thus, SNP_1 and SNP_7 are selected to be used as inputs to ANN.

§b GEARM-PE : In this second approach, the idea is to divide the data into different partitions and separately search for the association rules in each one. We call it ***Parallel Extraction (PE)***. As our application is classify case and control individuals, the data was divided into two partitions based on the sample class: one part contains all the case samples and the other contains all the control samples.

The main reason for this separation is to analyse the relation of between the SNPs in the same environment (either case or control). A set of association rules was generated for the SNPs of the control samples only, which means that the best features of this class was selected. Another set of rules was generated using the SNPs from the case samples. The best features for this class were also selected. The two sets of the best features were combined and used as inputs of the ANN.

Consider for instance the following rules extracted from the case samples:

- if SNP_1= AB and SNP_4= AB then SNP_2= AA
- if SNP_4= AA then SNP_3= BB and SNP_7= AB
- if SNP_8= AA and SNP_5= BB then SNP_6= AA

The selected features from this part will be: SNP_1, SNP_4, SNP_3, SNP_8 and SNP_5.

Consider also the following rules extracted from the control samples:

- if SNP_6= AA then SNP_2= AA
- if SNP_10= AB and SNP_4= BB then SNP_3= AA
- if SNP_15= BB then SNP_12= AA and SNP_3= BB

The selected features from this part will be SNP_6, SNP_10, SNP_4 and SNP_15.

Combining the two sets of selected features, the best SNPs that are considered as input for NN will be: SNP_1, SNP_4, SNP_3, SNP_8, SNP_5, SNP_6, SNP_10 and SNP_15.

5.3.3 NN-GEARM for Breast Cancer

Firstly, we have applied the combination of NN and GEARM on the Breast Cancer SNPs Dataset in order to test its performance and compare the two proposed techniques of rule extraction with FS: **GEARM-OE** and **GEARM-PE**. During the testing phase, several tests were performed by varying the parameters of GEARM and NN.

The NN-GEARM algorithm is represented in Figure 5.4 and consists of two parts:

- SNP selection with GEARM.
- Classification with the NN.

The fitness function used to evaluate each generated rule for these tests is presented in Equation 4.1 for each support and confidence greater than the minimum support and minimum confidence, respectively, where a and b are weights set by experimentation and their sum is equal to 1. We recall here this measure of fitness.

Fitness(R)= (a* sup(R)) + (b* conf(R))

We have used a Neural Network to perform the classification using the reduced set of SNPs as shown in the second part of Figure 5.4.

The NN is a two-layer feedforward network with a sigmoid transfer function used on both the hidden layer and the output layer. The selected SNPs from the previous stage represent the input of the network. The reduced data has been subdivided into three subsets: Training, Validation and Test data (Figure 5.4). Several tests were performed by increasing the number of neurons from 20 to 55 (with less than 20 neurons, the model had a weak experimental performance, and 55 neurons was set as a parameter to limit the test series of this contribution) for each set of input features. The results obtained by NN-GEARM on Breast Cancer SNP data are presented in Section 5.3.5.

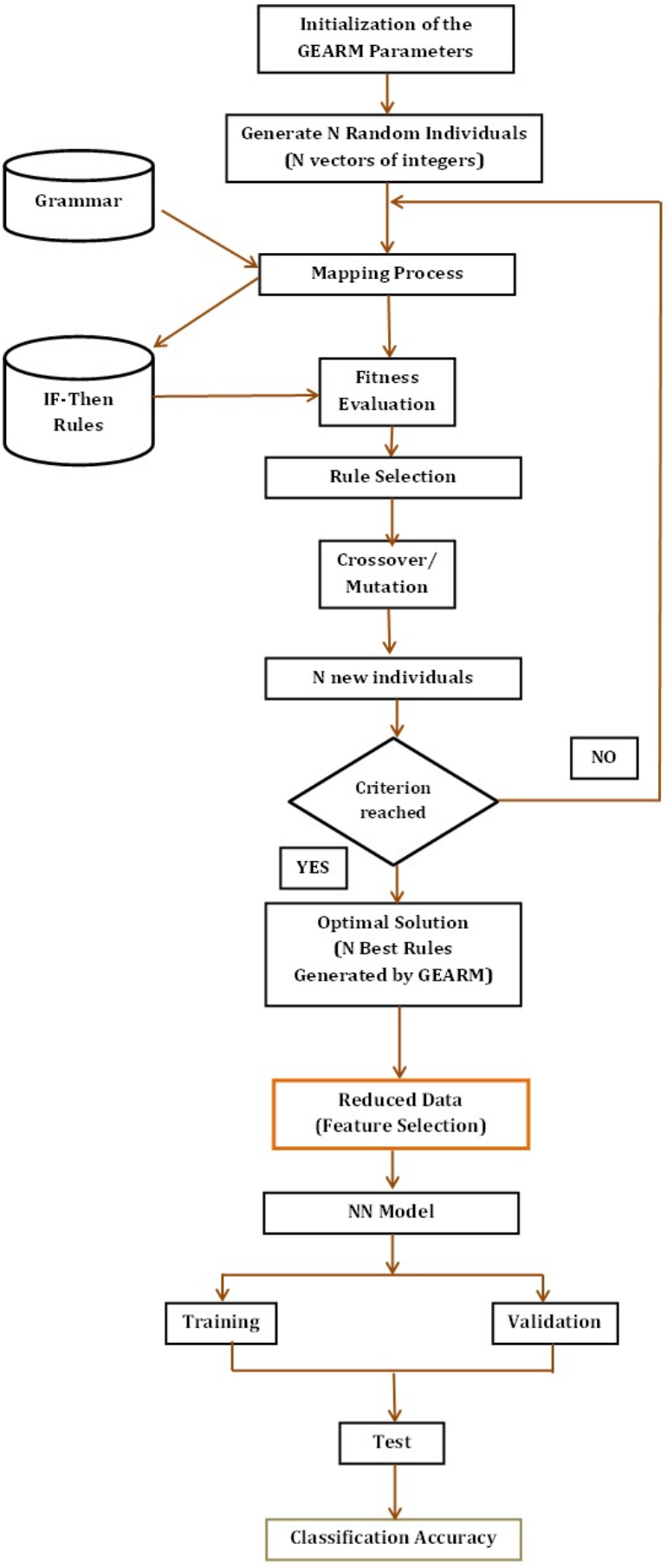


FIGURE 5.4: NN-GEARM Algorithm

5.3.4 GA-NN-GEARM for SNP Selection and Classification

Two important tasks in terms of analyses performed with SNP datasets are the individual assignment to the original class and the selection of the most informative markers (SNPs), which is summarized in the classification and feature selection tasks, respectively.

The purpose here is to build a hybrid paradigm for complex diseases, which is able to classify individuals as case or control samples based on large SNP datasets with the highest possible accuracy. In other words, we intend (1) to search for the relationship between the genotype and the phenotype of interest based on the relationships between SNPs, and (2) to identify their interactions more effectively than with the existing methods.

Feature selection being the key in pattern recognition, the GA-NN-GEARM algorithm consists of the following steps:

- Extracting AR between SNPs following the GEARM process in order to reduce the dimensionality of the data and select the most informative features for the classification;
- Performing an efficient classification of individuals based on the selected features using ANNs; and
- Setting the parameters of the combined techniques by a GA to increase the performance of the model.

We will go through these points one by one in detail, and explain all the steps of our proposed approach which is represented in Figure 5.5.

5.3.4.1 Feature Selection

In order to obtain the most informative features from the set of ARs, we perform parallel feature extraction (see Section §b). We divide the data into two parts based on the sample class: one part contains all the case samples and another all the control samples.

To evaluate the extracted rules, we take into consideration three interesting measures and set the *fitness* function for each *Support (Sup)*, *Confidence (Conf)* and *Conviction (Conv)* greater than *minimum support*, *minimum confidence* and *1*, respectively, as shown in Equation 5.3.

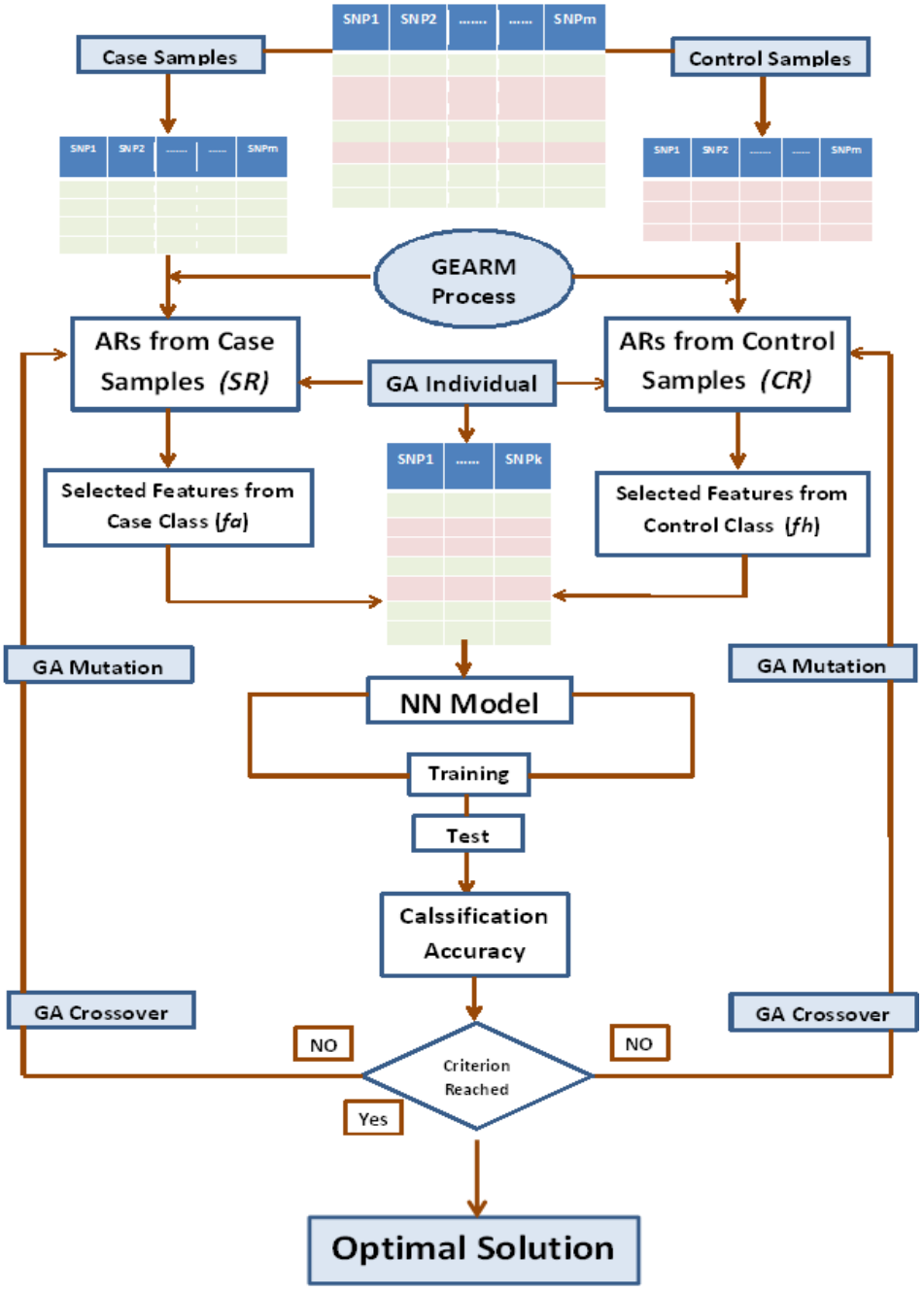


FIGURE 5.5: The GA-NN-GEARM Process

$$F(R) = \frac{(w1 * Sup(R)) + (w2 * Conf(R)) + (w3 * Conv(R))}{w1 + w2 + w3} \tag{5.3}$$

where *w1*, *w2* and *w3* are weights that are set empirically and which sum up to 1. The *Conviction* is an asymmetric measure that was proposed to tackle some of the weaknesses of *confidence*. We have considered it to evaluate the *rule* in order to deal with circular dependencies by selecting rules that have the highest degree of implication.

A set of association rules is first generated for the SNP of the control (healthy) samples. All the SNPs appearing in the antecedent part of the rules are selected as the best features of this class; let it be f_h. Then, another set of rules is generated using the SNPs from the case (affected) samples and, again, the SNPs appearing in the antecedents of the rules in this part are selected as the best features for this class; let it be f_a. The two sets of best features were combined in a single set, f_s, and used as inputs to the ANN; so $f_s = f_h \cup f_a$.

5.3.4.2 Neural-Network-Based Classification

In the broadest sense, machine learning research for a disease association study can be divided into two parts: how to choose significant SNPs and how to classify the selected SNPs to yield the most accurate prediction outcomes. To perform the classification of the reduced SNPs and understand phenotype-genotype correlations, we have used and compared three different neural networks: a two-layer feedforward network (**MLPNN**) with a sigmoid transfer function, a Radial Basis Network (**RBFNN**) with the Gaussian activation function on the hidden layer, and a Focused Time Delay Neural Network (**Focused Time Delay (FTD)NN**).

The selected SNPs from the previous stage represent the reduced input of the networks. The data has been subdivided into 70% for training and 30% for testing (Figure 5.5) to calculate the accuracy of the model. Based on the accuracy we reached, our method performs a modification of the architecture of each ANN so as to improve its performance. The number of features, the set of features, the number of hidden neurons and the maximum number of iterations will be modified several times, and different combinations will be tested to select the best ANN for each model.

5.3.4.3 Parameter-Setting using a GA

Choosing the best parameters for a specific algorithm which lead to the best performance is a crucial problem that can be solved by an EA. In the current study, GA is used to set the parameters of the combined techniques by generating I individuals evolving for G generations. We focus mainly on the architecture of the ANN.

Each individual of the GA is encoded as a vector of integers that represent four different parameters generated randomly within a specific range (set according to the acquired experience from the different tests performed in the previous contributions). The representation of the GA chromosome is shown in Table 5.6 and is explained as follows:

TABLE 5.6: The encoding of a GA individual

NB-Rules	Hidden-Neurons	Max-Epoch	Max-Iter

1. ***NB-Rules*** is the number of rules generated from each partition of the used dataset. *NB-Rules* $\in [10, 55]$
2. ***Hidden-Neurons*** gives the number of neurons used on the hidden layer of the NN. *Hidden-Neurons* $\in [20, 100]$
3. ***Max-Epoch*** is the maximum number of iterations used by the NN to converge. *Max-Epoch* $\in [1000, 1500]$
4. ***Max-Iter*** states the maximum number of times the NN will repeat the training process with different initial weight values with a given configuration of the NN. The weights that perform the highest accuracy are considered. *Max-Iter* $\in [5, 10]$

5.3.4.4 Model Evolution and Evaluation

Following the constituents of Evolutionary Algorithms, we step by step explain our solution which is presented in Algorithm 2.

