Artificial intelligence applied to medical genetics

PhD Thesis

Øivind Braaten

2010

Series of dissertations submitted to the
Faculty of Medicine, University of Oslo
No. 1113

ISBN 978-82-8072-938-5

Cover: Inger Sandved Anfinsen.
Printed in Norway: AIT Oslo AS.

Produced in co-operation with Unipub.
The thesis is produced by Unipub merely in connection with the
thesis defence. Kindly direct all inquiries regarding the thesis to the copyright
holder or the unit which grants the doctorate.

Øivind Braaten,
Department of medical genetics,
Ullevål University Hospital
0407 OSLO,
NORWAY
Phone: +47 22 11 98 60
FAX: +47 22 11 98 99
email address: oivind.braaten@medisin.uio.no

This thesis can be downloaded from
http://folk.uio.no/~oivindb/phdthesis

Contents

Acknowledgements **vii**

Preface **ix**

List of articles **xi**

I General introduction **1**

1 Theme and aims **3**
1.1 The thesis' theme . . . 3
1.2 Aims of the thesis . . . 3
1.2.1 Investigate the applicability of AI to medical genetics . 3
1.2.2 Procreate new AI methods . . . 3
1.2.3 Contribute to objective syndrome diagnosis . . . 3
1.2.4 Create search methods for uncharted parts of the genome 4
1.3 Overview . . . 4
1.3.1 The subject matter: Medical genetics applications . . . 5
1.3.1.1 Syndromology . . . 5
1.3.1.2 Bioinformatics . . . 5
1.3.2 The methods . . . 5
1.3.2.1 The genetic algorithm . . . 5
1.3.2.2 The ID3 identification tree . . . 5
1.3.2.3 The feature vector and the set method . . . 6

2 The articles **7**
2.1 Preamble . . . 7
2.2 Article summaries . . . 8
2.2.1 Article I . . . 8
2.2.2 Article II . . . 8

2.2.3 Article III 8
2.2.4 Article IV 9
2.3 Introduction 9
2.3.1 The syndromology studies 9
2.3.2 General principles of diagnostic value of clinical signs in syndromology 10
2.3.3 Strategies for using clinical signs 11
2.3.4 General overview of syndrome terminology 12
2.3.5 Approaches to syndrome classification 14
2.3.6 Measurement/ recording of clinical signs 16
2.3.7 Mathematical-statistical diagnostic approaches 17
2.3.7.1 Summing up on the mathematical-statistical approaches 20
2.3.8 Artificial intelligence diagnostic approaches 20
2.3.8.1 Identification trees 20
2.3.8.2 Case based reasoning, CBR 22
2.3.8.3 Other approaches 23
2.3.9 Conclusion 23
2.3.10 The bioinformatics studies 24
2.3.11 General approaches to DNA pattern searching 26
2.3.12 Genetic algorithm approaches to DNA pattern searching 28
2.4 Material and methods 28
2.4.1 Methods 28
2.4.1.1 Identification trees 28
2.4.1.2 The genetic algorithm 29
2.4.1.3 Feature vector/ set method 30
2.4.1.4 Cluster analysis 31
2.4.1.5 Naïve Bayes' calculations 31
2.4.1.6 MEME 31
2.4.1.7 Training set/ test set 31
2.4.1.8 The POSSUM patient records 32
2.4.1.9 Computer 32
2.4.1.10 Programming languages 32
2.4.2 Material 32
2.4.2.1 Artificially generated patients. 32
2.4.2.2 Artificially generated DNA sequences 33
2.4.2.3 Randomization 33
2.4.2.4 LDL receptor haplotypes 33
2.4.2.5 Human DNA sequences 34
2.5 Results 34
2.5.1 The syndromology studies 34

2.5.2 Bioinformatics studies 36
2.6 Discussion 37
2.6.1 Syndromology studies 37
2.6.2 Bioinformatics studies 41
2.7 Implications 44

II Articles 45
2.8 Article I: Artificial intelligence in pediatrics 47
2.9 Article II: Diagnosis: Human intuition or machine intelligence? 57
2.10 Article III: The genetic algorithm applied to haplotype data . 69
2.11 Article IV: Finding DNA patterns with the genetic algorithm . 79

III Appendices 109

A Artificial intelligence 111
A.1 Introduction 111
A.2 Expert systems 111
A.3 Neural nets 112
A.4 Machine learning 112
A.4.1 Supervised learning 113
A.4.2 Unsupervised learning 113
A.4.3 Data mining and text mining 113
A.5 Identification trees 113
A.6 AI search strategies 115
A.7 The genetic algorithm, general introduction 116
A.7.1 Review of biological terms and concepts 116
A.7.2 The initial population 119
A.7.3 Fitness 119
A.7.4 Mating 119
A.7.5 Recombination, mutation, and inversion 120
A.7.6 The population – the concept of generation 121
A.7.7 The fitness function 121
A.8 Genetic programming 123
A.9 Complex adaptive systems 123
A.10 Fuzzy logic 123
A.11 Case based reasoning 124
A.12 Data mining 124
A.13 The semantic web 124
A.14 The 'closed universe' 125

A.15 A summary history of artificial intelligence in medicine 125

B Elements of clinical epidemiology 131
B.1 Introduction 131
B.1.1 Concepts and indices of clinical epidemiology 131
B.1.1.1 Sensitivity 131
B.1.1.2 Specificity 131
B.1.2 False and true, two by two table 133
B.2 Receiver operating curve 134
B.3 Syndrome-diagnostic approaches 134
B.3.1 Consistency 134
B.3.1.1 Aspects of consistency 134
B.3.1.2 Bayes' formula 136
B.4 Using a diagnostic test in a different setting 137
B.5 Quality of data 138
B.5.1 Ascertainment bias 138
B.5.2 Other methodological issues 139
B.5.3 Changing definitions of disease entities 139
B.5.4 Clinical signs used in diagnosis 140

C Philosophical background 141
C.1 Introduction 141
C.2 Nominalism versus essentialism 142
C.3 Categories 142
C.4 Tacit knowledge 143

D Code: Lisp code for the genetic algorithm 145

E Sample output from the genetic algorithm 151
E.1 Output showing the first five individuals of a population . . . 151
E.2 Output showing part of a 'path' 154
E.3 Output showing the n-mer hash table 156
E.4 Output showing a 'pack' 157

F Scripts: Script code text 161
F.1 Perl code 161
F.2 Python code 167
F.3 'R' code 176

IV References **177**

Bibliography **179**

Acknowledgements

This thesis has come into being during my years at the Institute of medical genetics at the University of Oslo and the Department of medical genetics at Ullevål University Hospital.
I wish to express my heartfelt thanks to my supervisor professor Herman Ruge Jervell at the Institute of informatics, the University of Oslo.
I am also indebted to my other supervisor professor Kåre Berg, formerly at the Institute of medical genetics, former Head of department and senior clinical geneticist at the Department of medical genetics. Kåre Berg, sadly, died in January 2009.
They are both outstanding in their respective fields, - and patient men.
I thank my co-author Johannes Friestad and my co-authors Olaug Kristin Rødningen, Trond P. Leren and Inger Nordal.
I appreciate the collaboration with the members of the sadly short-lived bioinformatics group at the Department of medical genetics, Ullevål University Hospital; Inger Solberg, Morten Mattingsdal, Josef Thingnes, and later Geir Ivar Jerstad
I am grateful to collaborators at CBS (Center for Biological Sequence Analysis, The Technical University of Denmark/ Danmarks Tekniske Universitet (DTU)) and its director Søren Brunak who generously let me stay and work at the CBS for a month and at revisits, and who has a policy for letting former collaborators keep access to the computers at CBS. I thank Kasper Lage Hansen and Thomas Blicher in particular for making me feel welcome, and Kristoffer Rapacki and Peter Wad Sackett for expert help with programs and computers.
I thank Tore Fagerlund, my office mate for many years, with whom I have had many inspiring discussions.
I thank friends and colleagues at the Department of medical genetics and at the Institute of medical genetics.
And Kaia, Simen and Hilde; you make it worth while.

Preface

Medical artificial intelligence is an interdisciplinary subject. A fundamental difficulty with such an interdisciplinary subject matter, is that the 'medical' part may be accessible to a medical audience, who may find the 'artificial intelligence' part inaccessible; whereas an informatics or artificial intelligence audience may find the AI part understandable, but may not feel at home in the medical parts. To make this thesis more accessible to the potential reader, I have included background and Appendices that may be longer than what would be considered necessary in a text aimed at a homogeneous audience. I have tried to start out at a level that may not be totally fundamental, but hopefully understandable to an enlightened reader from either an informatics or a medical (or any other) background. The price to pay for such an approach is that the medical reader may feel the medical part is too basic, and the informatics reader may feel the informatics part is too basic. I hope that, on the contrary, some readers may find the Appendices useful, and that those readers who do not find them useful feel free to skip them.

The text was written in Emacs and typeset with the LaTeX document formatting and typesetting system.

Oslo, March 2010

Øivind Braaten

List of articles

Article I

Braaten Ø:
Artificial intelligence in pediatrics:
Important clinical signs in newborn syndromes.
Computers and Biomedical Research, 29:153 – 161, 1996.

Reprinted in Yearbook of medical informatics 1997, 566 – 574.

Article II

Braaten Ø and Friestad J:
Syndrome diagnosis: Human intuition or machine intelligence?
The Open Medical Informatics Journal, 2, 149 - 159, 2008

Article III

Braaten Ø, Rødningen OK, Nordal I and Leren TP:
The genetic algorithm applied to haplotype data at the LDL receptor locus.
Computer methods and programs in biomedicine, 61, 2000, 1-9.

Article IV

Braaten Ø:
Finding DNA patterns with the genetic algorithm: On the track of unknown functional elements?
Submitted, Artificial Intelligence in Medicine

Part I

General introduction

Chapter 1

Theme and aims

1.1 The thesis' theme

The theme of this thesis is the application of artificial intelligence methods to medical genetics. In a broader perspective, the theme is artificial intelligence in medicine.

1.2 Aims of the thesis

1.2.1 Investigate the applicability of AI to medical genetics

A major aim of this project was to investigate whether artificial intelligence could be applied to medical genetics to produce sensible and useful results.

1.2.2 Procreate new AI methods

Another major aim was to develop new artificial intelligence methods that could be applied to medical genetics; or to adapt existing methods to the task.

1.2.3 Contribute to objective diagnostic systems in syndromology

The lack of a criterion standard for syndrome diagnosis, and the need for objective diagnosis, are fundamental problems for syndromology, and by extension, for medicine as a whole. This project aimed to illuminate these problems, and to contribute objective methods of diagnosis.

1.2.4 Create search methods for uncharted parts of the genome

The ninety eight per cent or more of the human genome that is still termed 'junk DNA', is becoming a major research area. This project aimed at developing and testing methods of searching such unknown DNA sequences.

As an undercurrent of a unifying theme for the whole thesis, runs the search for patterns in very large search spaces.

1.3 Overview

There would be no point in applying artificial intelligence methods to a field such as medical genetics, unless it would imply improvement in some respect, i.e. doing things in a better way or doing things with less resources in terms of time, money or people. The applications in this thesis attain solutions that would have been difficult to achieve by other methods.

There are areas in medical genetics, and in medicine and biology in general, where research question arise that cannot easily be handled by conventional methods. An example of this is many of the sequence searching problems of bioinformatics. Some search problems are straightforward and readily solved by existing methods. Some searches, however, cannot be solved by either statistical-mathematical methods, or by all-encompassing (so called 'brute force') informatics algorithms.

This type of problem is addressed in the part of this thesis regarding the genetic algorithm.

Medical artificial intelligence is definitely an interdisciplinary field. When such new fields arise, often fundamental questions of philosophy of science are uncovered. The established sciences rarely question their basic assumptions. In burgeoning fields, researchers may have to confront fundamental dogmas in their science, to challenge such dogmas, and possibly to redefine what are valid research questions and what direction research should take.

On a more mundane level, examples of this would be the lack of a criterion standard of diagnosis in syndromology, and its consequences. Similarly, the lack of objective diagnosis in fields of medicine does have implications for medicine as a scientific discipline.

This type of problem is addressed in the part of this thesis regarding the ID3 algorithm and the feature vector.

1.3.1 The subject matter: Medical genetics applications

The artificial intelligence methods were applied to two areas of medical genetics: Syndromology, and bioinformatics sequence searching.

1.3.1.1 Syndromology

When a child is born with a set of clinical traits that appear abnormal, such traits often delineate a syndrome. Syndromes often do not have a criterion standard of diagnosis ('a gold standard'). There is therefore often no objectively verifiable diagnosis in syndromology. That is a practical and a philosophical problem.

1.3.1.2 Bioinformatics

Bioinformatics - the use of computers and algorithms to extract information and knowledge from biological data - is entwined with medical genetics. Searching biological sequences is fundamental in bioinformatics.

1.3.2 The methods

Artificial intelligence methods are exemplified in this thesis by two versions of the genetic algorithm; the ID3 - an identification tree algorithm; and two basic methods - a feature vector method and a set method.

1.3.2.1 The genetic algorithm

The genetic algorithm is a computer algorithm based on evolution in nature. It can usually search large search spaces, and may provide solutions to problems where other methods cannot cope. It is not guaranteed to find the best solution to a problem, but it will often find a good solution.

1.3.2.2 The ID3 identification tree

The ID3 is an algorithm that will divide a search space - e.g. all clinical signs used to diagnose certain syndromes - by the element (e.g. clinical sign) that succeeds best in dividing it into equal parts. It will proceed until all elements (clinical signs) have been used. In this process it creates a tree, called an identification tree or a decision tree. This tree can be used later as a guide through the search space.

1.3.2.3 The feature vector and the set method

The feature vector method and the set methods are two basic artificial intelligence methods. These are used to illustrate how objective diagnosis can be attained for a field such as syndromology.

Chapter 2

The articles

2.1 Preamble

Two of the articles that form part of this thesis apply artificial intelligence methods to the area of syndrome diagnosis. The two other articles deal with bioinformatics.

The two articles concerned with syndromes both originate from the lack of objective diagnosis in syndromology.

In article I, an identification tree is applied to a set of artificially generated patients. An identification tree is an objective method where the basic assumptions can be evaluated. It is also **transparent** in that all parts of the algorithm can be examined and adjusted if necessary. It is straightforward to repeat the runs, and with small to medium data sets the results are **repeatable**.

Article II is also focused on objective diagnosis. The methods applied here are basic methods. The advantage of such basic methods is that the methods are readily comprehensible. These methods are also - like the identification tree method - transparent and repeatable. Examples of methods that are not transparent, are for instance neural networks, and, to some degree, genetic algorithms. These methods are also only repeatable up to a certain point. I do argue in favour of non-deterministic methods in the sections on the genetic algorithm, where I hold that these methods should be judged by whether they attain good solutions. When it comes to objective diagnosis in an area such as syndrome diagnosis, however, it is a highly desirable quality that the methods are transparent and repeatable, since the problem is the very lack of objectivity.

The two bioinformatics articles both apply the genetic algorithm.

The genetic algorithm's strength is the ability to search large search spaces.

A quality that some find disquieting, is the fact that it will not necessarily produce the same solution in every run - the repeatability concern. At best, the genetic algorithm will find good solutions. but not necessarily the best solution - if one single 'best' solution exists. The search spaces in bioinformatics are often vast, and all-enumerating or all- examining methods break down. In some cases heuristics can limit the search space, but for many problems there simply are no alternatives.

2.2 Article summaries

The article numbering in roman numerals refers to the list of articles on page xi.

2.2.1 Article I

This study applies the identification tree (ID3) method to a set of artificially generated patients. The ID3 finds clinical signs that can be used in syndrome diagnosis. The ID3 is an example of an objective diagnostic method.

2.2.2 Article II

This study focuses on the lack of objective methods in syndrome diagnosis. It applies a vector method and a set method to artificially generated patients. A cluster analysis, a naive Bayes' calculation, and an identification tree method were run as controls, with good correlation to the basic methods. In this study, sets of clinical signs are elicited. These sets in principle must occur together to be diagnostic. The study concludes that consistent diagnoses are feasible with the use of basic artificial intelligence methods.

2.2.3 Article III

In this study, a genetic algorithm is applied to the problem of finding RFLP haplotypes in the LDL receptor gene. The method finds haplotypes associated with high cholesterol values. Runs with added 'noise' and mix-in of 'misleading' haplotypes show the genetic algorithm still can discern the high cholesterol haplotypes.

2.2.4 Article IV

This study was motivated by the wish to search unknown DNA sequences for patterns, without preformed notions of what are interesting patterns. It features a new genetic algorithm, with diploidy, parthenogenesis, 'paths' - to help give the genetic algorithm a memory, and 'packs' - to help the genetic algorithm individuals match longer sequences of input DNA. It shows a slightly inferior performance compared to the control program MEME, when searching for patterns in artificially generated DNA. It finds interesting sequences not found by MEME when applied to a set of four related human genes.

2.3 Introduction

In the following pages, the deliberation splits in two, according to the two application areas for the artificial intelligence methods; syndromology and bioinformatics.

2.3.1 The syndromology studies

The **syndromology studies**, articles I and II, were motivated by the **lack of a criterion standard of diagnosis** in syndromology, and the consequent **lack of objective diagnoses**.
'Proper syndromes' do not have a known biochemical, chromosomal or other genetic cause. A syndrome is a phenotype or disease condition defined only by its clinical signs [1]. As discussed in article II, this implies there is no external criterion standard for the diagnosis. This is a scientific problem, since syndrome diagnoses cannot be verified against an objective external standard. For a condition that is its own criterion standard, some of the indices discussed later, like sensitivity and specificity, lose their meaning. The consistency/ repeatability/ precision of syndrome delineation and diagnosis thus takes on even more importance.
The predictive value of any test in medicine - and any clinical sign in syndrome diagnosis - is strongly dependent on the prior probability of the diagnosis. Thus, an additional problem in syndrome diagnosis, as in medical diagnosis in general, is the fact that many diagnostic methods do not take into account the **effect that the prior probabilities of a diagnosis has on the predictive value of a clinical sign** (A more formal discussion of elements of clinical epidemiology is given in Appendix B.1 on page 131.)
A common syndrome with an unusual presentation, may be a more plausible diagnostic suggestion than an extremely rare syndrome with its

standard presentation.
These problems are well recognised and treated in clinical epidemiology texts [2], but not always recognised among clinicians. Among those who examine the effects on clinical performance of ignoring these principles, are Cahan et al [3] and Richardson [4]. Clinical experience does not ameliorate the problem of misjudging prior probabilities [5].
To use the concept of predictive value sensibly, the frequency of the syndrome has to be known. The frequency of a given clinical sign among those who have the syndrome, also has to be known. Snowman et al critically evaluate clinical assessments of frequency estimates, and conclude that the qualitative terms used to denote frequency by clinicians, are misleading [6].

2.3.2 General principles of diagnostic value of clinical signs in syndromology

A good clinical sign for diagnostic use is a sign that can be used to single out from all others the patients with a specific diagnosis. 'All others' are those with other diagnoses, plus those who are not affected.
A clinical sign that dramatically reduces the number of possible diagnoses, or that could even represent a one-step diagnosis, is often referred to as a 'handle' or a pivot sign.
A seemingly obvious way to find a good clinical sign would be to look for a sign present in many of the affected persons. However, this alone is not enough to be a good clinical sign, since the sign will be virtually useless if it is also found in many other conditions.

There are two main diagnostic situations:

- **The basic or first diagnosis**
- **The differential diagnosis**

Some underlying principles apply to both situations. A good clinical sign should be found in a large number of patients with the diagnosis in question. A sign found in ten per cent of patients might be useful when found, but would still be of limited practical value.
The most important point when searching for a clinical sign to discriminate groups is this: As few as possible of the others should have the sign. Again, 'others' are all normals, plus those with other diagnoses.
For **the first diagnosis**/ the basic situation, the above principles apply without modification.

A special situation of 'first diagnosis' is the **screening** test situation. The patients coming for a screening test are either low risk patients or persons who have no known risk, i.e. they are from the general population. For syndrome diagnosis, the general population may be the general population of newborns seen by a paediatrician doing a newborn examination. This situation carries the particular risk of false positives, i.e. misdiagnosing healthy newborns as having a syndrome. In a screening situation, there is a risk of diagnosing very few who are actually affected, and misdiagnosing a large number of healthy persons as affected. Thus, in a screening situation, it is essential to use signs that only very few of the 'others' have. (The particular risk of misdiagnosing healthy people as affected in screening, comes from the low prior probability of disease. To try to counteract this, signs should have a very high specificity. The concepts prior probability and specificity are discussed on page 131).
A high specificity in this situation means extremely few of the others should have the sign.
In **the differential diagnosis** situation, the number of possible diagnostic categories may be only two or three-four. The problem now is either to refute one or two less plausible diagnoses, or to confirm the number one diagnosis.
The 'others' group is different. A test that would have been previously discarded could now be valuable.
For **a confirmation diagnosis**, it is even more important that very few of the others have the sign for the diagnostician to accept one diagnosis with confidence.
To **refute a diagnosis**, conversely, a sign is needed that occurs reasonably frequently among the others, but rarely among those with the diagnosis in question (the diagnosis to be refuted). If this sign is found, the diagnosis can be refuted.

2.3.3 Strategies for using clinical signs

One strategy is to find clinical signs that can immediately lead to the diagnosis. Another strategy is one that resembles a binary search strategy. Such an approach would ideally halve the universe of possible diagnoses with each sign, e.g. if the patient does not have the clinical sign in question, about half the diagnoses can be ruled out and the other half is still under consideration. An identification tree approach [7] might yield signs of this type.
Ideally, such a strategy could always be used and would potentially be very effective. If one did not reach a diagnosis, a secondary goal would be to find

the optimal test to order. Consequently, clinical signs could be used that guided the diagnostician in the right direction as far as test ordering is concerned. Some clinical signs might point to a chromosomal aberration, other clinical signs might lead in the biochemical direction and so on. Certainly, some such signs do exist. However, methods such as the identification tree method might find clinical signs that would be more rewarding in this kind of directed diagnostic search.
This overview has centred on individual clinical signs. In practice, a diagnosis would be based on several co-occurring signs. Although this makes the situation more complicated, the basic principles outlined here still apply.

2.3.4 General overview of syndrome terminology

Most syndromes are rare, but there is a large number of syndromes. Syndromes will affect the patient, the family, and society. Thus, numerous people will be affected by a child born with a syndrome.
A syndrome diagnosis is important for prognosis, possible treatment and educational measures, in many cases for genetic counselling, and for the parents' need to know.
Syn means together and *dramein* means to run, i.e. a syndrome literally is something that runs together.
A number of mechanisms underlie the clinical signs that are observed in syndrome patients - the dysmorphological traits.
A **malformation** is a morphological defect of an organ, part of an organ, or a larger region of the body resulting from an intrinsically abnormal developmental process. Malformations thus occur if something has not been formed properly from the start.
A **deformation** is an abnormal form or position of a part of the body caused by non-disruptive mechanical forces. Deformations, then, are the result of mechanical forces acting on the fetus. Deformations need not be very extensive or serious. They are exemplified by development that has not been wrong from the start, but later modified. 'Club foot' or pes equinovarus is an example of a deformation.
Disruptions are examples of a development that was normal, but has then been destructively transformed. Amniotic bands may cause disruptions.
Dysplasia refers to an abnormal growth of a particular tissue (aplasia/ hypoplasia/ hyperplasia) or a disorganization of cells within a tissue.
A **polytopic field defect** involves distantly located anatomic structures that are developmentally related [8].

The word polytopic comes from *poly* many, and *topos* place. A single noxious agent may affect several separate tissues or organs. This may be effected by influence on a postulated 'developmental field', a set of tissues or organs being developed in concert.

An **association** is a non-random occurrence in two or more individuals of multiple anomalies not yet known to be a polytopic field defect, sequence or syndrome [8, 9, 10].

A 'proper syndrome' may be defined as a condition where diseases with a known etiologic [1] mechanism have been ruled out, and where the diagnosis is based upon physical signs. For a proper syndrome as opposed to an association though, there is implicitly a postulated etiologic mechanism, not yet discovered. More formally, a syndrome is 'a pattern of multiple anomalies thought to be etiologically related and not known to represent a single sequence or a polytopic field defect' [8].

A **Sequence** is a pattern of multiple anomalies from a single known or presumed prior anomaly or mechanical factors [8].

An example of this is the Potter sequence. The basic problem here is a lack of amniotic fluid (e.g. because the fetus has no kidneys, renal agenesis). The Potter sequence consists of a flat, 'squashed' face, a 'compressed' pattern of limb positioning defects, in severe cases undeveloped lungs, etc. The physical traits found in a baby with Potter sequence all result from the lack of amniotic fluid.

The mechanism behind syndromes is poorly understood. For some syndromes there is a known genetic defect, but often there is no obvious link between the diverse set of clinical signs in a syndrome and the genetic defect. The expression 'pleiotropic' is often used to denote that one gene affects several tissues and organs, and as an explanation of different disease manifestations in different people. Heterogeneity, on the other hand, is used for the situation where different 'causes', e.g. different genes, lead to the same clinical picture.

'Genetic field defects' are proposed as potential causes for syndromes. The 'field' is seen as governed by an organizing element that commands a set of successive operations during embryological development, and when the genetic organizer is deranged, a set of clinical and morphological consequences will follow.

Findings such as these are often explained by a 'final common pathway'. This, however, is more of a theoretical postulate than an explanation. In

[1] **Etiologic** here meaning 'pertaining to the cause of the disease', whereas **pathogenic** refers to the development of the disease, i.e. the pathogenesis succeeds the etiological factor(s)

fact, it might prove more fruitful to reason from clinical signs to genes instead of reasoning from gene to syndrome. This approach has been advocated by for example Brunner [11].
Some syndromes are the result of a noxious agent - a teratogen - exerting its influence on the embryo and the foetus.

2.3.5 Approaches to syndrome classification

The most basic form of classification of syndromes is regarding each syndrome as a single entity based on its clinical signs.
The traditional medical approach is considering a syndrome entity as consisting of clinical signs, symptoms, and a 'pathological' disturbance. Pathological in this connection is used in a broad sense, including possible genetic changes and/ or biochemical disturbances, or conditions involving several components of a metabolic system.
Exciting recent work in syndromology revolves around syndrome families. The concept of families of syndromes and genetic diseases is established in medical genetics. Brunner and van Driel systematise this, and takes it further by suggesting using the groups of syndromes to postulate which genes are involved in causing the syndrome [11]. Pondering the fact that different genes (or alleles of different genes) can cause very similar phenotypes, they consider similarity between proteins involved. Through common knowledge/ literature searches they conclude that such cases do exist. They move further on by considering not the protein as the basic unit, but the metabolic network that the protein participates in. In syndromes having dysfunctional proteins that are part of the same gene-protein-metabolic-interactomic network, may have the same clinical sign. Conversely, syndromes with the same clinical appearance may point to causation by malfunctioning of such common networks.
The construction of networks had already begun in bioinformatics, and has lead to an upsurge in the use of the artificial intelligence techniques of data mining and text mining [12, 13].
Oti and Brunner elaborated on the ideas of 'modules' of genetic diseases [14]. They identified possible 'modules' as for instance a multiprotein complex, a cell's organelle, or a metabolic pathway.
There are many ontologies created that appease bioinformatic research by standardising terms and concepts. To facilitate research on phenomes and networks, a Human Phenotype Ontology has been created [15]. The Human Phenotype Ontology contains 8000 terms, and is cross referenced with Online Mendelian Inheritance in Man.

Loscalzo and co-workers argued for a new disease classification in medicine based on biological networks [16]. Schadt et al argue along the same lines, and also extend the point of view to suggest that with data from the next generation genomic analysis, the network classification will be necessary to develop new therapeutic agents [17].
Jiang et al [18] provided more evidence for the existence of modules, adding data from HugeIndex of tissue-selective genes, and DrugBank, a Food and Drug Administration databank of 1700 drug targets.
The ideas of interrelatedness are taken even further by including diseases that were thought to be distinct in a landscape of human diseases [19, 20].
The natural extension to phenome networks have given rise to more elaborate networks that include different elements of information and knowledge, primarily data generated by bioinformatics, and to a large extent based on experimental and laboratory work.
Interesting examples of such networks are a phenome-interactome network of protein complexes studied by Lage et al [21]. The authors used a text mining tool (MetMapTransfer) that maps to the thesaurus of terms in UMLS (Unified Medical Language System), and parses OMIM (Online Mendelian Inheritance in Man). They linked the phenome information thus extracted to information about protein complexes. By using a Bayesian predictor they could detect possibly disease causing proteins and genes.
An earlier work by Oti et al [22] had used protein-protein interaction data to suggest candidate genes for genetic diseases and syndromes.
These are instances of building on the phenome networks. Since the subject of the syndromology part of this thesis is strictly clinical, the multitude of interesting network applications are not pursued further.
The causation of syndromes have been debated for years. Syndromes and syndrome conditions seem to be caused by chromosome aberrations, deletions/ duplications, mutated alleles of genes, teratogens, and physical forces during embryological development. As far as the genetic causes are concerned, it can sometimes be hard to conceive of how the multitude of diverse clinical consequences can result from a single deranged gene. One possibility is that a supergene is affected, a gene that controls the effect of a number of other genes that act in concert during development. This is true for homeobox genes, e.g. the DIx homeobox [23]. A deranged gene could also affect a whole system of body development. Again there are numerous examples of this, e.g. development of the pharyngeal arches [24]. Another traditional explanation is the 'developmental field' for instance as reviewed by Volpe et al for disorders of prosencephalic development [25]. With the advent of gene-interactome-phenome networks, both the developmental field and the systems of body development are explanations that may need

rethinking in light of the network approaches.