1. **Algorithm 2: GA-NN-GEARM**
2. **Input**: SNP dataset
3. **Output**: Optimal Solution
4. **Begin**
5. Generate I chromosomes */*(I vectors of parameters)*
6. **for** each chromosome (i) **do**:
7. $(SR, f_a) = GEARM\_Case\_Samples(NB\_Rules)$
8. $(CR, f_h) = GEARM\_Control\_Samples(NB\_Rules)$
9. $f_s = f_a \cup f_h$
10. $F(i) = ANN(f_s, Neurons_i, Epoch_i, Iter_i)$
11. **end for**
12. **while** Generation <= Max_Generation **do** :

13. (i_1, i_2) = Select best Chromosomes
14. $(i'_1, i'_2) = Crossover, Mutation(i_1, i_2)$
15. $(SR'_1, SR'_2, f_{a1}', f_{a2}') = Crossover(SR_1, SR_2)$
16. $(CR'_1, CR'_2, f_{h1}', f_{h2}') = Crossover(CR_1, CR_2)$
17. $f'_s = f'_a \cup f'_h$
18. $F(i')$= ANN (f'_s, $Neurons_{i'}$, $Epoch_{i'}$, $Iter_{i'}$)
19. **end while**
20. Optimal Solution = Best: i, SR, CR, f_s, ANN /* *according to the Accuracy*
21. Return (Optimal Solution)
22. **End**

The classification accuracy obtained by the NN, using the features extracted by GEARM and the specified parameters extracted by GA, is used as the fitness to evaluate each individual of the GA population. The best solutions are the vectors that get the highest accuracy values.

The process starts by generating I individuals. Each individual specifies the number of rules that will be generated by GEARM and the parameters that will be used by the ANN to classify the SNPs of the related extracted ARs. We obtain I different solutions evaluated according to their accuracy. The best solutions are then selected for crossover and mutation to produce a new generation of solutions. We use one point crossover between two vectors. For each GA individual we associate a set of ARs extracted from the control people partition that we name *Control Rules* (CR), and a set of ARs generated from the case people partition that we name *Case Rules* (SR), where the number of rules in each set is related to the *NB_Rules* component of the vector.

The crossover operation is not restricted to the parameters of the solution only, but we aim to extend it to the extracted rules, i.e. to the selected features. As such, we propose to perform the crossover operation between the sets of rules. We define a crossover between two AR sets *S1* and *S2* as the exchange of a specific number of rules selected randomly between the two sets. Accordingly, if the crossover is performed between *NB_Rules* of *S1* and *S2*, the crossover between their rules sets will be performed too, where *CR* of *S1* is crossed with *CR* of *S2*, and *SR* of *S1* is crossed with *SR* of *S2*. The number of exchanged rules between the sets is equal to half of the smallest set. The mutation operation is performed on the GA individuals. The value of *NB_Rules*

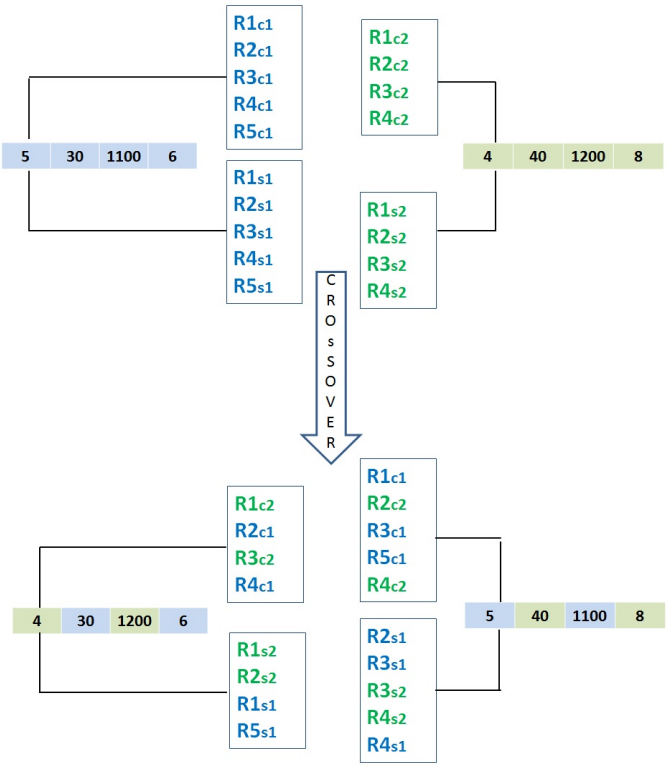


FIGURE 5.6: Crossover Operation for a GA Individual

will not undergo any change. Figure 5.6 illustrates the crossover operation for the GA individuals.

The new individuals will produce a new architecture of the NN for the new SNPs to perform the classification, and the process is iterated until the maximum number of generations is reached. At the end, the best features and the best architecture of the NN that yield the highest accuracy is selected as the optimal solution of GA-NN-GEARM.

5.3.5 Results and Comparison

In this section we present the results obtained using NN-GEARM on the Breast Cancer SNP dataset and GA-NN-GEARM on four SNP datasets: Autism, Mental Retardation, Colon Cancer and Breast Cancer all downloaded from the NCBI Gene Expression Omnibus. The comparisons and the discussion of the results are given in the following two subsections.

5.3.5.1 Evaluation of NN-GEARM

The first series of tests consisted in applying NN-GEARM (without using a GA) on Breast Cancer SNP data consisting of 11 samples and 500.000+ features. This is a

TABLE 5.7: The number of selected features (SNP) and the average fitness (AvrFit) obtained by GEARM-PE, and the corresponding accuracy obtained by NN (Accur_NN).

SNPs	**AvrFit_GEARM-PE**	**Accur_NN**
34	0,968	70,3 %
35	0,919	72 %
37	0,968	74 %
41	0,820	68,5 %
42	0,871	90 %
129	1	75 %

TABLE 5.8: The number of selected features (SNP) and the average fitness (AvrFit) obtained by GEARM-OE, and the corresponding accuracy obtained by NN (Accur_NN).

SNPs	**AvrFit_GEARM-OE**	**Accur_NN**
22	0,974	71,2 %
23	0,972	68 %
24	0,975	69,4 %
32	0,937	74,8 %
37	0,964	74 %
45	0,945	71 %

clear case of "curse of dimensionality" problem. The reason behind this series is to test the power of our Hybrid model to reach a high classification accuracy, and also to perform a comparison between the Overall and Parallel Rule extraction techniques that were proposed, where the best of them will be used for the rest of our tests for the other diseases. During the evaluation of NN-GEARM, several tests were performed by varying the parameters of GEARM and NN. Both **GEARM-OE** and **GEARM-PE** were applied on the set of data to select the best features.

Table 5.7 gives the average fitness of the extracted association rules after several executions of **GEARM-PE** for varying numbers of SNPs selected from each rule set, along with the classification accuracy obtained with NN for each set of selected features.

As shown in Table 5.7 the best classification performance for **GEARM-PE** was obtained with 42 SNPs and its accuracy is 90%.

Similarly, Table 5.8 gives the average fitness of the extracted association rules after several executions of **GEARM-OE** for varying numbers of SNPs selected from each rule set, along with the classification accuracy obtained with NN for each set of selected

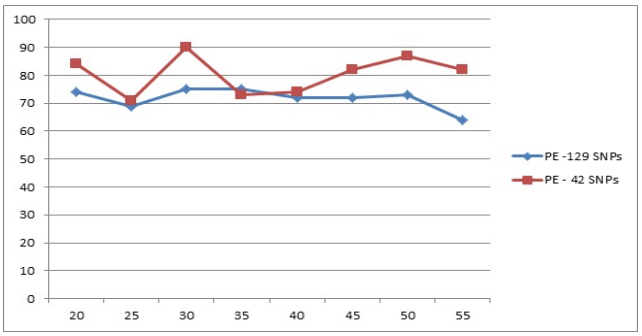


FIGURE 5.7: Comparison of the Accuracy of 129 and 42 selected features by GEARM-PE as a function of the number of neurons used in NN

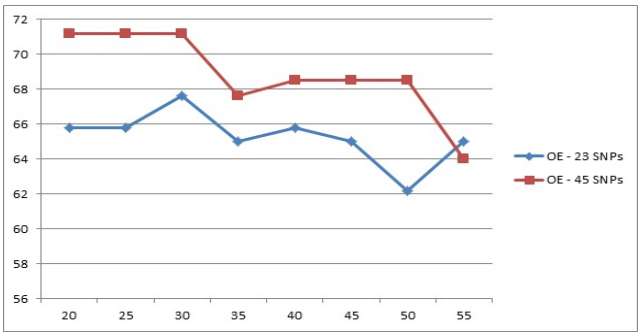


FIGURE 5.8: Comparison of the Accuracy of 23 and 45 selected features by GEARM-OE as a function of the number of neurons used in NN

features. The best classification performance for **GEARM-OE** was obtained with 32 SNPs and it is almost 75% (Table 5.8).

The values given in Table 5.7 and Table 5.8 represent the best results obtained after several executions. Figure 5.7 and Figure 5.8 show the variation in accuracy for some selected features with GEARM-PE and GEARM-OE, respectively, as a function of the number of neurons of the ANN model. We can clearly see that using 30 neurons for the classification ensures the highest accuracy for most of the selected features.

Comparing GEARM-PE and GEARM-OE on the basis of the experimental results, we can conclude that the Parallel extraction of features provides a better performance than the Overall one. This is further confirmed in Figure 5.9, where even for the same number of selected features (37 SNPs), the GEARM-PE accuracy is always better for any number of neurons except for 50 neurons, and this is due to the quality of the selected input SNPs.

Figures 5.10 and 5.11 summarize the average fitness for each set of discovered association rules as a function of the number of selected features, while Figures 5.12 and 5.13 indicate the execution time in seconds as a function also of the number of features used in GEARM-PE and GEARM-OE, respectively.

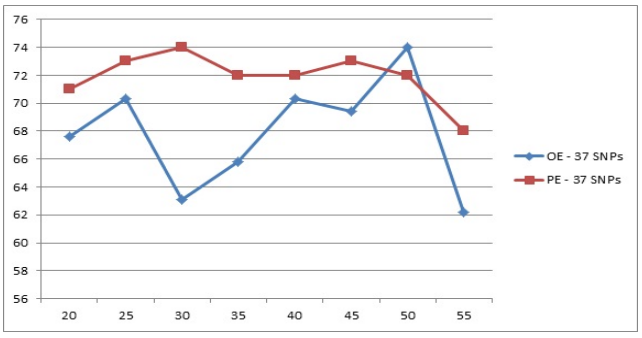


FIGURE 5.9: Comparison of the Accuracy of 37 selected features by GEARM-PE and GEARM-OE as a function of the number of neurons used in NN

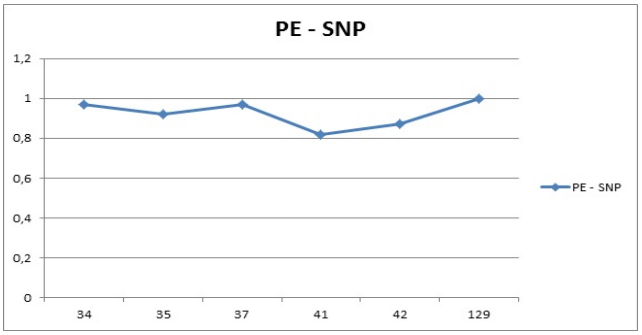


FIGURE 5.10: The average fitness of the generated rules as a function of the number of features selected by GEARM-PE

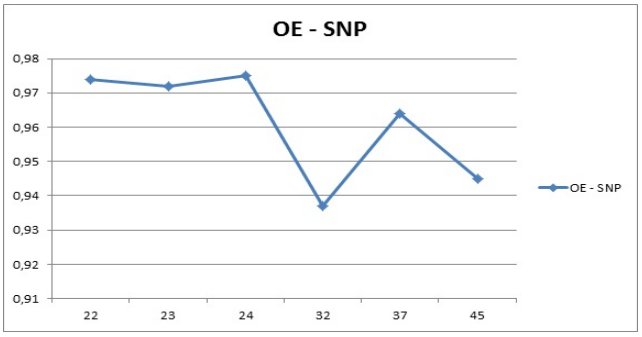


FIGURE 5.11: The average fitness of the generated rules as a function of the number of features selected by GEARM-OE

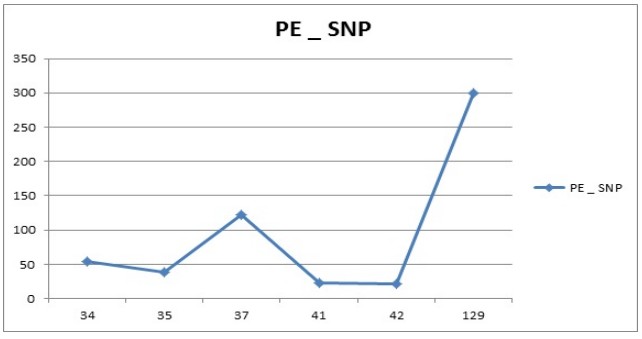


FIGURE 5.12: Time in seconds used by GEARM-PE to generate association rules as a function of the number of selected features

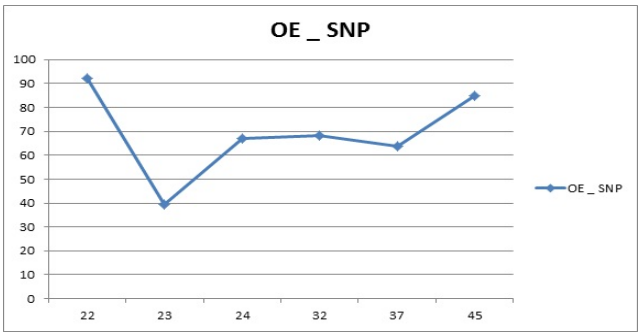


FIGURE 5.13: Time in seconds used by GEARM-OE to generate association rules as a function of the number of selected features

From the results we have just presented, we can see the NN-GEARM hybrid intelligent approach dealt with the high number of SNPs by selecting the best features and classifying the resulting ones with a classification accuracy of up to 90%. The accuracies we have reached using a NN have shown that GEARM-PE had a better performance than GEARM-OE, although both gave encouraging results and can be considered as useful techniques for feature selection. Based on these results, GEARM-PE will be the technique used for feature selection for the rest of the evaluation of the GA-NN-GEARM model tested on four different SNP datasets.

5.3.5.2 Evaluation of GA-NN-GEARM

Our name now is to evaluate our GA-NN-GEARM proposed approach on the four different SNP datasets we have been using. The model identified SNP associations. It was then validated in terms of accuracy and compared with other proposed methods for the same datasets.

Experimental results demonstrate that our approach (presented in bold in Tables 5.12 to 5.15) is better or at least comparable with previously proposed combination methods. In (Batnyam et al., 2013) the authors have used four algorithms: Feature Selection based on Distance Discriminant (FSDD), feature weight based ReliefF, R-value based Feature Selection (RFS) and an Algorithm based on Feature Clearness (CBFS), with multiplication (mult) and average (avg) methods of Future Fusion (FFM); and classified by the machine learning techniques: k-Nearest Neighbor (KNN/k=7) and Support Vector Machine (SVM); and an Artificial Gene Making method (AGM/Alpha). We have compared the results of GA-NN-GEARM with the best results presented in (Batnyam et al., 2013) in terms of the quality of the solutions obtained by each algorithm and for each dataset.

We review in this section the experimental results for the four datasets: GSE922 and GSE13117 two datasets of patients with mental disorders and their healthy parents; and, GSE16125 and GSE16619 two datasets of cancer, labelled case and control . We present for each series the number of the features that were selected along with the accuracy obtained with each used NN model and a comparison with the mentioned techniques in terms of the best accuracy and the number of selected features. In Tables 5.12 to 5.15, the techniques are presented under the format (feature selection method (Releif or RFS or FSDD or CBFS) + feature fusion method (Average or Multiple) + classification method (SVM or K-NN or Alpha). The best performance of the model would be with the highest accuracy and lowest number of SNPs.

Tables 5.9, 5.10 and 5.11 give a summary of the number of SNPs selected and the relative number of positive and negative rules extracted with the evaluation measures. In each case, for each rule set of each dataset we have calculated the average Support, Confidence, Conviction and Fitness, in addition to the accuracy and the corresponding number of neurons used in the hidden layer of the Multi Layer Perceptron (MLP), Radial Basis Function (RBF) and Focused Time Delay Neural Network (FTDNN), respectively. It is clear from these Tables that the proposed approach was able to reach a high accuracy with a low number of features. This is especially true for FTDNN.

Table 5.12 presents the comparison of GA-NN-GEARM with different combinations of feature selection and classification techniques for the GSE9222 series. The RFS algorithm had the worst performance, where its best accuracy was obtained when combined with the Alpha classifier and Multiplication method for feature fusion, and reached only 64% with 100 SNPs. Our hybrid technique however was able to reach an accuracy of 68% with only 9 SNPs for MLP and 71% for RBF and more than 96% with the dynamic NN (FTDNN) with only 7 SNPs. GA-NN-GEARM outperformed the best performance of all the other techniques for the Autism dataset except for CBFS+SVM which exceeds both MLP and RBF but it uses a high number of features.

For the GSE13117 series, the compared algorithms were able to classify the mental retardation dataset with high accuracy as shown in Table 5.13. The comparison in this case was performed on the number of selected features. A good accuracy was obtained with FSDD when using the SVM classifier and Multiplication method. GA-NN-GEARM selected the lowest number of features and gave the highest accuracy, better than ReliefF and RFS, and CBFS, and close to FSDD while using only 21 SNPs for MLP, and 15 SNPs for RBF. The FTDNN reached the highest accuracy of up to 92%, which proves the high performance of our proposed approach to extract the most relevant features.

A high accuracy was reached by GE-NN-GEARM for the Colon cancer dataset, 83% with RBF, 93% with MLP and 100% with FTDNN, all of which were obtained with

TABLE 5.9: Summary of the Number of extracted positive rules (NB-PR) and negative rules(NB-NR) with average support (Arv-S), average confidence (Avr-C), average conviction (Avr-V) and average fitness (Avr-Fit) along with the number of selected SNPs (#SNPs), the number of hidden neurons (NB-N) and the accuracy (ACC) of the **MLPNN**

Data	NB-PR	NB-NR	Avr-S	Avr-C	Avr-V	Avr-F	#SNPs	NB-N	ACC
GSE9222	12	12	0.24	0.79	1.56	0.93	9	25	68%
GSE13117	17	18	0.53	0.89	1.59	1.06	21	30	85 %
GSE16125	14	14	0.1	0.5	1.88	0.92	22	20	93%
GSE16619	16	20	0.63	0.83	1.47	1.02	42	30	100%

TABLE 5.10: Summary of the Number of extracted positive rules (NB-PR) and negative rules (NB-NR) with average support (Arv-S), average confidence (Avr-C), average conviction (Avr-V) and average fitness (Avr-Fit) along with the number of selected SNPs (#SNPs), the number of hidden neurons (NB-N) and the accuracy (ACC) of the **RBFNN**

Data	NB-PR	NB-NR	Avr-S	Avr-C	Avr-V	Avr-F	#SNPs	NB-N	ACC
GSE9222	12	12	0.24	0.79	1.56	0.93	9	60	71%
GSE13117	20	20	0.24	0.7	1.65	0.94	15	81	83.3 %
GSE16125	12	12	0.05	0.5	1.82	0.9	7	22	83%
GSE16619	13	12	0.35	0.8	1.6	0.97	27	50	100%

a small number of features, as shown in Table 5.14. Our proposed approach clearly outperformed the other techniques in terms of the solution quality for this GSE16125 series, except for CBFS that had a higher accuracy than RBF with a higher number of features.

The last series of experiments was performed on the breast cancer dataset (see Table 5.15). The CBFS was able to correctly classify all the samples using SVM for both the Multiplication and Average methods of feature fusion However, this high classification accuracy has required 70 SNPs, whereas GA-NN-GEARM selected a small number of SNPs for a fully correct classification by the three NN models. The Hybrid intelligent approach was able to reach total accuracy with this very small number of features, 14 only for FTDNN.

The Figure 5.14 presents the average accuracy for 10 different sets of selected SNPs, obtained by GA-NN-GEARM with MLP, RBF and FTDNN, for each series of dataset. The results proved the high performance of our proposed intelligent approach for feature selection and classification.

TABLE 5.11: Summary of the Number of extracted positive rules (NB-PR) and negative rules (NB-NR) with average support (Arv-S), average confidence (Avr-C), average conviction (Avr-V) and average fitness (Avr-Fit) along with the number of selected SNPs (#SNPs), the number of hidden neurons (NB-N) and the accuracy (ACC) of the **FTDNN**

Data	**NB-PR**	**NB-NR**	**Avr-S**	**Avr-C**	**Avr-V**	**Avr-F**	**#SNPs**	**NB-N**	**ACC**
GSE9222	12	12	0.04	0.58	1.57	0.81	7	22	97%
GSE13117	20	20	0.35	0.82	1.69	1.03	26	35	91.7 %
GSE16125	12	12	0.05	0.5	1.9	0.92	6	20	100%
GSE16619	10	10	0.27	0.7	1.52	0.9	14	22	100%

TABLE 5.12: Comparison of classification accuracy on the Autism (ASD) **GSE9222** dataset

Technique	**#SNPs**	**Accuracy**
ReliefF+ (Avg)+ SVM	10	0.64
RFS+ (Mult)+ Alpha	100	0.64
FSDD+ (Avg)+ SVM	10	0.64
CBFS+ (Avg)+ SVM	60	0.783
GA-NN-GEARM :		
MLPNN	**9**	**0.68**
RBFNN	**9**	**0.71**
FTDNN	**7**	**0.965**

TABLE 5.13: Comparison of classification accuracy on the Mental Retardation **GSE13117** dataset

Technique	**#SNPs**	**Accuracy**
ReliefF+ (Mult)+ SVM	20	0.807
RFS+ (Mult)+ SVM	30	0.775
FSDD+ (Mult)+ SVM	30	0.872
CBFS+ (Avrg)+ SVM	30	0.831
GA-NN-GEARM:		
MLPNN	**21**	**0.85**
RBFNN	**15**	**0.83**
FTDNN	**26**	**0.917**

5.4 Conclusion

We have presented in this chapter the success of our proposed Algorithm GEARM and its hybridization to other intelligent techniques in solving the dimensionality reduction and the classification problem based on SNPs datasets for complex diseases.

GEARM provides an efficient mechanism for the classification of individuals and the

TABLE 5.14: Comparison of classification accuracy on the Colon Cancer **GSE16125** dataset

Technique	**#SNPs**	**Accuracy**
ReliefF+ (Avg+R)+ SVM	40	0.708
RFS+ (Mult)+ SVM	50	0.708
FSDD+ (Mult+R)+ SVM	20	0.708
CBFS+ (Avg+R)+ Alpha	30	0.85
GA-NN-GEARM:		
MLPNN	**22**	**0.93**
RBFNN	**7**	**0.83**
FTDNN	**6**	**1**

TABLE 5.15: Comparison of classification accuracy on the Breast Cancer **GSE16619** dataset

Technique	**#SNPs**	**Accuracy**
ReliefF+ Alpha	30	0.66
RFS+ (Mult+R)+ KNN	10	0.61
FSD+ SVM	20	0.64
CBFS+(Mult/Avg)+ SVM	70	1
GA-NN-GEARM:		
MLPNN	**42**	**1**
RBFNN	**27**	**1**
FTDNN	**14**	**1**

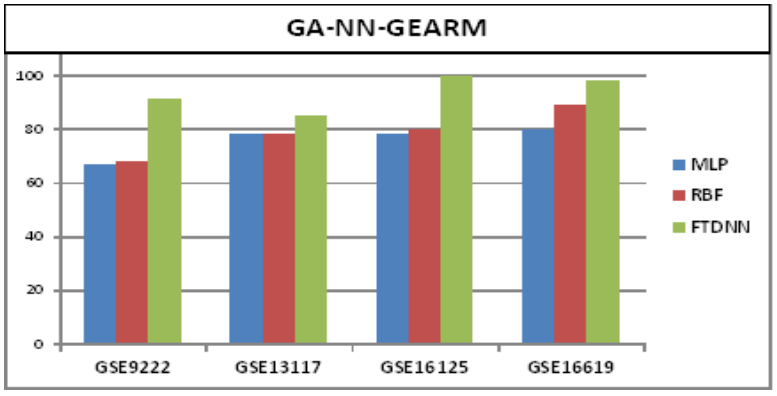


FIGURE 5.14: The average accuracy for 10 different sets of selected SNPs obtained by GA-NN-GEARM for the four series of datasets.

detection of gene-gene interactions in the presence or absence of main effects. It has been tested on simulated datasets of different Epistasis models. Our proposal has yielded a reduced set of association rules. Also, with this small association rule set, we have managed to cover all the SNPs in the dataset.

GEARM was then combined to a ANN for NN-GEARM and tested on Breast Cancer

SNP data which is a clear case of the "curse of dimensionality" problem. Our proposed approach dealt with the high number of SNPs by selecting the best features and classifying the resulting ones. A classification accuracy of up to 90% was reached. GEARM was used to ensure the best selection of the input vector, while ANN was used for classification. We have proposed two different ideas to reduce the dimensionality of the problem; they differ in their way of extracting association rules among the data. The first, that we called GEARM-OE, accomplishes the extraction of the rules on the entire data, while GEARM-PE extracts the rules from two separated sets according to the samples class. The accuracies we have reached using a ANN have shown that GEARM-PE had a better performance than GEARM-OE, although both gave encouraging results and can be considered as useful techniques for features selection.

To improve our results, NN-GEARM was combined to GA, and we created a new hybrid intelligent model for SNP selection and classification for complex diseases that we called GA-NN-GEARM. The approach is based on two principles: the first consists in selecting the best features by parallel extraction of ARs between SNP using GEARM-PE, since it gives the best performance, and the second in classifying the reduced dataset using three different types of ANNs: MLPNN, RBFNN and FTDNN. We have used a GA to specify the parameters of these techniques, focusing on the number of rules in the GEARM algorithm and the number of hidden neurons of the ANN.

GA-NN-GEARM has been applied on four different SNP datasets for four different diseases: Autism, Mental Retardation, Colon Cancer and Breast Cancer. It has been compared to different combined techniques from the literature that performed very well on this same data. The results have shown that using GA and three types of a ANN improved the accuracy for Breast Cancer SNP data, and our method has outperformed the other techniques in terms of quality of the solution: our hybrid intelligent method was able to select a small number of features and reached high accuracies of up to 100% in the different tests we have carried out. GA-NN-GEARM is able to deal with the curse of dimensionality problem with a high performance. We believe that we have attained the goal of reaching high classification accuracy for such challenging types of complex data for diseases diagnosis.

Chapter 6

Discovering Drug-Disease Associations for Drug Repositioning

6.1 Introduction

Because of the time consuming and costly process of discovering a new drug, Drug Repositioning or Drug Repurposing is considered as a viable strategy that can reduce the cycle time of drug discovery. This particularly appealing since the drug side effects have already been studied in a clinical environment. Drug repositioning is known as the 'old drug, new uses on diseases' paradigm. A main alternate strategy for the pharmacology industry is to find new applications for already approved drugs. The currently available computational power has been exploited to improve the effectiveness and efficiency of drug discovery. One of the most followed direction by these computational techniques is the identification of connections between drug and disease based on the connections between biological entities. In this chapter, the discovery of new repurposing of previously discovered drugs is addressed. The technique of ARM with GE is used to extract hidden relationships between a set of targets of Drug-Disease pairs in order to discover new pairs. Two types of target data are used: Genes and Pathways. For each Target (Genes/Pathways) a set of association rules is extracted and a set of Drug-Disease pairs is generated.

TABLE 6.1: (1)The binary matrix of n drug-target Genes/Pathways for each known pair of Drug-Disease (DR,DI)x. (2)The binary matrix of m disease-target Genes/Pathways for each known pair of Drug-Disease (DR,DI)x.

	Drug	Target	Genes/Pathways		Disease	Target	Genes/Pathways	
K Pairs	G_1/P_1	G_2/P_2	...	$G_n/P_{n'}$	G_1/P_1	G_2/P_2	...	$G_m/P_{m'}$
$(DR,DI)_1$	0	1	...	1	0	1	...	1
$(DR,DI)_2$	1	0	...	0	1	0	...	0
$(DR,DI)_3$	0	1	...	1	0	0	...	1
	...	...	...	...	...	...	...	...
$(DR,DI)_k$	0	0	...	1	0	1	....	0

6.2 Genes/Pathways-Drugs-Diseases Databases

During the data preparation phase, two kinds of datasets were collected:

- Firstly, rule mining based on genes target drugs and diseases was performed. Drug-Gene, Disease-Gene, and known Drug-Disease datasets were collected from a previous study (Zhao and Li, 2012). These datasets were originally taken from *DrugBank* (Law et al., 2014), *OMIM* (Amberger et al., 2009), and *CTD* (Davis et al., 2014) databases respectively.
- Secondly, rule mining was performed on Pathway-based drug repositioning. Pathways were extracted from *KEGG* (Kanehisa et al., 2014) database.

For the definition and description of each dataset, please refer to the Section 3.3. In terms of data preprocessing, two binary matrices for Genes/Pathways were prepared (in total four binary matrices):

1. A Drug-target matrix (one for Genes and one for Pathways), and
2. a Disease-target matrix (one for Genes and one for Pathways).

The two matrices (for each target: Genes and Pathways) are merged and illustrated in Table 6.1. Each row represents a drug and its corresponding disease respectively (known pairs of drug-disease), and columns are all the Genes/Pathways of interest. Each entry in the table indicates the existence of an association where *1* means the Gene/Pathway is a target of the corresponding drug (disease), and *0* otherwise. Note that the sets of Genes (Gx)/Pathways (Px') in the columns of the two matrices are not necessarily the same. However, the two matrices have the same number of rows since a row represents a known Drug-Disease (DR,DI) association.