2.3.6 Approaches to measurement and recording of clinical signs

A syndrome diagnosis is in many cases based both on clinical and molecular biological grounds. As discussed in article II, this is a problem when patients are diagnosed differently with the two approaches. An example is Marfan syndrome which is diagnosed clinically by a set of clinical criteria. It is diagnosed by molecular biology methods by finding mutations in the FBN1 gene. Exact correlations between genotype and phenotype prove hard to establish [26, 27]. Similarly, the phenotypic findings may not be consistent for submicroscopic deletions or duplications found by array Comparative Genome Hybridization (aCGH), e.g. as described by Dhar et al for the 22q13.3 syndrome [28].

People change with age, and so do syndrome patients. However, the clinical signs that were diagnostic at a young age, may change in such a way as to make the clinical sign much less conspicuous, or even useless as a finding to discriminate between those who have the syndrome in question and others. Garavelli et al [29] give a description of how the facial dysmorphology of patients with Mowat-Wilson syndrome changes.

A problem in diagnostic work is when different diagnosticians do not use the same definition of the diagnostic entity. This is also a problem in syndromology. Breugem asked a number of experts about their definition of the Pierre-Robin sequence. Sixty five experts gave 29 different descriptions of the condition [30]. The problem is probably universal.

Similarly, diagnosticians may disagree as to whether a specific clinical sign is present or not. An important further step in the work concerning international standardization of terms and definitions in syndromology was done through the work of Judith Hall [31] and the article series in the American journal of Medical Genetics in 2009 [32, 33].

Anthropometric exact measurements, often done from pictures of patients, are a way to minimise interobserver and intraobserver disagreement. It also makes it easier to make consistent definitions of syndromes. A seminal early work in this area was done by Cohen, e.g. [34]. (There is a further general discussion of inter- and intra-observer variation in Appendix B.1 on page 135.)

Ward et al [35] studied 278 individuals with different syndromes where craniofacial signs are apparent (although not being 'craniofacial syndromes' in a stricter sense). They found that a summary score of anthropometric

measurements agreed with clinicians' evaluations.

Douglas, Mutsvangwa et al used standardised facial image analysis, doing morphometric analyses on pictures of children with fetal alcohol syndrome [36, 37]. They also compared pictures taken at different ages (at 5 and 12 years of age, respectively), and found the clinical features less distinct and with less predictive value at the older age. A problem with their approach was that they compared children with fetal alcohol syndrome with normals only. In this way they avoided the potential real life problem of similar syndromes to the fetal alcohol syndrome making the diagnosis more difficult.

Hammond et al in 2005 studied localised facial features with a so called 'dense surface model' [38]. This study made use of three dimensional computer generated models of syndrome patient faces, a project started a few years earlier (e.g. [39]). The face models were based on 3D images. To distinguish between the phenotypes of four different syndromes, and a normal control group, they used three classifier algorithms, e.g. a support vector machine (SVM). Again, a problem with this approach is that few syndromes were included. Prior probability - outside of the confined universe of the study - were not taken into account.

Hammond summarises this work in [40]. Of course, from an artificial intelligence point of view, it is an exciting possibility to combine database models of craniofacial - or whole body - three dimensional models of a large number of patients with syndrome conditions. Thus, a 3D real time image taken of a patient with a suspected syndrome, could be compared to the database models and possibly render a diagnosis automatically.

The previous approaches are in line with the objectivity criterion advocated in article I and II of this thesis. Anthropometric measurements will make an objective diagnostic method all the more powerful, since it lowers the risk of observer variation. It will increase consistency, although it will not have an effect on the problem of no external criterion standard of diagnosis. It will not in itself take care of the problems posed by not taking into account prior probabilities of the syndromes.

2.3.7 Mathematical-statistical diagnostic approaches to syndrome diagnosis

In the search for objective methods of diagnosis, mathematical-statistical approaches would seem to be good candidates. The next pages give an account of such methods in syndromology.

Cluster analysis Cluster analysis is a statistical method of classifying or partitioning a set of elements by its features, for example classifying a number of syndromes by the clinical signs found in the syndrome. The hierarchical cluster analysis procedure will start with a number of elements, with a set of features to each element. It will classify the elements into groups, based on their features. The groups will have a distance calculated to the other groups that are formed. The most common is hierarchical agglomerative clustering, starting with single elements and joining them into clusters. The closeness of two elements is decided by the distance measure. In classifying people by height, this could simply be the difference in height in centimeters between two people. The cluster analysis uses a procedure for joining either of several clusters of elements to an existing cluster. The most commonly used method is single linkage or 'nearest neighbour', where the cluster is considered closest that has the closest element to the existing cluster. The furthest neighbour method, or complete linkage, considers the two clusters closest that have the shortest distance between the two features that are furthest apart. Average linkage uses more information in calculating the average of distances between all pairs of features in the two clusters.
As in any statistics or other classifying scheme, it is important which variables or features of the elements are selected for inclusion in the analysis.
In medicine and biology cluster analysis is attractive because it does not rest on as many basic assumptions as do the other multivariate methods.
Discriminant analysis
Discriminant analysis will group a set of elements according to a set of variables or features. To use discriminant analysis, the group membership must be known. Discriminant analysis can identify variables or features that can be used in predicting group membership for new, unknown cases. Group membership is a categorical variable (group1, group2 etc). Discriminant analysis will perform optimally only with elements taken from multivariate normal populations. The input variables/ features to discriminant analysis should be continuous. For categorical variables discriminant correspondence analysis can be used. Ross used a discriminant analysis approach to establish neurocognitive profiles in Turner syndrome [41], Loesch used it on carriers of fragile X syndrome [42], Moore et al [43] applied anthropometric measurements analysed by discriminant analysis to the task of diagnosing fetal alcohol syndrome. They found the method could group the patients into fetal alcohol syndrome, partial fetal alcohol syndrome and normals. Discriminant analysis of children with fetal alcohol syndrome was also the subject of Astley's study [44]. Murdoch-Kinch and

Ward used discriminant analysis to data from measurements of metacarps in people with Crouzon syndrome.

Principal component analysis and **factor analysis** both find a set of factors that explain the group membership of the elements. This set of factors ideally is smaller than the set of features or variables in the original data set. Principal component analysis has been applied to syndromes, e.g. [45, 46]. Volk et al found subgroups among ADHD patients using principal component analysis [47, 48]. Haley used factor analysis to find syndromic disease entities in groups of soldiers who participated in the Gulf war. The appropriateness of the approach is discussed in [49].

Multiple regression

Multiple regression is the basis method among the multivariate methods. It models the relationship between a dependent variable and one or more independent variables. It rests on the assumption of multinomial normal distributions.

Loesch analysed body shape in fragile X patients using multivariate analysis [50]. Preus was one of the pioneers of so called numerical taxonomy and the use of multivariate methods [51, 52, 53, 54]. One of the early proponents of objective methods and numerical analysis was Verloes [55].

Logistic regression

Logistic regression predicts an event (group membership, or occurrence of the event) from a set of variables. The variables may be categorical, in contrast to several of the other multivariate methods. Basically, it predicts one event. The extension to predict several groups or events is called multinomial logit modelling.

Non-negative matrix factorization

Non-negative matrix factorization is related to principal component analysis. It takes the data set, a matrix of variables and values, and transforms it into other matrices, a feature matrix and a weight matrix, which are transposed matrices made from the input. By matrix multiplication and transposing, the input set of data may be reduced to a smaller set of features that still explains the data. The goal is thus to reduce the input set of observations (e.g. clinical signs found in syndrome patients) to a smaller set that captures the common features.

Zhang et al [56] extended a non-negative matrix factorization method to what they called a topology preserving NMF or TPNMF. They found they could recognize faces taken from a database of 1200 face images, where the images were taken under different conditions of lighting, facial expression and pose. Face recognition is a separate research area, not involved with diagnosis in any way, but does hold obvious resemblances to recognising traits of facial dysmorphology.

Non-negative matrix factorization is a more recent method than the classical statistical methods. Some consider it an artificial intelligence method [57].

2.3.7.1 Summing up on the mathematical-statistical approaches

Two problems make it difficult to apply the multivariate statistical methods to syndrome diagnosis. First, most of the methods are based on the assumption of multinomial normal distributions. This implies binary variables are not readily acceptable, although the various methods show different degrees of vulnerability to violation of this assumption. In many syndrome diagnostic situations, the diagnostician will record a clinical sign as present or not present, i.e. not on a quantitative scale.
Secondly, these methods are often vulnerable when it comes to missing values for variables. Many syndromes have a list of signs that may be present or not. A data set from a syndrome survey will typically contain signs present, signs not present, and signs not recorded as present or not present - i.e. missing values.
Several of these mathematical-statistical methods are used in machine learning, and the boundary between artificial intelligence and statistics may be blurred. Non-negative matrix factorization may be considered an artificial intelligence method.
The artificial intelligence techniques used in the artificial intelligence tool support vector machines, bear a resemblance to the methods discussed above [58].

2.3.8 Artificial intelligence diagnostic approaches to syndrome diagnosis

This section refers some studies that use of the artificial intelligence techniques utilised in the syndromology articles, or revolves around syndrome diagnosis by artificial intelligence.

2.3.8.1 Identification trees: ID3/ C4.5

The identification tree algorithm used in the syndromology articles, called ID3, has evolved into the program C4.5 [7], and later into C5.0. The first versions were freeware programs, from C5.0 the program is commercial. (The term 'decision tree' that is often used, is a bit unfortunate, since it is also used about the trees made from a decision analysis, a means to assist in choosing between alternatives, with weighting the eventual outcomes).

C4.5 has the ability to handle continuous variables (by creating a threshold and splitting the data into those above and those below the threshold), and also features automatic pruning of the resultant identification tree.
CART - classification and regression trees - are identification trees that use a multivariate approach [59, 60]. The algorithm produces smaller identification trees, but is slower and the trees are harder to interpret.
Podgorelec et al concluded that identification trees in medicine showed a high classification accuracy [61]. There are few applications of identification trees to syndromology.
There are quite a few applications, though, of identification trees in other branches of medicine.
Forsström [62] used an ID3 algorithm on a data set from patients with thyroid illness. The patients classified differently by clinicians and the ID3 algorithm were reported back to clinicians. These patients appeared to be 'special cases' and the reporting helped clinicians in quality control.
Forsström [63] further applied an ID3 algorithm to patients suspected of having thyroid illness. The accuracy of classification result were considered good, but the performance of the algorithm deteriorated when a number of patients with missing values were included.
In forensic medicine, it may be desirable to determine gender from radiological measurements. McBride et al [64] used an ID3 algorithm repeatedly on a sample of data with 31 attributes from radiographic measurements. They left different variables/ attributes out in different runs, to determine which attributes could be ignored. They found a correct classification rate of 93 % , and an agreement between the ID3 and radiological experts of 90 %. The number of variables/ attributes could be reduced to three, still with a high correct classification rate.
Lamy et al [65] applied a C4.5 algorithm to a decision support system for clinical guidelines, to save extensive testing otherwise required.
Attempting to predict chronic fatigue syndrome based on genetic data (SNPs, single nucleotide polymorphisms), Huang et al [66] applied both C4.5, a support vector machine, and a naive Bayes' method to their data set. They used these methods as supplied by the Weka Machine Learning Workbench [59]. They found the naive Bayes' approach with a 'wrapper based feature selection', that is, a program that first selects a subset of the features, outperformed the other programs. The comparatively low overall performance of the methods - sensitivity of about 65 % and specificity of about 50 - 55 % , may stem from the complexity of their data set.
Tanner et al [67] used a C4.5 algorithm to differentiate between dengue fever patients and other patients with fever. They quote a diagnostic accuracy of 84.7 % . The authors state that the algorithm can be used in

other locations (with other prior probabilities of dengue fever), but this does not seem sufficiently substantiated.
A problem with the applications of identification tree algorithms referred to above, is that the prior probability of diagnosis is not taken into account. This may not be a problem if the system developed will be used at the institution where it was developed. Implicitly, one would then assume the prior probability of the condition would not change considerably. The prior probability would then be built into the system by induction. If employed at another institution, however, the algorithm would be expected to perform at a lower level if prior probabilities of the conditions were different.

2.3.8.2 Case based reasoning, CBR

Case based reasoning (see also page 124) is based on the simple idea that if a problem has previously been successfully solved, that solution can be used the next time a similar problem arises. The seeming simplicity is disturbed by the formalism surrounding how to decide whether two cases are alike, and to what extent differences will make the previous solution inappropriate for the present problem.
Evans and Winter applied case based reasoning to syndrome diagnosis [68, 69]. Their aim was to go beyond the earlier systems (BUSCA, GenDiag, and others) that were computer based diagnostic aids, to actual computer diagnostic systems. They used a weighting system from a (then future) version of London Dysmorphology Database, where clinical signs with a high specificity were given a high weight (e.g. 'severe rib shortening') and less specific signs were given a lower weight. They used three such classes. The algorithm traverses a tree of signs, passing high level signs first. If a match is made with several signs from the case to be diagnosed, the system enters the subtree through that node. If a sufficiently similar case is met, a diagnosis is made. In addition, the new case is entered and becomes part of the diagnostic tree. The authors found the algorithm performed well in subsets of syndrome diagnoses, for instance diagnoses falling in the 'acrocephalosyndactyly family'. They concluded that there were still problems to overcome as far as general syndrome diagnosis was concerned. Interobserver non-conformity was one of the difficulties, as was the lack of a clear hierarchical model of syndrome diagnoses.
Loos et al [70] used gray-scaled 2D pictures of people with syndrome diagnoses as the raw material for their study. They used a method from general face recognition, called the bunch graph matching algorithm, which is based on graphs of lines between defined landmarks in the human face. The authors state that clinical geneticists who were shown the photographs

classified the syndrome patients correctly in 62 % of cases. This implies there existed an external criterion standard against which a geneticist's diagnosis could be deemed correct or not. The criterion standard was agreement on diagnosis by two independent geneticists. In general, claiming agreement between two examiners, will lower sensitivity and increase specificity, when one considers the two examinations as one 'test'. This will lead to a selection bias, where untypical cases tend to be excluded.
Based on all landmarks, the system attained an overall correct recognition rate of 58 % . Reanalysis after keeping only the most predictive landmarks, gave an overall recognition rate of 76 % .
Hammond et al's study using dense surface models of 3D models of faces [39], utilised thousands of points in the face landscape. They used a cross-validation against a set of other statistical and artificial intelligence methods: nearest mean, C5.0 decision trees, neural networks, logistic regression, and support vector machines. There was a reasonable agreement between all the methods used. The authors examined a select group of syndrome diagnoses, and did not adjust for prior probabilities of diagnoses.

2.3.8.3 Other approaches

Schorderet was one of the first to use a computer in syndrome diagnosis [71], alongside the clinical databases POSSUM and the London Dysmorphology Database. Schorderet's used a pseudo-bayesian algorithm, and attained a high correct classification rate. This may have been judged too leniently, since the computer diagnosis was considered correct if the correct diagnosis appeared among the top three suggestions by the algorithm.
Douglas, Mutsvangwa et al in their studies using morphometric analyses from pictures of children with fetal alcohol syndrome [36, 37], applied generalized Procrustes analysis, as well as regression and discriminant function analysis. They included 34 subjects, 17 with fetal alcohol syndrome. They compared children with fetal alcohol syndrome only with normal children.

2.3.9 Conclusion

Two major problems in the study of syndrome diagnosis is the lack of a criterion standard of diagnosis, and not taking into account the effect of the prior probability of diagnosis on the predictive value of clinical signs.
Recent work in syndromology has reintroduced the concept of syndrome families. This has lead to the construction of bioinformatics based networks, such as phenome-interactome-gene networks.

Syndromology faces some problems as a scientific discipline, among which are inter- and intraobserver variation, non-repeatable rendering of clinical signs, disagreement over phenotype definitions, and non-congruent classification of the same syndrome based on clinical or laboratory methods. These problems have recently been addressed.
Mathematical-statistical methods are one solution to the question of objective syndrome diagnosis. This is problematic because basic assumptions for the use of such methods are often violated in syndrome diagnostic studies.
The artificial intelligence methods applied to syndrome diagnosis often do not take the effect of prior probability into account. The studies are often done only on a select group of patients, in many cases not including normal controls. The criterion standard of diagnosis is often not explicitly stated.

2.3.10 The bioinformatics studies

The **bioinformatics studies**, articles III and IV, both confronted the challenge of large search spaces in DNA searches, searches that are hard to accomplish with most methods.
If new DNA pattern search methods could be found, one might gain new insights about DNA function.
The RFPL haplotype problem of article III, must be considered against the backdrop of the existing limited success of correlating phenotypes with RFLPs at that time. (RFLP meaning restriction fragment length polymorphism. A restriction enzyme binds to a certain short DNA sequence, typically 4-6 bases. If this sequence is present the restriction enzyme will cut, and a given length of DNA will be cut in two shorter fragments. If the sequence recognised by the restriction enzyme is not present, the restriction enzyme will not cut, and the DNA sequence will remain as one uncut sequence. There is thus a potential for telling the difference between different DNA sequences, based on whether they have a restriction enzyne cutting site or not. This is a polymorphism, in this case a restriction fragment length polymorphism or RFLP).
One was searching for DNA markers such as RFLPs to predict levels of serum cholesterol in people with familial hypercholesterolemia and normals. Single RFLPs had proven inadequate, partly because linkage disequilibrium in the region of the LDL receptor locus meant there were a number of uninformative markers [72].
Statistical methods had not given satisfactory results. Sing et al [73] had tried cladistic analysis but this had also proven inadequate.
There was thus a need for new methodological approaches.

Instead of using only single RFLPs, one wanted to try haplotypes, or sets of RFLPs.
Genetic algorithms do not rest upon basic assumptions that some statistical methods do. A genetic algorithm can handle a haplotype, and can also treat a haplotype as a regular expression, that is it can for instance let presence of a cutting site be represented by 1, absence by 0, and if the presence or absence is of no consequence, let this be represented by e.g. a 'X' as was chosen here.
The DNA search problem of article IV is more sophisticated.
The motivation for this study was the desire to construct a search tool that may find patterns in the parts of DNA that is 'unknown'. Finding such patterns might lead to new insights about function of DNA that at present has no known function.
Large parts of DNA are termed 'junk DNA', and has been considered unfunctional. In recent years years it has been recognised that the assumedly non-functional DNA contains genes or putative genes for non coding RNA [74] non coding conserved regions [75, 76], highly conserved non coding regions (HCNs)/ genomic regulatory blocks (GRBs)[77], and so-called pyknons [78, 79, 80], as well as repeating patterns in 'disease genes' [81]. There is therefore reason to believe that not all of the DNA that has been termed 'junk' actually is non-functional.
One fundamental way of trying to extract meaning from the seemingly meaningless, is to search for patterns.
The goal of article IV was to construct a search tool that would be independent of existing knowledge about what are considered interesting DNA sequences.
This, of course, blocks such a tool from using very much of the amassed knowledge about DNA. There is a priori reason to believe that such a tabula rasa approach will put the search tool at a disadvantage, and slow down the search considerably compared to search tools that do use knowledge about binding sites, repeats, known motifs, gene structure, etc. The potential reward - finding new and unexpected patterns - was considered so great that this approach was chosen, regardless of the obvious disadvantages.
There are a host of DNA search programs, each serving a different purpose. Only a limited number of those are possible contenders against the genetic algorithm based search program presented in article IV, since their purpose is to solve different search problems.
In the following pages, an expose is given first of some search programs that search for patterns, some multiple alignment programs, and some DNA motif searching programs.

Then, some existing genetic algorithm based multiple alignment programs are reviewed.
Obviously, pattern matching is not the only way to find novel features of DNA. One traditional method, but still an exciting area of research, is to try to find conserved regions. Several projects, for instance, utilise a comparative genomics approach for tracing conserved elements in non coding regions [82, 83].
Searching for patterns in medicine/biology in general, many of the basic artificial intelligence methods become crucial. It is often not possible in biology to crisply divide data into two or more distinct groups. This is an area where so called fuzzy sets can be very useful. Fuzzy sets will assign a degree of membership to a group. A man of 40 may be assigned to 'middle aged' with a value of .5, to 'young' with a value of .25, and to 'old' with a value of .25. Khatibi and Motazer used fuzzy sets in pattern recognition/ similarity evaluation on a problem of bacteria recognition [84]. Information theory/ entropy calculation, which is used e.g. in identification trees, was employed by [85]. Using their 'entropic profiler' they could detect over-represented and under-represented segments of DNA sequences.

2.3.11 General approaches to DNA pattern searching

A basic division among DNA pattern searching, and also between multiple sequence alignment programs, are between those using exact methods, and those applying some sort of heuristic. A heuristic is a method to disregard part of the search space, to make the search faster. There is a cost with this however, as the use of the heuristic may lead the search to a local optimum, and failure to find the globally optimal solution. However, a strict exact method is so computationally intensive as to preclude its use in all but the simplest cases.
One example of an exact method is the use of suffix trees [86].
A Gibbs sampler (a type of a Markov chain Monte Carlo model) is a technique built into many programs. This method requires a large number of input sequences to build the model. It has a propensity to end up at local optima [87].
A large group of programs that apparently do the same kind of search as the genetic algorithm of article IV, are motif finding programs. These are often more limited in scope, however, in being aimed at finding comparatively short sequences of less than 10 - 20 bases.
Chan et al [88] introduce a de novo motif finding program for transcription factor binding sites, named GALF-G. It uses a genetic algorithm for the subproblem of finding several overlapping motifs at the same time.

Multiple sequence alignment programs try to find high similarity subsequences among a set of input sequences. Usually, the multiple sequence alignment programs will expect the input sequences to have a high degree of similarity, and the successful output will often have aligned the input sequences for large proportions of their lengths.

Though multiple sequence alignment is not pattern finding, the aligned parts are similar subsequences shared by a number of input sequences - which amounts to a pattern.

Multiple sequence alignment was recently reviewed by Kemena and Notredame [89]. They state the fact that many multiple sequence alignment programs are based on comparing protein sequences, either direct protein-protein comparisons, or DNA translated to proteins, and back-translated to DNA. This may pose a problem when comparing sequences that either cannot or would not naturally be translated into proteins. They further argue that multiple sequence alignment programs have become faster and better because more information about DNA structure is incorporated into the algorithms. This is undoubtedly true. With new techniques using both protein and DNA data - such as ChIp-chip (chromatin immuno-precipitation with micro array technology, ChIp/ chip respectively) multiple sequence alignment becomes more powerful. These are not necessarily valid arguments, however, when using multiple sequence alignment for the specialised purpose of detecting novel patterns, especially not when applying the algorithms to sequences of unknown DNA.

Multiple sequence alignment algorithms can be classified based on the heuristics they use.

Clustal [90] and T-Coffee[91] implement the progressive method of Feng and Doolittle [92]. It starts with two by two comparisons and thus finds starting points for possible elongation of aligned subsequences. The problem is that once underway, this process cannot be reversed. This means the algorithm may find a local optimum.

The exact method of Lipman and Altschul [93] will find the optimal solution, but is in practice to computationally intensive.

The iteration based method, is the one used e.g. in searches based on hidden Markov models. The iterative method proposes an alignment and applies successive attempts at improving the original alignment.

2.3.12 Genetic algorithm approaches to DNA pattern searching

Two multiple alignment programs based on genetic algorithms are referred to in article IV, SAGA [94, 95] and MSA-GA' [96]. Article IV argues against the elitist strategies employed by SAGA (overlapping generations where the best individuals are kept, and fitness proportionate selection). Such strategies in general would be expected to lead to premature convergence, possibly to a local optimum [97].
'MSA-GA', [96] operates on an objective function of alignment scoring in the same manner that SAGA uses an objective function.
Both programs use sequences as their genetic algorithm individuals. It might have been desirable to separate the actual genetic algorithm and the input data/ the sequences.
Article IV also comments on a program using evolutionary programming techniques to perform multiple sequence alignment [98] (Chellapilla and Fogel). Applied to dissimilar sequences, this program attained better results than ClustalW. In this program tournament selection was chosen in place of the elitist fitness proportionate selection of SAGA and MSA-GA.

2.4 Material and methods

2.4.1 Methods

2.4.1.1 Identification trees: ID3/ C4.5

Identification trees organise data by building a tree. The data set consists of instances or elements, all having characteristics (also called features or attributes). It finds the attribute that will split the instances in two groups that are as homogeneous as possible. The identification tree will split the data set according to values of the characteristics, and will assign the elements of the original data set to the leaves, i.e. the end points of the branches of the tree. When applied to a database of artificially generated syndrome patients, it finds the clinical sign that divides the patients into two groups that are as homogeneous as possible. When applied to syndromes, it builds a tree with syndromes as the end points or 'leaves' of the tree.
Several methods can be used to calculate how homogeneous the subgroups of an identification tree is. The CART algorithm [59, 60] uses what is called a Gini index or a Gini measure of impuritiy [57]. The ID3 algorithm of

article I and II, uses a measure of entropy, which minimizes the degree of 'disorder' in the subgroups, i.e. how mixed the subgroup is.

2.4.1.2 The genetic algorithm

The genetic algorithm is a general search and optimization method, based on evolution. Its advantage is that it can usually perform searches with extensive search spaces, that is, problems of such a magnitude that many search algorithms cannot find a solution. The genetic algorithm is non-deterministic and will not necessarily return the same solution in every run. It will often find a good solution, but not necessarily the supposedly best solution.
The genetic algorithm used in article III and article IV are the same program, but the article IV version is considerably enhanced.
This program is 5000 lines of code, written in the Lisp programming language. It is written by the author of this thesis, without use of any modules or plug-ins. It is conceptually based on the ideas behind genetic algorithms as laid down in general texts [99, 100, 101, 102], and conforms with this tradition, but is an independently written program.
This genetic algorithm uses diploid individuals, thus segregation is an important operator to assure diversity of the populations. It also uses mutation and recombination. Since the algorithm uses diploid individuals, recombination is actual recombination between chromosomes/ genes of a pair, not the recombination between different haploid individuals used in the 'simple genetic algorithm' (SGA). This genetic algorithm also uses dominance. In a diploid algorithm, some mechanism has to decide which of the two genes should be expressed. This was solved by a dominance operator contained in the gene. The dominance operator was also subject to mutation, so a 'recessive' gene could change to a 'dominant' gene.
Selection is by 'tournament selection' instead of the more elitist fitness proportionate selection.
The genetic algorithm in its basic form has no memory. An individual containing a good solution may perish, and that particular solution may never come into existence again.
One remedy for this is to keep some individuals through two or more generations. This has been solved by overlapping generations. This is rooted in the nature that inspired the genetic algorithm, but the analogy breaks down if individuals persist for a number of generations. My solution to this problem was to let a small number of individuals procreate by parthenogenesis ('virgin birth', one parent only), making the offspring individual an exact copy of the parent individual. Overlapping generations

and parthenogenesis are both elitist strategies.
Another solution introduced in article IV, was the use of 'paths'. This does not amount to anything more than keeping a record of the short DNA sequences found by the genetic algorithm individuals (n-mers, e.g. a sequence of 8 bases). It also kept record of where in the input sequence such a subsequence n-mer was found. Thus, it would implicitly keep a record of where different individuals had found adjacent subsequences/ n-mers. In the 'natural' terminology of genetic algorithms, this was called a 'path', since it would contain records of where more individuals had found subsequences - as in an 'oft-trodden path'.
Finally, the concept of a 'pack' was introduced. Again, this is not a change in any way to the genetic algorithm. It is a means to join the subsequences/ n-mers into longer sequences. A pack combined their individually found subsequences into one or more areas, a subset of adjacent sequences. The nature analogue is that of individuals hunting together - a 'pack'. To join a pack an individual will have to have found a subsequence that is adjacent to the collective sequence of a pack, or being close. What 'close' is, is set in the *close* parameter, 30 bases in these runs.
(More details of the programming of the genetic algorithm are given in section A.7 on page 116.)
The genetic algorithm was used to search through haplotypes at the LDL receptor locus in article III and to search for recurring DNA patterns in article IV.

2.4.1.3 Feature vector/ set method

The feature vector method (article II) is presented with a patient and a database of patients. It calculates the 'distance' between the patient and each patient in the database. The distance is the number of clinical signs the patient and each database patient do not have in common ('the exclusive or'). Ties were resolved by assigning the case to the most prevalent of the two diagnoses in the database when two diagnoses were deemed equally close to the case.
The set method is applied to the database of artificially generated patients. For each syndrome the set method finds a list of signs common to a diagnostic group - a syndrome. It does this by finding the intersection of clinical signs for all pairs of patients, then the intersection of these intersections of clinical signs again. This produced sets of lists of clinical signs for each syndromes. These were then searched to produce prototypes. These prototypes were the result presented by the set method.

2.4.1.4 Cluster analysis

Two of the articles used cluster analysis. Cluster analysis does not require basic assumptions of for example multinomial normal distributions. Neither does it require knowledge of classification groups before analysis.
Different distance measures can be used for 'closeness' of clusters, as discussed on page 18.
In article I the distance measure was the cosines of differences between corresponding variables. In this study complete linkage was used.
In article II a binary distance measure was used (Sokal and Sneath 5), and average linkage.
The reason for this difference is that in article I the unit of the cluster analysis was a syndrome, whereas in article II the basic unit was a clinical sign. The variables for syndromes were continuous (e.g. 32 per cent of patients with a certain syndrome might have a certain clinical sign), and the variables for clinical signs in individual syndrome patients were binary (the patient either had or did not have the clinical sign).
The cluster analyses in article I and II both used the SPSSX statistical package.

2.4.1.5 Naïve Bayes' calculations

A naive Bayes' calculation was also used as a reference method in article II. This will take into account both the sensitivity and the specificity, as well as the prior probability of the syndrome. The 'naive' in 'naive Bayes" implies the algorithm assumes the clinical signs are independent.

2.4.1.6 MEME

MEME [103] was used as a comparison program in article IV. MEME is a motif finding program. The web server based program has limitations in the allowable length of the motif, the acceptable length of the input sequences etc. Therefore a local installation was done on the laptop used for the genetic algorithm runs. MEME is based on the expectation maximization algorithm, which is an iterative method. By using an heuristic for finding the starting point for the EM algorithm, it performs a greedy search for motifs [104, 105].

2.4.1.7 Training set/ test set

The ID3 algorithm of article I created an expert system, based on the ID3 tree. This expert system could be constructed using a training set of

artificially generated patients, and validated against a test set. The test set and training sets were generated anew for each run with the expert system.

2.4.1.8 The POSSUM patient records

In article I, a database of patients was used for comparison. The database POSSUM (Pictures of Standard Syndromes and Undiagnosed Malformations, now at www.possum.net.au) lists a number of patients with different syndromes, and their clinical findings.

2.4.1.9 Computer

The programs of article I and III were run on a Sun workstation
The programs of article II were run on a dual core 1.73 GHz T2250 laptop computer with 2Gb of RAM and 2Gb of swap, running the Linux/ Ubuntu operative system.
The genetic algorithm of article IV was run on a laptop computer with duo 2.53 GHz T9440 processors, 3 GB of RAM, under Linux/ Ubuntu 9.10.

2.4.1.10 Programming languages

The genetic algorithm of article III and IV was programmed in Lisp (by Braaten Ø), the programming language used in a large number of artificial intelligence programs. Both the Austin Kyoto Common Lisp and the CMUCL (Carnegie Mellon Common Lisp, versions 19a and 19f) interpreters/ compilers were used. Both conform to the ANSI common Lisp standard.
The feature vector and sets programs of article II were programmed in Lisp by the co-author of article II, Friestad J.
Some short scripts used as convenient helper programs for article IV were programmed in Perl and Python (by Braaten Ø). The Perl module for approximate string comparison (StringApprox) was downloaded from cpan.org.
Some summary statistics and figure drawings were performed using the 'R' statistical package.

2.4.2 Material

2.4.2.1 Artificially generated patients.

Artificially generated patients were generated by using figures for prevalence of each syndrome and frequency of each clinical sign for each

syndrome, from 'Birth Defects Encyclopedia'. The figure for prevalence for a given disease in Birth Defects Encyclopedia was converted to a fraction, and multiplied by an arbitrary figure, e.g. 100 000. This would give a number of patients of this type.

The listed frequency of presence of clinical signs for each syndrome was used to randomize whether a sign would be present in an artificially generated patient or not.

An artificial syndrome patient in these runs would consist of a label for the syndrome name, and a list of signs present and signs not present. The signs were present or not present were represented by a list of ones and zeroes, thus FAS 1 1 0 1 0 ...

In the runs of article II, six thousand artificial patients were generated.

2.4.2.2 Artificially generated DNA sequences

Artificially generated DNA was generated by generating strings of A, C, G and T. The bases were generated with uniform probabilities. In article IV, the genetic algorithm (and the comparison program MEME) was tested against two kinds of artificial DNA sequences. First five sequences each 10 000 bases long had a sixty base stretch of A's introduced with an editor in three of the sequences. Secondly, seven sequences of 10 000 random bases were generated. Into sequence 1, 2, 5 and 7, a similar sequence of 150 bases was introduced in each. This 150 base sequence had five per cent of its bases (i.e. 8 bases) randomly changed in sequence 2, 5 and 7, so that each of these four 150 base sequences was slightly dissimilar from the others.

2.4.2.3 Randomization

The randomization procedure was not considered crucial, neither when generating artificially generated patients nor artificially generated DNA. Although there is an extensive literature on randomization, no such sophisticated randomization procedures were used.

2.4.2.4 LDL receptor haplotypes

The LDL receptor haplotypes of article III resulted from RFLP haplotyping of 114 people from families with familial hypercholesterolemia, and 61 normals. Altogether, this gave 175 people and 350 haplotypes.

2.4.2.5 Human DNA sequences

In addition to the randomly generated DNA, four human DNA sequences were chosen as test sequences. These were chosen because they are known to contain so called kringle structures [106, 107]. These sequences thus contain patterns of DNA that a pattern searching program would be expected to find. These DNA sequences used in article IV were downloaded from EMBL, build 36.

2.5 Results

2.5.1 The syndromology studies

The **syndromology studies**, article I and II, found clinical signs that may be used by clinicians, or in machine diagnosis.
Article I presented the syndromes as the leaves of an identification tree, and as a dendrogram from a cluster analysis.
Article II presented the results as sets of clinical signs. The cluster analysis in article II also provided a dendrogram of clinical signs.
There was a good correspondence between the artificial intelligence methods employed in article II - the feature vector method and the set method - and the reference methods - the ID3 identification tree, the cluster analysis, and the naive Bayes' calculation.
The methods of article II are robust and can be applied to a large number of patients, with a large number of clinical signs.
These studies show that artificial intelligence methods can provide objective diagnostic methods in syndromolgy.
In article I, the ID3 tree was compared to a cluster analysis as a reference method. Groups found by the ID3 algorithm were also found in the cluster analysis. Both grouped FAS (fetal alcohol syndrome), Williams syndrome and de Lange syndrome together, as they did the Noonan, Klippel-Feil and Turner syndromes. The Prader-Willi, Zellweger and Beckwith-Wiedemann syndromes were also grouped together by both algorithms, though the ID3 also included other syndromes in this group.
The clinical signs long philtrum, short palpebral fissures, low set ears and hepatosplenomegaly were found high in the identification tree, indicating these clinical signs would be best at dividing the universe of syndromes under consideration into two different groups that would be the most homogeneous.
Rerunning the ID3 a number of times with slightly different values for prevalence of the different syndromes did not change the identification tree

significantly. Neither did small changes to the frequencies of the clinical signs change the resultant tree.
The ID3 algorithm would produce an expert system. When using this on a training set first and then checking against a test set, the medians of the correct classification rates were 92.1 % - 98.1 % , in 20 runs each against 419, 836, and 4180 patients. The higher correct classification rates were attained with more patients in the sets.
The expert system produced was tested against clinical cases from POSSUM. This achieved a low correct classification rate, consistently about 20% .
In article II the main results were that the feature vector method had a low diagnsotic error rate, and that the set method attained a high predictive value for most of the sets of clinical signs.
There was a high degree of correspondence between the artificial intelligence methods and the reference methods.
The feature vector method attained predictive values of between 94.9 and 100 % , the lowest being fragile X syndrome, which has few discerning clinical signs in the newborn period. The Smith-Lemli-Opitz syndrome had a low sensitivity of 69.6 % as did Zellweger syndrome (86.7%).
The feature vector method only made diagnoses, and did not report the clinical signs used.
The set method reported sets of clinical signs to diagnose specific syndromes. These lists of signs were pruned, so that overly long lists are not reported, although they may have had higher predictive values.
With the set method, several signs are reported that must be present simultaneously. Theoretically, this should lower sensitivity and increase specificity. The sets of sign all had high specificities and lower sensitivities, when compared to the other methods. The sensitivities were still acceptable, but with Smith-Lemli-Opitz syndrome at a low of 15 % in sensitivity. This set of signs would therefore not be very useful in diagnosis.
Among the reference methods, the naive Bayes calculation attained the lowest global error rate, although the difference to the other methods was not great.
The cluster analysis grouped clinical signs. This grouping was consistent with what would be expected clinically. Again, Smith-Lemli-Opitz syndrome seemed to be difficult to single out with the data set used here.
There was a good correspondence between comparable methods, when comparing test methods and reference methods.
When comparing the vector method versus the set method versus ID3, it was found that they did equally well, as judged from the global error rate. The methods performed on an equal level in all syndrome groups.

Cluster analysis is a well documented statisitcal method. It was therefore chosen as a comparison method against the set method. The outputs from the two methods are not identicaL. The set method produced lists of clinical signs, the cluster analysis grouped clinical signs. The grouping of signs from the cluster analysis has a qualifying element in that the distance along the axis leading to each clinical sign indicates how closely related the signs are. The cluster analysis only renders the grouping, and does not name syndromes. Given these restrictions, the sets and groups matched reasonably well. The set method listed short palpebral fissures and midface hypoplasia as signs indicating fetal alcohol syndrome. The cluster analysis grouped short palpebral fissure, long philtrum and midface hypoplasia. The set method found flat face, upslanting palpebral fissures, and flat occiput in Down syndrome. The cluster analysis found a tight group of flat face, upslanting palpebral fissures, and flat occiput.

2.5.2 Bioinformatics studies

In the **bioinformatics studies**, article III shows that the genetic algorithm found RFLPs in the LDL receptor that were associated with high cholesterol values.
The genetic algorithm found the restriction sites 1, 2 and 7, and to a lesser degree, site 3 to be associated with the highest cholesterol values.
It demonstrated that the algorithm could find the RFLPs in human DNA as well as in artificially generated haplotypes. The genetic algorithm still found the RFLPs when presented with an artificially generated data set with noise and 'false leads' added.
In article IV, the genetic algorithm was compared to the program MEME [103, 104, 105, 108], which performs an iterative search based on the expectation maximization algorithm. MEME is a mature, well established program.
Both the genetic algorithm and MEME found the stretch of 60 A's introduced into three of five 10 000 base sequences of randomly generated DNA. Next, the search problem was to find slightly dissimilar subsequences in four of seven sequences, each 10 000 bases long. These subsequences were 150 bases long. The 150 bases were originally identical, but three of the four sequences had been changed in five per cent of the bases, randomly chosen. Both MEME and the genetic algorithm found these subsequences. The genetic algorithm did not find them in all runs. Thus, the genetic algorithm performed slightly inferiorly to the comparison program MEME when it comes to finding the patterns in artificially generated DNA, and was also generally slower.

When searching DNA sequences from the four human genes lipoprotein(a), hepatocyte growth factor, plasminogen, and macrophage stimulating factor, it found longer, and more, sequences than the control program. The findings were corroborated by the database searching programs BLAT and Paralign.
These studies - especially the study of article IV - contributed novel features to the genetic algorithm. It introduced the path - which gives the genetic algorithm a memory, and the pack - which increases the ability of the genetic algorithm to search longer sequences.

2.6 Discussion

2.6.1 Syndromology studies

In the **syndromology studies**, article I and II, there are two main findings. First, the studies found clinical signs to be used by clinicians or in machine diagnosis. Secondly, the studies demonstrate that methods from artificial intelligence can provide objective diagnosis in the sense that diagnoses are consistent. Given an input database of real patients, artificial intelligence methods will provide objective diagnoses, limited only by the quality of the input data.
A main premise for these articles is the lack of a criterion standard of diagnosis for syndromes. In fact, quite a few syndromes now have a criterion standard in the sense that other medical diagnoses have criterion standards. The criterion standard may be a chromosomal aberration or a DNA alteration, such as a mutation or a deletion. There are a large number of diagnoses, however, where the problem persists. It may also be argued that in practice the problem of no criterion standard of diagnosis extends to medicine in general. If, for example, the criterion standard is a post mortem examination by a pathologist, this will not be available during the course of the illness while the patient is alive.
In both article I and II, artificially generated patients are used. There are obvious arguments against using artificially generated patients.
However, given the rarity of most syndromes [109], two aspects of using 'real' patients are apparent. A prospective study will be close to infeasible. For a syndrome with a birth prevalence of 1/ 20 000, even in a national study, in a country such as Norway, it would take 5 - 15 years to recruit enough patients. The clinical signs of syndrome patients and, not least important, the normals who might have similar looking features, would be recorded by health personnel not trained in dysmorphology. Some newborns

may die shortly after birth, and the diagnosis may not be recorded.
In retrospective studies, the typical study is the one-syndrome-review [110]. Patients included will tend to have the typical, classical form of the syndrome. There is thus a selection bias operating. For syndromes defined by its clinical signs, a patient may be included in the study on the basis of having clinical signs A, B and C. The study may conclude that patients with this syndrome are characterised by having clinical signs A, B and C. For syndromes with a universally accepted external criterion of diagnosis, one may feel the situation is different. However, some of the patients who have the syndrome in question may not have the typical mutation or deletion that is routinely checked in the laboratory. These patients may be considered not having the syndrome. Some patients may look so atypical for the syndrome that the investigation is never done. In a syndrome review study, these patients will not be included. Again, there is a selection bias operating.
These difficulties do not imply that one should not strive to improve the quality of patient registries, but it is a case for not writing off the use of artificially generated patients.
The choice of syndromes taken from the Birth Defects Encyclopedia, could be critisised. For instance, a syndrome such as Smith-Magenis might have been included. This was not included, being a 'chromosomal' condition. This decision is not consistent however, since trisomies were actually included. Such a critique would affect the findings of clinical signs to be used in syndrome diagnosis. It would not however, be a valid critique against the objective methods themselves.
In the same way, a strong argument could be made for including normals. Some normal children will have one or more clinical signs that could be confused with those of children with syndromes. The main reason for not including normals, is that data are even more difficult to find than data on syndrome children. The inclusion of normals would have had an effect on the prior probabilities, but not on the relative prior probabilities between the syndromes.
The set of syndromes chosen constitutes what is called a 'closed universe'. If all other syndromes are ruled out, the diagnosis will have to be the last syndrome in the set. This may lead to diagnostic errors when using clinical signs that are found, on real patients. In the field of artificial intelligence, this problem is often discussed explicitly. In other situations, the problem may be present, but is not explicitly discussed. One syndromology example, is the one-syndrome-review, which is an extreme form of the 'closed universe'.
Other artificial intelligence methods than those used, might have been

considered for the syndromology studies. Three such methods are artificial neural nets, support vector machines, and Bayesian classifiers.
Neural nets have been extensively used in medical applications. Grossi et al have used neural nets in classification of dyspepsia [111] and for finding the initial symptoms of Alzheimer's disease [112]. Eken applied neural nets to patients with renal colic [113], and Joo [114] used neural nets in diagnosis of solid breast tumors. Pace reviewed the use of neural nets in gastroenterology [115]. Neural nets could certainly be used in classification and diagnosis of syndromes. A major problem with neural nets, is that they are 'black boxes' [116, 117]. The artificial neural network will accept input, it will produce output, but it does not report how the classification was done, or how the diagnosis was reached. This may or may not be a problem. The same objection applies to the genetic algorithm, although not to the same degree since intermediary results can be examined. The identification tree/ ID3 and cluster analysis, on the other hand, are transparent, and the process can be retraced. Especially in syndrome diagnosis, it is desirable to have methods that are transparent.
In some cases, one wants the option of being able to use incremental learning. Incremental learning means that once a classification of a data set has been done, a new small data set may be fed to the classifier, and these data are included in the classification. A method that cannot learn incrementally, would have to add the new data set to the old data set, and perform the classification anew. A neural network is an example of a system that can learn incrementally. An ID3 algorithm would have to do the classification a second time.
Like for the genetic algorithm, there are no hard and fast rules about how to set the parameters of adjusting a neural network. This means that using a neural net system requires experimenting with the setting of the parameters. (Often, a genetic algorithm is used to as a helper program to choose the appropriate parameters.) The problem of parameter setting is also evident for a genetic algorithm. Thus, neural networks as well as genetic algorithms are both an art and a craft. For this reason, many scientists would prefer traditional, reliably repeatable methods where that is possible. However, these methods have their place when the problem is beyond the scope of other methods.
One method that could have been used for the syndromology studies, is the support vector machine (SVM). This is a method that builds a predictive model by finding the dividing line between two categories. If no such straight dividing line can be found, the support vector machine can use some form of polynomial transformation [59]. With transformations, and more groups, the dividing line will be a hyperplane in multidimensional

space.
The support vector machine is a fast and effective algorithm, with many applications in medicine and bioinformatics [118]. Laurikkala et al [119] used support vector machines, along with other methods, such as a genetic algorithm, to classify urinary incontinence in women. Huang et al [120] examined ion channel proteins to predict potential drug targets.
Three problems in my opinion make the support vector machine less appropriate than the methods chosen in the syndromology articles.
The support vector machine is a black box method - it will not reveal the data behind the choice of dividing lines. With transformed data, this might also have been very difficult to comprehend.
The support vector machine may need different transformations for different problems, and the search for a proper transformation can be time consuming.
Finally, a support vector machine works better with large data sets, and less well with small data sets. In some cases, this will be a problem in syndromology applications.
A Bayesian classifier - such as the one used in article II - could be used in place of the identification tree or a cluster analysis. A Bayesian classifier, however, cannot deal with interdependent clinical signs. Since many of the clinical signs found in syndrome patients are dependent on each other, this makes the Bayesian classifier less appropriate.
Compared to these other possible choices of simple methods, the identification tree and the feature vector methods' main advantages are that they are easy to interpret, and they can handle large data sets.
A main result of article I were that the signs long philtrum, short palpebral fissures, hepatosplenomegaly and low set ears were found high in the identification tree. The importance of the first two clinical signs (in fetal alcohol syndrome) is corroborated by the findings of Douglas and Mutsvangwa [36, 37]. Many clinicians would hold that e.g. low set ears is a too unspecific sign to be of diagnostic value. However, since this sign is found in many patients, it is a sign that may be used to divide the universe of possible diagnoses in two. This actually is a powerful search method, similar to the one used in binary searches - and identification trees. A finding of a sign such as short palpebral fissures is consistent with Bayesian theory. This is a sign that is frequent in a syndrome with a high prevalence.
There was poor agreement between the expert system produced by the ID3 algorithm and the POSSUM database. This may reflect problems with the signs found by the ID3 algorithm. It may, however, also reflect problems with distorted prior probabilities caused by selection bias in the database.
As a main result of article II, it was found that the methods advantages

were robustness, simplicity, power, and scalability. The robustness means that the methods do not require normal distributions of variable values, they do not require statistical independence of signs, they can handle binary variables, and they can handle missing values. The methods are simple in that they are easy to understand. The methods are powerful in being able to handle large amounts of data compared with competitior methods. They are also very fast. These methods possess scalability in that they can handle tens of thousands of features from hundreds of thousands of patients.
An important issue is the question of validation of the methods. A premise of these articles was that 'proper' syndromes do not have an external criterion standard of diagnosis. This does not hold true for all the syndromes included. Still, in the set up chosen here, with a large number of artificial patients, it would not have been possible to validate the findings against some criterion standard.
It would have been possible to compare the findings to what has been found in other syndrome investigation studies. Findings were indeed corroborated by e.g. [36, 37]. Having critisised this type of studies for possible selection bias, for not using objective diagnostic methods, and for misleadingly consider only one syndrome at a time, it appears inconsequential to use such data as an external criterion standard.
Missing an external objective means the important point of validation becomes the internal consistency between the objective methods. The methods used did display a high degree of internal consistency, although no quantitative measure of consistency was used. One possibility might have been to use the kappa measure used to compare human observers.

The **main contribution to the field of syndromology by articles I and II** was emphasizing the importance of objective methods, and the demonstration that it is feasible to apply objective methods from artificial intelligence to syndromology. The search for objective diagnostic measures advocated in these articles can in retrospect be seen to be part of a tradition from the early attempts at a numerical nosology [55] and systematization of the description of syndromes [31] [34], to recent attempts at standardising measurements and descriptions [32, 33] and Orphanet's effort to produce reliable figures for the prevalence of rare conditions [109].

2.6.2 Bioinformatics studies

In the **bioinformatics studies**, in article III the main findings were the RFLP haplotypes associated with high cholesterol. In article IV, the main findings was that the genetic algorithm could find patterns in unknown

DNA on par with a widely used program such as MEME. The other important point was the study's contribution to the genetic algorithm (the path and the pack).
The genetic algorithm is a non-deterministic algorithm, and may not produce the same results in every run. The advantage of the genetic algorithm is its ability to search large search spaces.
The genetic algorithm used in articles III and IV is a diploid genetic algorithm. This is not a new invention, and not a rarity in the genetic algorithm literature [100, 101], but many applications use some form of the simple genetic algorithm, which is haploid.
Uyar and Harmanci [121] investigated the consequences of using a diploid genetic algorithm. They found the diploid genetic algorithm to be superior over a set of test functions. These authors found the diploid set up adds diversity to the genetic pool. They stress the effect of recessive genes that can survive in the population to be expressed later, preventing traits that may prove to be useful, from being lost.
It is interesting to consider the parallels between natural systems and artificial systems such as the genetic algorithm. Rice and Chippindale [122] wanted to test the theoretic claim that sexual recombination will increase the power of selection. They set up a D melanogaster model system. They found that recombination increased the selection in the population.
In article III, the genetic algorithm found the restriction sites 1, 2 and 7, and to a lesser degree, site 3, to be associated with the highest cholesterol values.
This draws on the genetic algorithms propensity for searching for 'schemata'. It will search for combinations of sites present or not present, without regard to the actual haplotype. Thus, also a multiplicative model would be expected to be discovered by the genetic algorithm, say, if the combination of restriction site 1 and 7 would give a much higher cholesterol value than expected from values for 1 and 7 individually.
At the time of the study, both statistical methods and cladistics had been applied to this problem without success. There was therefore no obvious external validation of the genetic algorithm findings at hand (apart from what could be read from an inspection of the haplotypes and cholesterol values of the raw data).
The genetic algorithm was run against artificial data containing noise and false leads, and extracted the correct RFLP's even with a high degree of noise and false leads. The algorithm proved robust, and gave consistent results in repeated runs.
In article IV, the genetic algorithm and the comparison program MEME found subsequences in artificially generated DNA. This showed the genetic

algorithm could consistently pick out subsequences from a set of input sequences. This amounts to a validation of the genetic algorithm's ability to find such patterns or subsequences. Though this was a set of randomly generated DNA, the search is not trivial.
These random sequences were made with a uniform probability random DNA generator. A choice was made not to use e.g. a random generator based on a hidden Markov model, since the genetic algorithm was intended to be able to search sequences of DNA of unknown function. Using a hidden Markov model based on well known/ functional DNA might therefore be misleading.
Next, the genetic algorithm was applied to four sequences of human DNA; lipoprotein(a), plasminogen, hepatocyte growth factor and macrophage stimulating factor. These are known to contain kringle domains.
The reason the genetic algorithm did not perform as well as the comparison program MEME, may be that these input sequences contained too many similar subsequences. Thus, the MEME program may have had an advantage.
The main point in this study, was that the genetic algorithm actually performed on par with MEME, one of the most widely used programs for motif/ pattern finding. At the same time, some of the potential limitations of MEME, such as problems with starting anew if an elongation from an initially found short sequence did not prove optimal, does not hold to the same degree for the genetic algorithm.
Full scale testing of the genetic algorithm against real sequences of DNA of unknown function would of course be interesting. That was, however, considered beyond the scope of this primarily methodological study.
There is an argument for testing the genetic algorithm against a benchmark database of test sequences, such as Balibase [123]. However, such databases are made for a specific purpose - as far as Balibase is concerned, multiple sequence alignment - and it would be inappropriate to draw conclusions from such tests.
There are several examples of genetic algorithms [124, 125] having been applied to DNA searches - notably the SAGA program [126] and MSA-GA [96]. These programs, however, are often adjunct programs to other search programs, use other varieties of genetic algorithm, and perform the searches differently. They are therefore not directly comparable to the genetic algorithm used in this project.
Some of the problems of bioinformatics - such as the one presented here, of finding unknown patterns in unknown DNA - present an algorithm with very large search spaces. Although a number of DNA search programs exist, few are fit for the task of finding such patterns. The genetic

algorithm introduced here, may be an important supplement to existing search algorithms.
Exciting development along the lines of thinking behind the genetic algorithm, are algorithms based on the behaviour of the social insects, such as ants. These algorithms use aspects of the ant hill, the behaviour of ants, pheromone tracks etc to search through large search spaces. These methods are being applied in medicine and bioinformatics [127, 128][129].
The **main contributions of articles III and IV** was to point to the need for new search methods for finding patterns in DNA - especially in the sequences of the human genome that are unknown or deemed 'uninteresting' by common consent - and to propose a genetic algorithm with some novel fetaures that could be one such tool.

2.7 Implications

The **syndromology studies** of this project may contribute to an increased focus on problems with the lack of a criterion standard of diagnosis and of objective diagnoses. These problems may be more acute in syndrome diagnosis, but extends to medical diagnosis in general. Objective, consistent diagnoses are necessary both in bedside diagnostic work by clinicians, and in machine diagnostic systems or computer aided diagnosis.
It could be argued that it is hubris on the part of the international research community to write off ninety eight per cent of our genome as 'junk'. This is changing, but calls for a new set of mind - and it calls for new techniques.
The **bioinformatics studies** of this project have furnished the genetic algorithm as a tool for searching for unknown patterns in DNA sequences. It would be valuable if the study could also contribute to the prevalent trend of curiosity and interest in the 'uninteresting' parts of DNA.

Part II

Articles

2.8 Article I: Artificial intelligence in pediatrics

Reprinted in
Yearbook of medical informatics

2.9 Article II: Diagnosis: Human intuition or machine intelligence?

Syndrome Diagnosis: Human Intuition or Machine Intelligence?

Øivind Braaten* and Johannes Friestad

Department of Medical Genetics, Ullevål University Hospital, Oslo, and Institute of Medical Genetics, University of Oslo, Norway

Institute of Informatics, University of Oslo, Norway

Abstract: The aim of this study was to investigate whether artificial intelligence methods can represent objective methods that are essential in syndrome diagnosis. Most syndromes have no external criterion standard of diagnosis. The predictive value of a clinical sign used in diagnosis is dependent on the prior probability of the syndrome diagnosis. Clinicians often misjudge the probabilities involved. Syndromology needs objective methods to ensure diagnostic consistency, and take prior probabilities into account. We applied two basic artificial intelligence methods to a database of machine-generated patients - a 'vector method' and a set method. As reference methods we ran an ID3 algorithm, a cluster analysis and a naive Bayes' calculation on the same patient series. The overall diagnostic error rate for the the vector algorithm was 0.93%, and for the ID3 0.97%. For the clinical signs found by the set method, the predictive values varied between 0.71 and 1.0. The artificial intelligence methods that we used, proved simple, robust and powerful, and represent objective diagnostic methods.

Keywords: Artificial intelligence, diagnosis, computer-assisted, classification, diagnostic errors, syndrome.

1. INTRODUCTION

This study aims to investigate whether artificial intelligence methods can represent objective methods in syndrome diagnosis. Such methods are essential because most syndromes lack a criterion standard of diagnosis, and because clinicians often misjudge the effect that prior probabilities have on the predictive value of diagnostic handles, such as clinical signs.

When a child is born with malformations, it is devastating for the parents. To quickly find a diagnosis is important for possible treatment, prognosis, and for the parents' need to know.

The child's malformations may represent a syndrome. But syndrome diagnosis is beset with difficulties, e.g. the lack of an external validation of the diagnosis for most syndromes.

We argue that objective methods are essential in syndrome diagnosis, and, indeed, necessary in all forms of clinical diagnosis.

We show that simple artificial intelligence (AI) methods may be such objective methods, capable of establishing diagnostic criteria in syndrome diagnosis.

1.1. Syndromes: No Criterion Standard of Diagnosis

In this article the word 'syndrome' means 'congenital malformation syndrome'. (For example Table **6** in the results section gives examples of syndromes and the associated clinical signs or features). A syndrome is a clinical delineation based on the presence of a set of clinical signs. The standard method in clinical syndrome diagnosis is the 'pattern recognition' method where the clinician looks for the clinical signs that make up a certain syndrome.

For most syndromes, there is no 'gold standard' or 'criterion standard' of diagnosis. There may thus be no biochemical, radiological, DNA diagnostic or chromosomal investigation to verify the diagnosis. The accuracy (validity, 'correctness') of the diagnosis may for many syndromes have to be relinquished because of this lack of a criterion standard of diagnosis. Still, the *sine qua non* of scientific method - **consistency** - remains a fundamental goal.

1.2. The Effect of Prior Probability on Predictive Value Confuses the Issue

The predictive value of clinical signs is strongly dependent on how common the syndrome is, the 'prior probability'. Tables **1**, **2** and **3** show the striking effect of the prior probability on a clinical sign's worth as a diagnostic measure. Clinicians do not always estimate the prior probability of a disease correctly [1-3] -- the standard prevalence figures do not necessarily apply in a differential diagnostic situation. This often leads to confusion about the diagnostic value of a particular diagnostic sign.

Table 1. Clinical Indices

	Syndrome Present		Syndrome Not Present	
Positive test	TP	**a**	**b**	FP
Negative test	FN	**c**	**d**	TN

TP, true positives, FN, false negatives, FP, false positives, TN, true negatives. In the context of this article, positive test means clinical sign present, and negative test means clinical sign not present. Sensitivity is a/a+c, the probability of having the clinical sign, given that you have the disease, specificity is d/ b+d, the probability of not having the clinical sign, given that you do not have the disease. Predictive value is a/ a+b, the probability of having the disease, given that you have the clinical sign.