6.3 GEARM for Drug Repositioning by Mining Genes/-Pathways Associations

A number of studies have shown that target binding of a drug often affects not only the intended disease-related Genes/Pathways, leading to unexpected outcomes. Thus, if the affected Genes/Pathways are related to other diseases, then this will allow the repositioning of an existing drug. The aim is to find hidden relations between Drug targets and Disease-related Genes/Pathways so as to find new hypotheses of new Drug-Disease pairs (Figure 6.1). We have applied *GEARM* to 288 Drugs and 267 Diseases, forming 5018 known Drug-Disease pairs. The process of extraction of associations for each type of data (Genes/Pathways) is described and the results we have reached with the generated Drug-Disease pairs are presented bellow.

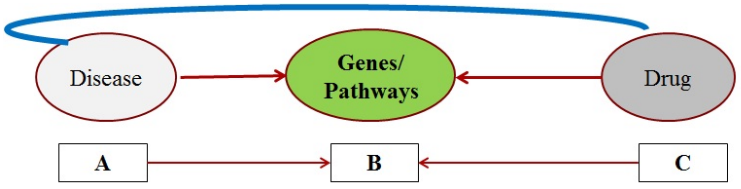


FIGURE 6.1: Finding heading relationship between A and C based on the intermediate B

6.3.1 Extracting Genes/Pathways Associations

To generate an interpretable set of rules, a suitable grammar must be defined which specifies the antecedent and the consequent of each rule both as combinations of Target (either Genes or Pathways). Then each extracted rule is evaluated according to a fitness function to decide on its efficiency.

6.3.1.1 Grammar for Targets Associations

For the adaptation of ARs to deal with the data at hand, the set of known Drug-Disease pairs (DR,DI) represents the set of transactions (the transaction database) D= { $(DR, DI)_1$; $(DR, DI)_2$; ...;$(DR, DI)_k$ }, while the Targets (Genes/Pathways) (T) represent the items (I):

- For the Genes : $I = \{G_1, G_2, ..., G_x\}$.
- For the Pathways : $I = \{P_1, P_2, ..., P_{x'}\}$.

Each antecedent and consequent is a subset of I depending on whether the Genes/Pathways are Drug or Disease targets.

A Context-Free Grammar is used to represent the association rules that can be generated between Drug and Disease targets. The general form of the grammar for the association rules between Drug and Disease Targets *(T)* is defined as follows:

$G = \{S, N, T, P\}$
$S = \{Rule\}$
$N = \{Rule, Antecedent, Consequent, Target\_Ant, Target\_Cons\}$
$T = \{Ta_1, Ta_2, ..., Ta_n, Tc_1, Tc_2, .., Tc_m\}$
$P = \{< Rule >::=< Antecedent > SEP < Consequent >$
$< Antecedent >::=< Target\_Ant > | < Target\_Ant >< Antecedent >$
$< Consequent >::=< Target\_Cons > | < Target\_Cons >< Consequent >$
$< Target\_Ant >::= Ta_1|Ta_2|...|Ta_n$
$< Target\_Cons >::= Tc_1|Tc_2|...|Tc_m\}$

The Ta_i and Tc_j used in the grammar are either Genes or Pathways. Ta_i is used to denote a Gene/Pathway used in the antecedent and Tc_j to denote a Gene/Pathway used in the consequent.

All the non terminal symbols will be substituted by terminals by selecting an appropriate production rule using the Genotype-Phenotype mapping function (*MOD*) to transform each vector of integers (Genotype) to an Association Rule (Phenotype) (see Chapter 4) :

Rule = (*Codon*) *MOD* (*Number of Alternatives of the production rule*)

6.3.1.2 Evaluation of the Extracted Rules

The data set is randomly divided into four parts, and for each one used *3/4* were used for the training and the remaining *1/4* for the tests.

Each rule R with Support and Confidence higher than the minimum support (*MinSup*) and minimum confidence (*MinConf*), respectively, is first evaluated on the training dataset by calculating its fitness using Equation (4.1).

- $Fitness(R) = (a * Support(R)) + (b * Confidence(R))$

The weights a and b are two parameters decided empirically by selecting the values that gave the best results after executing the program many times with different values of a and b ; we have ended up with $a = b = 0.5$ The best set of rules is then checked against the test data set to measure its accuracy which is defined using the True Positive (TP), True Negative (TN), False positive (FP) and False Negative as indicated in Equation (**??**)

- $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

6.3.2 GEARM Process for Genes/pathways Associations

The *GEARM* process (as presented in Chapter 4) starts by defining specific parameters. The parameters of the genetic algorithm were chosen based on different combinations of population and generation size. Below is the set of values that have led to the best performance:

- population size (N) = 200 individuals; generation size (NB) = 500;
- crossover rate = 0.9; mutation rate = 0.01.

The GE parameters have also been defined on the basic of the performance of the method. The chosen value for the minimum chromosome size ensures that at least one Gene/Pathway will appear in the antecedent and one in the consequent, so that no part of the rule will be empty. On the other hand, a maximum chromosome size limits the number of Genes/Pathways in the antecedent and the consequent to 3 Genes/Pathways maximum. The codon size is related to the number of Genes/Pathways used.

- wrap count = 2;
- codon = 1850;
- minimum chromosome size = 4;
- maximum chromosome size = 12.

A set of ARs is generated for each kind of dataset used. Each rule is evaluated according to the fitness function (Equation 4.1) and the best rules are selected for crossover and mutation which are performed on the chromosome level (Vector of integers). The process is repeated for a maximum number of interactions. The accuracy of each generated rule in the final set is calculated according to Equation **??**

Here is an example of the form of the generated association rules of the Disease-Drug Targets:

- For the *Genes* dataset : *If* G_1 *and* G_3 *and* G_5 *Then* G_3
- For the *Pathways* dataset : *If* P_2 *and* P_6 *Then* P_1 *and* P_3

It is important to mention that any Gene/Pathway can be common between a drug and a disease, i.e. related to a specific disease and target a specific drug, like the G_3 in the first rule of the previous example.

The two rules given in the example represent the association between the considered biological entities to represent the association between Drug and Disease in general (i.e. it is not specified which Genes/Pathways are related to disease and which Genes/Pathways are targets drug for each rule).

6.3.3 Types of Drug-Disease Association Rules

Two possible types of rules are identified depending on the Genes/Pathways that appear in the antecedent and the consequent part.

6.3.3.1 Disease → Drug Rules

In this type of AR the Genes/Pathways of the antecedent are disease targets and the Genes/Pathways of the consequent are drug targets. The rule means that if the $Ta_i(Ga_i/Pa_i)$ is related to a specific disease, then the corresponding drug is applied on the Tc_j (Gc_j/Pc_j).

6.3.3.2 Drug → Disease Rules

In this type of AR the Genes/Pathways of the antecedent are drug targets and the Genes/Pathways of the consequent are disease targets. The rule means that if the drug is applied on $Ta_i(Ga_i/Pa_i)$, then the corresponding disease has affected the Tc_j (Gc_j/Pc_j) .

6.3.4 Extracting (Drug, Disease) Pairs from Target Association Rules

After generating all the sets of rules, the aim is to find the (Drug, Disease) pairs related to each rule. To that end, the minimum number of Drugs (Diseases) related to each

Gene/Pathway in the antecedent and the consequent of the rule is taken into consideration as shown in Figure 6.2.

Let us illustrate the idea with an example.
Given the rule of Genes data which is related to the type $Disease \longrightarrow Drug$:

if G_1 *and* G_2 *and* G_3 *Then* G_4 *and* G_5
In order to find the (Drug, Disease), pairs the steps are as follow:

- First, we look for diseases that are affect the three genes of the antecedent at once. Let us suppose they are DI_1 and DI_2.
- Then we look for the drugs that target the two genes of the consequent part also at once. Let us suppose it is DR_1.

The pairs that can be deduced should combine all the found drugs and diseases. Thus the generated pairs will be (DR_1, DI_1) and (DR_1,DI_2).

If (DR_1, DI_1) happens to be already known, then (DR_1,DI_2) is suggested as a new pair.
Let us take another example for Pathway data and the type of rule related to $Drug \longrightarrow Disease$.

if P_2 *and* P_5 *Then* P_5
In order to find the (Drug, Disease) pairs, the steps are as follow:

- First we look for drugs that target the two Pathways of the antecedent part at once. Let us suppose they are DR_2, DR_5 and DR_8.
- Then we look for the diseases that affect the Pathway of the consequent part. Let us suppose it is DI_1 and DI_3.

The pairs that can be deduced should combine all the found drugs and diseases. Thus the generated pairs will be (DR_2, DI_1), (DR_2,DI_3), (DR_5, DI_1), (DR_5, DI_3), (DR_8, DI_1) and (DR_8, DI_3).

If supposing that (DR_2,DI_3) and (DR_8, DI_1) are known, then the suggestd new pairs will be (DR_2, DI_1), (DR_5, DI_1), (DR_5, DI_3), and (DR_8, DI_3).

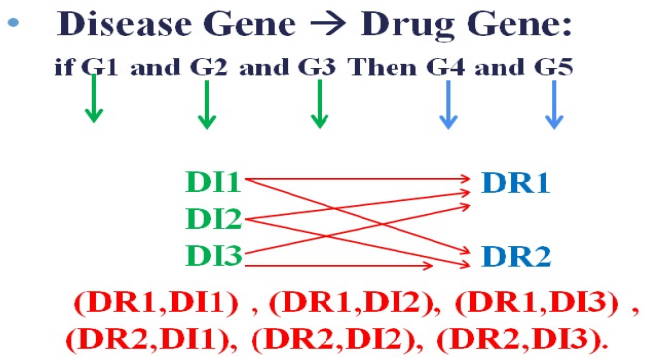


FIGURE 6.2: Predicting new pairs (Drug, Disease) based on Genes (G)

6.3.5 Tests and Results

The GEARM approach was applied to predict new (Drug, Disease) combinations to benefit from the power of ARM in solving the drug repositioning problem.

Table 6.2 gives a summary of the evaluation of the generated sets of rules. It shows the number of rules (*Nb_Rules*), the average of the Fitness values (*Avr_Fit*) and the average of the Accuracy values (*Avr_Accr*). All this information is given for each type of rules and for both Pathway and Gene datasets.

Figures 6.3 and 6.4 show the graphical representation of some generated rules with the highest accuracy for Genes and Pathways, respectively. It is expected that each rule can have at most three elements on the Left Hand Side and three elements on the Right Hand Side. The arrow *in* represents the antecedent of the rule while, the *out* one represents the consequent.

Table 6.3 gives the number of generated drug-disease pairs. The number of known pairs found (Nb KnownP) and the number of unknown pairs (NB UnknownP) for the two types of rules target Genes and Pathways are given. It is clear from analysing the obtained results that GEARM was able to find some known pairs and suggest some new ones.

From the generated rule sets, it was possible to find some already known pairs and to discover others that are unknown. From the results, we can mention a few found pairs.

The known pairs *(Icosapent, Breast Cancer) and (Icosapent, Colorectal Cancer)* of (drug, disease) were found from the rule *(if PIK3CA and TP53 then PTGS2)* of the Gene data related to *Disease* $\longrightarrow$ *Drug*. This same rule can lead to discover new pairs by combining all the diseases and drugs target Genes. It is known that:

TABLE 6.2: The results of the evaluation of the generated rules using GEARM: Number of Rules (Nb_Rules), Average Fitness (Avr_Fit), Average Accuracy (Avr_Accr).

	Nb_Rules	Avr_Fit	Avr_Accr
DI_Genes $\longrightarrow$ DR_Genes	200	0.389	0.921
DR_Genes $\longrightarrow$ DI_Genes	200	0.270	0.959
DI_Pathways $\longrightarrow$ DR_Pathways	52	0.222	0.960
DR_Pathways $\longrightarrow$ DI_Pathways	170	0.200	0.961

- Genes *(PIK3CA)* and *(TP53)* are related to *Breast Cancer* and *Colorectal Cancer* diseases.
- Gene *(PTGS2)* is targeted by the *Icosapent* drug and *Dihomo-linolenic acid* drug .

In addition to the known pairs that have been found, we have discovered two other pairs *(Dihomo-linolenic acid, Breast Cancer)* and *(Dihomo-linolenic acid, Colorectal Cancer)* as new pairs of drug-disease.

The rule *(if ADRB1 and VEGFA then IL6)* related to *Drug* $\longrightarrow$ *Disease*, can find three different pairs, two of which are known, *(Carvedilol, Inflammatory Bowel)*, *(Carvedilol, Rheumatoid Arthritis)*, and a new one is *(Carvedilol, Diabetes Mellitus)*. The pairs have been discovered by combining all the drugs and diseases that target all the genes present in the generated rule.

- Genes *(VEGFA)* and *(ADRB1)* target the drug *Carvedilol*
- Gene *(IL6)* related to *Inflammatory Bowel*, *Rheumatoid Arthritis* and *Diabetes Mellitus* diseases.

Obviously, the pairs generated by GEARM and which are not already known can represent a gold mine for pharmacologists to be validated experimentally.

TABLE 6.3: The generated Drug-Disease pairs: Number of known pairs (Nb_KwnP), Number of unknown pairs (NB_UnkwnP).

	Nb_KwnP	NB_UnkwnP	Accuracy
DI_Genes ⟶ DR_Genes	156	634	92 %
DR_Genes ⟶ DI_Genes	261	1151	92 %
DI_Pathways ⟶ DR_Pathways	18	708	92 %
DR_Pathways ⟶ DI_Pathways	348	2321	91 %

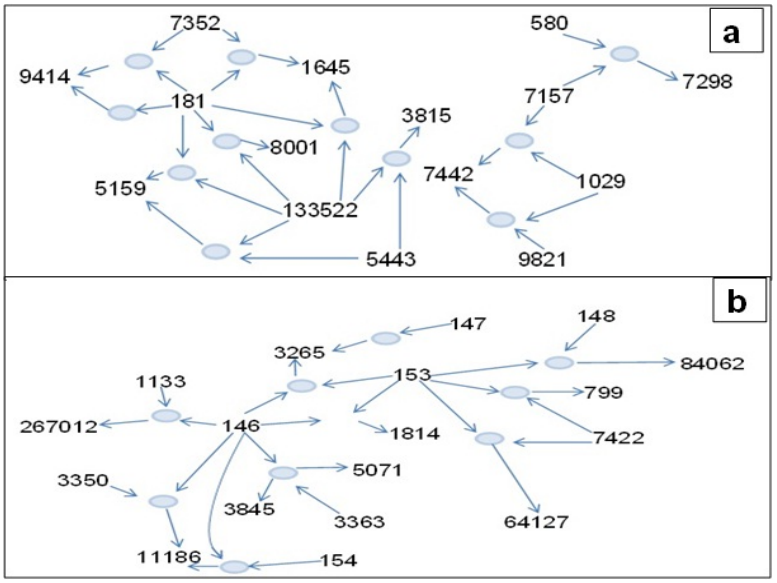


FIGURE 6.3: Graphical representation of some generated association rules with accuracy 0.9 for the **Genes** datasets.
(a): the graph represents Disease → Drug. Example of a rule from the graph *if 1029 and 7157 then 7442.*
(b): the graph represents Drug → Disease, example of rule from the graph *if 146 and 3363 then 5071 and 3845* .

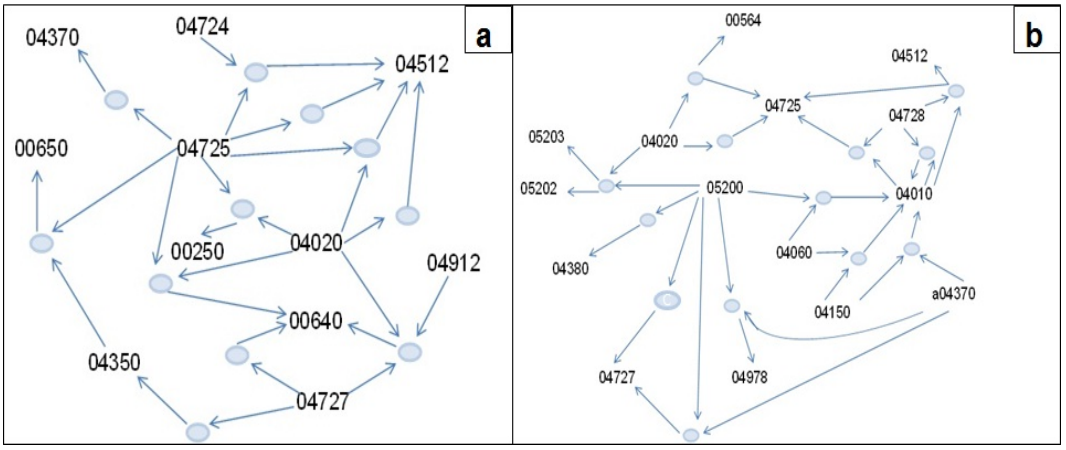


FIGURE 6.4: Graphical representation of some generated association rules with accuracy 0.9 for **Pathways** datasets.
(a): Disease → Drug rules. **(b)**: Drug → Disease rules.

6.4 Conclusion

In this chapter, we have presented the application of GEARM to the problem of drug repositioning using two different types of datasets: Genes and Pathways. In order to find new pairs from the learned set of ARs, two types of rules were considered $Drug \longrightarrow Diseases$ and $Disease \longrightarrow Drug$, both of which were investigated rule by rule and a set of drug-disease pairs was generated.

The sets of rules that have been generated were based on the disease target Genes/Pathways and the drug target Genes/Pathways. The results we have reached have shown the power of GEARM to find some existing pairs of drug-disease and to discover new, unknown pairs which can be a new indication for a drug. The proposed intelligent model combines the advantages of two main strategies used to predict new indications of existing drugs in a unified framework.

One strategy is similarity-based methods which exploit the similarities between drugs and between diseases independently, and assume that similar drugs are likely to combat similar diseases. The other is to discover the underlying biological mechanisms between drugs and their corresponding diseases and use the drug-disease relationships to introduce drug repositioning frameworks.

Using association rules, similar drugs that have the same targets are likely to treat the same diseases. Moreover, it is possible to hypothesise about Genes or Pathways that underlay each pair of new drug and disease. This will advance our understanding and lead to testing new drugs in the lab.

Part IV

General Conclusion and Future Work

Conclusion

This thesis has covered a range of proposals for hybrid intelligent models that were created as powerful techniques able to deal with the complexity of the Biological databases. The developed models have been successfully used on different benchmark datasets and provided promising results for Genetic Diseases and Drug Repositioning.

The thesis began by presenting the problem addressed in this work, as well as the motivation behind getting into this field of Computational Biology.

A Biological Overview that covered the basic concepts of Biology that are needed to understand and tackle the problem at stake was then given. Followed by a presentation on the two main pillars of the thesis problem statement. First, we described the concept of *Gene-Gene* interaction or *Epistasis* and its relationship with traits, based on SNPs of complex diseases. Four different multi-factorial diseases susceptible to genetic variation were taken into consideration: *Autism*, *Mental Retardation*, *Colon Cancer* and *Breast Cancer*. Second, Drug Repositioning concept was presented by showing its important role in saving time and money that lead to saving patient's life by getting the treatment earlier. A list of its benefits on the industrial, academic and clinical levels was as well provided.

All the Computational Intelligent Techniques used in the current contributions were, also, presented, including: Association Rule Mining , Classification using *ANN* with three different types: *MLP*, *RBF* and *FTD*, two Evolutionary Algorithms: *GE* and *GA*, and feature selection techniques that were used for the comparison study: *Filter*, *wrapper* and *Embedded* methods. We followed the techniques definitions by a state of the art of Hybrid ARM- and ANN- based EA. By stating these coupling methods, we were able to acquire a good knowledge about the usefulness of such hybrid techniques as used in different areas in the literature. This gave us a high motivation to design improved models based on the mentioned techniques for complex biological problems.

The coupling of Biological problems and Computational Techniques has given birth to a new field known as Bioinformatics or Computational Biology. The definition of this new area as well as some datasets developed to store the huge amount of data generated by Biology in Computational format were addressed too. The powerlessness of statical methods to deal with such huge complex databases was explained and the need for Data Mining and Machine Learning was covered. DM and ML methods indeed need to be

developed to computationally model the relationships that exist between: first, combinations of SNPs and disease susceptibility, second, biological entities related to drug and disease for Drug Repositioning. This is so because traditional statistical approaches have limited power for modelling high-order non-linear interactions and Associations that are likely to be important in the etiology of complex diseases and Drugs. A detailed state of the art and survey about using computational techniques for GWAS and Drug Repositioning was presented. Based on the literature survey, our aim is to create an intelligent model for the extraction of associations among biological entities in order to reach a better performance for genetics relationships in complex diseases and Drug Repositioning.

The thesis aimed at developing a new intelligent algorithm for Association Rule extraction. A technique named GEARM was developed to extract ARs from any kind of database and for any problem. To evaluate the GEARM algorithm and prove its effectiveness, it was tested on various datasets frequently used in data mining, and compared to different optimized and exact approaches. The results showed that our paradigm outperformed some existing optimized algorithms in terms of solution quality. As well as, it was proved the performance of the optimized techniques to gain more in time compared to the exact approaches.

Since the main focus of this thesis is the extraction of genomic relationships, GEARM was first applied on a simulated SNP dataset that represented Epistasis models in order to tackle the problem of gene-gene interaction. Compared to GEDT, the results indicated that while GEARM and GEDT can both detect gene-gene interactions, GEARM could do it more efficiently and had higher power to detect two-locus interactions. Also, from the results, we could see that GEARM provided an efficient mechanism for the classification of individuals and the detection of gene-gene interactions in the presence or absence of main effects. the proposal has yielded a reduced set of association rules. Also,with the small association rules set that was generated, we have managed to cover all the SNPs in the dataset.

Next, a hybrid GEARM with ANN was developed to classify SNP data for Breast Cancer on 500.000+ SNPs against 111 samples, a clear case of "curse of dimensionality" problem. The proposed approach dealt with the high number of SNP by selecting the best features and classifying the resulting ones. A classification accuracy of up to 90% was reached. GEARM was used to ensure the best selection of the input vector, while a was used for classification. We have proposed two different ideas to reduce the dimensionality

of the problem; they differ in their way of extracting association rules among the data. The first, that we called *Grammatical Evolution Association Rule Mining for Overall Extraction (GEARM-OE)*, achieves the extraction of the rules on the entire data, while *Grammatical Evolution Association Rule Mining for Parallel Extraction (GEARM-PE)* extracts the rules from two separated sets according to the samples class. The accuracies we have reached using ANN have shown that *GEARM-PE* had a better performance than *GEARM-OE*, although both gave encouraging results and can be considered as useful techniques for features selection.

To improve the Neural Network-Grammatical Evolution Association Rule Mining (NN-GEARM) model and apply it for other diseases, we have optimised this proposed method with a Genetic Algorithm to create a new hybrid intelligent model *Genetic Algorithm-Neural Network- Grammatical Evolution Association Rule Mining (GA-NN-GEARM)* which was applied on four different SNPs datasets: Autism, Mental Retardation, Colon Cancer, and Breast cancer for complex disease SNP selection and classification. GEARM was used for dimensionality reduction, which is the key for high classification accuracy, by parallel extraction rules to select the best SNPs. The set of best features that were selected was used as input to the ANN to perform the classification task. GA was used to set the parameters of the model, finding the best architecture for the ANN (hidden neurons, number of iterations, ...) and the number of rules to be extracted by GEARM. We have used three types of ANNs: Multi Layer Perceptron, Radial Basis Function and Focused Time Delay a specific kind of Dynamic ANN. The results proved the high performance of our proposed approach for complex disease based on SNPs interaction, compared to other feature selection and classification techniques. The proposed hybrid model reached an accuracy of 100% for such challenging database, proving its high performance for such complex problems.

Beside genetic diseases, finding a new use for an already approved drug was another problem addressed by this thesis. *GEARM* was also used to tackle of the problem of Drug Repositioning which is defined as finding new uses for a known drug. We have used tow kinds of datasets, Genes and Pathways. GEARM extracted a set of ARs between Genes/Pathways targets drug and Genes/Pathways related to disease in order to find new pairs of drug-disease. Two kinds of rules were considered, rules where Genes/Pathways targets drug represent the antecedent part and the Genes/Pathways related to disease represent the consequent part, and rules where the Genes/Pathways related to disease represent the antecedent part and the Genes/Pathways targets drug represent the consequent part. A set of *(Drug-Disease)* pairs was successfully found for each kind of extracted rule and each type of data targets. Thus we aim to find

hidden relations between drug targets and disease-related Genes/Pathways in order to reprofile existing drugs. The technique was applied to 288 drugs and 267 diseases, forming 5018 known drug-disease pairs. Based on the learned rule sets representing hidden relationships among Gene/Pathways targets, the method discovered interesting pairs of drugs and diseases. The results produced by this combination showed a high accuracy of up to 95 % for the extracted rules. Likewise, the suggested approach was able to discover interesting pairs of drugs and diseases with an accuracy of 92 %. Some of these pairs have previously been reported in the literature while others can serve as new hypotheses to be explored.

Future Work

The thesis has presented a rule mining technique using GE and a number of hybrid approaches aimed at extracting genetic association for complex diseases by using ANN and GA and to repurpose already known drugs. In keeping with this line of research in hybrid intelligent system for biological and medical problems, we give some possible avenues for further work.

GEARM was proposed to optimise the extraction of ARs using GE as a sequential Algorithm. The approach has to be improved for dealing with larger datasets especially in better time. Indeed, when the number of transactions becomes too large, the evaluation of the solution requires the exploration of the entire transaction database and hence the execution time increases drastically. To address this problem, it will be interesting to use a parallel version of the approach using a Graphics Processing Unit (GPU), for example. As a perspective of this contribution we have to investigate a parallel implementation of GEARM.

To avoid generating redundant information in the set of ARs, methods for pruning and combining the extracted ARs is useful to be involved in the GEARM algorithm. We consider the redundancy that some items can appear in the antecedent of a rule and in the consequent of another rule in the same set. This situation indicates the existence of a relationship between the two rules based on the shared items. In one of our contributions, AR was used for feature selection, where all the features appearing in the antecedent of the generated rules with a good fitness value are considered. A filter technique is about to be proposed to face any possible information redundancy under a well defined logic.

The thesis has presented different NN types used with GA-NN-GEARM on complex disease SNP data which gave promising results. However, a further work needs to be done to improve this model. In addition to the selection of the parameters and the best architecture of the ANN, a selection of the training and test subsets is not less important. For the challenging datasets that represent a clear case of "curse of dimensionality", and with the small number of the available samples, the way of splitting the data into two subsets has a significant influence on the efficiency of the model. A better splitting subsets leads to a better training process, which in its turn leads to a better performance of the model. We aim to develop a new approach to divide the set of data, with a good selection of the instances, into two groups used in the training and test phases. As an initial idea, the proposed approach is based on an optimization technique just as TWIST uses genetic algorithm to perform the data splitting. Also, a series of comparisons with powerful techniques of feature selection is needed especially for those that gave promising results in the medical field like the TWIST algorithm.

For the drug repositioning aspect, one important direction to follow, is to develop a hybrid model based on ARM to find interesting links between Drugs and Diseases based on several biological entities. In another word, different kinds of data that provide important information about the disease development and the drug application will be integrated in the same framework. The drug can be characterized by many features like chemical compound; side effects; genes, pathways, protein targets; ontologies, and others. Also, a disease is characterised by different features such as phenotype; disease ontologies; genes, pathways, protein targets; and others. Instead of evaluating the generated rule between drug and disease based only on the association between the genes, or the pathways separately, it will take into account all the previous mentioned entities in the same measure of fitness function. A series of comparisons with promising drug repositioning techniques is necessary in order to situate our approach in term of efficiency for this young promising research area.

Finally, and as any performed research work in Bioinformatics and Computational Biology, the aim is to arrive to the state of the clinical trial and medical validation for the reached results, and the real use of the proposed techniques in our medical centres and hospitals. The the noble objective behind the interest in this area is to contribute in the improvement of the humans health which leads to the improvement of life.

Bibliography

Abrahams, B. S. and Geschwind, D. H. (2008). Advances in autism genetics: on the threshold of a new neurobiology. *Nature Reviews Genetics*, 9(5):341–355.

ADN – Evolutions, A. (2015). *LEvolution liée à lidentification génétique*. http://www.police-scientifique.com/adn/evolutions/, Consulted: March 2016.

Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *In: Proceedings of the 1993 ACM SIGMOD international conference on Management of data. New York, NY, USA*, volume 22, pages 207–216. ACM.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499.

Aguiar, V., Seoane, J. A., Freire, A., and Guo, L. (2010). Ga-based data mining applied to genetic data for the diagnosis of complex diseases. *Soft Computing Methods for Practical Environment Solutions: Techniques and Studies*, pages 219–239.

Ahmad, A. M., Khan, G. M., Mahmud, S. A., and Miller, J. F. (2012). Breast cancer detection using cartesian genetic programming evolved artificial neural networks. In *In: Proceedings of the 14th annual conference on Genetic and evolutionary computation GECCO '12.New York, NY, USA*, pages 1031–1038. ACM.

Ahmadizar, F., Soltanian, K., AkhlaghianTab, F., and Tsoulos, I. (2015). Artificial neural network development by means of a novel combination of grammatical evolution and genetic algorithm. *Engineering Applications of Artificial Intelligence*, 39:1–13.

Altman, R. B. (2004). Editorial: Building successful biological databases. *Briefings in bioinformatics*, 5(1):4–5.