*Address correspondence to this author at the Department of Medical Genetics, Ullevål University Hospital, Kirkeveien 166, 0407 Oslo, Norway; E-mail: oivind.braaten@medisin.uio.no

1874-4311/08

Table 2. Predicitive Value, High Prevalence

	Syndrome Present	Syndrome Not Present	
Positive test	95	10	
Negative test	5	90	
	100	100	200

Sensitivity 0.95, specificity 0.90, prevalence 0.50. Positive predictive value 95/ 95 + 10 = 0.90, i.e. the probability that the patient has the syndrome if this sign is present, is ninety per cent.

Table 3. Predictive Value, Low Prevalence

	Syndrome Present	Syndrome Not Present	
Positive test	95	990	
Negative test	5	8910	
	100	9900	10000

Sensitivity 0.95, specificity 0.90, prevalence 0.01. Positive predictive value 95/ 95 + 9900 = 0.087, i.e. the probability that the patient has the syndrome if this sign is present, is still less than nine per cent.

1.3. The Philosophical-Scientific Issue

The lack of objective methods has a philosophical-scientific, and a practical aspect.

The question may seem a problem of marginal importance, of interest to those involved in the mathematical side of medicine. On the contrary, it is a major, though not much recognised problem. Objective methods are necessary in the reductionist philosophy of science that medicine claims to be a part of. The question is at the foundation of medicine as a scientific discipline. If diagnoses cannot be validated against a criterion standard, and are not even consistent, it is not possible to consider medicine a scientific discipline.

It could be argued that the problem of 'no criterion standard' of diagnosis extends to virtually all areas of medicine. Both clinical diagnosis and laboratory diagnosis may vary from one medical practitioner to another. Even for diseases such as diabetes, hypertension or peptic ulcer, doctors may differ in what the definition of the disease is. Although professional bodies establish diagnostic criteria, these may not be congruent with what an individual doctor uses. For clinical diagnoses there may be no agreed-upon diagnostic criteria. Since a diagnosis links to information about prognosis and treatment, vague diagnostic criteria may be harmful both in medical practice and in medicine as science.

1.4. The Consequences of Diagnostic Errors

A false positive diagnosis may lead to the patient receiving unnecessary and potentially harmful treatment. It may mean fear and worry for the patient and her or his relatives.

A false negative diagnosis may mean the patient will forgo life-saving or disease modifying treatment, or important educational measures.

Depending on the situation, both false positive and false negative diagnoses may lead to further unnecessary, potentially harmful, and costly investigations.

1.5. Objective Methods are Needed to Establish Diagnostic Criteria

It is obviously important to avoid the diagnostic errors and their consequences. The prevalent intuitive pattern recognition approach to syndrome diagnosis is open to misdiagnoses. Objective methods can act as a corrective to the intuitive approach and help remedy some of its shortcomings.

We approach this by trying to establish diagnostic criteria to be used by clinicians.

1.6. Objective Methods: Mathematical--Statistical Approaches

Mathematical-statistical methods might represent methods that could establish diagnostic criteria. But there are problems with using statistical methods, primarily because basic assumptions often are not met. A number of statistical classification methods have been applied to syndromology, such as factor analysis/ principal component analysis [4,5], discriminant analysis [6-10], log-linear analysis [11], latent class analysis [12], and cluster analysis [13,14].

Most multivariate statistical methods are parametric, and require multinomial normal distributions of the variables, as well as continuous variable values. These basic assumptions can rarely be met. Missing values for one or more variables is often an additional problem.

1.7. Objective Methods: Sophisticated AI Methods

Several artificial intelligence and informatics methods could be used to tease out the clinical signs with the highest predictive value in syndrome diagnosis. Neural nets, support vector machines, and non-negative matrix factorization [15,16] are examples of such methods.

Case based reasoning [17,18] and the ID3 algorithm [19] have previously been tried as alternatives to statistical methods.

Problems with the more sophisticated AI methods are that they may seem so complex and unfamiliar as to alienate clinicians who would be the ones to use the results of the analyses. Especially with small data sets there is also the problem of using too much sample specific information and not getting generalizable results, i.e. overfitting.

1.8. Objective Methods: Our Approach

We hold that some fundamental artificial intelligence techniques can successfully be applied to the problem of establishing diagnostic criteria.

We introduce a feature vector method, a set method, and also apply other artificial intelligence methods.

The techniques we propose are variants of known methods rather than basically new. What we argue is that the situation in syndrome diagnosis warrants objective methods, i.e. these methods are a necessity, and the methods we propose represent a possible practical solution. The application

of these methods to syndrome diagnosis is new, and, in our opinion, an example of a type of approach that is necessary.

1.9. Conclusion

In syndrome diagnosis there is often no criterion standard of diagnosis.

In syndrome diagnosis as in medical diagnosis in general there is a need to be alert to the strong effect of prior probability on the predictive value of diagnostic indicators, such as clinical signs. Objective methods can help counteract the misdiagnoses that can be caused by neglecting this. The human intuitive approach is not very good at estimating and taking into account the probabilities involved.

Objective methods are warranted as a corrective to the intuitive approach to syndrome diagnosis.

Using mathematical-statistical approaches entails problems with the basic assumptions of these methods.

Using the more sophisticated AI methods may also violate basic assumptions. The complexity of these methods may alienate clinicians.

We apply two simple informatics/ artificial intelligence methods to see whether these methods can help establish diagnostic criteria for syndromes.

2. MATERIAL AND METHODOLOGY

We created a database of machine-generated patients.

We applied 'the vector method' and the set method as well as one artificial intelligence reference method - the ID3 -, and two mathematical reference methods -- cluster analysis and the naive Bayes -- to this patient series.

The Birth Defects Encyclopedia (BDE) [20] -- a classical catalogue of clinical syndromes -- lists the occurrence (prevalence or incidence) of syndromes along with the clinical signs found in the syndrome. It also lists the frequency of these clinical signs in each syndrome.

In this study, we included syndromes with a listed occurrence of one per fifty thousand or more. Some conditions were excluded, such as isolated neural tube defect, as well as several groups of syndromes, for example the arthrogryposes.

We generated 'artificial patients' based on the BDE.

The data from the BDE was transformed into artificial patients in the following manner: For each syndrome the figure for occurrence, e.g. 1/ 20 000, was multiplied by a common arbitrary figure, e.g. 100 000. This gave the number of artificial patients, in this case five artificial patients. For each artificial patient, the algorithm had to decide whether each clinical sign was to be present or not. For this, it used the listed frequency of the clinical sign for this syndrome. For each sign, a random number between zero and one was generated. If the random number was smaller than the listed frequency of the sign, it was decided that this sign would be present in this particular artificial patient. If the random number was larger than the listed frequency of the sign, it was decided that this sign would not be present in this artificial patient.

Each artificial patient therefore consisted of a syndrome name and a list of signs present ('1') or not present ('0').

We generated six thousand artificial patients. This gave a reasonable number of patients even for the least common syndromes.

The list of artificial patients had the syndromes in 'true proportion' to their occurrence as given in the BDE. The clinical signs had the same overall frequency as listed in the BDE. Any non-random co-existence of clinical signs was lost by the randomization process.

2.1. The 'Vector Method'

The vector method algorithm starts with a set of patients with known diagnoses on the one hand and a patient to be diagnosed on the other hand.

In our context, the database of patients with known diagnoses was the artificially generated patients.

When presented with a new case - the patient to be diagnosed - the main procedure of the vector method algorithm compared the new case to all existing cases. For each individual case in the database, it calculated the 'distance' between the patient to be diagnosed and the database case. The algorithm assigned a new case to the syndrome diagnosis where the 'distance' was smallest.

Basically, the 'distance' is the number of clinical signs that two patients do not have in common, i.e. those signs that either of the patients has and the other does not have.

The algorithm calculated this distance by finding the 'exclusive or' for a pair of patients, i.e., the signs present in one syndrome patient but not the other.

This represents the difference or the dissimilarity or the distance between the two cases.

2.1.1. Ties

In some instances, two cases or more in the database had equally small distances to the case that was to be diagnosed. In this situation, the new case was assigned the diagnosis of the database case belonging to the most prevalent of the syndromes with the same distance.

2.2. The Set Method

The vector method algorithm would diagnose a new patient, but did not give information about which signs were used in diagnosis.

To present such a list of clinical signs, we applied a set method to the database of artificial patients as well.

The approach of the set method is similar to the one used by the vector method algorithm, but with the set method there were no individual patients to be diagnosed. The set method finds a list of clinical signs common to each syndrome group - a 'feature vector'.

The algorithm first found the intersection of the lists of clinical signs for all pairs of patients for a given diagnostic group. We thus got all sets of features common to at least two patients. The algorithm then proceeded by intersecting all pairs of these sets again, producing sets of clinical signs

common to at least four patients, and so on. We repeated this cycle until no more feature sets were produced. In this way, we found the most common sets of features for each syndrome.

However, the most common set of features may not be the most predictive. The clinical signs that are common in one syndrome, may be common in another syndrome as well. This set of features then cannot be used to distinguish between diagnostic groups.

We therefore searched for prototypes - feature vectors which were common to a large number of the patients in a given diagnostic group, but which differed from common feature vectors of other diagnostic groups. The algorithm also identified subclasses within diagnostic groups. If a large subclass existed within a syndrome, the algorithm rendered the feature vector for that subclass.

2.2.1 Computer and Programming Language

We used the Lisp programming language. The programs were run on a PC with the Linux operating system.

2.3. The Reference Methods

As a reference for basic artificial intelligence methods, we used the ID3 algorithm, cluster analysis, as well as a 'naive Bayes' 'calculation'.

2.3.1. The ID3 Identification Tree Method

The ID3 starts by dividing the patients into two subgroups, where each subgroup is as homogeneous as possible. Homogeneous in our context means that the patients have the same clinical signs. After the first division into subgroups, each subgroup is subdivided into two new subgroups, and so on. This procedure builds a tree, where the original group is the root/ trunk, subgroups are branches, subsubgroups are twigs, and the basic unit of analysis is a leaf. The basic unit of analysis is e.g. an individual patient or a syndrome. The signs used to discriminate between groups, are the branching points in the tree.

To decide how homogeneous a group is, the ID3 algorithm uses an information theory formula:

$$\sum_b \left(\frac{n_b}{n_t}\right) \times \left(\sum_c -\frac{n_{bc}}{n_b}\ log_2\ \frac{n_{bc}}{n_b}\right)$$

where nb is the number of instances in branch b, nt is the total number of instances in the whole tree, and nbc is the total of instances in branch b of type c. In our context, 'type c' stands for 'syndrome patients who have a certain clinical sign'. At each branching point in the tree, the remaining syndrome patients are divided into two groups, those who have the clinical sign and those who do not have the clinical sign.

2.3.2. Cluster Analysis

With the cluster analysis runs, we used the same data sets as we used for the runs using the basic artificial intelligence methods.

We ran cluster analyses using average linkage between groups, and nearest neighbour as the clustering method. Since our data were binary, we used a binary measure of similarity ('Sokal and Sneath 5', the squared geometric mean of conditional probabilities of positive and negative matches). Clinical signs were used as the basic unit of analysis.

2.3.3. 'Naive Bayes' Calculations

Theoretically, the optimal way of finding which clinical signs have the largest predictive value, is using a calculation based on Bayes' formula. This formula takes into account the sensitivity as well as the specificity of the clinical sign, and the prior probability of the syndrome.

There are two problems with using 'Bayes' formula. First, it assumes that clinical signs are independent. This does not always hold true. 'Upward slanting palpebral fissures' as a sign clearly is not independent mathematically from 'downward slanting palpebral fissures'. 'Low set ears' and 'upward slanting palpebral fissures' probably occur together more often than expected by chance, etc.

Secondly, the figures that go into Bayes' formula are often not readily available.

2.4. Runs Using Artificial Intelligence Methods

In these runs, the results presented for the vector method algorithm, the set method, the ID3 and the 'naive Bayes' ' are all averages of ten runs with six thousand artificial patients in each run. The vector method algorithm was directly applied to ten consecutive batches of six thousand patients, i.e. with no training phase. The ID3 and set methods were first trained on a set of six thousand patients, and then tested with the ten batches of six thousand patients each. Each batch of six thousand patients for the test runs was new, in that it was generated anew. However, the batches were all made using the same procedure for generating patients.

3. RESULTS

3.1. General Observations

3.1.1. Feasibility of the Artificial Intelligence Approach

The vector method had a low diagnostic error rate. This holds true for the global error rate, as well as for the error rates of the individual syndromes. The set method attained a high predictive value for most of the sets of clinical signs.

These basic artificial intelligence methods were easy to implement, rapid, and showed consistent results in repeated runs.

3.1.2. Correspondence Between Artificial Intelligence Methods and Reference Methods

There was a good correspondence between comparable methods.

The set method on the one hand, and the cluster analysis using clinical signs as the basic unit on the other hand, gave signs or groups of signs that match.

3.2. The Artificial Intelligence Methods

3.2.1. The Vector Method Algorithm

With the vector method algorithm, there was no learning phase. This algorithm directly diagnosed the patients.

As seen in Table **4**, the predictive values were high, with the lowest being 94.9 for fragile X syndrome. Fragile X syn-

drome does not have many distinguishing features in the newborn period.

Table 4. Vector Method/ Nearest Neighbour Run

Syndrome Name	No of Cases	Sensitivity	Specificity	Predictive Value
FAS	3597	99.9	99.5	99.7
Trisomy 21	702	100.0	100.0	100.0
Fragile X	355	99.4	99.7	94.9
Noonan	299	99.7	100.0	99.7
Congenital CMV	221	94.6	99.8	95.4
Trisomy 18	208	99.0	99.8	95.8
Turner	123	94.3	100.0	98.3
Trisomy 13	93	90.3	100.0	98.8
deLange	81	97.5	100.0	100.0
Williams	66	97.0	100.0	98.5
Beckwith	56	96.4	100.0	100.0
Prader-Willi	55	100.0	100.0	98.2
Meckel	38	94.7	100.0	100.0
Cri du chat (5p-)	30	100.0	100.0	100.0
Zellweger	30	86.7	100.0	100.0
Klippel-Feil	23	95.7	100.0	100.0
SLOS	23	69.6	100.0	100.0

FAS, fetal alcohol syndrome, SLOS, Smith-Lemli-Opitz syndrome. Average of ten runs of 6000 artificial patients in each run. On average correctly diagnosed 5944, global error rate 0.93%.

The global error rate is satisfactory. However, quite low sensitivities were observed for some syndromes, with Smith-Lemli-Opitz syndrome (SLOS) at a low of 69.6%.

The vector method did not produce any output other than diagnoses. Thus, the algorithm did not have to make any concessions for the sake of readability. The algorithm could therefore use all available information without doing any pruning. ('Pruning' here means removing twigs on an ID3 tree, or parts of other search results which do not cover many cases, but which contribute to making it more complex). The nearest neighbour algorithm attained very high specificities, at one hundred per cent, or close to a hundred per cent.

3.2.2. The Set Method

The set method table (Table **6**) lists the sets of signs found by the set method, along with their clinical indices. It should be stressed that these are sets of signs, i.e. either the full set listed is present, or it is not. This theoretically should have the effect of lowering sensitivity and increasing specificity. The impression from the tables is definitely that the specificity is higher than is usual, in many instances 100%.

Yet, the sensitivity does not seem to be dramatically lowered, though e.g. Smith-Lemli-Opitz syndrome (SLOS) with the set method is down to a 15% sensitivity. Although the predictive value is very good, this particular set will therefore not be a very useful set of signs in diagnosis.

Table 5. ID3run

Syndrome Name	No of Cases	Sensitivity	Specificity	Predictive Value
FAS	3597	99.7	99.5	99.7
Trisomy 21	702	100.0	100.0	100.0
Fragile X	355	93.2	98.9	98.2
Noonan	299	100.0	100.0	100.0
Congenital CMV	221	94.6	99.7	98.3
Trisomy 18	208	100.0	100.0	99.5
Turner	123	94.3	99.8	90.6
Trisomy 13	93	98.9	100.0	100.0
deLange	81	100.0	100.0	100.0
Williams	66	97.0	99.8	87.7
Beckwith	56	100.0	100.0	100.0
Prader-Willi	55	100.0	100.0	100.0
Meckel	38	100.0	100.0	100.0
Cri du chat (5p-)	30	100.0	100.0	100.0
Zellweger	30	100.0	100.0	100.0
Klippel-Feil	23	95.7	99.9	88.0
SLOS	23	100.0	100.0	100.0

FAS, fetal alcohol syndrome, SLOS, Smith-Lemli-Opitz syndrome.
Average of ten runs of 6000 artificial patients in each run. On average correctly diagnosed 5942, global error rate 0.97%.

The lists of signs found by the set method have been pruned to make them more accessible to a human reader. We have tried to strike a balance between two concerns. The lists of clinical signs that we present are few per syndrome, and fairly short, in some instances the list of signs is just one single clinical sign. The sensitivity and specificity are still in general quite acceptable. The signs found make sense from a clinical point of view. The most cumbersome diagnoses are trisomy 13 and Zellweger syndrome. In trisomy 13 four lists of three clinical signs each are presented. In Zellweger syndrome, the longest list has four clinical signs that have to be present simultaneously.

On the other hand, ten of the seventeen syndromes have fairly predictive lists of only one sign.

The clinical sign 'short palpebral fissures' has a predictive value of one hundred per cent. It has a sensitivity of 89%, so this is a useful clinical sign.

3.3. The Reference Methods

3.3.1. The ID3

As seen in Table **5**, the global error rate is low for the ID3 run, at about the same level as the vector method.

3.3.2. Cluster Analysis

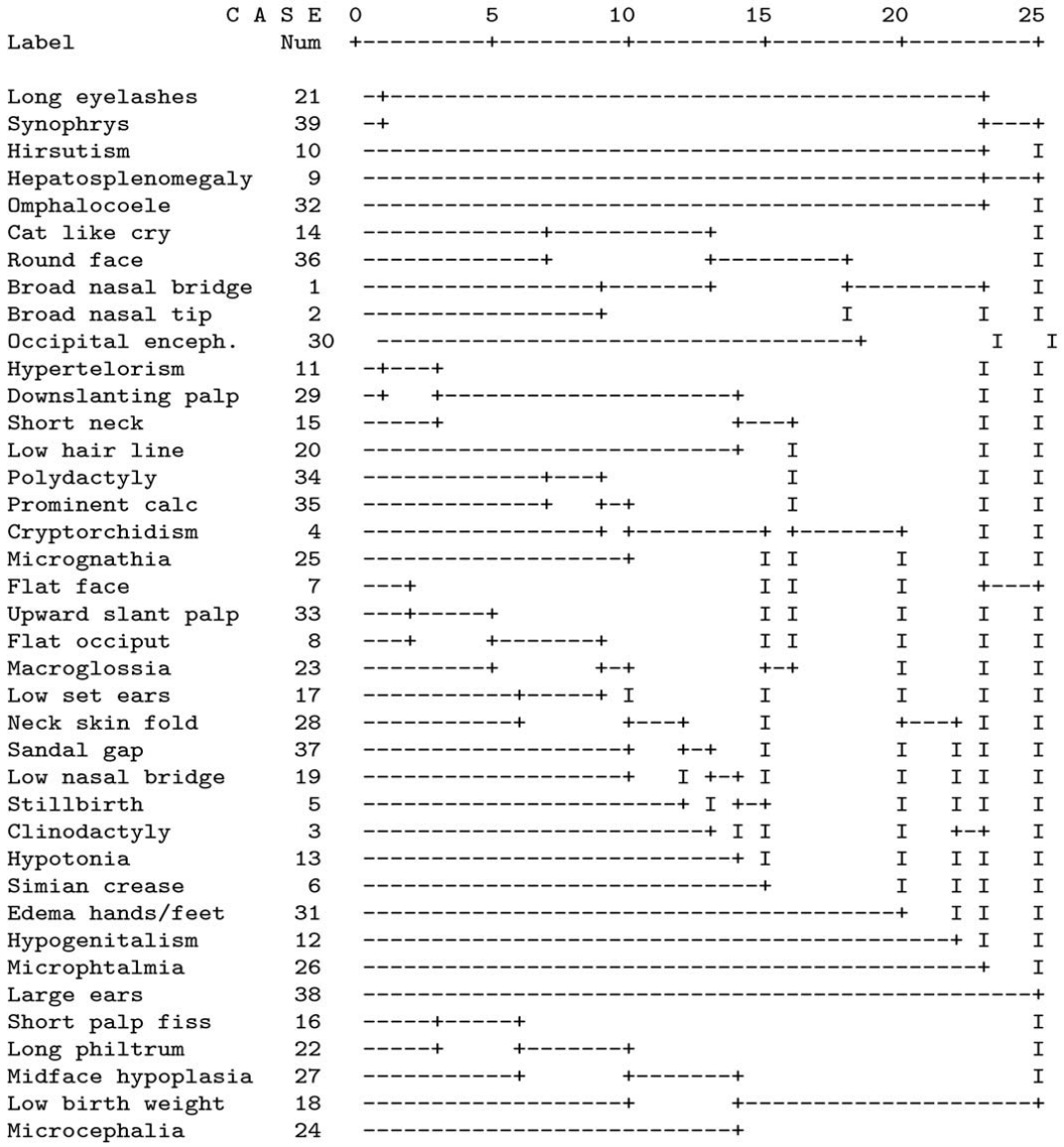
Table **7** shows a dendrogram, after a cluster analysis has been run, where the clinical signs were used as the basic measure of analysis.

Table 6. Sets of Clinical Signs Versus Syndromes, 'Set Method' Results

Syndrome Name Set of clinical signs	Sensitivity	Specificity	Predictive Value
FAS			
Short palpebral fissures	0.89	1.0	1.0
Midface hypoplasia	0.79	0.98	0.98
Trisomy 21			
Flat occiput	0.77	1.0	1.0
Upward slanting palpebral fissures	0.79	0.99	0.97
Flat face	0.90	0.99	0.93
Fragile X			
Large ears	0.88	0.98	0.76
Noonan			
Down slanting palpebral fissures	0.96	0.99	0.87
Hypertelorism Low set ears	0.87	0.98	0.74
Congenital CMV			
Hepatosplenomegaly	0.89	0.99	0.71
Trisomy 18			
Large ears Cryptorchidism	0.32	0.99	0.78
Prominent calcaneus Cryptorchidism	0.65	0.99	0.78
Polydactyly Cryptorchidism	0.86	0.99	0.71
Polydactyly	0.86	0.98	0.65
Turner			
Oedema of hands and feet	0.38	1.0	1.0
Micrognathia Low hair line	0.61	0.99	0.70
Trisomy 13			
Hypertelorism Polydactyly Cryptorchidism	0.69	1.0	1.0
Polydactyly Microcephaly Cryptorchidism	0.59	1.0	1.0
Hypertelorism Microcephaly Cryptorchidism	0.56	1.0	1.0
Hypertelorism Simian crease Cryptorchidism	0.66	0.99	0.95
deLange			
Synophrys	0.82	1.0	1.0
Long eyelashes	0.76	1.0	1.0
Long philtrum Clinodactyly	0.58	1.0	1.0
Williams			
Broad nasal tip	0.74	1.0	1.0
Broad nasal bridge Long philtrum	0.61	1.0	1.0
Broad nasal bridge	0.59	0.82	0.99
Beckwith			
Macroglossia Midface hypoplasia	0.84	1.0	1.0
Macroglossia Cryptorchidism	0.79	1.0	1.0
Macroglossia Hepatosplenomegaly	0.73	1.0	1.0
Prader-Willi			
Flat face Cryptorchidism	0.69	1.0	1.0
Hypogenitalism	0.97	0.99	0.92
Meckel			
Polydactyly Hepatosplenomegaly Stillbirth	0.86	1.0	1.0
Occipital encephalocoele Stillbirth	0.86	1.0	1.0
Cri du chat (5p-)			
Cat like cry	1.0	1.0	1.0
Zellweger			
Hepatosplenomegaly Hypotonia	0.80	1.0	1.0
Low BW Hypotonia Upward slant palp fissures	0.5	0.99	0.83
Micrognathia Low BW Hypotonia Cryptorchidism	0.25	0.99	0.73
Klippel-Feil			
Short neck Low hairline Microcephaly	0.29	1.0	1.0
SLOS			
Polydactyly Microcephaly Micrognathia Low BW	0.15	1.0	1.0

FAS, fetal alcohol syndrome, SLOS, Smith-Lemli-Opitz syndrome. Low BW, low birth weight, Upward slant palp fissures, upward slanting palpebral fissures.

Table 7. Dendrogram from a Hierarchical Cluster Analysis Using Single Linkage, Showing the Relationship Between Clinical Signs



Distance along the axis is a relative measure of dissimilarity. Occipital enceph, occipital encephalocoele, Downslanting palp, downslanting palpebral fissures, Prominent calc, prominent calcaneus, Upward slant palp, upward slanting palpebral fissures, Short palp fiss, short palpebral fissures.

The triad of long eyelashes, synophrys and hirsutism is characteristic of Cornelia de Lange syndrome. Hepatosplenomegaly and omphalocoele are less distinctive, but point to Beckwith-Wiedeman syndrome. The hepatosplenomegaly alone also fits with Zellweger syndrome or congenital cytomegalovirus infection (CMV). Though not syndrome specific, this sign could be seen as pointing to this group of syndromes. 'Cat like cry' and round face are hallmarks of 'Cri du chat' or monosomy of the short arm of chromosome five. Williams syndrome is characterized by broad nasal tip/ broad nasal bridge. The single sign occipital encephalocoele is a strong pointer to Meckel syndrome.

Three syndromes have several similarities as far as clinical signs are concerned: Noonan syndrome, Turner syndrome and Klippel-Feil syndrome. The next group of clinical signs, hypertelorism, downward slanting palpebral fissures, short neck, and low hairline fit these syndromes. It can be seen from the arbitrary scale of the dendrogram that hyperelorism and downward slanting palpebral fissures are closely related, and in comparison stand apart from short neck and low hairline. This may distinguish Noonan syndrome from Turner syndrome and Klippel-Feil syndrome. Turner syndrome patients when newborn also have edema of hands and feet, found as a single clinical sign at line nine from the bottom of the dendrogram.

A large group of clinical signs, from polydactyly to simian crease, denote the trisomies (trisomy 21, 18 and 13). The first and smallest subgroup of this group fits trisomy 13 and 18, with the signs polydactyly, prominent calcaneus, cryptorchidism and micrognathia. The second, larger subgroup of clinical signs here is consistent with trisomy 21 (Down syndrome).

Because of the relatively high prevalence of the trisomies, some clinical signs seem to have been 'stolen' from the less prevalent syndromes. An example of this is the Prader-Willi syndrome (hypotonia, cryptorchidism).

No individual syndrome springs to mind for hypogenitalism as a single sign. In this context, however, hypogenitalism would strongly suggest Prader-Willi syndrome. Similarly, large ears strongly indicate Fragile X/ Martin-Bell syndrome.

Short palpebral fissures, long philtrum, and midface hypoplasia define fetal alcohol syndrome.

This leaves the signs microphtalmia, low birth weight and microcephaly as signs with no associated syndrome.

The syndromes that have not been taken into account are Smith-Lemli-Opitz syndrome (SLOS) and to a certain degree Zellweger syndrome and congenital cytomegalovirus infection. Smith-Lemli-Opitz syndrome (SLOS) seems to be difficult to diagnose for several of the methods with the data used here.

3.3.3. 'Naive Bayes' Calculations

The results for the 'naive Bayes' 'calculations are listed in Table **8**. Although the difference is not large, the naive Bayes' calculations attain the lowest global error rate of diagnosis. Like in the vector method runs, the naive Bayes' calculation uses all available information, and does not have to compromise to satisfy a demand for human readability.

3.4. Comparing the Methods

3.4.1. The Vector Method Versus the Set Method Versus ID3

These three methods did roughly equally well as judged by the overall error rate. None of the methods did very badly in any of the syndrome groups. (It would have been possible to have a good overall performance, even with a poor performance in the smaller syndrome groups).

Small variations in specificity could lead to relatively large variations in predictive value.

Table 8. 'Naïve Bayes' Calculation

Syndrome Name	No of Cases	Sensitivity	Specificity	Predictive Value
FAS	3597	99.9	99.5	99.9
Trisomy 21	702	100.0	100.0	100.0
Fragile X	355	99.4	99.7	95.7
Noonan	299	100.0	100.0	100.0
Congenital CMV	221	95.5	99.9	97.7
Trisomy 18	208	100.0	99.9	98.1
Turner	123	95.9	100.0	99.2
Trisomy 13	93	95.7	100.0	100.0
deLange	81	100.0	100.0	100.0
Williams	66	97.0	100.0	100.0
Beckwith	56	100.0	100.0	100.0
Prader-Willi	55	100.0	100.0	98.2
Meckel	38	97.4	100.0	100.0
Cri du chat (5p-)	30	100.0	100.0	100.0
Zellweger	30	100.0	100.0	100.0
Klippel-Feil	23	95.7	100.0	100.0
SLOS	23	100.0	100.0	95.8

FAS, fetal alcohol syndrome, SLOS, Smith-Lemli-Opitz syndrome. Average of ten runs of 6000 artificial patients in each run. On average correctly diagnosed 5971, global error rate 0.48%.