Amberger, J., Bocchini, C. A., Scott, A. F., and Hamosh, A. (2009). Mckusick's online mendelian inheritance in man (omim®). *Nucleic acids research*, 37(suppl 1):D793–D796.

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., and Hamosh, A. (2015). Omim. org: Online mendelian inheritance in man (omim®), an online catalog of human genes and genetic disorders. *Nucleic acids research*, 43(D1):D789–D798.

American Speech-Language-Hearing Association, A. (2016). Autism. http://www.asha.org/PRPSpecificTopic.aspx?folderid=8589935303&section=Overview, Consulted: March 2016.

Andrade-Vieira, R., Xu, Z., Colp, P., and Marignani, P. A. (2013). Loss of lkb1 expression reduces the latency of erbb2-mediated mammary gland tumorigenesis, promoting changes in metabolic pathways. *PloS one*, 8(2):e56567.

Anekboon, K., Lursinsap, C., Phimoltares, S., Fucharoen, S., and Tongsima, S. (2014). Extracting predictive snps in crohn's disease using a vacillating genetic algorithm and a neural classifier in case–control association studies. *Computers in Biology and Medicine*, 44:57–65.

Angelini, D. (2009). Bio 110: General biology-course syllabus.

Antonie, L. and Bessonov, K. (2011). Classifying microarray data with association rules. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, pages 94–99. ACM.

Ashburn, T. T. and Thor, K. B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. *Nature reviews Drug discovery*, 3(8):673–683.

Association, D.-. A. P. (2013). Diagnostic and statistical manual of mental disorders. *Arlington: American Psychiatric Publishing*.

Attwood, T., Gisel, A., Bongcam-Rudloff, E., and Eriksson, N. (2011). *Concepts, historical milestones and the central place of bioinformatics in modern biology: a European perspective*. INTECH Open Access Publisher.

Avery, O. T., MacLeod, C. M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *The Journal of experimental medicine*, 79(2):137–158.

Baldassarre, D., Grossi, E., Buscema, M., Intraligi, M., Amato, M., Tremoli, E., Pustina, L., Castelnuovo, S., Sanvito, S., and Gerosa, L. (2004). Recognition of patients with cardiovascular disease by artificial neural networks. *Annals of Medicine*, 36(8):630–640.

Barrett, T. and Edgar, R. (2006). Gene expression omnibus: Microarray data storage, submission, retrieval, and analysis. *Methods Enzymology*, 411:352–369.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., and Holko, M. (2013). Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research*, 41(D1):D991–D995.

Batnyam, N., Gantulga, A., and Oh, S. (2013). An efficient classification for single nucleotide polymorphism (snp) dataset. In *Computer and Information Science*, pages 171–185. Springer.

Baumann, K. (2003). Cross-validation as the objective function for variable-selection techniques. *TrAC Trends in Analytical Chemistry*, 22(6):395–406.

Biology Notes for IGCSE, B. (2014). *Chromosomes, DNA, genes and alleles*. http://igbiologyy.blogspot.com/2014/03/chromosomes-dna-genes-and-alleles.html, Consulted: March 2016.

Biomedecine, B. (2012). *Repurposing drugs*. http://www.davidfunesbiomed.eu/2016/02/136-repurposing-drugs.html, Consulted: March 2016.

Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Advanced Texts in Econometrics. Clarendon Press. https://books.google.fr/books?id=T0S0BgAAQBAJ.

Bishop, C. M. (1999). Pattern recognition and feed-forward networks. *The MIT encyclopedia of the cognitive sciences. MIT, Cambridge*, pages 629–632.

Bodmer, W. F. and McKie, R. (1997). *The book of man: the Human Genome Project and the quest to discover our genetic heritage*. Oxford University Press on Demand.

Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J., and Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. *Information Sciences*, 282:111–135.

Boolell, M., Allen, M., Ballard, S., Gepi-Attee, S., Muirhead, G., Naylor, A., Osterloh, I., and Gingell, C. (1996). Sildenafil: an orally active type 5 cyclic gmp-specific phosphodiesterase inhibitor for the treatment of penile erectile dysfunction. *International journal of impotence research*, 8(2):47–52.

Booth, B. and Zemmel, R. (2004). Prospects for productivity. *Nature Reviews Drug Discovery*, 3(5):451–456.

Brin, S., Motwani, R., Ullman, J. D., and Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. In *SIGMOD '97 Proceedings of the 1997 ACM SIGMOD international conference on Management of data. New York, NY, USA*, volume 26, pages 255–264. ACM.

Buscema, M. (2004). Genetic doping algorithm (gend): theory and applications. *Expert Systems*, 21(2):63–79.

Buscema, M., Breda, M., and Lodwick, W. (2013). Training with input selection and testing (twist) algorithm: a significant advance in pattern recognition performance of machine learning. *Journal of Intelligent Learning Systems and Applications*, 5(1):10, DOI:10.4236/jilsa.2013.51004.

Buscema, M., Consonni, V., Ballabio, D., Mauri, A., Massini, G., Breda, M., and Todeschini, R. (2014). K-cm: A new artificial neural network. application to supervised pattern recognition. *Chemometrics and Intelligent Laboratory Systems*, 138:110–119.

Buscema, M., Grossi, E., Capriotti, M., Babiloni, C., and Rossini, P. (2010). The ifast model allows the prediction of conversion to alzheimer disease in patients with mild cognitive impairment with high degree of accuracy. *Current Alzheimer Research*, 7(2):173–187.

Buscema, M., Terzi, S., and Breda, M. (2006). Using sinusoidal modulated weights improve feed-forward neural network performances in classification and functional approximation problems.. *WSEAS Transactions on information science and applications*, 3(5):885–893.

Buscema, M., Vernieri, F., Massini, G., Scrascia, F., Breda, M., Rossini, P. M., and Grossi, E. (2015). An improved i-fast system for the diagnosis of alzheimer's disease from unprocessed electroencephalograms by using robust invariant features. *Artificial intelligence in medicine*, 64(1):59–74, DOI:10.1016/j.artmed.2015.03.003.

Bush, W. S., Motsinger, A. A., Dudek, S. M., and Ritchie, M. D. (2005). Can neural network constraints in gp provide power to detect genes associated with human disease? In *Applications of Evolutionary Computing*, pages 44–53. Springer.

Buxbaum, J. D. (2009). Multiple rare variants in the etiology of autism spectrum disorders. *Dialogues Clin Neurosci*, 11(1):35–43.

Cai, R., Hao, Z., Wen, W., and Huang, H. (2010). Kernel based gene expression pattern discovery and its application on cancer classification. *Neurocomputing*, 73(13):2562–2570.

Carlborg, Ö. and Haley, C. S. (2004). Epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics*, 5(8):618–625.

Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi, Q., Kruglyak, L., and Nickerson, D. A. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for

association analyses using linkage disequilibrium. *The American Journal of Human Genetics*, 74(1):106–120.

Cawley, G. C. and Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11:2079–2107.

Chaste, P. and Leboyer, M. (2012). Autism risk factors: genes, environment, and gene-environment interactions. *Dialogues Clin Neurosci*, 14(3):281–92.

Chawla, S. (2010). Feature selection, association rules network and theory building. In *FSDM*, pages 14–21.

Chen, S., Cowan, C. F., and Grant, P. M. (1991). Orthogonal least squares learning algorithm for radial basis function networks. *Neural Networks, IEEE Transactions on*, 2(2):302–309.

Chen, S.-H., Sun, J., Dimitrov, L., Turner, A. R., Adams, T. S., Meyers, D. A., Chang, B.-L., Zheng, S. L., Grönberg, H., and Xu, J. (2008). A support vector machine approach for detecting gene-gene interaction. *Genetic epidemiology*, 32(2):152–167.

Chen, Y., Abraham, A., and Yang, B. (2006). Feature selection and classification using flexible neural tree. *Neurocomputing*, 70(1):305–313.

Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., Zhou, W., Huang, J., and Tang, Y. (2012). Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput Biol*, 8(5):e1002503.

Chiang, A. P. and Butte, A. J. (2009). Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. *Clinical pharmacology and therapeutics*, 86(5):507.

Cho, S.-B. and Ryu, J. (2002). Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features. *Proceedings of the IEEE*, 90(11):1744–1753.

Cho, S.-B. and Won, H.-H. (2003). Machine learning in dna microarray analysis for cancer classification. In *APBC '03 Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics. Darlinghurst, Australia.*, volume 19, pages 189–198. Australian Computer Society, Inc.

Christin, C., Hoefsloot, H. C., Smilde, A. K., Hoekman, B., Suits, F., Bischoff, R., and Horvatovich, P. (2013). A critical assessment of feature selection methods for biomarker discovery in clinical proteomics. *Molecular & Cellular Proteomics*, 12(1):263–276.

Colditz, G. A., Kaphingst, K. A., Hankinson, S. E., and Rosner, B. (2012). Family history and risk of breast cancer: nurses' health study. *Breast cancer research and treatment*, 133(3):1097–1104.

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

Collins, F. S. and Mansoura, M. K. (2001). The human genome project. *Cancer*, 91(S1):221–225.

Coppedè, F., Grossi, E., Buscema, M., and Migliore, L. (2013). Application of artificial neural networks to investigate one-carbon metabolism in alzheimer's disease and healthy matched individuals. *PloS one*, 8(8):e74012.

Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human molecular genetics*, 11(20):2463–2468.

Craft, J. (2013). Genes and genetics: the language of scientific discovery.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314.

Daily, D. K., Ardinger, H. H., and Holmes, G. E. (2000). Identification and evaluation of mental retardation. *American family physician*, 61(4):1059–67.

Davis, A. P., Grondin, C. J., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky, D., King, B. L., Wiegers, T. C., and Mattingly, C. J. (2014). The comparative toxicogenomics database's 10th year anniversary: update 2015. *Nucleic acids research*, page gku935.

Davis, A. P., Grondin, C. J., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky, D., King, B. L., Wiegers, T. C., and Mattingly, C. J. (2015). The comparative toxicogenomics database's 10th year anniversary: update 2015. *Nucleic acids research*, 43(D1):D914–D920.

Delpisheh, E. (2010). *Two new approaches to evaluate association rules*. PhD thesis, Lethbridge, Alta.: University of Lethbridge, Dept. of Mathematics and Computer Science, c2010.

Demuth, H., Beale, M., and Hagan, M. (2008). Neural network toolbox^{TM} 6. *User's guide*.

Demuth, H., Beale, M., and Works, M. (1992). *MATLAB: Neural Network Toolbox: User's Guide*. Math Works.

Deutsch, J. (2003). Evolutionary algorithms for finding optimal gene sets in microarray prediction. *Bioinformatics*, 19(1):45–52.

DeWeerdt, S. E., Culliton, B. J., and Gibbs, M. S. (2003.). *What's a Genome?* Genomenewsnetwork.org. http://www.genomenewsnetwork.org/resources/whats_a_genome/Chp1_1_1.shtml, Consulted: February 2016.

Dickson, M. and Gagnon, J. P. (2009). The cost of new drug discovery and development. *Discovery Medicine*, 4(22):172–179.

DiMasi, J. A., Hansen, R. W., Grabowski, H. G., and Lasagna, L. (1991). Cost of innovation in the pharmaceutical industry. *Journal of health economics*, 10(2):107–142.

Disability, N. (2011). Intellectual disability. *Center for Parent Information and Resources.* http://www.parentcenterhub.org/repository/intellectual/.

Djenouri, Y., Drias, H., and Habbas, Z. (2014). Bees swarm optimisation using multiple strategies for association rule mining. *International Journal of Bio-Inspired Computation*, 6(4):239–249.

Drenos, F., Grossi, E., Buscema, M., and Humphries, S. E. (2015). Networks in coronary heart disease genetics as a step towards systems epidemiology. 10(5):DOI:10.1371/journal.pone.0125876.

Dreyfus, G. (2005). *Neural Networks: Methodology and Applications*. Springer Berlin Heidelberg. https://books.google.fr/books?id=-feS13CoyS4C.

Dudek, S. M., Motsinger, A. A., Velez, D. R., Williams, S. M., and Ritchie, M. D. (2005). Data simulation software for whole-genome association and other studies in human genetics. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 499–510.

Eldridge, L. (2015). Genetic predisposition. http://lungcancer.about.com/od/glossary/g/geneticrisk.htm, Consulted: February 2016.

Elumalai, A. and Eswaraiah, M. C. (2013). Review on application of bioinformatics. *Journal of science Bioinformatics*, 3(1):21–27.

Espejo, P. G., Ventura, S., and Herrera, F. (2010). A survey on the application of genetic programming to classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 40(2):121–144.

Feero, W. G., Guttmacher, A. E., Feero, W. G., Guttmacher, A. E., and Collins, F. S. (2010a). Genomic medicine—an updated primer. *New England Journal of Medicine*, 362(21):2001–2011.

Feero, W. G., Guttmacher, A. E., Rotimi, C. N., and Jorde, L. B. (2010b). Ancestry and disease in the age of genomic medicine. *New England Journal of Medicine*, 363(16):1551–1558.

Floreano, D., Dürr, P., and Mattiussi, C. (2008). Neuroevolution: from architectures to learning. *Evolutionary Intelligence*, 1(1):47–62.

Fodor, I. K. (2002). A survey of dimension reduction techniques.

Fukuoka, Y., Takei, D., and Ogawa, H. (2013). A two-step drug repositioning method based on a protein-protein interaction network of genes shared by two diseases and the similarity of drugs. *Bioinformation*, 9(2):89–93.

Gage, M., Wattendorf, D., and Henry, L. (2012). Translational advances regarding hereditary breast cancer syndromes. *Journal of surgical oncology*, 105(5):444–451.

Genetics course at UW-Madison, G. (2014). *Hereditary Fructose Intolerance (HFI) and the Aldolase B Gene*. http://fordmadelinegen564s14.weebly.com/, Consulted: March 2016.

Genetics Home Reference, G. What are complex or multifactorial disorders? *in Citing Medicine: The NLM Style Guide for Authors, Editors, and Publishers*. https://ghr.nlm.nih.gov/handbook/mutationsanddisorders/complexdisorders.

Genetics Home Reference, G. What is a gene mutation and how do mutations occur? *in Citing Medicine: The NLM Style Guide for Authors, Editors, and Publishers*. http://ghr.nlm.nih.gov/handbook/mutationsanddisorders/genemutation.

Genetics Home Reference, G. (2016a). What does it mean to have a genetic predisposition to a disease? *in Citing Medicine: The NLM Style Guide for Authors, Editors, and Publishers*. http://ghr.nlm.nih.gov/handbook/mutationsanddisorders/predisposition.

Genetics Home Reference, G. (2016b). What is dna? *in Citing Medicine: The NLM Style Guide for Authors, Editors, and Publishers*. http://ghr.nlm.nih.gov/handbook/basics/dna, Consulted: March 2016.

Gibson, G. and Muse, S. V. (2009). *primer of genome science*. Sinauer Associates.

Gillespie, J. H. (2010). *Population genetics: a concise guide*. JHU Press.