3.4.2. The Set Method Versus Cluster Analysis

These two methods are comparable in that they both rendered lists or clusters of clinical signs. We chose cluster analysis as a reference method since it is a mainstream mathematical method. The cluster analysis with clinical signs as the basic unit is most appropriate for comparison with the set method. This analysis did not name syndromes, it just grouped clinical signs. Given this restriction, the clinical signs grouped by the cluster analysis, and the sets of signs found by the set method match reasonably well. For example, Table **6** shows, from the top, that FAS (fetal alcohol syndrome) according to the set method has the signs short palpebral fissures, and midface hypoplasia. Table **7**, the cluster anaysis, shows in line 5, 4 and 3 from the bottom, that short palpebral fissures, long philtrum, and midface hypoplasia are grouped closely together. Next, for trisomy 21 (Down syndrome) in Table **6** the set method found the signs flat occiput, upward slanting palpebral fissures, and flat face. In Table **7** (the cluster analysis) in the middle of the figure finds a narrow grouping of flat face, upward slanting palpebral fissures, and flat occiput.

4. DISCUSSION

The principal aim of this study was to demonstrate that our vector method and other basic artificial intelligence methods represent objective methods that are essential in establishing diagnostic criteria in syndromology.

4.1. The Artificial Intelligence Methods

The vector method attained high rates of correct diagnoses. The set method did find a set of clinical signs for each syndrome diagnosis. These findings were corroborated by

the results of the cluster analysis. The clinical signs teased out by the set method are also reasonable from a clinical point of view.

In contrast to many other studies, our study had a data set with correct proportions between the different syndrome diagnoses.

Thus, the study has dealt with the problem of prior probabilities.

If the artificial intelligence methods can successfully be applied to data from artificially generated patients, it seems valid to infer that they could be used on data from real patients.

The algorithms found clinically useful signs, signs that may be used both by clinicians, and for machine diagnosis.

The **vector method and set method's main advantages** are

- **Robustness**

These methods are robust in that:

- They do not require normal distributions of variable values.
- They do not require statistical independence of signs.
- They can handle binary variables.
- They can handle missing values.

- **Simplicity**

- The methods are basic and easy to understand.

- **Power**

- The methods are powerful in that they can handle larger amounts of data than most of its competitor methods. They are also very fast.

- **Scalability**

- Some methods which are useful with a small number of cases/ patients do not scale up to large numbers. The vector method should be able to manage tens of thousands of features and hundreds of thousands of patients. In practice this means the limiting factor will be how many patients the researcher is able to collect.

The term predictive value used for the vector method algorithm is to a certain degree a misnomer, since there was no clinical sign or set of clinical signs that could be evaluated for predictive value. The 'predictive value' here is calculated *post hoc*. The term has been kept for consistency.

The time used by the vector method algorithm increases linearly with the number of cases (O(n)), while the time increases as the square of the number of cases for the set method (O(n*n)).

4.2. Cluster Analysis

In this study, we used cluster analysis as a control, to see if the findings by the set method could be substantiated. The cluster analysis lends support to the set method findings.

4.3. Details of Our Study -- Discussion of Validity of Results

4.3.1. General Considerations

Using Randomly Generated Patients

Doctors as well as informaticians often prefer using 'real patients' to e.g. machine generated patients. Syndromes are rare, so it would in practice be a prohibitive task to find a representative number of patients for each syndrome group. Furthermore, biases may be introduced when using selected groups of 'real patients', e.g. by the inclusion of only the 'classical cases' in the patient series. Thus, it may actually be the better option to use machine-generated patients.

In a situation with no criterion standard, it would be potentially misleading to directly compare the performance of the artificial intelligence methods with clinicians' performance. If either approach - AI or clinical - were chosen as the reference standard, that approach by definition would outperform the other.

There are overwhelming practical and methodological problems with doing a prospective study encompassing all syndromes to establish the frequency of clinical signs in each individual syndrome and in the patient group at large.

Our primary goal was to demonstrate that the artificial intelligence methods could be used to pick out the most predictive clinical signs in syndrome diagnosis. We were not concerned with diagnosis of individual syndrome patients. We therefore chose the scheme described using figures from the Birth Defects Encyclopedia, and randomly generated artificial patients.

Our randomization procedure generated a small number of 'patients' with very few clinical signs just by chance. Since it was set up to generate a clinical sign in an individual patient with a probability of 0.9 if 90% of patients were listed in BDE to have the sign, 1 in 10 would not have the clinical sign in question. The probability that a given artificially generated patient would lack both of two such signs, would be 0.1*0.1, or one in a hundred. When a large number of patients were generated, the occasional patient would have very few signs altogether.

This will obviously make the diagnostic task more difficult, for an artificial intelligence method, as well as for any other method.

Any co-existence of clinical signs would be lost by the randomization procedure. This may be a source of error when the methods are applied to artificially generated patients, but the first order predictive value of signs is probably greater than the second order or combined effect of two individual clinical signs.

4.3.2. The Set Method

Pruning and Prototyping

The original lists of clinical signs found by the set method are obviously the best to use to arrive at a diagnosis. The set method, though, may also be counter-intuitive,

stating that the patient should have all the signs listed. Pruning and prototyping will simplify matters for a clinician as the less important signs are removed, and the remaining list is more manageable. We have arbitrarily pruned by removing lists of clinical signs that contain more than 3-4 signs.

When it comes to machine diagnosis, however, pruning is unnecessary and will only lower the diagnostic performance.

4.4. General Considerations in Syndrome Diagnosis with Respect to our Study

4.4.1. Accept old Diagnoses or form New Ones?

Most studies on syndrome diagnosis accept established diagnoses. Diagnoses in single patients may be questioned, but the diagnostic groups themselves are often considered untouchable.

Using methods such as the vector method or cluster analysis, it is an option to challenge the existing diagnostic groups. Set up in this way, it is conceivable that the nearest neighbour algorithm could suggest lumping or splitting of diagnostic groups, that new groups with different boundaries should be formed, or that totally new groups should be established.

'New' syndromes As far as establishing new diagnoses is concerned, an objective method has advantages compared to the pattern recognition method.

The pattern recognition method would be dependent on a single clinician seeing enough cases of a new syndrome to realize it was actually a new syndrome.

The syndromologist would then have to report it, other syndromologists would have to read the report and recognize the syndrome themselves. This obviously works in many cases, since new syndromes are regularly reported.

It is a disturbing fact, though, that we cannot know how many syndromes are **not** reported. It is reasonable to think that an international central database of syndromes would be useful for awareness to detect new syndromes. One important group would be teratogenic syndromes, e.g. possibly caused by the mother living close to a nuclear plant, caused by estrogen-like pollutants in the environment, caused by maternal drug abuse etc.

4.4.2. Using One Sign Versus Using a Set of Signs

The solution provided by the set method is a set of clinical signs that have to be present simultaneously.

This is different from the single-sign method, where one sign, when found, increases the probability of the syndrome, the next sign may increase or decrease the probability etc.

In general, the requirement for several signs to be present at the same time, increases specificity and decreases sensitivity.

This is reflected in the tables of the Results section, where several lists of signs have a very high specificity, often one hundred per cent. Once found, these clinical signs (the set of clinical signs) will be better predictor variables.

A very long list of signs that have to be present simultaneously, may not be of value to a human diagnostician. Such a list would make perfect sense in machine diagnosis, though.

4.4.3. Using the Sign to Find a Diagnosis Versus Using the Sign to Partition the Universe of Possible Diagnoses

The most common approach with syndrome diagnosis based on clinical signs, is to use single signs to get closer to a diagnosis. With other methods, like the ID3 method, one partitions the 'universe' of possible diagnoses and thus continually circles in the few diagnoses that remain. In artificial intelligence, this way of searching is common, whereas in clinical thinking it may not seem that natural (although many diagnosticians use this way of thinking, perhaps unconsciously).

4.4.4. The 'Closed World Assumption'

In artificial intelligence, it is common to make explicit the concept of the 'closed world'. Many studies make this assumption, but do not state it explicitly. In the closed world of our study, there were only seventeen syndromes. Thus, if sixteen of the syndromes could be ruled out, the diagnosis would have to be the seventeenth syndrome. This may be unrealistic in a real-world situation.

4.4.5. Inclusion of Negative Signs

Syndromologists often speak of 'handles', i.e. clinical signs with a high positive predictive value. We kept to this standard approach of using positive signs, i.e. signs present.

Of course, signs not present may help single out diagnosis just as effectively. Conversely, a sign may have a high negative predictive value, i.e. if this sign is present, the diagnosis becomes much less likely.

4.4.6. Clinical Phenotype or DNA Based Diagnosis?

DNA diagnosis and diagnosis based on the clinical phenotype could either give the same result, or different results. In some cases the problem is small, since there is no alternative to clinical classification and diagnosis.

In other cases, one might ask which would be the 'correct' classification.

The clinical classification may be more practical. The DNA diagnosis is easier, more clear cut, and may have a higher status [21].

However, clinical classification is not outdated, and never will be. What is of interest, is ultimately the phenotype, the human being. If the overlap between a phenotypic classification and a DNA classification is little, so is the interest in the DNA 'defect'.

5. CONCLUSION

For most syndromes there is no criterion standard of diagnosis.

In many cases, one will therefore have to forgo an accurate diagnosis. It is therefore of paramount importance to have a consistent set of diagnostic criteria. Thus, there is a need for objective methods of diagnosis. Traditionally, these have been various statistical methods. However, statistical methods have certain weaknesses, e.g. they require basic assumptions that often cannot be met.

The vector method and the set method used here, are objective methods that are robust, simple and powerful. This study has shown they can successfully be applied to a database of clinical signs and syndrome diagnoses. In this study, we used these basic methods to elicit objective clinical signs with high predictive value; signs that can be used by clinicians.

These methods may also be used in computer assisted diagnostic systems.

In conclusion, the two basic methods used here, can embody the objective methods that are mandatory in syndrome diagnosis, and necessary in all forms of medical diagnosis.

SUPPLEMENTARY MATERIAL

This article is accompanied by an overview slide presentation and it can be viewed at www.bentham.org/open/tominfoj

REFERENCES

[1] Cahan A, Gilon D, Manor O, Paltiel O. Probabilistic reasoning and clinical decision-making: do doctors overestimate diagnostic probabilities? QJM 2003; 96: 763-9.

[2] Cahan A, Gilon D, Manor O, Paltiel O. Clinical experience did not reduce the variance in physicians' estimates of pretest probability in a cross-sectional survey. J Clin Epidemiol 2005; 58: 1211-6.

[3] Richardson WS. Five uneasy pieces about pre-test probability. J Gen Intern Med 2002; 17: 882-3.

[4] Haley RW, Kurt TL, Hom J. Is there a Gulf War Syndrome? Searching for syndromes by factor analysis of symptoms. JAMA 1997; 277: 215-22.

[5] Kosaki K, Jones MC, Stayboldt C. Zimmer phocomelia: delineation by principal coordinate analysis. Am J Med Genet 1996; 66: 55-9.

[6] Ross JL, Kushner H, Zinn AR. Discriminant analysis of the Ullrich-Turner syndrome neurocognitive profile. Am J Med Genet 1997; 72: 275-80.

[7] Loesch DZ, Scott D. Application of the anthropometric discriminant functions in estimation of carrier probabilities in Martin-Bell syndrome. Clin Genet 1989; 36: 145-51.

[8] Loesch DZ, Wilson SR. Multivariate analysis of body shape in fragile X (Martin-Bell) syndrome. Am J Med Genet 1989; 33: 200-8.

[9] Murdoch-Kinch CA, Ward RE. Metacarpophalangeal analysis in Crouzon syndrome: additional evidence for phenotypic convergence with the acrocephalosyndactyly syndromes. Am J Med Genet 1997; 73: 61-6.

[10] Astley SJ, Clarren SK. A fetal alcohol syndrome screening tool. Alcohol Clin Exp Res 1995; 19: 1565-71.

[11] Neel JV, Julius S, Weder A, *et al*. Syndrome X: is it for real? Genet Epidemiol 1998; 15: 19-32.

[12] Volk HE, Henderson C, Neuman RJ, Todd RD. Validation of population-based ADHD subtypes and identification of three clinically impaired subtypes. Am J Med Genet B Neuropsychiatr Genet 2006; 141B: 312-8.

[13] Preus M. The Williams syndrome: objective definition and diagnosis. Clin Genet 1984; 25: 422-8.

[14] Verloes A. Numerical syndromology: a mathematical approach to the nosology of complex phenotypes. Am J Med Genet 1995; 55: 433-43.

[15] Kotsia I, Zafeiriou S, Pitas I. A novel discriminant non-negative matrix factorization algorithm with applications to facial image characterization problems. IEEE Trans Forensic Security 2007; 2: 588-95.

[16] Pascual-Montano A, Carmona-Saez P, Chagoyen M, *et al.* bioNMF: a versatile tool for non-negative matrix factorization in biology. BMC Bioinform 2006; 7: 366.

[17] Evans CD. A case-based assistant for diagnosis and analysis of dysmorphic syndromes. Med Inform (Lond) 1995; 20: 121-31.

[18] Evans CD, Winter RM. A case-based learning approach to grouping cases with multiple malformations. MD Comput 1995; 12: 127-36.

[19] Braaten O. Artificial intelligence in pediatrics: important clinical signs in newborn syndromes. Comput Biomed Res 1996; 29: 153-61.

[20] Buyse M. Birth Defects Encyclopedia. Dover: Center for Birth Defects Information Services, Inc 1990.

[21] Biesecker LG. Lumping and splitting: molecular biology in the genetics clinic. Clin Genet 1998; 53: 3-7.

Received: August 21, 2008 Revised: September 21, 2008 Accepted: October 21, 2008

2.10 Article III: The genetic algorithm applied to haplotype data

2.11 Article IV: Finding DNA patterns with the genetic algorithm

Part III

Appendices

Appendix A

Artificial intelligence

A.1 Introduction

This appendix is a brief overview of some artificial intelligence methods and techniques. The emphasis is on methods used in medicine, with the focus especially on the methods used in this thesis, the identification tree algorithm and the genetic algorithm.

A.2 Expert systems

Many of the artificial intelligence system that have been used in medicine, are expert systems. This was especially true of the early years of medical artificial intelligence, but expert systems are still advocated [130].
Expert systems in artificial intelligence traditionally are built from a number of rules. These rules are usually of the form

> **if** some antecedent condition(s) is/ are met
> **then** some consequent action(s) is/ are carried out

Such rules are known as if-then-rules, production rules or antecedent-consequent-rules.
To find the appropriate rules for a certain problem area ('domain'), one relies on the process of 'knowledge acquisition', see also section C.4 on page 143 on 'tacit knowledge'. This basically consists of extended interviews with a person knowledgeable in the area in question – a 'domain expert'. Such a knowledge extraction process may run into what is classically known as 'the bottleneck problem': On one side the domain expert with her/ his knowledge, on the other side a 'knowledge engineer' ready to build an expert system – and between them the bottleneck of knowledge acquisition.

Partly to overcome this problem, one turned to induction. Using inductive learning from a set of examples, a system would learn the expert knowledge indirectly.
An ID3 tree is an example of induction. Conversely, an ID3 tree can be transformed into an expert system.

A.3 Neural nets

Neural nets have been widely used in medical applications of artificial intelligence [131, 117] for a wide range of problems in medicine; from selecting markers for Alzheimer's disease [112] to diagnosing gastroenterologic disease [115] to evaluating images in cancer diagnosis [114, 132]. The reason for this appeal to the biomedical research community may be because its biological basis - neurons and synapses - is familiar to biologists and doctors.
A neural net consists of layers of neurons that are interconnected. The neural networks are most often used for problem solving, but could also be used to gain insight into biological neural systems.
Neurons in an artificial neural network are organized in layers, where each neuron in one layer has a number of connections to neurons in the next layer. The first layer is usually a layer capable of perceiving, e.g. 'seeing' an image. The neurons or nodes in a neural network will change their characteristics depending on input from the previous layer. A neuron in an artificial neural network will fire - send an impulse to neurons in the next layer - if the excitatory input exceeds a pre-set threshold. In some cases a genetic algorithm is used to train the neural net and set the optimal level of thresholds for firing.
Typically, a neural network is dedicated to learning, e.g. to recognise a written letter. During the learning phase, the neural network minimises a 'cost function'. The cost function is a measure of how far away the present solution is from an optimal solution.
Although the approach is different, this has a number of analogies to the genetic algorithm.

A.4 Machine learning

Machine learning is a broad topic in artificial intelligence, and there are numerous methods and algorithms of machine learning in use in medical artificial intelligence. These methods are applied in bioinformatics

[133, 134, 135] as well as in clinical medicine [136].
The area of machine learning can be divided along different lines, for instance between deductive and inductive learning. A common division is between supervised and unsupervised learning.

A.4.1 Supervised learning

In supervised learning, the learning algorithm is presented with a set of exemplars or examples, as well as the classification of these exemplars. In the next phase, the algorithm's task is to classify new, unknown specimen.

A.4.2 Unsupervised learning

In unsupervised learning, there are no known classification - such as for example diagnoses - and the learning algorithm has to learn by induction.

A.4.3 Data mining and text mining

Two areas that may be considered part of machine learning, are data mining and text mining. Like machine learning in general, these are subjects with a long tradition, in an artificial intelligence context. Both data mining and text mining apparently are being used to an increasing degree in biomedicine[137, 138, 139, 140].

A.5 Identification trees

The identification tree, or induction tree algorithm ID3 was introduced in artificial intelligence by Quinlan [7]. These trees are also called decision trees. [1]
The ID3 algorithm, like other inductive methods, requires a set of examples or exemplars to start with.
The algorithm uses the examples to build an identification tree. The tree building proceeds by partitioning the elements into two groups, based upon one of the characteristics.
The raw material
The raw material/ the raw data for an ID3 algorithm to work on, consists of individual cases or instances with common features. These features or characteristics can take on different values. In the syndrome diagnosis

[1] I prefer the term identification trees, since decision trees can be confused with the trees of decision analyses.

problem, the individual cases would be newborns with a syndrome. The characteristics would be the clinical signs found. In syndrome diagnosis the values would typically be 'present' and 'not present'. The algorithm assumes discrete classes, i.e. no instance belongs to several classes. A newborn cannot be classified as having two or more syndromes.

The algorithm

The most common algorithm, and the one used in this project, tries to divide the instances into two groups, where each group is as homogeneous as possible; then to divide each of these subgroups into two homogeneous groups again, and so on. To this end, it uses information theory.

In information theory, the log to the base of 2 is used to find the number of bits necessary to represent a piece of information.

Less mathematically, one could consider two bowls of a thousand objects each. One bowl contains 1000 identical 1 cm in diameter white marble balls. In the other bowl, the first object is a red tetrahedron with 1 mm black stripes, the second is a 1 cm in diameter black plastic ball, the third is a die, white with black spots, etc. Clearly, a description of the first set of instances - the 1000 white marbles - will represent far less information than when the other bowl is to be described. Put another way, the instances - or white marble balls - of the first bowl as a whole are far more homogeneous. To decide how homogeneous a group is, the ID3 algorithm uses an information theory formula:

$$\sum_b (\frac{n_b}{n_t}) \times (\sum_c -\frac{n_{bc}}{n_b}\ log_2\ \frac{n_{bc}}{n_b})$$

In this formula n_b is the number of instances in branch b, n_t is the total number of instances in the whole tree, and n_{bc} is the total of instances in branch b of type c.

Advantages of identification tree approaches are that they are simple, especially simple to interpret. Most types of data can be used without transformations. The identification trees are robust.

The algorithm has developed into C4.5, and later C5.0 (which is a commercial program, and not open source).

A problem with identification trees is that the approach is open to fragmentation. This happens when a characteristic can take on a large number of values. (A continuous characteristic or variable does not necessarily lead to this problem, if the values can be grouped e.g. into 'high', 'medium' and 'low').

To counteract these problems, an extension of identification trees/ decision trees is the decision graph.

A similar approach to an identification tree is a regression tree. For regression trees the outcome is a figure, e.g. number of days of a hospital stay. The CART - Classification and regression tree - combines the two methods.

A.6 AI search strategies

As artificial intelligence came into being, informatics already had a set of techniques to solve search problems. Therefore one tried to cast problems in the form of searches.

The most basic search strategies in artificial intelligence are traditional informatics searches.

A search is often envisaged as a traversing of a search tree. Such a tree is either balanced or unbalanced (i.e. symmetric or unsymmetric). A binary tree is a tree where all branches to the left of the root have elements that are of smaller values than the root, and all elements to the right are of larger value. This knowledge makes it far simpler to find the element one wants to find. In a binary tree with the elements A-B-C-D-E-F-G, D would be the root. The left branch would have B as the new root or main branch, and B would have A as the left branch and C as the right branch.

A **breadth first search** search would go through or traverse the tree with the main branches first: D-B-F, and then the next level, A C E G

Conversely, a **depth first search** would go to the bottom of the search tree along one set of branches, then backtrack and go down the next branch, D-B-A, etc.

A **search space** can be envisioned as a landscape with peaks, ridges, and valleys. In a search space such as this, the peaks will be good solutions, i.e. points where certain variable values are at an optimum. The highest peak will represent the best solution.

The algorithm called **'hill climbing'** will move in this landscape, always going in the direction that is 'steepest'. In an otherwise flat landscape - meaning there is one and only one solution - this approach will find the highest peak, or optimum solution. In a landscape with several peaks - several good solutions - it might get stuck at the second or third highest peak, i.e. a local optimum. The hill climbing algorithm would then not find the highest peak - the global optimum.

This way of thinking applies directly also to the genetic algorithm.

A.7 The genetic algorithm, general introduction

The most commonly used term now is evolutionary computation[2], which includes a number of methods and techniques inspired by evolution, among those are genetic algorithms, genetic programming, evolutionary programming, plus hybrid systems. An example of hybrid systems is GANN, genetic algorithm/ neural net, where a genetic algorithm is used to optimise parameter settings for the neural net.
Evolution has been a success in producing ever better-fit individuals/ species. The basic idea underlying the genetic algorithm is to extract the basic principles of evolution, model the algorithm based upon these principles, and apply the algorithm in computer-based problem solving. Thus, the genetic algorithm is an abstraction that mimics nature: One considers a **population** of individuals, each **individual** having a certain genetic make-up. The individuals struggle for survival in an **environment**, and succeed or fail according to their **fitness**, relative to the other individuals in the population. Those individuals with the highest fitness reproduce. During **reproduction** there is possibly a **recombination** or **crossing over** of the **chromosomes**. Occasionally a **mutational event** takes place.

A.7.1 Review of biological terms and concepts

> It may be said that natural selection is daily and hourly scrutinising, throughout the world, every variation, even the slightest; rejecting that which is bad, preserving and adding up all that is good; silently and insensibly working; whenever and wherever opportunity offers, at the improvement of each [...] being in relation to its organic and inorganic conditions of life. *Charles Darwin,* page 133, The Origin of Species [141]

Fitness in natural genetics is defined as the probability of transmitting one's genes to the next generation. One simple way to calculate fitness is to take the number of offspring produced on average by some subgroup of the population and divide this by the average number of offspring in the 'normal' population.
It is important to realize that 'survival of the fittest' means survival of the fittest **genes**.

[2]The GECCO 2009, Genetic and Evolutionary Computational Conference, combines the 18th International Conference on Genetic Algorithms (ICGA), and the 14th Annual Genetic Programming conference (GP) (www.sigevo.org/gecco-2009).

The **gene** is the unit of heredity. For genetic algorithm purposes, it is enough to think of a gene as part of a **chromosome**. Each individual carries several chromosomes.
All chromosomes taken together constitute the material of heredity for the individual in question.
Genetic algorithms often use **haploid organisms** , i.e. organisms with one set of chromosomes. These genetic algorithms rely heavily on recombination to maintain population diversity.
In a **diploid organism** there are two sets of chromosomes, consisting of pairs of homologous (not identical) chromosomes. When producing sex cells, the diploid organism will send along a haploid set of chromosomes, randomly choosing one chromosome from each of the pairs of homologous chromosomes. This is the process of **segregation**. Thus, in diploid organisms there is a shuffling, or random selection of chromosomes from one individual pairing up with a random selection of chromosomes from another individual to form the chromosome set of an offspring individual.
Segregation to a large part is responsible for keeping up population diversity in diploid organisms.
Additional genetic operators are **recombination**, **mutation**, and **inversion**.
When two homologous chromosomes pair up in meiosis (reduction division to create sex cells) they may both break at equivalent points, and exchange material. This is called **recombination** (or **'crossing over'**).
The smallest unit of DNA are the bases C, G, A, and T. The change of, say, an A to a T represents a (point) **mutation**. This is a random, lasting change to DNA (and consequently to the gene and to the chromosome).
If a single chromosome breaks, and the broken part turns on itself, and then reattaches to the same chromosome, this is known as **inversion**.
The real-world phenomena of evolution need to be transformed into an abstraction. The basic units to be put into an abstract form are the gene and the chromosome.
A chromosome is often represented as a binary string, an ordered list of 0's and 1's.
A gene is a sequence of a certain length of this list.
An individual is seen primarily as a 'container' for the chromosomes.
For each of the two individuals involved, a mating will consist of taking a random pick of a chromosome from each of its chromosome pairs, going through the set of chromosomes. A copy of a chromosome from one individual is joined with a copy of a chromosome from another individual, the process repeating along the chromosome sets of each of these parent individuals to make a new diploid chromosome set for an offspring

individual.

A generation could be limited by e.g. a pre-set number of matings. To keep a population stable, one would let the number of matings (i.e. the number of offspring) equal the number of individuals in the present generation.

This genetic algorithm differs from the mainstream variety in that it deals with diploid organisms.

In the CLOS (the Lisp programming language, the object oriented version, Common Lisp Object System) setting it seemed natural to let the individuals be represented as objects. The two chromosome sets are two slots in an individual. The chromosome sets are lists of lists. The 'inner' lists are the individual chromosomes.

A general genetic algorithm could be written as the following simple piece of pseudocode:

```
Set the generation counter to zero
Make a random population
Evaluate fitness
Test for termination criterion
     While not terminated do
       Increase the generation counter by one
       Perform recombination, mutation and inversion
       Select individuals to reproduce (based on fitness)
       Perform mating, build the new population
       Let the present population 'die',
         let the new population become the present population
       Evaluate fitness
     end of while-loop
end of genetic algorithm
```

The first step in the genetic algorithm run is the creating of a random population of individuals, each individual coding for a solution to the real-world problem by way of its chromosome string set-up. Some of these problem solutions will be inferior, quite a few will be mediocre, and a few will be superior.

'Individuals' coding for superior solutions will mate more often. Parts of their chromosome sets will appear in the subsequent generation.

Through several generations the best chromosomes/ chromosome parts will be retained, inferior ones will be lost. After many generations there will be, relatively speaking, few inferior individuals; there will be some mediocre individuals, and some superior and exceptional individuals.

A.7.2 The initial population

The first generation of the genetic algorithm is made by generating a pre-determined number of individuals. Each individual carries a set of chromosomes, also made by generating random numbers within the range decided. As the chromosomes code for a problem solution, this solution can be evaluated, and fitness apportioned.
Each individual with its chromosomes can be seen as a starting point for a search in solution space.

A.7.3 Fitness

Fitness in the genetic algorithm is in a way the reverse of natural genetic fitness. In nature, an individual or group of individuals have a higher fitness the more offspring they produce, no matter what the reason for the fecundity might be.
In the genetic algorithm fitness is assigned on the basis of how well an individual performs in the artificial world. Based on this, it is decided how many chances of begetting offspring the individual will have.

A.7.4 Mating

Based on fitness, individuals are allowed to mate. High-fitness individuals mate more often than low-fitness individuals. A simple way to arrange this is by the so-called **'fitness proportionate' or 'roulette wheel' selection:**
All fitnesses are summed up. Each individual is allotted a slice of the total fitness, in proportion to its individual fitness. A random number between zero and total fitness sum is generated. If that random number falls in the interval of the running sum of the individual fitnesses allotted to a specific individual, that individual is allowed to mate.Roulette wheel selection is thus a fitness proportional selection. A problem with the fitness proportional selections is that as many individuals in the population attains comparable fitnesses, there is little drive to cover the final distance up to the optimum solution.
In **tournament selection** some individuals are drawn from the population. A fitness-based 'tournament' is arranged between them, and the best individual is allowed to mate. This favours the best individuals, although they may have only marginally higher fitnesses than their contenders.

It is desirable to prevent in-breeding (sibling matings) (not to mention self-matings: An individual mating with itself). This could be done by keeping a list of each individual's parents (and, if deemed necessary, grandparents) and reject matings by individuals descending from the same parent/ ancestor.

On the other hand, it may be desirable to assure that exceptional individuals be kept in the population of the next generation. This can be achieved by **parthenogenesis** ('virgin birth'), where an individual is a direct replica of one single parent individual.

A.7.5 Recombination, mutation, and inversion

Two basic type pitfalls to the genetic algorithm are **'Greedy exploitation'** and **premature convergence**. Greedy exploitation, i.e. setting up the algorithm to go for the apparent solution too fast, may give a suboptimal solution (local maximum).

Premature convergence results when all individuals close in on the same solution early in the run. To prevent premature convergence, recombination, mutation, and inversion are used. This will disrupt some of the stable solutions building up in the population.

Over-using these genetic operators will result in chaos, with stable solutions never showing up.

Figure, recombination

Chromosome A	Chromosome B		Chromosome A	Chromosome B
1	0		0	1
1	0		0	1
1	0	⟶	0	1
1	0		1	0
1	0		1	0
1	0		1	0
1	0		1	0

Figure A.1: Recombination

Figure, inversion

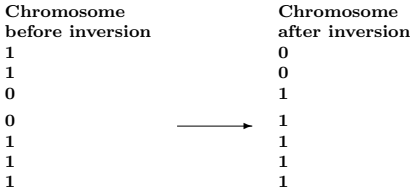


Figure A.2: Inversion

Figure, mutation

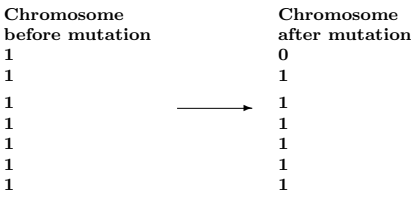


Figure A.3: Mutation

A.7.6 The population – the concept of generation

There are several ways of deciding what constitutes a population.
The simple method used for this genetic algorithm is known as the generational method.
This consists in making a new population with the same number of individuals each time. No mating is allowed across generations.

A.7.7 The fitness function

The fitness function is the genetic algorithm equivalent of natural genetics fitness. Natural genetics 'fitness', of course, is only a concept used to describe the reproductive ability of a certain group of individuals.

The genetic algorithm fitness function is more of a driving force. It transforms an ability of an artificial individual to solve a real world problem into a chance for mating. The high performance individual will be given a high probability of mating.

The fitness function of the genetic algorithm consists of the following steps:

1. Decode the individual's chromosome set, rendering it into a problem solution

2. Compare the decoded solution to real-world solution
   OR let the decoded version try to solve the real-world problem

3. Assign fitness based on the quality of the solution

Assigning fitness is usually done by some mathematical method. The method could be either 'fair' or biased. An example of a method that is not 'fair', is the so-called 'elitist' strategy. This will give extra credit to the best individuals of the population (in addition to the fitness fairly allotted to them for being the best). Such an elitist strategy may speed up finding the best solution, but may fall prey to the 'greedy exploitation' dangers by finding a local maximum in solution space.

Choosing a reasonable fitness function requires knowledge of the real-world problem to be solved. Obviously, one has to know the problem area to be able to tell a good solution from a bad one. Besides, it is important to be familiar with problem area background information e.g. to tell whether a certain fitness scaling would make sense.

Schema theorem

The **schema theorem** has been proposed as a theoretical explanation for the success of genetic algorithms. A schema is a part of a chromosome in a genetic algorithm. This could be a contiguous part, such as '1 1 0 0 0 1'. It could also be defined as a subset of this part, such as '1 1 * * * 1'. The schema would code for a part solution. This might help select for the schema itself, irrespective of the intermediary parts of the chromosome (the * * * in '1 1 * * * 1'). Schemas would be broken up by the genetic operators, such as mutation, inversion, recombination. It is possible to quantify this with probabilistic methods. The schema theorem has been the subject of long debates, and although the debate has centred on the simple genetic algorithm, doubt has been cast on the schema theorem in general.

A.8 Genetic programming

Genetic programming (www.gp-field-guide.org.uk) [142] is a area that has evolved to represent a large part of the whole field that is called 'evolutionary programming' , i.e. programming based on the principles of natural evolution.
In genetic programming the individual does not contain a set of chromosomes, but is a short actual program. The program is usually represented as a tree consisting of operators that can be read as program text. A basic example might be the program to find a solution to $5x^2$. This might be e.g. (2 + 3) * x * x. The tree would then have e.g. the multiplication operator * as its root, and a left branch operator +, with the leaves 2 and 3. The right branch operator would be *, with the leaves x and x. Such a program tree mutates by removing, shifting or implanting branches and twigs on the tree. The first left branch might mutate to (10 / 2), giving the same result.

A.9 Complex adaptive systems

The study of so-called complex adaptive systems [143] builds on the ideas of genetic algorithms.
Complex adaptive systems are formed by a number of autonomous elements working in concert to create a system that has qualities that would not be obvious from study of the individual elements.
Examples of such systems are cells growing to form an organism, ants in a colony, the stock market, a flock of birds, or the immune system.
Such systems show emergent behaviour or emergence, the behaviour that the whole system demonstrates and that transcends that of the individual element. An example is an ant hill, which is not planned by any individual ant, but that is still built by the ants together. A related term is swarm intelligence.
A prerequisite for this emergence is the self organization of the elementary units of the system.
Artificially created systems of this type are used similarly to the genetic algorithm in problem solving [127].

A.10 Fuzzy logic

A basic tenet of classical predicate logic is that a phenomenon cannot both be A and not A.

Fuzzy logic, on the contrary, holds that a 'truth value' can vary between 1 and 0. Though it does have similarities to probabilistic reasoning, fuzzy logic is more intuitive. Fuzzy logic seems well fit for many medical applications[144, 145], partly because medical systems have to confront the problem of uncertain knowledge.

A.11 Case based reasoning

Case based reasoning (CBR) is founded on the idea that if a similar problem has been solved before, one could re-use the solution for a new problem. There are numerous applications of case based reasoning [146, 68, 69].
This seemingly simple idea entails two difficulties: Deciding to what extent two cases are similar, and having a large enough case base.
To resolve the similarity issue, one could use for instance the basic nearest neighbour methods advocated elsewhere in this thesis.
As far as the case base is concerned, there is a need to cover all possible cases in the 'domain area'. Thus, it is necessary to include even rare cases. However, this must not lead to a situation where the rarities appear to be general cases.
Again, this is analogous to the line of thought concerning prior probabilities and predictive diagnostic value of clinical signs.

A.12 Data mining

Data mining is long established in artificial intelligence. Data mining algorithms extract information from data, either in databases, structured text, or from raw text. There seems to be an upsurge for data mining in bioinformatic research [135, 136, 147]. One data mining application is text mining [138, 140, 139, 148], both classification and cross-referencing of research articles and databases, and attempts at natural language processing, i.e. attempting to get directly to the meaning of the written article.

A.13 The semantic web

The semantic web [149, 150, 151] is the term used for the internet or the World Wide Web, when seen as a medium for search and interchange of data, information, and knowledge. To facilitate this, definitions and ontologies for the content are necessary. In bioinformatics the GO or gene

ontology (www.geneontology.org) is an important example, with its comprehensive system of terms, definitions and relationships between the entities. In medicine in general, UMLS (the Unified Medical Language System) (www.nlm.nih.gov/research/umls) serves a similar purpose with its set of vocabularies and thesauri. There are an increasing number of XML-derived projects. XML - eXtensible Markup Language - by describing the content of a document, for example a web page, is essential in the transformation of the web for research purposes. A number of the XML projects that exist are within biomedicine, such as Systems Biology Markup Language, SMBL (www.smbl.org), Cell Markup Language, CellML (www.CellML.org) and MicroArray Gene Expression Markup Language, MAGE-ML (www.mged.org/Workgroups/MAGE/mage-ml.html).

A.14 The 'closed universe'

Many real world problems are open ended, there are no confines within which the solution to a problem has to exist.
As opposed to this, many artificial intelligence methods operate within a 'closed universe'. First, a set of possible solutions are defined, then the search for the optimal solution is launched.
Many expert systems assume a so-called 'closed universe', a static world where things do not change over time. This is not true of syndromology. On the contrary, new syndromes are 'discovered' all the time. This obviously has consequences for a diagnostic system. Either it has to be dynamic and incorporate new information, or its performance will deteriorate.

A.15 A summary history of artificial intelligence in medicine

Artificial intelligence seems to have been steadily increasing in medical research since 1990, as shown by the increasing number of publications, figure A.4 and figure A.5.
One remarkable trait when one considers the development over the years of artificial intelligence in medicine, is to what degree the same subjects are represented. That does not mean there is no development. But still there is a tradition where the early methods do not die out, and where the beginnings of the new methods seem to have been there for a long time.
More poetically put [152]:

> computational methods are edging into higher-level interpretation of clinical data and into diagnosis. We may have thought this would happen with a sensational breakthrough. Instead, it seems to be happening slowly, everywhere, all the time, like the tide rising.

In artificial intelligence in general there were the 'AI winters', starting around 1974 and 1987 respectively. After periods of enthusiasm there were periods of drought.

That does not dampen the impression of a steady rise in artificial intelligence in medicine, as illustrated in the figures of Medline publications.

The artificial intelligence methods most often applied are artificial neural networks [131, 115, 112, 111, 153, 117, 114] genetic algorithms [95, 154, 124, 126, 125, 129] and hybrid systems with these two methods [155, 132]. Almost all artificial intelligence techniques are being used in medicine, noteworthy examples are case based reasoning (CBR) [68, 146] and fuzzy systems [144, 145].

The most general trend is the change from the knowledge intensive applications, often standalone diagnostic systems such as expert systems installed on a PC, to the data driven, data intensive systems of today [156]. The current systems and computers are networked, involved in data collection, information extraction and interpretation.

Machine learning, and text mining, apparently are gaining momentum in medical artificial intelligence [147, 136, 135]. Both data mining [137] and text mining [140, 138] are growing because of the need to handle the ever expanding body of text and information both in PubMed itself, and in databases and data repositories reachable through the internet. Databases hold bioinformatics information, clinical information (though usually with restricted access), and text, often in the form of medical publications and books.

With the possibilities of collecting data from accessible data sources, there was a need to either collect information in a standardised form, or to understand text that was not standardised. The existence of different classifications, different terms used for the same entities, and different diagnoses for the same conditions, hampered progress.

Standardization, and later the development of ontologies [157] remedied the situation. Ontologies then prepared the ground for the most recent development, the networks [158, 159]. Networks are gaining ground especially in bioinformatics, but the nets are encompassing also the clinical areas of medicine.

The other data collecting force is text mining, and natural language processing (NLP).

Data mining, text mining and network construction, are the triad of present day artificial intelligence in medicine.

The programs of the last two of the biannual artificial intelligence in medicine conferences, AIMED'07 and AIMED'09 [160, 161], show the most prevalent subjects are machine learning, guidelines and work flow systems, decision support systems, natural language processing, agent based systems and ontologies.

Machine learning can be mediated by so called intelligent agents [162], programs or 'software entities' that can react to what it finds or senses in the environment. On the internet, machine learning can be through specialised web crawlers, spiders or robots; small pieces of program code that are sent out on the net to retrieve specific pieces of information.

Networks - in the sense of interconnected computers - and networks - in the sense of mash-ups of internet information collected through data mining - are in the forefront. But traditional artificial intelligence methods are still being used extensively, and are important in medicine.

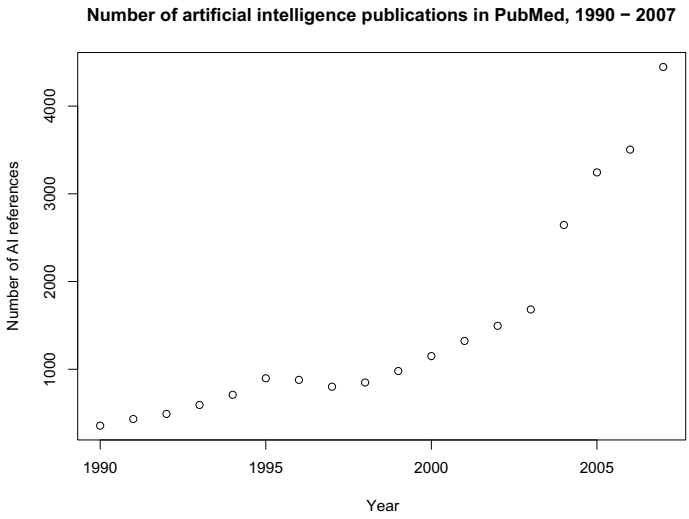


Figure A.4: A plot showing the number of artificial intelligence references in PubMed 1990-2007. (Figure made with the "R" statistical package.)

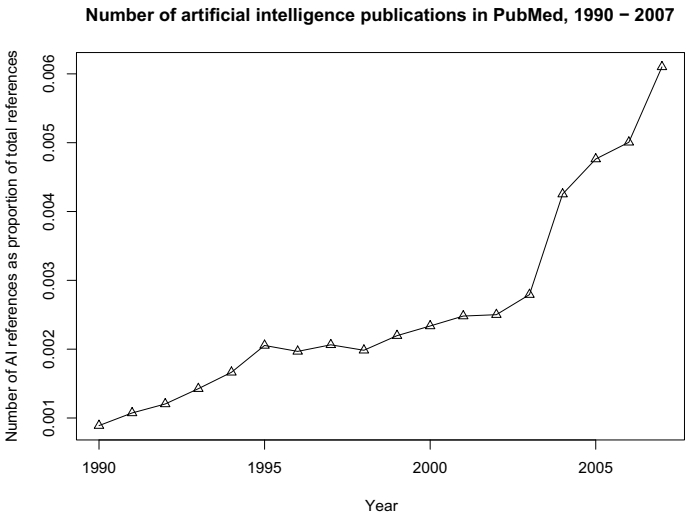


Figure A.5: A plot showing the number of artificial intelligence references in PubMed 1990-2007, as proportion of total number of references.

Appendix B

Elements of clinical epidemiology

B.1 Introduction

This appendix introduces some concepts of clinical epidemiology, with a focus on the predictive value of clinical signs, based on Bayes' theorem. There is also a brief discussion of quality of data in syndromology. The discussion relates to syndromology, although the principles are generally applicable to medicine.
A major point made here, is that the theory developed for biochemical tests, is also valid for for clinical 'tests', i.e. the finding or eliciting of clinical signs, and the clinical history or anamnesis.

B.1.1 Concepts and indices of clinical epidemiology

B.1.1.1 Sensitivity

The **sensitivity** is the probability that a diseased person has a positive test result. It can also be seen as the proportion of diseased people who have a positive test.

B.1.1.2 Specificity

The **specificity** is the probability that a non-diseased person has a negative test result. (Non-diseased would be not having the disease in question, either being disease-free, or having another disease).
The specificity is also the proportion of non-diseased people with a negative test.

The specificity and sensitivity are useful when evaluating a test. These clinical indices or test indices, however, do not predict the probability of the person tested being ill.

Predictive value To arrive at the predictive value, the number of affected persons with a positive test (e.g. the clinical sign is present) is divided by the total number of persons with a positive test.

The predictive value is heavily influenced by the prior probability of disease.

The prior probability of disease is the probability that the patient has the disease, before the test is done. In a screening situation, or in general practice, this probability may be low. At a local hospital the prior probability is higher, because some dubious cases have been filtered out and are not admitted to hospitals. At a third or fourth line specialist referral centre the prior probability is typically high, since several lines of doctors have seen the patients, and only those likely to have the disease remain.

Likelihood ratio

Likelihood ratios are used to calculate the risk that a patient has the disease, after a test has been performed. The likelihood ratio itself is not dependent on the prior probability of disease.

A desirable quality of the likelihood ratio is that it can take the post-test probability (the posterior probability) of one test as its pre-test (prior) probability for a second test. In this way several test results and likelihood ratios can be multiplied to give the end result of a set of tests.

A technical inconvenience with the likelihood ratio is that pre-test probabilities have to be converted to odds, and the post-test odds that result from a chain of probability and likelihood ratio multiplications, have to be converted back to probabilities.

Example of the effect of prior probability

In the following example, the prior probability is set at 40% , and it is assumed that 1000 people are 'tested', i.e. examined for a clinical sign.

	Syndrome present	Syndrome not present	
Positive test			
Negative test			
	400	600	1000

If the sensitivity is 0.9 and the specificity is 0.95, the figures in the table below follow:

	Syndrome present	Syndrome not present	
Positive test	360	30	
Negative test	40	570	
	400	600	1000

	Syndrome present	Syndrome not present	
Positive test	360	30	390
Negative test	40	570	
	400	600	1000

To arrive at the predictive value, the figure representing patients who have the clinical sign, and who are affected, is divided by the figure representing all those with the clinical sign, affecteds and unaffecteds alike. In this example, the predictive value becomes 360/ 390 = 0.92. Thus, in this example, if a person has a positive test (clinical sign present) there is a 92 per cent or 0.92 that he has the disease.
To re-examine this example using likelihood ratios, the prior probability is converted to a prior odds. The prior odds are 0.4/ 0.6 = 0.67. (The prior probability for being affected is 0.40, the prior probability for not being affected is 1 - 0.40 = 0.60, and the prior odds are 0.4/ 0.6 = 0.67).
The prior odds are multiplied by the likelihood ratio. The likelihood ratio is sensitivity/ (1 - specificity), or 0.9 / 0.05 = 18. So, to get the posterior odds, the prior odds are multiplied by the likelihood ratio, 0.67 * 18 = 12. Finally, to get the posterior probability the formula odds / 1 + odds is used: 12/ 1 + 12 = 0.92.

B.1.2 False and true, two by two table

Two by two table showing TP, FN, FP, TN

	Affected			Not affected
Positive test	True positives	TP	**a**	**b** FP False positives
Negative test	False negatives	FN	**c**	**d** TN True negatives

'Signal to noise ratio'
Another way to consider the false negatives/ false positives, is the signal to noise ratio. This index contrasts true positives (the signal) with false positives (the 'noise'). As proportions this is the sensitivity or true positive rate divided by 1 minus specificity (the false positive rate).

The receiver operating curve generalises this approach by plotting possible values of false positive rates against true positive rates.

B.2 Receiver operating curve

The concept of **ROC curves** or receiver operating curve is taken from engineering. With a radio transmitter, there is a signal that the receiver wants, and noise that obscures the signal.
To receive the signal in the clearest way, one strives to maximize the signal to noise ratio. The 'signal' in medical diagnosis is the true positives. The noise is the false positives. A receiver operating curve can be constructed by plotting the true positives versus the false positives. An ROC curve will tell where the mathematically best cut-off point is, if the goal is to maximize the number of true positives, while minimizing the number of false positives. In medical diagnosis the consequences of false negative and false positive diagnoses also have to be taken into account.
An example might be suspected child abuse. A false positive diagnosis of child abuse implies falsely accusing parents of mistreating their own child. A false negative diagnosis means the child abuse is not detected, and that abuse of the child might continue.

B.3 Syndrome-diagnostic approaches

B.3.1 Consistency

Two further aspects of clinical signs are:

- It should be easy to recognize
- Clinicians should agree as to whether the sign is present or not (consistency, inter-observer and intra-observer variation)

B.3.1.1 Aspects of consistency

Accuracy For high accuracy one often uses the metaphor of being centred on the target. Though individual hits may be off mark, as a group the hits centre around the bull's eye.
Precision Similarly, precision means that hits closely centre on a point, though possibly off target.

Inter-observer variation Clinicians - or, more generally, observers - do not necessarily agree. One clinician may consider a sign present, another may say it is not.
If this type of disagreement between observers is substantial, the clinical sign loses its value. It is conceivable that signs that are measured (as opposed to being classified as present/ not present) would be more reliable.
Intra-observer variation
Intra-observer variation would give the same kind of problems as inter-observer variation. In intra-observer variation, the observer disagrees with what she or he found on a previous occasion. (It is assumed that the clinical sign is not a sign that changes).
Kappa
Obviously, different situations may give rise to different degrees of disagreement between observers. This can be quantified by the measure Kappa. If two observers both find a clinical sign in most of the patients, there is less scope for disagreement between them.
Kappa is an attempt at quantifying the degree of disagreement beyond that expected by chance.
Kappa is the proportion of agreement observed minus the proportion of agreement expected by chance, divided by one minus the proportion of agreement expected by chance alone. Complete agreement results in Kappa = 1. For agreement beyond that expected by chance, Kappa is greater than zero.
The coincidental sign
Some clinical signs are common. This means such a sign may be present by coincidence in a child with a syndrome where this signs supposedly is not present. In this way, coincidental signs may 'pollute' a clear clinical picture.
Negative signs
Both in the clinic and in machine diagnosis, it would be possible to use negative signs.
'Negative signs' could be interpreted as the negative predictive value of signs that are present.
It could also mean the predictive effect of signs that are not present.
More than one sign
In practice, a clinician will rely on more clinical signs.
If more than one sign has to be present for a diagnosis to be made, fewer of the affected will fill the criteria. Thus, sensitivity falls. Conversely, the number of false negatives will increase.
On the other hand, the false positive rate will fall as well, and since the true negative rate is 1 - false positive rate; the true negative rate (and specificity) will go up.

Repeated testing

When many tests are performed, the probability that one or more test will be falsely positive, increases. A false positive rate of 0.05 or five per cent, means the risk it will falsely show that an unaffected patient is diseased, is five per cent. For more tests, it is less intricate to calculate the probability of not having a false positive. If there is a five per cent risk of a false positive result, there is a ninety five per cent probability of not having a false positive test.

The probability that two tests will not give a false positive result is 0.95*0.95.

With, for example, thirteen tests, the probability that no false positive tests occur, is 0.95*0.95*0.95..., or 0.95 to the power of thirteen. With thirteen tests there is a risk that one or more of the tests is a false positive, is given by 1 - 0.51 = 0.49.

Cut off points

The cut off point is a value that divides between affecteds and not affecteds. This concept fits best with continuous measurements, but in principle applies to all clinical situations.

If one sets the cut off point to minimize e.g. the number of false positives, this will give an increased number of false negatives, other things being equal.

Receiver operating curves can aid in choosing the best cut off points. Often these considerations on a mathematical basis have to be adjusted by which consequences the false positives and the false negatives will have. It may not be a catastrophe to falsely diagnose someone with allergy, but falsely diagnosing cancer may be.

Prior probabilities in rare diseases

In a differential diagnostic situation with three very rare diseases, the prior probability for each may be 0.33.

If, however, there might be another possible diagnosis, this might severely change the prior probabilities. Since the three diagnoses were very rare, the other contender diagnosis might attain a prior probability which were much higher.

Non-mathematically, this should bring to mind that in a diagnostic situation with very rare diseases one should also consider a rare phenotype of a more common condition.

B.3.1.2 Bayes' formula

Bayes' formula deals with conditional probabilities. In medicine, it is often used in a form that considers two possibilities, affected and not affected.

The Bayes' formula reads:

$$P(syndrome|sign) = \frac{P(sign|syndrome) \times P(syndrome)}{P(sign)}$$

P = probability, '|' = given that, on the condition that,
thus the whole formula reads: The probability that the child does have the syndrome if he has this particular clinical sign, is the probability of having this sign if one has the syndrome, multiplied by the probability of having the syndrome, divided by the probability (in the general population of children) of having the sign.

B.4 Using a diagnostic test in a different setting

A diagnostic system developed in a specific setting, need not perform well when applied in a different setting.
If the system is developed at a tertiary centre - with high prior probabilities - it may fail in a setting with low prior probabilities and a different spectrum of diseases.
An illustration of the effect of probability can be given by considering two situations, one with a medium and one with a high prior probability, in both cases using a clinical sign with a sensitivity of 0.95 and a specificity of 0.99.

- Prior probability **0.50**
- Prior odds 0.50/ (1 - 0.5) = 1
- Likelihood ratio 0.95/ (1 - 0.99) = 95
- Posterior odds 1 * 95 = 95
- Posterior probability 95/ 1 + 95 = **0.98**96

Thus there has been a huge increase in the probability for the disease, doing the test has given a lot of information.
Then, one might use the same sign in a diagnostic situation where the prior probability of disease is already very high, like in a tertiary care centre.

- Prior probability **0.95**
- Prior odds 0.95/ (1 - 0.95) = 19
- Likelihood ratio 0.95/ (1 - 0.99) = 95
- Posterior odds 19 * 95 = 1805
- Posterior probability 1805/ 1 + 1805 = **0.99**94

At the tertiary care centre, the probability has been moved a mere four - five per cent. The decisions to treat may not have been changed by the test. The reason to test in this situation would be if an extreme degree of certainty was necessary, e.g. if the treatment were potentially very harmful.
Problems with using the Bayes' approach in syndromology are that the prior probabilities of syndromes may be just estimates, that the frequency of a sign in a certain syndrome or in the general population may not be known with certainty, and that signs are not independent.
An extension of a Bayes' formula approach, is a Bayesian network. In a Bayesian network, the nodes are interconnected and if one node is updated, this spreads throughout the network.
Training data for methods such as neural nets and case based reasoning can have cases where the numbers sum up to correct prior probabilities. Such a system would still have to be retrained to handle a new environment.

B.5 Quality of data

B.5.1 Ascertainment bias

Ascertainment is a problem in medical research in general. It is often a problem in medical genetics.
The problem with ascertainment is how to handle the affecteds who brought a family to the attention of the health care system or the researcher, and avoid an exaggerated number of affecteds.
Additionally, the chance of finding a family with more affected children is greater than the chance of finding a family with fewer affecteds.
Ascertainment bias may lead to risk figures and prevalence figures that are incorrect, and may distort the input that is to be used in designing and training an artificial intelligence system.

B.5.2 Other methodological issues

Selection bias
It is less complicated to search for patients at for example a clinic or an outpatient clinic, than in the general population. If looking for patients with cystic fibrosis at a lung clinic or a hepatology department, one may find a presumably falsely elevated figure for the frequency of severe lung or liver problems. This is a problem in purely clinical research. It is also a problem for artificial intelligence systems that build on such figures.

B.5.3 Changing definitions of disease entities

Diagnoses may be considered fixed. In fact there is often in practice a change of diagnostic criteria, e.g. when a new test is introduced. This may give rise to problems of classification.
The general case
In general, if A represents the patients classified by the old definition, B represents the patients by the new definition, and C represents the patients where the old and the new definitions agree, the following holds:

Figure B.1: Venn diagram: 'Old disease' A, 'new disease' B, agreement between 'old' and 'new' disease definition C (A $\cap$ B)

- **A**: Had the disease by the old definition, does not have it now ('Old disease')
- **B**: Did not have the disease by the old definition, does have it now ('New disease')
- **C**: Agreement between old and new disease definition

If A and B are both small, practical problems may be insignificant. If A or B is large, this may cause problems. A diagnosis is a tag, a carrier of information. If the disease is no longer the same, the information will be wrong.

B.5.4 Clinical signs used in diagnosis

An artificial intelligence system will need information on what clinical signs to use in diagnosis.
This information may come from various sources.

The prospective investigation

Ideally, one would want to do a prospective investigation, examining all newborns in an area over a certain period. This would give figures both for the frequency (birth prevalence) of the different syndromes, and for the clinical signs associated with these syndromes. However, such an investigation would probably prove prohibitive both in terms of time and labour.
Even for common syndromes with a birth prevalence of one in ten thousand, an investigation would have to last for years in most hospitals to have numbers large enough to give statistical significance.

The one-syndrome review

The one-syndrome review is undertaken by investigators who searches a large uptake area for cases of a single syndrome.
A potential problem with the one-syndrome review is the circular argument that follows from the lack of an external validation. The investigator may decide on a set of clinical signs as inclusion criteria for a syndrome research project. The findings are recorded, and the conclusion is that in this syndrome, these very clinical signs frequently occur.
For siblings with genetic syndromes, it would seem highly reasonable to assign all siblings to the same diagnostic group, even if one or two of the siblings lacked one clinical sign. This may lead to a bias, though, when comparing with a sporadic or non-genetic condition.
The one-syndrome investigation does not say anything about the frequency of a particular clinical sign in the general population of newborns, or in groups of patients with syndromes that might be considered in the differential diagnosis.
One-syndrome investigations are likely to mix syndrome patients of different age groups. This may be a problem with clinical signs hat change with time.
One-syndrome investigations in addition are prone to selection bias.

Appendix C

Philosophical background

C.1 Introduction

This appendix briefly presents a few of the philosophical issues that underlie syndrome diagnosis, and to some extent, medical diagnosis in general.
Relative to scientific fields such as physics, artificial intelligence is a very young science.
Because of its lack of an established philosophical foundation, philosophical issues that more established disciplines have left behind them, often emerge to the surface. The new field also faces and has to come to grips with, philosophical questions that older disciplines such as medicine take for granted, and does not consider topics for discussion. An example of this is the lack of a criterion standard of diagnosis. In everyday clinical work, this may seem unproblematic. As discussed in earlier sections, however, it does have consequences that are far reaching, and eventually has consequences for medicine's standing as a scientific discipline.
Clinical syndrome diagnosis is based upon the assumption that signs found by physical examination of a newborn could be used to assign the newborn to the appropriate diagnostic group.
Diagnoses are labels, and carriers of links to information about prognosis and treatment.
In modern medicine, many consider the distinction between disease, illness and sickness reasonable. The disease is seen as a natural phenomenon. The illness is what the disease does to a person, the clinical manifestations. The sickness is how the sick person experiences his illness.
In this distinction, the disease is a 'universal', which has a set of qualities that is always seen. This concept is based upon the 'ideas' or 'forms' described by Plato.

C.2 Nominalism versus essentialism

Essentialism and an essentialist view of disease, is congruent with this line of thinking. Essentialism holds that there is an 'essence', a part of Nature, that researchers can discover. 'Truth' can be disclosed more or less perfectly, but that does not change truth, which exists irrespective of what we do or believe. Diagnosticians can diagnose a disease correctly, and in that case they have found the disease or diagnosis [163, 164]
Nominalism In nominalism, diseases do not exist as separate entities. Disease diagnoses are labels that are attached to people who exhibit a set of clinical signs [165, 166], to facilitate communication about diagnoses, and the managing of sick people.