Gillham, N. (2011). *Genes, Chromosomes, and Disease: From Simple Traits, to Complex Traits, to Personalized Medicine*. Ft Press Science Series. FT Press. https://books.google.fr/books?id=tmYNayPcf_sC.

Gironi, M., Borgiani, B., Farina, E., Mariani, E., Cursano, C., Alberoni, M., Nemni, R., Comi, G., Buscema, M., and Furlan, R. (2015). A global immune deficit in alzheimer's disease and mild cognitive impairment disclosed by a novel data mining process. *Journal of Alzheimer's disease: JAD*, 43(4):1199–1213.

Giugno, R., Pulvirenti, A., Cascione, L., Pigola, G., and Ferro, A. (2013). Midclass: Microarray data classification by association rules and gene expression intervals. *PloS one*, 8(8):e69873.

Goethals, B. and Zaki, M. (2003). Frequent itemset mining implementations repository. *This site contains a wide-variety of algorithms for mining frequent, closed, and maximal itemsets, http://fimi. cs. helsinki. fi*.

González-Recio, O., Rosa, G. J., and Gianola, D. (2014). Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. *Livestock Science*, 166:217–231.

Gottlieb, A., Stein, G. Y., Ruppin, E., and Sharan, R. (2011). Predict: a method for inferring novel drug indications with application to personalized medicine. *Molecular systems biology*, 7(1):496.

Grossi, E. and Buscema, M. (2007). Introduction to artificial neural networks. *European journal of gastroenterology & hepatology*, 19(12):1046–1054.

Grossi, E., Mancini, A., and Buscema, M. (2007). International experience on the use of artificial neural networks in gastroenterology. *Digestive and liver disease*, 39(3):278–285.

Grossi, E., Podda, G. M., Pugliano, M., Gabba, S., Verri, A., Carpani, G., Buscema, M., Casazza, G., and Cattaneo, M. (2014). Prediction of optimal warfarin maintenance dose using advanced artificial neural networks. *Pharmacogenomics*, 15(1):29–37.

Guttmacher, A. E. and Collins, F. S. (2003). Welcome to the genomic era. *New England Journal of Medicine*, 349(10):996–998.

Guvenir, H. A. and Uysal, I. (2000). Bilkent university function approximation repository.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182.

Halakou, F., Eftekhari, M., and Esmailizadeh, A. (2015). Comparison of hybrid and filter feature selection methods to identify candidate single nucleotide polymorphisms. *Journal of Computing and Security*, 1(3).

Hamosh, A., Scott, A. F., Amberger, J., Bocchini, C., Valle, D., and McKusick, V. A. (2002). Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 30(1):52–55.

Han, J., Kamber, M., and Pei, J. (2011). *Data mining: concepts and techniques*. Elsevier.

Han, J., Pei, J., Yin, Y., and Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery*, 8(1):53–87.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12.

He, H., Oetting, W., Brott, M., and Basu, S. (2010). Pair-wise multifactor dimensionality reduction method to detect gene-gene interactions in a case-control study. *Human heredity*, 69(1):60.

He, J., Hu, H.-j., Chen, B., Tai, P. C., Harrison, R., and Pan, Y. (2008). Rule extraction from svm for protein structure prediction. In *Rule Extraction from Support Vector Machines*, pages 227–252. Springer.

Health-Conditions (2013). Colorectal cancer. https://mobile.bluekc.com/HealthFitness/content/major/hw198266.html, Consulted: March 2016.

Helt, M., Kelley, E., Kinsbourne, M., Pandey, J., Boorstein, H., Herbert, M., and Fein, D. (2008). Can children with autism recover? if so, how? *Neuropsychology review*, 18(4):339–366.

Hendrick, R. E. (2010). Radiation doses and cancer risks from breast imaging studies 1. *Radiology*, 257(1):246–253.

Hershey, A. D. and Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. *The Journal of general physiology*, 36(1):39–56.

Hogeweg, P. (2011). The roots of bioinformatics in theoretical biology. *PLoS Comput Biol*, 7(3):e1002021.

Hong, J.-H. and Cho, S.-B. (2006). Efficient huge-scale feature selection with speciated genetic algorithm. *Pattern Recognition Letters*, 27(2):143–150.

Howlin, P., Goode, S., Hutton, J., and Rutter, M. (2004). Adult outcome for children with autism. *Journal of Child Psychology and Psychiatry*, 45(2):212–229.

Hurle, M., Yang, L., Xie, Q., Rajpal, D., Sanseau, P., and Agarwal, P. (2013). Computational drug repositioning: from data to therapeutics. *Clinical Pharmacology & Therapeutics*, 93(4).

Ian, H. and Eibe, F. (2005). Data mining: Practical machine learning tools and techniques. *Morgan Kaufmann, San Francisco*.

Indira, K. and Kanmani, S. (2012). Performance analysis of genetic algorithm for mining association rules.

Islam, S. and Islam, S. (2015). Dealing with intellectually disabled children. *Northern International Medical College Journal*, 7(1):91–93.

Jiawei, H. and Kamber, M. (2001). Data mining: concepts and techniques. *San Francisco, CA, itd: Morgan Kaufmann*, 5.

Jing, Y., Dong, H., and Liang, G. (2012). Study on characteristic of fractional master-slave neural network. In *Computational Intelligence and Design (ISCID), 2012 Fifth International Symposium on*, volume 2, pages 498–501. IEEE.

Jirapech-Umpai, T. and Aitken, S. (2005). Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. *BMC bioinformatics*, 6(1):148.

John Lu, Z. (2010). The elements of statistical learning: data mining, inference, and prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(3):693–694.

Kadota, M., Sato, M., Duncan, B., Ooshima, A., Yang, H. H., Diaz-Meyer, N., Gere, S., Kageyama, S.-I., Fukuoka, J., and Nagata, T. (2009). Identification of novel gene amplifications in breast cancer and coexistence of gene amplification with an activating mutation of pik3ca. *Cancer research*, 69(18):7357–7365.

Kanehisa, M. (1997). A database for post-genome analysis. *Trends in genetics: TIG*, 13(9):375.

Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2011). Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, page gkr988.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in kegg. *Nucleic acids research*, 42(D1):D199–D205.

Kantardzic, M. (2011). *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons.

Karabatak, M. and Ince, M. C. (2009). An expert system for detection of breast cancer based on association rules and neural network. *Expert Systems with Applications*, 36(2):3465–3469.

Ke, X., Taylor, M. S., and Cardon, L. R. (2008). Singleton snps in the human genome and implications for genome-wide association studies. *European Journal of Human Genetics*, 16(4):506–515.

Keiser, M. J., Setola, V., Irwin, J. J., Laggner, C., Abbas, A. I., Hufeisen, S. J., Jensen, N. H., Kuijer, M. B., Matos, R. C., and Tran, T. B. (2009). Predicting new molecular targets for known drugs. *Nature*, 462(7270):175–181.

Kesharaju, M. and Nagarajah, R. (2015). Feature selection for neural network based defect classification of ceramic components using high frequency ultrasound. *Ultrasonics*, 62:271–277.

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., and Neveu, V. (2011). Drugbank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research*, 39(suppl 1):D1035–D1041.

Knudsen, S. (2011). *A Biologist's Guide to Analysis of DNA Microarray Data*. Wiley. https://books.google.fr/books?id=V-f17YO5Jr8C.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324.

Koo, C. L., Liew, M. J., Mohamad, M. S., and Mohamed Salleh, A. H. (2013). A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology. *Biomed Res. Int.*, 2013(2013):13.

Kooperberg, C. and Ruczinski, I. (2005). Identifying interacting snps using monte carlo logic regression. *Genetic epidemiology*, 28(2):157–170.

Kuo, R. and Shih, C. (2007). Association rule mining through the ant colony system for national health insurance research database in taiwan. *Computers & Mathematics with Applications*, 54(11):1303–1318.

Larose, D. T. (2014). *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons.

Lavine, B. and Rayens, W. (2014). Classification: basic concepts. *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis. Vol*, 3.

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., and Neveu, V. (2014). Drugbank 4.0: shedding new light on drug metabolism. *Nucleic acids research*, 42(D1):D1091–D1097.

Lea, D. H. (2009). Basic genetics and genomics: A primer for nurses. *OJIN: The Online Journal of Issues in Nursing*, 14(2).

Li, J., Zheng, S., Chen, B., Butte, A. J., Swamidass, S. J., and Lu, Z. (2015). A survey of current trends in computational drug repositioning. *Briefings in bioinformatics*, page bbv020.

Li, R., Holzinger, E. R., Dudek, S. M., and Ritchie, M. D. (2014a). Evaluation of parameter contribution to neural network size and fitness in athena for genetic analysis. In *Genetic Programming Theory and Practice XI*, pages 211–224. Springer.

Li, S., Kang, L., and Zhao, X.-M. (2014b). A survey on evolutionary algorithm based hybrid intelligence in bioinformatics. *BioMed research international*, 2014.

Li, W. and Reich, J. (2000). A complete enumeration and classification of two-locus disease models. *Human heredity*, 50(6):334–349.

Lisboa, P. J. (2002). A review of evidence of health benefit from artificial neural networks in medical intervention. *Neural networks*, 15(1):11–39.

Lisboa, P. J. and Taktak, A. F. (2006). The use of artificial neural networks in decision support in cancer: a systematic review. *Neural networks*, 19(4):408–415.

Liu, B., Cui, Q., Jiang, T., and Ma, S. (2004). A combinational feature selection and ensemble neural network method for classification of gene expression data. *BMC bioinformatics*, 5(1):136.

Lowe, D. (1995). *Radial Basis Function Networks, in Arbib M. A, The Handbook of Brain Theory and Neural Networks, A Bradford Book.* MIT press, Cambridge, Massachusetts, London, England.

Mahdevar, G., Zahiri, J., Sadeghi, M., Nowzari-Dalini, A., and Ahrabian, H. (2010). Tag snp selection via a genetic algorithm. *Journal of biomedical informatics*, 43(5):800–804.

Manning, T., Sleator, R. D., and Walsh, P. (2013). Naturally selecting solutions: the use of genetic algorithms in bioinformatics. *Bioengineered*, 4(5):266–278.

Manning, T., Sleator, R. D., and Walsh, P. (2014). Biologically inspired intelligent decision making: a commentary on the use of artificial neural networks in bioinformatics. *Bioengineered*, 5(2):80–95.

Marshall, C. R., Noor, A., Vincent, J. B., Lionel, A. C., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D., and Ren, Y. (2008). Structural variation of chromosomes in autism spectrum disorder. *The American Journal of Human Genetics*, 82(2):477–488.

Mata, J., Alvarez, J., and Riquelme, J. (2001). Mining numeric association rules with genetic algorithms. In *Artificial Neural Nets and Genetic Algorithms*, pages 264–267. Springer.

Mata, J., Alvarez, J.-L., and Riquelme, J.-C. (2002). An evolutionary algorithm to discover numeric association rules. In *Proceedings of the 2002 ACM symposium on Applied computing*, pages 590–594. ACM.

McKinney, B. A., Reif, D. M., Ritchie, M. D., and Moore, J. H. (2006). Machine learning for detecting gene-gene interactions. *Applied bioinformatics*, 5(2):77–88.

McMullan, D. J., Bonin, M., Hehir-Kwa, J. Y., de Vries, B., Dufke, A., Rattenberry, E., Steehouwer, M., Moruz, L., Pfundt, R., and de Leeuw, N. (2009). Molecular karyotyping of patients with unexplained mental retardation by snp arrays: a multicenter study. *Human mutation*, 30(7):1082–1092.

Mitchell, M. (1998). *An introduction to genetic algorithms*. MIT press.

Mohamad, M., Omatu, S., Yoshioka, M., and Deris, S. (2009). A cyclic hybrid method to select a smaller subset of informative genes for cancer classification. *International Journal of Innovative Computing Information and Control*, 5(8):2189–2202.

Moody, G. (2004). *Digital code of life: How bioinformatics is revolutionizing science, medicine, and business*. John Wiley & Sons.

Mooney, S. (2005). Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Briefings in bioinformatics*, 6(1):44–56.

Moore, J. H., Asselbergs, F. W., and Williams, S. M. (2010). Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, 26(4):445–455.

Moore, J. H. and Hill, D. P. (2015). Epistasis analysis using artificial intelligence. In *Epistasis*, pages 327–346. Springer New York.

Moore, J. H. and Ritchie, M. D. (2004). The challenges of whole-genome approaches to common diseases. *Jama*, 291(13):1642–1643.

Moore, J. H. and Williams, S. M. (2002). New strategies for identifying gene-gene interactions in hypertension. *Annals of medicine*, 34(2):88–95.

Moore, J. H. and Williams, S. M. (2005). Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *Bioessays*, 27(6):637–646.

Moslehi, P., Bidgoli, B. M., Nasiri, M., and Salajegheh, A. (2011). Multi-objective numeric association rules mining via ant colony optimization for continuous domains without specifying minimum support and minimum confidence. *International Journal of Computer Science Issues (IJCSI)*, 8(1):34–41.

Motsinger, A. A., Dudek, S. M., Hahn, L. W., and Ritchie, M. D. (2006). Comparison of neural network optimization approaches for studies of human genetics. In *Applications of Evolutionary Computing*, pages 103–114. Springer.

Motsinger, A. A. and Ritchie, M. (2008). Neural networks for genetic epidemiology: past, present, and future. volume A.

Motsinger-Reif, A. A., Deodhar, S., Winham, S. J., and Hardison, N. E. (2010). Grammatical evolution decision trees for detecting gene-gene interactions. *BioData mining*, 3(1):1.

Motsinger-Reif, A. A., Dudek, S. M., Hahn, L. W., and Ritchie, M. D. (2008a). Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. *Genetic epidemiology*, 32(4):325–340.

Motsinger-Reif, A. A., Fanelli, T. J., Davis, A. C., and Ritchie, M. D. (2008b). Power of grammatical evolution neural networks to detect gene-gene interactions in the presence of error. *BMC research notes*, 1(1):65.

Musani, S. K., Shriner, D., Liu, N., Feng, R., Coffey, C. S., Yi, N., Tiwari, H. K., and Allison, D. B. (2007). Detection of gene× gene interactions in genome-wide association studies of human population data. *Human heredity*, 63(2):67–84.

Mutalib, S., Abdul-Rahman, S., and Mohamed, A. (2014). Mining frequent patterns for genetic variants associated to diabetes. In *2014 25th International Workshop on Database and Expert Systems Applications (DEXA). Munich.*, pages 28–32. IEEE.

Myers, S. M. and Johnson, C. P. (2007). Management of children with autism spectrum disorders. *Pediatrics*, 120(5):1162–1182.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, 44(1):190–204.

Napolitano, F., Zhao, Y., Moreira, V. M., Tagliaferri, R., Kere, J., D'Amato, M., and Greco, D. (2013). Drug repositioning: a machine-learning approach through data integration. *J. Cheminformatics*, 5:30.

National Human Genome Reseach Institute, N. (2011). Deoxyribonucleic acid. *from National Human Genome Reseach Institute:*. http://www.genome.gov/25520880, Consulted: February 2016.