C.3 Categories

From the roots of Aristotelian philosophy, categories have been important in Western philosophy. In Nature no two things are identical, not even two grains of sand. Categorization is an attempt at bringing order to chaos.
A syndromes is recognised as a pattern by a clinician, and is categorized by being given a name. However, it is not obvious that the syndrome is something that 'exists in Nature'. Later others may think this may be nearly the same as another syndrome, that the new syndrome should actually be split into two sub-syndromes, or that there are intermediate forms between this and another syndrome.
In the words of philosopher-novelist Robert Pirsig:

> ... there is a knife moving here.
>
> A very deadly one; an intellectual scalpel so swift and so sharp you sometimes don't see it moving. You get the illusion that all those parts are just there and are being named as they exist. But they can be named quite differently and organized quite differently depending on how the knife moves. *Robert M Pirsig* [167]

Several artificial intelligence methods do not rely on predefined categories, e.g. neural nets, identification tree techniques [168]. Likewise, some mathematical methods like cluster analysis do not need predefined categories.

C.4 Tacit knowledge

As discussed on page 111, the process of knowledge acquisition tries to extract information through interviews with an expert, for example a syndromlogist. What will often happen is that the expert does not render expert advice. She may even fall back on what she was taught herself as a novice.
Polanyi, in his book 'The tacit dimension' [169], suggests this is because the expert does not know what she knows, she is not aware of what her expert knowledge consists in. Such 'tacit knowledge', then, according to Polanyi, is knowledge that an expert masters without being conscious of it.

Appendix D

Code: Lisp code for the genetic algorithm

I had intended to include all the code for the genetic algorithm Lisp program.
However, the program's 5000 lines of code would have added an extra 100 - 150 pages to this thesis, even with smaller than normal font. It was therefore decided not to include the program code.
The program consists of modules in different files. The different program parts are two small files for loading and compiling the other files, and the program proper. The modules of the genetic algorithm program are the following: There is a population module for building the random first generation, and for taking care of mating. An 'individual module' creates and maintains individuals of the genetic algorithm, and applies mutation, inversion and recombination. The DNA module handles the input DNA sequences and builds the hash table of short (n-mer) DNA sequences. A module called 'area' keeps track of the areas that are being formed, and that consist of he joined n-mer sequences from the genetic algorithm individuals. The pack and path modules handle their respective parts of the program. The rawdna module is used as a container for the input sequences to the program. There is a utilities module for smaller utility functions used in other parts of the program. What is included below, is the first few lines of the program.

```
;=============================================================
;===               GENETIC ALGORITHM PROGRAM                 ===
;===                                                         ===
;===                         AUTHOR                          ===
;===                      OIVIND BRAATEN                     ===
;===               ULLEVAL UNIVERSITY HOSPITAL               ===
;===                          OSLO                           ===
;===                         NORWAY                          ===
;=== =========================================================


;;; ============================================================
;;;         'GLOBAL' VARIABLES, PROCEDURE MAIN               ;;;
;;; ============================================================


(defvar *number-of-generations* 0)
(defvar *halfway-through-generations* 0)
(defvar *verbose* nil)
(setf *verbose* nil)
(defvar *no-of-runs*)
(defvar present-no-of-runs)


(setf *gc-verbose* '())


(defun initialize ()


(load "load-files.lsp")

(activate-dna-module)
```

```
(activate-individual-module)

(activate-population-module)

(activate-pack-module)

(activate-utilities-module)

(activate-paths-module)



(setf *number-of-generations* 16)

(setf *halfway-through-generations*
       (floor (/ *number-of-generations*  2)))
                                              )



(defun main ()

(print-time)

(print-settables)


(dotimes (outer-counter *number-of-generations*)

      (generational-report outer-counter)
      (fitness-calculation *population*)
     (if (>= outer-counter *halfway-through-generations*)
      (progn
      (form-packs *population*)
      (expand-packs *population*)
      (calculate-fitness-all-packs *population*)
      (check-merge-all-packs *population*)
      (go-through-packs-update-paths *population*)
                                                    )) ; progn & if
```

```
      (print-fitnesses-file present-no-of-runs)

      (mutation)
      (recombination)
      (inversion)
      (mating *population*)
      (setf *population* *new-population*)
      (setf *new-population* '() )

                                                          ) ; dotimes

;     (compute-fitness-and-print-population *population*)
      (print-paths)
      (print-sequences)
      (print-sequences-file)

      (print-time)
                                                           ) ; main



(defun generational-report (gen-ctr)

   (format t "~& Entering generation number ~a , run number ~a ~3% "
                                    (+ gen-ctr 1) present-no-of-runs)
                                                                    )









;;; ============================================================ ;;;
;;; =================     MAIN RUN       ======================= ;;;
;;; ============================================================ ;;;


(setf *no-of-runs* 1)
(setf present-no-of-runs 1)
```

```
(dotimes (cnt *no-of-runs*)




(format t ";;; ================================================ ;;;")
(format t "~& Run no ~a  ~%" (+ cnt 1))
(format t ";;; ================================================ ;;;")


(initialize)
(main)



(setf present-no-of-runs (+ present-no-of-runs 1))

                                                    )




(print-seq-counts-multiple-runs)
(end-of-run-text)




;;; ============================================================ ;;;
;;; ==============     END OF MAIN RUN       ==================== ;;;
;;; ============================================================ ;;;



⋮
```

```
(format t "~& ===================================== ~%")
(format t "~& LOADING FILES OF THE GENETIC ALGORITHM ~%" )
(format t "~& ===================================== ~3%")

    (load "rawdna")
    (load "utilities")
    (load "dna")
    (load "ind")
    (load "pop")
    (load "area")
    (load "pack")
    (load "path")
```

Appendix E

Sample output from the genetic algorithm

E.1 Output showing the first five individuals of a population

```
===== THE POPULATION =====


Population id: 0

Individual-ID: 0
Parents: (1222 468)
Generation number: 5
Chromosome set 1: ((0 1 0 0 0 1 1 0 1 0 1 0 1 0 1 1 1)
                   (1 1 1 0 0 0 0 1 1 1 0 0 1 0 0 0 0))
Chromosome set 2: ((0 0 1 0 1 1 1 0 1 0 0 0 1 0 1 0 1)
                   (1 0 0 1 1 0 0 0 1 1 1 1 0 0 1 0 1))
Sequence position: 656
Sequence identifier: HGF
All sequence positions:
  HGF : (656 2423)
  PLA : (376 496 1012 1291 1318)
```

```
  LPA : (240 267 609 951 1293 1635 1977 2319 2661 3003 3345 3687 4029 4371
         4713 5055 5397 5739 6081 6423 6765 7107 7449 7791 8133 8475 8817 9159
         9501 9843 10185 10527 10842 10869 11211 11529 11871 12186 12213 12528
         12555)
Short DNA strings: ((A A C T A C))
Fitness: 143

Individual-ID: 1
Parents: (2 1054)
Generation number: 5
Chromosome set 1: ((1 0 1 0 1 0 1 0 1 1 1 1 1 1 0 1 0)
                   (0 0 1 0 0 0 0 0 1 0 0 1 1 0 0 1 1))
Chromosome set 2: ((1 0 1 0 0 0 1 1 1 0 1 0 0 1 1 0 0)
                   (1 1 0 0 1 0 1 1 0 0 1 1 1 0 0 0 1))
Sequence position: 9774
Sequence identifier: LPA
All sequence positions:
  HGF : (3323)
  PLA : (1249)
  LPA : (198 540 882 1224 1566 1908 2250 2592 2934 3276 3618 3960 4302 4644
         4986 5328 5670 6012 6354 6696 7038 7380 7722 8064 8406 8748 9090 9432
         9774 10116 10458 10800 12144 12486 13360)
Short DNA strings: ((T G G T C A))
Fitness: 110

Individual-ID: 2
Parents: (815 1347)
Generation number: 5
Chromosome set 1: ((0 1 1 1 1 1 0 1 0 0 0 0 1 0 1 1 0)
                   (1 0 1 0 0 0 1 1 1 0 0 0 1 1 1 0 1))
Chromosome set 2: ((0 1 1 0 1 0 1 1 1 0 0 0 0 0 0 1 0)
                   (0 1 0 1 1 1 0 0 1 0 0 1 0 1 0 0 0))
Sequence position: 1311
Sequence identifier: HGF
All sequence positions:
  HGF : (1311 5243)
  PLA : (1077 1383 1748)
  LPA : (332 674 1016 1358 1700 2042 2384 2726 3068 3410 3752 4094 4436 4778
         5120 5462 5804 6146 6488 6830 7172 7514 7856 8198 8540 8882 9224 9566
         9908 10250 10592 10934 11936 12278 12620 12985)
Short DNA strings: ((G T G G G A))
Fitness: 122
```

```
Individual-ID: 3
Parents: (142 1313)
Generation number: 5
Chromosome set 1: ((0 1 0 1 1 0 1 0 0 1 1 0 1 1 1 1 1)
                   (1 1 0 1 1 1 0 1 1 1 1 1 0 1 0 1 0))
Chromosome set 2: ((0 1 0 0 1 0 1 0 1 0 1 0 0 1 0 0 1)
                   (1 1 0 0 1 1 0 0 0 0 0 0 0 1 1 0 0))
Sequence position: 12158
Sequence identifier: LPA
All sequence positions:
  HGF : (601 850 1550 2308 2848)
  PLA : (1225 1263 1363)
  LPA : (212 231 554 896 1238 1580 1922 2264 2606 2948 3290 3632 3974 4316
         4658 5000 5342 5684 6026 6368 6710 7052 7394 7736 8078 8420 8762 9104
         9446 9788 10130 10472 10814 11156 11474 11493 11816 12158 12500)
Short DNA strings: ((A C C A C A))
Fitness: 140

Individual-ID: 4
Parents: (882 1373)
Generation number: 5
Chromosome set 1: ((0 0 1 0 1 0 1 1 0 1 0 1 1 1 1 0 1)
                   (0 1 1 1 1 0 1 1 1 0 1 0 1 0 0 1 1))
Chromosome set 2: ((0 0 0 0 0 0 0 1 1 1 0 1 1 0 1 1 1)
                   (1 0 1 1 1 0 0 0 0 0 0 1 0 1 0 1 0))
Sequence position: 1261
Sequence identifier: PLA
All sequence positions:
  HGF : NIL
  PLA : (794 1282 1370 1597)
  LPA : (163 505 573 847 915 1189 1257 1531 1599 1873 1941 2215 2283 2557 2625
         2899 2967 3241 3309 3583 3651 3925 3993 4267 4335 4609 4677 4951 5019
         5293 5361 5635 5703 5977 6045 6319 6387 6661 6729 7003 7071 7345 7413
         7687 7755 8029 8097 8371 8439 8713 8781 9055 9123 9397 9465 9739 9807
         10081 10149 10491 10765 10833 12177 12519 12607)
Short DNA strings: ((A C T C G A) (A C T C C A) (A C C C G A) (A C C C C A))
Fitness: 134
```

E.2 Output showing part of a 'path'

```
----- ----- ----- ----- -----
  THE SEQUENCE: PLA
 ----- ----- ----- ----- -----


 Sequence: PLA Base: G     Base  no   1166:  315 *********************************
 Sequence: PLA Base: G     Base  no   1168:  316 *********************************
 Sequence: PLA Base: G     Base  no   1169:  320 *********************************
 Sequence: PLA Base: T     Base  no   1170:  324 *********************************
 Sequence: PLA Base: C     Base  no   1171:  325 *********************************
 Sequence: PLA Base: C     Base  no   1172:  327 **********************************
 Sequence: PLA Base: A     Base  no   1173:  328 **********************************
 Sequence: PLA Base: G     Base  no   1174:  328 **********************************
 Sequence: PLA Base: G     Base  no   1175:  329 **********************************
 Sequence: PLA Base: A     Base  no   1176:  332 **********************************
 Sequence: PLA Base: C     Base  no   1177:  333 **********************************
 Sequence: PLA Base: T     Base  no   1178:  335 ***********************************
 Sequence: PLA Base: G     Base  no   1179:  337 ***********************************
 Sequence: PLA Base: C     Base  no   1180:  338 ***********************************
 Sequence: PLA Base: T     Base  no   1181:  339 ***********************************
 Sequence: PLA Base: A     Base  no   1182:  339 ***********************************
 Sequence: PLA Base: C     Base  no   1183:  340 ***********************************
 Sequence: PLA Base: C     Base  no   1184:  341 ***********************************
 Sequence: PLA Base: A     Base  no   1185:  341 ***********************************
 Sequence: PLA Base: T     Base  no   1186:  341 ***********************************
 Sequence: PLA Base: G     Base  no   1187:  342 ***********************************
 Sequence: PLA Base: G     Base  no   1188:  342 ***********************************
 Sequence: PLA Base: T     Base  no   1189:  342 ***********************************
 Sequence: PLA Base: G     Base  no   1190:  343 ***********************************
 Sequence: PLA Base: A     Base  no   1191:  342 ***********************************
 Sequence: PLA Base: T     Base  no   1192:  344 ***********************************
 Sequence: PLA Base: G     Base  no   1193:  344 ***********************************
 Sequence: PLA Base: G     Base  no   1194:  344 ***********************************
 Sequence: PLA Base: A     Base  no   1195:  344 ***********************************
 Sequence: PLA Base: C     Base  no   1196:  344 ***********************************
 Sequence: PLA Base: A     Base  no   1197:  344 ***********************************
 Sequence: PLA Base: G     Base  no   1198:  344 ***********************************
 Sequence: PLA Base: A     Base  no   1199:  344 ***********************************
 Sequence: PLA Base: G     Base  no   1200:  343 ***********************************
 Sequence: PLA Base: C     Base  no   1201:  343 ***********************************
```

```
Sequence: PLA Base: T     Base  no   1202:  343 **********************************
Sequence: PLA Base: A     Base  no   1203:  344 **********************************
Sequence: PLA Base: C     Base  no   1204:  344 **********************************
Sequence: PLA Base: C     Base  no   1205:  345 **********************************
Sequence: PLA Base: G     Base  no   1206:  345 **********************************
Sequence: PLA Base: A     Base  no   1207:  345 **********************************
Sequence: PLA Base: G     Base  no   1208:  345 **********************************
```

E.3 Output showing the n-mer hash table

The output shows the n-mer (here 8 bases), the sequence names, and the start point of that n-mer (that 8 base sequence) in that sequence.

```
(A C G C A A A A) = ((ENSG00000180509) (ENSG00000053918 294501 12860)
                     (ENSG00000183873) (ENSG00000145362 238759 112415)
                     (ENSG00000008086) (ENSG00000169057) (ENSG00000055118))
(C C G C A A A A) = ((ENSG00000180509) (ENSG00000053918 28941)
                     (ENSG00000183873) (ENSG00000145362 304411)
                     (ENSG00000008086 117230) (ENSG00000169057 22795)
                     (ENSG00000055118))
(G C G C A A A A) = ((ENSG00000180509) (ENSG00000053918)
                     (ENSG00000183873 72406) (ENSG00000145362)
                     (ENSG00000008086) (ENSG00000169057) (ENSG00000055118))
(T C G C A A A A) = ((ENSG00000180509) (ENSG00000053918) (ENSG00000183873)
                     (ENSG00000145362 50245) (ENSG00000008086 80403)
                     (ENSG00000169057) (ENSG00000055118))
(A G G C A A A A) = ((ENSG00000180509 31550 14676 12499)
                     (ENSG00000053918 388660 359647 324708 323314 307619
                      208949 204518)
                     (ENSG00000183873 72732 71144 14052 7030)
                     (ENSG00000145362 304625 303714 287732 263295 159985
                      132442 67911 51205 49937 1324)
                     (ENSG00000008086 218181 134845 127535 122634)
                     (ENSG00000169057 33594) (ENSG00000055118))
(C G G C A A A A) = ((ENSG00000180509) (ENSG00000053918 338367)
                     (ENSG00000183873) (ENSG00000145362 122073)
                     (ENSG00000008086 178328) (ENSG00000169057)
                     (ENSG00000055118))
```

E.4 Output showing a 'pack'

The output shows the 'areas' found in the long DNA sequence, and the individuals that make up the pack.

```
===== THE PACK =====


Pack id: 88
Pack fitness: 9.0
Common DNA strings: ((T C A G C C C C) (G G T A A A A T) (T A A A A T T G))
--- --- --- --- ---
Areas:

AREA-ID: AREA
Sequence type: SEKV5
Start: 31
Stop: 48
Area fitness: 9/25
Sequences: (31 46 48)
The SEKV5 sequence from 31 to 56 :
(T C A G C C C C A A A A G A A G G T A A A A T T G C)



AREA-ID: AREA
Sequence type: SEKV4
Start: 31
Stop: 48
Area fitness: 9/25
Sequences: (31 46 48)
The SEKV4 sequence from 31 to 56 :
(T C A G C C C C A A A A G A A G G T A A A A T T G C)
```

```
AREA-ID: AREA
Sequence type: SEKV3
Start: 60
Stop: 77
Area fitness: 9/25
Sequences: (60 75 77)
The SEKV3 sequence from 60 to 85 :
(T C A G C C C C A A A A G A A G G T A A A A T T G C)



AREA-ID: AREA
Sequence type: SEKV2
Start: 60
Stop: 77
Area fitness: 9/25
Sequences: (60 75 77)
The SEKV2 sequence from 60 to 85 :
(T C A G C C C C A A A A G A A G G T A A A A T T G C)



AREA-ID: AREA
Sequence type: SEKV1
Start: 60
Stop: 77
Area fitness: 9/25
Sequences: (60 75 77)
The SEKV1 sequence from 60 to 85 :
(T C A G C C C C A A A A G A A G G T A A A A T T G C)


--- --- --- --- ---

Individual-ID: 127
Parents: (270 469)
Generation number: 11
Chromosome set 1: ((0 0 0 1 1 0 0 0 0 1 0 0 0 1 0 1 0)
                   (0 1 0 0 1 0 0 1 1 0 0 0 1 0 1 0 1))
Chromosome set 2: ((1 1 0 1 1 1 0 1 1 1 0 0 0 0 1 1 1)
```

```
                    (0 1 1 0 0 0 0 0 0 0 0 1 0 1 0 0 0))
Sequence position: 48
Sequence identifier: SEKV5
All sequence positions:
  SEKV5 : (48)
  SEKV4 : (48 908 8159)
  SEKV3 : (77)
  SEKV2 : (77)
  SEKV1 : (77)
Short DNA strings: ((T A A A A T T G))
Fitness: 34

Individual-ID: 609
Parents: (685 667)
Generation number: 11
Chromosome set 1: ((1 1 0 1 1 1 0 1 1 1 0 0 0 0 1 1 1)
                    (0 1 1 1 1 0 0 0 1 0 0 1 0 1 1 1 0))
Chromosome set 2: ((1 1 0 0 0 1 0 0 1 1 0 1 0 0 1 1 1)
                    (1 1 1 1 0 0 0 1 1 1 1 0 1 0 0 1 0))
Sequence position: 1193
Sequence identifier: SEKV5
All sequence positions:
  SEKV5 : (46 1193)
  SEKV4 : (46)
  SEKV3 : (75)
  SEKV2 : (75)
  SEKV1 : (75)
Short DNA strings: ((G G T A A A A T))
Fitness: 29

Individual-ID: 750
Parents: (469 822)
Generation number: 11
Chromosome set 1: ((1 1 0 1 1 1 0 1 1 1 0 0 0 0 1 1 1)
                    (0 0 0 0 0 0 1 1 1 1 0 0 1 0 1 1 1))
Chromosome set 2: ((0 0 1 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1)
                    (1 0 0 1 0 0 0 0 0 0 0 0 0 1 1 0 1 0))
Sequence position: 31
Sequence identifier: SEKV5
All sequence positions:
  SEKV5 : (31)
```

```
  SEKV4 : (31)
  SEKV3 : (60)
  SEKV2 : (60)
  SEKV1 : (60 2054)
Short DNA strings: ((T C A G C C C C))
Fitness: 29
```

Appendix F

Scripts: Script code text

F.1 Perl code

```
#! /usr/bin/perl

use strict;
use warnings;

use String::Approx  'amatch', 'aindex';

use DBI;

my $database = 'oivindb_private';
my $server = 'mysql.cbs.dtu.dk';
my $user = 'oivindb';
my $password = '$ENV{HOME}/.my.cnf';

my $oivindb_private = DBI->connect("dbi:mysql:$database:$server",
                                   $user, $password, {RaiseError=>1});


###  The variables that must be set for each run:
my $ClinSign = "Retinitis pigmentosa%";
my $ClinSignName = "RetPigmentosavsDNA";
###



my $sqlcmd ;
my $sql ;
```

```
 $sqlcmd = qq{CREATE TABLE $ClinSignName (GeneCode VarChar(20) NOT NULL,
              DNASeq longtext NOT NULL)  };

$sql = $oivindb_private->prepare($sqlcmd);

$sql->execute();


 $sqlcmd = qq{INSERT INTO $ClinSignName (GeneCode, DNASeq)
                                        SELECT DISTINCT gene, sequence
        FROM mim2gene AS a
        INNER JOIN OBClinSign AS b
        ON (a.mim=b.Syndrome)
        WHERE b.ClinSign LIKE "$ClinSign%"};  ### This expression 'ClinSign'
                                              ### must be set manually
                                              ### on occasion

$sql = $oivindb_private->prepare($sqlcmd);

$sql->execute();


 $sqlcmd = qq{select GeneCode, DNASeq FROM $ClinSignName};

$sql = $oivindb_private->prepare($sqlcmd);

$sql->execute();


open (OUTFILE, ">FromSQLRaw");

while (my $row = $sql->fetchrow_arrayref) {
print OUTFILE join("\t", @$row), "\n";
}

close OUTFILE;

$oivindb_private->disconnect;


### Making two files by transforming from the raw format: .fa and .lsp,
### a fasta file and a lisp format file respectively.
```

```
###  The input file is the file made from the SQL search in the database.

open (OUTFILE, ">fil.fa");
 open (INFILE, "<FromSQLRaw");
 while (my $line=<INFILE>)  {

# if (s/([ACGT])/$1 /g)
chomp($line);
$line =~s/([ACGT])/$1 /g;
$line =~s/(ENSG)\s+(\d+)/>$1$2 \n/g ;
# {print OUTFILE  "$line  \n ";}
print OUTFILE  "$line \n";

}
close INFILE;
close OUTFILE;



open (OUTFILE, ">fil.lsp");
 open (INFILE, "<fil.fa");
 while (my $line=<INFILE>)  {

# if (s/([ACGT])/$1 /g)
$line =~s/>(ENSG\d+)/\)\) \($1 \( /g ;
# {print OUTFILE  "$line  \n ";}
print OUTFILE  "$line \n";

}
close INFILE;
close OUTFILE;

# s/([ACGT])/$1 /g

# s/(ENSG)\s+(\d+)/\)\) \($1$2 \( \n/g

# Legge inn topp- og bunntekst fra .lsp-fil


exit;


### End of the file transforming operation
```

```
#! /usr/bin/perl

use strict;
use warnings;
use String::Approx  'amatch', 'aindex';


my @inputs;



# @matches = aindex("xyzzy", @inputs);

#print "@matches \n ";

#  my @catches = amatch("plugh", ['2'], @inputs);

# use LWP;

&get_seqs;    # Calling the procedure get_seqs
              #  which will go through the
              # results file and pick out
              # sequences and place them in @inputs

# print "@inputs \n ";


### Calling procedures from the package StringApprox
###  These procedures will match strings that are not
###  identical, but with a given percentage of dissimilarity

# my @matches = amatch("TTTTTTTTTTGAGA", ['5%'], @inputs);
# my @matches = amatch("TTTTTTTTTTGAGA", ['10%'], @inputs);
my @matches = amatch("TTTTTTTTTTGAGA", ['10%'], @inputs);
# my @indmatches = aindex("CCTGGGTGACAGAGCGAGACTCTGTCTCAA", @inputs);
# print "@indmatches \n ";


# print "The matches found:  @matches \n ";
```

```
### Print out the matches found by the approximate
###  search for sequences

print "The individual matches: \n ";
foreach my $match (@matches) {print "Match: $match  \n ";};

my $matchlength=@matches;
print "The number of matches: $matchlength \n ";


############ procedure get_seqs  ##################

sub get_seqs {

 open (RESFILE, "<resRecUVI");  ## Example results file
 while (<RESFILE>)  {
 if (/\(([ACGT]+)\)/g)# {print "Found sequence: $1 \n ";#}
                       {push (@inputs, $1);}
}
close RESFILE;
}

############ procedure get_seqs  ##################


### Print matches to file

open (UTFIL, ">sekv.fa");
# print UTFIL "> \n ";
# foreach my $match (@matches) {print UTFIL " $match ";};
foreach my $match (@matches) {$match =~ s/([ACGT])/$1 /g};
foreach my $match (@matches) {print UTFIL ">Sekv \n $match  \n \n";};

# print UTFIL "> \n ";
# foreach my $match (@matches)  {print UTFIL " $match  \n ";};
close UTFIL;
```

F.2 Python code

```
#! /usr/bin/env/ python

import  os


### Run the lisp program


lisp = os.popen('lisp -dynamic-space-size 1500 > GALogg', 'w')

lisp.write("""
(load "comp.lsp")
(load "main")
""")


lisp.write("""
(quit)
""")

lisp.close()

print '\n\n Finished lisp run \n'



#####################################################
# OR:
# lisp.write("""
# (load "detcmds.lsp")
# """)
#
# WHERE detcmds.lsp says
#
# (load "comp.lsp")
# (load "main")
#
#####################################################
```

```
### Use Python to rearrange the output file
###  by running a Python command file

pycmd = 'python rearrfitns.py'
os.system(pycmd)


### Get the Python module for the 'R' statistical package

from rpy import *


### Use Python to run an 'R' command file
###  to make boxplots, and turn the boxplot figure
###  into an encapsulated postscript file

r.source('rcom.R')
```

```
#! /usr/bin/env/ python

import re

# Open the DNA sequence input file

rawfile=open('rawdna.lsp','r')
# rawfile=open('longQT.fasta','r')
# rawfile=open('SeqCntsMultRuns','r')
#rawfile=open('Results/lpaIVSeqs','r')

raw=rawfile.read()

# print raw


# Compile the pattern to search for in the input file

seq= re.compile(
#r'\([ACGT\s]*(G T T T G T T T)[ACGT\s]*\)',re.MULTILINE
# r'(A T T A C T G C C G A A A T C C A G A T G)'
r'(A [ACGT] T [ACGT])'  # Test case
)

# Collect the short sequences found

# print re.findall(seq,raw)

seqfnd = []

seqfnd = re.findall(seq,raw)



# Enter the sequences found into
# a dictionary/ hash table, add
# a count for each different sequence variation
# (if any variation exists)

seqcnt = {}
```

```
for elt in seqfnd:
  seqcnt.setdefault(elt,0)
  seqcnt[elt]+=1


# Print out the sequences found, and their counts

for elt in seqcnt: print elt, ' ', seqcnt[elt]
```

```
#! /usr/bin/env/ python


import  os, re

##############################################################

#             Calling Lisp functions

##############################################################

# sequence length, no of sequences

lisp = os.popen('lisp -dynamic-space-size 1200 ', 'w')


##### Parameters to random-dna-seq:
# Length of individual (randomly generated) DNA sequence
# and the Number of such sequences

lisp.write("""
(load "rndDNA.lsp")
(random-dna-seq 10000 5)
""")


lisp.write("""
(quit)
""")

lisp.close()

print '\n\n Finished lisp part \n'


##############   End, calling Lisp functions ############


##############################################################
```

```
#             Pyhton proper

#############################################################

innfil= open("randDNA.fasta",'r')

innfiltxt = innfil.read()

pattern= re.compile(
r'[()]',re.MULTILINE
)

###### Remove parentheses

rmvparthns = re.sub(pattern, ' ', innfiltxt)


innfil.close()

import time

utfil = open("frandDNA.fasta",'w')

utfil.write(rmvparthns)


utfil.close


########### Write to rawdna.lsp

utfil = open("tmprawdna.lsp",'a')


### Write header info

header = "\n\n;;;; Random DNA sequences, file created on " \
 + str(time.localtime()[2]) + ' ' + str(time.localtime()[1])+ ' '
+ \ str(time.localtime()[0]) \
+  " at " + str(time.localtime()[3])
+ str(time.localtime()[4]) + " ;;;;; \n\n\n"
utfil.write(header)
```

```
innfil= open("startlspraw",'r')

innfiltxt = innfil.read()


utfil.write(innfiltxt)

innfil.close()

### Write sequences

innfil= open("randDNA.lsp",'r')

innfiltxt = innfil.read()

utfil.write(innfiltxt)

innfil.close()


### Write tail


innfil= open("endlspraw",'r')

innfiltxt = innfil.read()


utfil.write(innfiltxt)

innfil.close()

utfil.close()
```

```
#! /usr/bin/env/ python


import re


RunGen = []; fitns = []   # store data pairs in lists x and

### Open output file from the genetic algorithm

fitnfile=open('fitness.dat','r')

### Compile patterns to match variable names
###  and values

pattern = re.compile(r'(?P<Var>Run\d+Gen\d{1,3})')
pattern2 = re.compile(r'(?:(?P<Ftns>-?\d{1,4}) )+')


### Read output of fitnesses from the genetic algorithm
###  and look for vaiable names and fitness values
###  Save what is found

while True:
    line = fitnfile.readline()
    if not line: break
    match = re.search(pattern,line)
    match2 = re.search(pattern2,line)
    if match:
       RunGen.append(match.group('Var'))
       fitns.append(match2.group())


cfitns = []

for elt in fitns:
    fftn = elt.split()
    cfitns.append(fftn)
```

```
fitnfile.close()


outcnt =  len(cfitns[1])
incnt  =  len(RunGen)

### Open file for writing rearranged variable names
###  and fitness values

out=open('rfitness.dat','w')


### Write the variables ad values to file,
###  now with variables as column headers
###  and values as columns

for name in RunGen:
    out.write(name)
    out.write("\t")

out.write("\n")

for inner in xrange(outcnt):
   out.write("\n")
   for outer in xrange(len(RunGen)):
       out.write(cfitns[outer][inner])
       out.write("\t\t")


out.close()
```

F.3 'R' code

```
ftndata = read.table("rfitness.dat", header=T)
attach(ftndata)
postscript(file= 'RPlotFitnesses.eps',
             horizontal=TRUE, paper='a4',
             title= "BoxPlot of fitnesses", width = 8.0,
             height = 6.0)
boxplot(ftndata,xlab='Generations',ylab='Fitness')
dev.off()
```

Part IV

References

Bibliography

[1] Khoury MJ, Moore CA, and Evans JA. On the use of the term 'syndrome' in clinical genetics and birth defects epidemiology. *Am J Med Genet*, 49:26 – 8, 1994.