National Human Genome Research Institute, N. (2015). Biological pathways. http://www.genome.gov/27530687.

Naulaerts, S., Meysman, P., Bittremieux, W., Vu, T. N., Berghe, W. V., Goethals, B., and Laukens, K. (2015). A primer to frequent itemset mining for bioinformatics. *Briefings in bioinformatics*, 16(2):216–231.

Neuman, R. J., Rice, J. P., and Chakravarti, A. (1992). Two-locus models of disease. *Genetic epidemiology*, 9(5):347–365.

Niel, C., Sinoquet, C., Dina, C., and Rocheleau, G. (2015). A survey about methods dedicated to epistasis detection. *Frontiers in genetics*, 6.

North, B., Curtis, D., Cassell, P., Hitman, G., and Sham, P. (2003). Assessing optimal neural network architecture for identifying disease-associated multi-marker genotypes using a permutation test, and application to calpain 10 polymorphisms associated with diabetes. *Annals of human genetics*, 67(4):348–356.

Nowlan, S. J. and Hinton, G. E. (1992). Simplifying neural networks by soft weight-sharing. *Neural computation*, 4(4):473–493.

Nunkesser, R., Bernholt, T., Schwender, H., Ickstadt, K., and Wegener, I. (2007). Detecting high-order interactions of single nucleotide polymorphisms using genetic programming. *Bioinformatics*, 23(24):3280–3288.

Olmo, J. L., Luna, J. M., Romero, J. R., and Ventura, S. (2011). Association rule mining using a multi-objective grammar-based ant programming algorithm. In *Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on*, pages 971–977. IEEE.

O'Neill, M. and Ryan, C. (2001). Grammatical evolution. *IEEE Transactions on Evolutionary Computation*, 5(4):349–358.

O'Neill, M. and Ryan, C. (2003). *Grammatical Evolution: Evolutionary Automatic Programming in an Arbitrary Language*. Kluwer Academic Publishers, Norwell, MA, USA.

Panchal, G., Ganatra, A., Shah, P., and Panchal, D. (2011). Determination of over-learning and over-fitting problem in back propagation neural network. *International Journal on Soft Computing*, 2(2):40–51.

Paolo, G. (2003). Applied data mining: Statistical methods for business and industry.

Parakhia, M. (2009). *Molecular Biology and Biotechnology: Microbial Methods*. New India Publishing Agency.

Park, S., Reyes, J. A., Gilbert, D. R., Kim, J., and Kim, S. (2009). Prediction of protein-protein interaction types using association rule based classification. *BMC bioinformatics*, 10(1):1.

Park, S. H. and Kim, S. (2012). Pattern discovery of multivariate phenotypes by association rule mining and its scheme for genome–wide association studies. *International journal of data mining and bioinformatics*, 6(5):505–520.

Pasche, B. (2010). Cancer genetics (cancer treatment and research).

Pass My Exams Biology, P. (2016). *Sex determination - Is it a boy or a girl?* http://www.passmyexams.co.uk/GCSE/biology/sex-determination.html, Consulted: March 2016.

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., and Schacht, A. L. (2010). How to improve r&d productivity: the pharmaceutical industry's grand challenge. *Nature reviews Drug discovery*, 9(3):203–214.

Peirson, B. (2013). Wilhelm johannsen's genotype-phenotype distinction. *Embryo Project Encyclopedia*.

Penco, S., Grossi, E., Cheng, S., Intraligi, M., Maurelli, G., Patrosso, M., Marocchi, A., and Buscema, M. (2005). Assessment of the role of genetic polymorphism in venous thrombosis through artificial neural networks. *Annals of Human Genetics*, 69(6):693–706.

Per, K. (2001). Biological databases: why? *Stockholm Bioinformatics Center, SBC: Lecture notes*. http://www.avatar.se/strbio2001/databases/why.html, Consulted: February 2016.

Persidis, A. (2011). The benefits of drug repositioning. *Drug Discov World*, 12:9–12.

Pujol, A., Mosca, R., Farrés, J., and Aloy, P. (2010). Unveiling the role of network and systems biology in drug discovery. *Trends in pharmacological sciences*, 31(3):115–123.

Rapin, I. and Tuchman, R. F. (2008). Autism: definition, neurobiology, screening, diagnosis. *Pediatric Clinics of North America*, 55(5):1129–1146.

Reid, J. F., Gariboldi, M., Sokolova, V., Capobianco, P., Lampis, A., Perrone, F., Signoroni, S., Costa, A., Leo, E., and Pilotti, S. (2009). Integrative approach for prioritizing cancer genes in sporadic colon cancer. *Genes, Chromosomes and Cancer*, 48(11):953–962.

Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. *The Journal of Machine Learning Research*, 3:1371–1382.

Ridley, M. (2013). *Genome: The Autobiography of a Species in 23 Chapters*. Harper-Collins. https://books.google.fr/books?id=7O775xBxfMoC.

Ritchie, M. D., Coffey, C. S., and Moore, J. H. (2004). Genetic programming neural networks as a bioinformatics tool for human genetics. In *Genetic and Evolutionary Computation–GECCO 2004*, pages 438–448. Springer.

Ritchie, M. D., White, B. C., Parker, J. S., Hahn, L. W., and Moore, J. H. (2003). Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. *BMC bioinformatics*, 4(1):28.

Rivero, D., Dorado, J., Rabuñal, J., and Pazos, A. (2010). Generation and simplification of artificial neural networks by means of genetic programming. *Neurocomputing*, 73(16):3200–3223.

Rosenblatt, F. (1961). Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, DTIC Document.

Rotondano, G., Cipolletta, L., Grossi, E., Koch, M., Intraligi, M., Buscema, M., Marmo, R., and on Upper Gastrointestinal Bleeding (Progetto Nazionale Emorragie Digestive, I. R. (2011). Artificial neural networks accurately predict mortality in patients with nonvariceal upper gi bleeding. *Gastrointestinal endoscopy*, 73(2):218–226.

Ruiz, R., Riquelme, J. C., and Aguilar-Ruiz, J. S. (2006). Incremental wrapper-based gene selection from microarray data for cancer classification. *Pattern Recognition*, 39(12):2383–2392.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., and Willey, D. L. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822):928–933.

Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517.

Sardana, D., Zhu, C., Zhang, M., Gudivada, R. C., Yang, L., and Jegga, A. G. (2011). Drug repositioning for orphan diseases. *Briefings in bioinformatics*, 12(4):346–356.

Sargent, D. J. (2001). Comparison of artificial neural networks with other statistical approaches. *Cancer*, 91(S8):1636–1642.

Schalock, R. L., Borthwick-Duffy, S. A., Bradley, V. J., Buntinx, W. H., Coulter, D. L., Craig, E. M., Gomez, S. C., Lachapelle, Y., Luckasson, R., and Reeve, A. (2010). *Intellectual disability: Definition, classification, and systems of supports (11th Edition)*. ERIC.

Schuebel, K. E., Chen, W., Cope, L., Glöckner, S. C., Suzuki, H., Yi, J.-M., Chan, T. A., Van Neste, L., Van Criekinge, W., and Van den Bosch, S. (2007). Comparing the dna hypermethylome with gene mutations in human colorectal cancer. *PLoS Genet*, 3(9):e157.

Schwender, H. and Ickstadt, K. (2008). Identification of snp interactions using logic regression. *Biostatistics*, 9(1):187–198.

Science Learning, S. (2011). Genotype and phenotype. http://sciencelearn.org.nz/Contexts/Uniquely-Me/Science-Ideas-and-Concepts/Genotype-and-phenotype.

Seo, M. and Oh, S. (2012). Cbfs: high performance feature selection algorithm based on feature clearness. *PloS one*, 7(7):e40419.

Shameer, K., Readhead, B., and T Dudley, J. (2015). Computational and experimental advances in drug repositioning for accelerated therapeutic stratification. *Current topics in medicinal chemistry*, 15(1):5–20.

Shardlow, M. (2009). An analysis of feature selection techniques. https://studentnet.cs.manchester.ac.uk/pgt/COMP61011/goodProjects/Shardlow.pdf.

Sharma, A. and Tivari, N. (2012). A survey of association rule mining using genetic algorithm. *Int J Comput Appl Inf Technol*, 1:5–11.

Sharma, P. and Kaur, M. (2013). Classification in pattern recognition: A review. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(4):3.

Shaughnessy, A. F. (2011). Old drugs, new tricks. *BMJ*, 342:d741.

Shen, R., Yang, Y., and Shao, F. (2014). Intelligent breast cancer prediction model using data mining techniques. In *2014 Sixth International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC). Hangzhou.*, volume 1, pages 384–387. IEEE.

Silva, A. M., Faria, A. W. C., Rodrigues, T. d. S., Costa, M. A., and Braga, A. d. P. (2013). Artificial neural networks and ranking approach for probe selection and classification of microarray data. In *2013 BRICS Congress on Computational Intelligence*

and 11th Brazilian Congress on Computational Intelligence (BRICS-CCI& CBIC). Ipojuca, pages 598–603. IEEE.

Silverman, C. (2008). Fieldwork on another planet: social science perspectives on the autism spectrum. *BioSocieties*, 3(03):325–341.

Singh, N. (2012). *A Comparative Analysis of Machine Learning Algorithms for Genome Wide Association Studies*.

Skapura, D. M. (1996). *Building neural networks*. Addison-Wesley Professional.

Slack, J. (2014). *Genes: A Very Short Introduction*. Oxford University Press.

Sleigh, S. H. and Barton, C. L. (2010). Repurposing strategies for therapeutics. *Pharmaceutical Medicine*, 24(3):151–159.

Society, A. C. (2008). *Cancer facts & figures*. The Society.

Soltanian, K., Tab, F. A., Zar, F. A., and Tsoulos, I. (2013). Artificial neural networks generation using grammatical evolution. In *2013 21st Iranian Conference on Electrical Engineering (ICEE). Mashhad.*, pages 1–5. IEEE.

Somorjai, R. L., Dolenko, B., and Baumgartner, R. (2003). Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics*, 19(12):1484–1491.

Sousa, T., Silva, A., and Neves, A. (2004). Particle swarm based data mining algorithms for classification tasks. *Parallel Computing*, 30(5):767–783.

Stages, I., II, I., and Operable, I. (2016). Breast cancer treatment (pdq). http://www.cancer.gov/types/breast/hp/breast-treatment-pdq#section/all, Consulted: February 2016.

Stefanatos, G. A. (2008). Regression in autistic spectrum disorders. *Neuropsychology review*, 18(4):305–319.

Stewart, B. and Wild, C. P. (2015). World cancer report 2014. *World*.

Sumathi, S. and Sivanandam, S. (2006). *Introduction to data mining and its applications*, volume 29. Springer.

Susan E, L., David S, M., and Robert T, S. (2009). Autism. *Lancet*, 374(9701):1627—-1638.

Tang, Y., Jin, B., and Zhang, Y.-Q. (2005). Granular support vector machines with association rules mining for protein homology prediction. *Artificial intelligence in medicine*, 35(1):121–134.

Tomita, Y., Tomida, S., Hasegawa, Y., Suzuki, Y., Shirakawa, T., Kobayashi, T., and Honda, H. (2004). Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma. *BMC bioinformatics*, 5(1):120.

Turner, S. D., Dudek, S. M., and Ritchie, M. D. (2010). Grammatical evolution of neural networks for discovering epistasis among quantitative trait loci. In *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, pages 86–97. Springer Berlin Heidelberg.

Upstill-Goddard, R., Eccles, D., Fliege, J., and Collins, A. (2013). Machine learning approaches for the discovery of gene–gene interactions in disease data. *Briefings in bioinformatics*, 14(2):251–260.

Wang, G. and Song, Q. (2012). Selecting feature subset via constraint association rules. In *Advances in Knowledge Discovery and Data Mining*, pages 304–321. Springer Berlin Heidelberg.

Wasserman, P. D. and Schwartz, T. (1988). Neural networks. ii. what are they and why is everybody so interested in them now? *IEEE Expert*, 3(1):10–15.

Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids. *Nature*, 171(4356):737–738.

Watson, J. D. and Crick, F. H. (1993). Genetical implications of the structure of deoxyribonucleic acid. *JAMA*, 269(15):1967–1969.

Wilmshurst, L. (2012). *Clinical and Educational Child Psychology: An Ecological-transactional Approach to Understanding Child Problems and Interventions*. John Wiley & Sons.

Winkler, H. (1920). Verbreitung und ursache der parthenogenesis im pflanzen-und tierreiche. *G. Fischer*, page 252.

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006). Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34(suppl 1):D668–D672.

World Health Organization, W. (2012). Breast cancer: prevention and control. *World Health Organization.[Online]*.

Wray, N. and Visscher, P. (2008). Estimating trait heritability. *Nature Education*, 1(1):29.

Wu, C., Gudivada, R. C., Aronow, B. J., and Jegga, A. G. (2013). Computational drug repositioning through heterogeneous network clustering. *BMC systems biology*, 7(Suppl 5):S6.

Xie, J., , J., and Qian, Q. (2009). Feature selection algorithm based on association rules mining method. In *Eighth IEEE/ACIS International Conference on Computer and Information Science 2009 (ICIS 2009). Shanghai.*, pages 357–362. IEEE.

Xu, S. and Chen, L. (2008). A novel approach for determining the optimal number of hidden layer neurons for fnn's and its application in data mining. *In: 5th International Conference on Information Technology and Applications (ICITA 2008), 23-26 June, Cairns, Queensland, Australia.*, pages 683–686.

Yan, X., Zhang, C., and Zhang, S. (2009). Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support. *Expert Systems with Applications*, 36(2):3066–3076.

Yang, J., Li, Z., Fan, X., and Cheng, Y. (2014). Drug–disease association and drug-repositioning predictions in complex diseases using causal inference–probabilistic matrix factorization. *Journal of chemical information and modeling*, 54(9):2562–2569.

Yang, L. and Agarwal, P. (2011). Systematic drug repositioning based on clinical side-effects. *PloS one*, 6(12):e28025.

Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J.-F., and Hua, L. (2012). Data mining in healthcare and biomedicine: a survey of the literature. *Journal of medical systems*, 36(4):2431–2448.

Yu, X. and Gen, M. (2010). *Introduction to evolutionary algorithms*. Springer Science & Business Media.

Zhang, K., Qin, Z. S., Liu, J. S., Chen, T., Waterman, M. S., and Sun, F. (2004). Haplotype block partitioning and tag snp selection using genotype data and their applications to association studies. *Genome Research*, 14(5):908–916.

Zhang, Z., Zhang, S., Wong, M.-Y., Wareham, N. J., and Sha, Q. (2008). An ensemble learning approach jointly modeling main and interaction effects in genetic association studies. *Genetic epidemiology*, 32(4):285–300.

Zhao, S. and Li, S. (2012). A co-module approach for elucidating drug–disease associations and revealing their molecular basis. *Bioinformatics*, 28(7):955–961.

Zhao, Z. and Liu, H. (2009). Searching for interacting features in subset selection. *Intelligent Data Analysis*, 13(2):207–228.