[2] DL Sackett, RB Haynes, and P Tugwell. *Clinical epidemiology*. Little, Brown and Company, Boston, MA, USA, 1985.

[3] Cahan A, Gilon D, Manor O, and Paltiel O. Probabilistic reasoning and clinical decision-making: Do doctors overestimate diagnostic probabilities? *Q J Med*, 96:763 – 769, 2003.

[4] Richardson WS. Five uneasy pieces about pre-test probability. *J Gen intern Med*, 17(11):882 – 883, 2002.

[5] Cahan A, Gilon D, Manor O, and Paltiel O. Clinical experience does not reduce the variance in physicians' estimates of pretest probabilities in a cross-sectional survey. *Journal of clinical epidemiology*, 58:1211 – 1216, 2005.

[6] Cari Snowman and Angela Scheuerle. Qualitative descriptors of disease incidence: commonly used and frequently muddled. *Am J Med Genet A*, 149A(7):1460–1462, Jul 2009.

[7] R Quinlan. *C4.5 Programs for machine learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[8] Cohen Jr MM. *The child with multiple birth defects*. Raven Press, New York, 1982.

[9] Opitz JM. Associations and syndromes: Terminology in clinical genetics and birth defects epidemiology. *Am J Med Genet*, 49:14 – 20, 1994.

[10] Lubinsky MS. Properties of associations: Identity, nature and clinical criteria, with a commentary on why CHARGE and Goldenhar are not associations. *Am J Med Genet*, 49:21 – 5, 1994.

[11] Brunner HG and van Driel MA. From syndrome families to functional genomics. *Nature Reviews Genetics*, 5:545 – 551, 2004.

[12] T. K. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet*, 28(1):21–28, May 2001.

[13] Marc A van Driel, Jorn Bruggeman, Gert Vriend, Han G Brunner, and Jack A M Leunissen. A text-mining analysis of the human phenome. *Eur J Hum Genet*, 14(5):535–542, May 2006.

[14] M. Oti and H. G. Brunner. The modular nature of genetic diseases. *Clin Genet*, 71(1):1–11, Jan 2007.

[15] Peter N Robinson, Sebastian Khler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet*, 83(5):610–615, Nov 2008.

[16] Joseph Loscalzo, Isaac Kohane, and Albert-Laszlo Barabasi. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Mol Syst Biol*, 3:124, 2007.

[17] Stephen Friend and Eric Schadt. Something wiki this way comes. interview by Bryn Nelson. *Nature*, 458(7234):13, Mar 2009.

[18] Xingpeng Jiang, Bing Liu, Jiefeng Jiang, Huizhi Zhao, Ming Fan, Jing Zhang, Zhenjie Fan, and Tianzi Jiang. Modularity in the genetic disease-phenotype network. *FEBS Lett*, 582(17):2549–2554, Jul 2008.

[19] Andrey Rzhetsky, David Wajngurt, Naeun Park, and Tian Zheng. Probing genetic overlap among complex human phenotypes. *Proc Natl Acad Sci U S A*, 104(28):11694–11699, Jul 2007.

[20] Martin Oti, Martijn A Huynen, and Han G Brunner. Phenome connections. *Trends Genet*, 24(3):103–106, Mar 2008.

[21] Kasper Lage, E. Olof Karlberg, Zenia M Strling, Pll I Olason, Anders G Pedersen, Olga Rigina, Anders M Hinsby, Zeynep Tmer, Flemming Pociot, Niels Tommerup, Yves Moreau, and Søren Brunak.

A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol*, 25(3):309–316, Mar 2007.

[22] M. Oti, B. Snel, M. A. Huynen, and H. G. Brunner. Predicting disease genes using protein-protein interactions. *J Med Genet*, 43(8):691–698, Aug 2006.

[23] Petra Kraus and Thomas Lufkin. Dlx homeobox gene control of mammalian limb and craniofacial development. *Am J Med Genet A*, 140(13):1366–1374, Jul 2006.

[24] Maria Rita Passos-Bueno, Camila C Ornelas, and Roberto D Fanganiello. Syndromes of the first and second pharyngeal arches: A review. *Am J Med Genet A*, 149A(8):1853–1859, Aug 2009.

[25] P. Volpe, G. Campobasso, V. De Robertis, and G. Rembouskos. Disorders of prosencephalic development. *Prenat Diagn*, 29(4):340–354, Apr 2009.

[26] L. Faivre, G. Collod-Beroud, B. Callewaert, A. Child, B. L. Loeys, C. Binquet, E. Gautier, E. Arbustini, K. Mayer, M. Arslan-Kirchner, A. Kiotsekoglou, P. Comeglio, M. Grasso, C. Beroud, C. Bonithon-Kopp, M. Claustres, C. Stheneur, O. Bouchot, J. E. Wolf, P. N. Robinson, L. Ads, J. De Backer, P. Coucke, U. Francke, A. De Paepe, C. Boileau, and G. Jondeau. Pathogenic fbn1 mutations in 146 adults not meeting clinical diagnostic criteria for marfan syndrome: further delineation of type 1 fibrillinopathies and focus on patients with an isolated major criterion. *Am J Med Genet A*, 149A(5):854–860, May 2009.

[27] Svend Rand-Hendriksen, Lena Tjeldhorn, Rigmor Lundby, Svein Ove Semb, Jon Offstad, Kai Andersen, Odd Geiran, and Benedicte Paus. Search for correlations between fbn1 genotype and complete ghent phenotype in 44 unrelated norwegian patients with marfan syndrome. *Am J Med Genet A*, 143A(17):1968–1977, Sep 2007.

[28] S. U. Dhar, D. del Gaudio, J. R. German, S. U. Peters, Z. Ou, P. I. Bader, J. S. Berg, M. Blazo, C. W. Brown, B. H. Graham, T. A. Grebe, S. Lalani, M. Irons, S. Sparagana, M. Williams, J. A. Phillips, A. L. Beaudet, P. Stankiewicz, A. Patel, S. W. Cheung, and T. Sahoio. 22q13.3 deletion syndrome: clinical and molecular analysis using array cgh. *Am J Med Genet A*, 152A(3):573–581, Mar 2010.

[29] L. Garavelli, M. Zollino, P. Cerruti Mainardi, F. Gurrieri, F. Rivieri, F. Soli, R. Verri, E. Albertini, E. Favaron, M. Zignani, D. Orteschi, P. Bianchi, F. Faravelli, F. Forzano, M. Seri, A. Wischmeijer, D. Turchetti, E. Pompilii, M. Gnoli, G. Cocchi, L. Mazzanti, R. Bergamaschi, D. De Brasi, M. P. Sperandeo, F. Mari, V. Uliana, R. Mostardini, M. Cecconi, M. Grasso, S. Sassi, G. Sebastio, A. Renieri, M. Silengo, S. Bernasconi, N. Wakamatsu, and G. Neri. Mowat-wilson syndrome: facial phenotype changing with age: study of 19 italian patients and review of the literature. *Am J Med Genet A*, 149A(3):417–426, Mar 2009.

[30] C. C. Breugem and A. B. Mink van der Molen. What is 'Pierre Robin sequence'? *J Plast Reconstr Aesthet Surg*, 62(12):1555–1558, Dec 2009.

[31] Judith Hall, Judith Allanson, Karen Gripp, and Anne Slavotinek. *Handbook of Physical Measurements*. Oxford university press, UK, 1998/ 2007.

[32] Judith E Allanson, Leslie G Biesecker, John C Carey, and Raoul C M Hennekam. Elements of morphology: introduction. *Am J Med Genet A*, 149A(1):2–5, Jan 2009.

[33] Judith E Allanson, Christopher Cunniff, H. Eugene Hoyme, Julie McGaughran, Max Muenke, and Giovanni Neri. Elements of morphology: standard terminology for the head and face. *Am J Med Genet A*, 149A(1):6–28, Jan 2009.

[34] M. M. Cohen. Syndromology: an updated conceptual overview. x. references. *Int J Oral Maxillofac Surg*, 19(2):89–96, Apr 1990.

[35] R. E. Ward, P. L. Jamison, and J. E. Allanson. Quantitative approach to identifying abnormal variation in the human face exemplified by a study of 278 individuals with five craniofacial syndromes. *Am J Med Genet*, 91(1):8–17, Mar 2000.

[36] Tinashe E M Mutsvangwa, Ernesta M Meintjes, Dennis L Viljoen, and Tania S Douglas. Morphometric analysis and classification of the facial phenotype associated with fetal alcohol syndrome in 5- and 12-year-old children. *Am J Med Genet A*, 152A(1):32–41, Jan 2010.

[37] Tania S Douglas and Tinashe E M Mutsvangwa. A review of facial image analysis for delineation of the facial phenotype associated with

fetal alcohol syndrome. *Am J Med Genet A*, 152A(2):528–536, Feb 2010.

[38] Peter Hammond, Tim J Hutton, Judith E Allanson, Bernard Buxton, Linda E Campbell, Jill Clayton-Smith, Dian Donnai, Annette Karmiloff-Smith, Kay Metcalfe, Kieran C Murphy, Michael Patton, Barbara Pober, Katrina Prescott, Pete Scambler, Adam Shaw, Ann C M Smith, Angela F Stevens, I. Karen Temple, Raoul Hennekam, and May Tassabehji. Discriminating power of localized three-dimensional facial morphology. *Am J Hum Genet*, 77(6):999–1010, Dec 2005.

[39] Peter Hammond, Tim J Hutton, Judith E Allanson, Linda E Campbell, Raoul C M Hennekam, Sean Holden, Michael A Patton, Adam Shaw, I. Karen Temple, Matthew Trotter, Kieran C Murphy, and Robin M Winter. 3D analysis of facial morphology. *Am J Med Genet A*, 126A(4):339–348, May 2004.

[40] Peter Hammond. The use of 3D face shape modelling in dysmorphology. *Arch Dis Child*, 92(12):1120–1126, Dec 2007.

[41] JL Ross, H Kushner, and AR Zinn. Discriminant analysis of the Ullrich-Turner syndrome neurocognitive profile. *Am J Med Genet*, 72(3):275 – 80, 1997.

[42] DZ Loesch and D Scott. Application of the anthropometric discriminant functions in estimation of carrier probabilities in Martin-Bell syndrome. *Clin Genet*, 36:145 – 51, 1989.

[43] E. S. Moore, R. E. Ward, P. L. Jamison, C. A. Morris, P. I. Bader, and B. D. Hall. The subtle facial signs of prenatal exposure to alcohol: an anthropometric approach. *J Pediatr*, 139(2):215–219, Aug 2001.

[44] SJ Astley and SK Clarren. A fetal alcohol syndrome screening tool. *Alcohol Clin Exp Res*, 19(6):1565 – 71, 1995.

[45] K Kosaki, MC Jones, and C Stayboldt. Zimmer phocomelia: Delineation by principal coordinate analysis. *Am J Med Genet*, 66:55 – 9, 1996.

[46] JV Neel, S Julius, A Weder, M Yamada, SL Kardia, and MB Haviland. Syndrome X: Is it for real? *Genet Epidemiol*, 15(1):19 – 32, 1998. Log-linear analysis.

[47] Volk HE, Henderson C, Neuman RJ, and Todd RD. Validation of population-based ADHD subtypes and identification of three clinically impaired subtypes. *Am J Med Genetics*, 141(3):312 – 318, 2006.

[48] Heather E Volk, Alexandre A Todorov, David A Hay, and Richard D Todd. Simple identification of complex adhd subtypes using current symptom counts. *J Am Acad Child Adolesc Psychiatry*, 48(4):441–450, Apr 2009.

[49] RW Haley, TL Kurt, and J Hom. Is there a Gulf war syndrome? Searching for syndromes by factor analysis of symptoms. *JAMA*, 277(3):215 – 22, 1997.

[50] DZ Loesch and SR Wilson. Multivariate analysis of body shape in fragile X (Martin-Bell) syndrome. *Am J Med Genet*, 33:200 – 8, 1989.

[51] M Preus and B MacGibbon. An application of numerical taxonomy to the classification of syndromes. *Birth defects: Original article series, Volume XIII Number 3A*, pages 31 – 8, 1977.

[52] M Preus and S Ayme. Formal analysis of dysmorphism: Objective methods of syndrome definition. *Clin Genet*, 23:1 – 16, 1983.

[53] M Preus. Differential diagnosis of the Williams and the Noonan syndromes. *Clin Genet*, 25:429 – 34, 1984.

[54] M Preus. The Williams syndrome: objective definition and diagnosis. *Clin Genet*, 25:422 – 8, 1984.

[55] A Verloes. Numerical syndromology: a mathematical approach to the nosology of complex phenotypes. *Am J Med Genet*, 55(4):433 – 43, 1995.

[56] T. Zhang, B. Fang, Y. Y. Tang, G. He, and J. Wen. Topology preserving non-negative matrix factorization for face recognition. *IEEE Trans Image Process*, 17(4):574–584, Apr 2008.

[57] Tony Segaran. *Programming collective intelligence*. O'Reilly, 2007, Sebastopol, CA, USA.

[58] Boaz Vigdor and Boaz Lerner. Accurate and fast off and online fuzzy ARTMAP-based image classification with application to genetic abnormality diagnosis. *IEEE Trans Neural Netw*, 17(5):1288–1300, Sep 2006.

[59] Ian H Witten and Eibe Frank. *Data Mining - practical machine learning tools and techniques*. Morgan Kaumann, CA, USA, 2005.

[60] Jiawei Han and Micheline Kamber. *Data mining - concepts and techniques*. Morgan Kaufmann, CA, USA, 2nd edition, 2006.

[61] Vili Podgorelec, Peter Kokol, Bruno Stiglic, and Ivan Rozman. Decision trees: an overview and their use in medicine. *J Med Syst*, 26(5):445–463, Oct 2002.

[62] J. Forsström, P. Nuutila, and K. Irjala. Using the ID3 algorithm to find discrepant diagnoses from laboratory databases of thyroid patients. *Med Decis Making*, 11(3):171–175, 1991.

[63] J. Forsström. Inductive learning of thyroid functional states using the ID3 algorithm. the effect of poor examples on the learning result. *Int J Biomed Comput*, 30(1):57–67, Jan 1992.

[64] D. G. McBride, M. J. Dietz, M. T. Vennemeyer, S. A. Meadors, R. A. Benfer, and N. L. Furbee. Bootstrap methods for sex determination from the os coxae using the ID3 algorithm. *J Forensic Sci*, 46(3):427–431, May 2001.

[65] Jean-Baptiste Lamy, Anis Ellini, Vahid Ebrahiminia, Jean-Daniel Zucker, Hector Falcoff, and Alain Venot. Use of the c4.5 machine learning algorithm to test a clinical guideline-based decision support system. *Stud Health Technol Inform*, 136:223–228, 2008.

[66] Lung-Cheng Huang, Sen-Yen Hsu, and Eugene Lin. A comparison of classification methods for predicting chronic fatigue syndrome based on genetic data. *J Transl Med*, 7:81, 2009.

[67] Lukas Tanner, Mark Schreiber, Jenny G H Low, Adrian Ong, Thomas Tolfvenstam, Yee Ling Lai, Lee Ching Ng, Yee Sin Leo, Le Thi Puong, Subhash G Vasudevan, Cameron P Simmons, Martin L Hibberd, and Eng Eong Ooi. Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness. *PLoS Negl Trop Dis*, 2(3):e196, 2008.

[68] CD Evans and RM Winter. A case-based learning approach to grouping cases with multiple malformations. *MD-Comput*, 12:127 – 36, 1995.

[69] CD Evans. A case-based assistant for diagnosis and analysis of dysmorphic syndromes. *Med Inform*, 20(2):121 – 31, 1995.

[70] Hartmut S Loos, Dagmar Wieczorek, Rolf P Wrtz, Christoph von der Malsburg, and Bernhard Horsthemke. Computer-based recognition of dysmorphic faces. *Eur J Hum Genet*, 11(8):555–560, Aug 2003.

[71] D. F. Schorderet. Diagnosing human malformation patterns with a microcomputer: evaluation of two different algorithms. *Am J Med Genet*, 28(2):337–344, Oct 1987.

[72] Leitersdorf E, Chakravarta A, and Hobbs HH. Polymorphic DNA haplotypes at the ldl receptor locus. *Am J Hum Genet*, 44:409, 1989.

[73] C. F. Sing, M. B. Haviland, K. E. Zerba, and A. R. Templeton. Application of cladistics to the analysis of genotype-phenotype relationships. *Eur J Epidemiol*, 8 Suppl 1:3–9, May 1992.

[74] Szymanski M, Erdmann VA, and Barciszewski J. Non-coding RNAs database (ncRNAdb). *Nucleic Acids Research*, 35:162 – 164, 2007.

[75] Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, and Haussler D. Ultraconserved elements in the human genome. *Science*, 304(5675):1321 – 1325, 2004.

[76] McEwen GK, Woolfe A, Goode D, Vavouri T, Callaway H, and Elgar G. Ancient duplicated conserved noncoding elements in vertebrates: A genomic and functional analysis. *Genome Res*, 16:451 – 465, 2006.

[77] Xianjun Dong, David Fredman, and Boris Lenhard. Synorth: exploring the evolution of synteny and long-range regulatory interactions in vertebrate genomes. *Genome Biol*, 10(8):R86, 2009.

[78] Rigoutsos I, Huynh T, Miranda K, Tsirigos A, McHardy A, and Platt D. Short blocks from the noncoding parts of the human genome have instances within nearly all known genes and relate to biological processes. *Proc Natl Acad Sci*, 103(17):6605 – 6610, 2006.

[79] Meynert A and Birney E. Picking pyknons out of the human genome. *Cell*, 125(5):836 – 838, 2006.

[80] Gennadi V Glinsky. Human genome connectivity code links disease-associated snps, micrornas and pyknons. *Cell Cycle*, 8(6):925–930, Mar 2009.

[81] Lopez-Bigas N, Blencowe BJ, and Ouzounis CA. Highly consistent patterns for inherited human diseases at the molecular level. *Bioinformatics*, 22(3):269 – 277, 2006.

[82] Jagath C Rajapakse, Chunxi Chen Pooja, and Sy-Loi Ho. Comparative genomic workflow discovery of conserved noncoding DNA patterns. *IEEE engineering in medicine and biology magazine*, July/ August:19–24, 2009.

[83] Fabian A Buske, Mikael Bodn, Denis C Bauer, and Timothy L Bailey. Assigning roles to DNA regulatory motifs using comparative genomics. *Bioinformatics*, Feb 2010.

[84] Vahid Khatibi and Gholam Ali Montazer. Intuitionistic fuzzy set vs. fuzzy set application in medical pattern recognition. *Artif Intell Med*, 47(1):43–52, Sep 2009.

[85] Francisco Fernandes, Ana T Freitas, Jonas S Almeida, and Susana Vinga. Entropic profiler - detection of conservation in genomes using information theory. *BMC Res Notes*, 2:72, 2009.

[86] Eleazar Eskin and Pavel A Pevzner. Finding composite regulatory patterns in DNA sequences. *Bioinformatics*, 18 Suppl 1:S354–S363, 2002.

[87] Modan K Das and Ho-Kwok Dai. A survey of DNA motif finding algorithms. *BMC Bioinformatics*, 8 Suppl 7:S21, 2007.

[88] Tak-Ming Chan, Gang Li, Kwong-Sak Leung, and Kin-Hong Lee. Discovering multiple realistic tfbs motifs based on a generalized model. *BMC Bioinformatics*, 10:321, 2009.

[89] Carsten Kemena and Cedric Notredame. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics*, 25(19):2455–2465, Oct 2009.

[90] J. D. Thompson, D. G. Higgins, and T. J. Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–4680, Nov 1994.

[91] C. Notredame, D. G. Higgins, and J. Heringa. T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1):205–217, Sep 2000.

[92] D. F. Feng and R. F. Doolittle. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol*, 25(4):351–360, 1987.

[93] D. J. Lipman, S. F. Altschul, and J. D. Kececioglu. A tool for multiple sequence alignment. *Proc Natl Acad Sci U S A*, 86(12):4412–4415, Jun 1989.

[94] C. Notredame and D. G. Higgins. Saga: sequence alignment by genetic algorithm. *Nucleic Acids Res*, 24(8):1515–1524, Apr 1996.

[95] Fogel GB and Corne DW. *Evolutionary computation in bioinformatics*. Morgan Kaufmann, CA, USA, 2003.

[96] C. Gondro and B. P. Kinghorn. A simple genetic algorithm for multiple sequence alignment. *Genet Mol Res*, 6(4):964–982, 2007.

[97] Riccardo Leardi. Genetic algorithms in chemistry. *Journal of chromatography*, 1158:226 – 233, 2007.

[98] Kumar Chellapilla and Gary B Fogel. Multiple sequence alignment using evolutionary programming. In *Proceedings of the 1999 conference on evolutionay computation*, pages 445 – 452. IEEE, 1999.

[99] Holland J. *Adaption in natural and artificial systems*. MIT Press, Cambridge, Massachusetts, USA, second edition, 1992.

[100] Goldberg DE. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison Wesley, Reading, Massachusetts, USA, 1989.

[101] Melanie Mitchell. *An introduction to genetic algorithms*. MIT Press, Cambridge, Massachusetts, USA, 1996.

[102] AE Eiben and JE Smith. *Introduction to evolutionary computing*. Springer, Berlin, Germany, 2nd printing 2007.

[103] Timothy L Bailey, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. Meme suite: tools for motif discovery and searching. *Nucleic Acids Res*, 37(Web Server issue):W202–W208, Jul 2009.

[104] Bailey TL and Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pages 28 – 36, Menlo Park, California, 1994. AAAI Press.

[105] Bailey TL, Williams N, Misleh C, and Li WW. Meme: Discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research*, 34:W369–W373, 2006. Web server issue.

[106] A. Mihalich, P. Magnaghi, L. Sessa, M. Trubia, F. Acquati, and R. Taramelli. Genomic structure and organization of kringles type 3 to 10 of the apolipoprotein(a) gene in 6q26-27. *Gene*, 196(1-2):1–8, Sep 1997.

[107] C. Thry and C. D. Stern. Roles of kringle domain-containing serine proteases in epithelial-mesenchymal transitions during embryonic development. *Acta Anat (Basel)*, 156(3):162–172, 1996.

[108] Timothy L Bailey, Nadya Williams, Chris Misleh, and Wilfred W Li. Meme: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res*, 34(Web Server issue):W369–W373, Jul 2006.

[109] Annie Olry. Prevalence of rare diseases: Bibliographic data. Orphanet Report Series, Rare Diseases collection 1, Orphanet, November 2009.

[110] Tinashe Mutsvangwa and Tania S Douglas. Morphometric analysis of facial landmark data to characterize the facial phenotype associated with fetal alcohol syndrome. *J Anat*, 210(2):209–220, Feb 2007.

[111] Andriulli A, Grossi E, Buscema M, Pilotto A, Festa V, and Perri F. Artificial neural networks can classify uninvestigated patients with dyspepsia. *Eur J Gastroenterol Hepatol*, 19(12):1055–1058, 2007.

[112] Grossi E, Borroni B, Zimmermann M, Marcello E, Colciaghi F, Gardoni F, Intraligi M, Padovani A, and Buscema M. Artificial neural networks allow the use of simultaneous measurements of alzheimer disease markers for early detection of the disease. *J Transl Med*, 3:30, 2005.

[113] Cenker Eken, Ugur Bilge, Mutlu Kartal, and Oktay Eray. Artificial neural network, genetic algorithm, and logistic regression applications for predicting renal colic in emergency settings. *Int J Emerg Med*, 2(2):99–105, 2009.

[114] Joo S, Yang YS, Moon WK, and Kim HC. Computer-aided diagnosis of solid breast nodules: use of an artificial neural network based on multiple sonographic features. *IEEE Trans Med Imaging*, 23(10):1292–1300, 2004.

[115] Pace F and Savarino V. The use of artificial neural network in gastroenterology: the experience of the first 10 years. *Eur J Gastroenterol Hepatol*, 19(12):1043–1045, 2007.

[116] Winston PH. *Artificial Intelligence*. Addison-Wesley, MA, USA, 3rd edition, 1992.

[117] Krogh A. What are artificial neural networks? *Nat Biotechnol*, 26(2):195–197, 2008.

[118] Fumiko Hoeft, Amy A Lightbody, Heather Cody Hazlett, Swetapadma Patnaik, Joseph Piven, and Allan L Reiss. Morphometric spatial patterns differentiating boys with fragile x syndrome, typically developing boys, and developmentally delayed boys aged 1 to 3 years. *Arch Gen Psychiatry*, 65(9):1087–1097, Sep 2008.

[119] J. Laurikkala, M. Juhola, S. Lammi, and K. Viikki. Comparison of genetic algorithms and other classification methods in the diagnosis of female urinary incontinence. *Methods Inf Med*, 38(2):125–131, Jun 1999.

[120] Chen Huang, Ruijie Zhang, Zhiqiang Chen, Yongshuai Jiang, Zhenwei Shang, Peng Sun, Xuehong Zhang, and Xia Li. Predict potential drug targets from the ion channel proteins based on svm. *J Theor Biol*, 262(4):750–756, Feb 2010.

[121] Sima E. Uyar and A. Emre Harmanci. Investigation of new operators for a diploid genetic algorithm. In *Poceedings of The International Society for Optical Engineering.*, 2003.

[122] W. R. Rice and A. K. Chippindale. Sexual recombination and the power of natural selection. *Science*, 294(5542):555–559, Oct 2001.

[123] Julie D Thompson, Patrice Koehl, Raymond Ripp, and Olivier Poch. Balibase 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, 61(1):127–136, Oct 2005.

[124] Zhang C and Wong AK. A genetic algorithm for multiple molecular sequence alignment. *Comput Appl Biosci*, 13(6):565 – 81, 1997.

[125] Namkung J, Nam JW, and Park T. Identification of expression quantitative trait loci by the interaction analysis using genetic algorithm. *BMC Proc*, 11:S69, 2007.

[126] Notredame C and Higgins DG. SAGA: Sequence alignment by genetic algorithm. *Nucleic Acids Res*, 24(8):1515 – 24, 1996.

[127] Qin L, Pan Y, Chen L, and Chen Y. An improved ant colony algorithm with diversified solutions based on the immune strategy. *BMC Bioinformatics*, 7 Suppl 4:53, 2006.

[128] Yulan He and Siu Cheung Hui. Exploring ant-based algorithms for gene expression data analysis. *Artif Intell Med*, 47(2):105–119, Oct 2009.

[129] Zhao F, Zhao F, Li T, and Bryant DA. A new pheromone-trail based genetic algorithm for comparative genome assembly. *Nucleic Acids Res*, 36(10):3455 – 3462, 2008.

[130] Soman S, Zasuwa G, and Yee J. Autmation, decision support and expert systems in nephrology. *Advances in Chronc Kidney Disease*, 15(1):42 – 55, 2008.

[131] Grossi E and Buscema M and. Introduction to artificial neural network. *Eur J Gastroenterol Hepatol*, 19(12):1046–1054, 2007.

[132] Jefferson MF, Pendleton N, Lucas SB, and Horan MA. Comparison of a genetic algorithm neural network with logistic regression for predicting outcome after surgery for patients with nonsmall cell lung carcinoma. *Cancer*, 79(7):1338–1342, 1997.

[133] Larranaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armananzas R, Santafe G, Perez A, and Robles V. Machine learning in bioinformatics. *Brief Bioinform*, 7(1):86–112, 2006.

[134] Huttenhower C, Schroeder M, Chikina MD, and Troyanskaya OG. The sleipnir library for computational functional genomics. *Bioinformatics*, 24(13):1559–1561, 2008.

[135] Pirooznia M, Yang JY, Yang MQ, and Deng Y. A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics*, 9 Suppl 1:S13, 2008.

[136] Sajda P. Machine learning for detection and diagnosis of disease. *Annu Rev Biomed Eng*, 8:537–565, 2006.

[137] Han J and Kamber M. *Data Mining - concepts and techniques*. Morgan kaufmann, CA,USA, 3rd edition, 2006.

[138] Zweigenbaum P, Demner-Fushman D, Yu H, and Cohen KB. Frontiers of biomedical text mining: current progress. *Brief Bioinform*, 8(5):358–375, 2007.

[139] Kim JJ, Pezik P, and Rebholz-Schuhmann D. Medevi: retrieving textual evidence of relations between biomedical concepts from medline. *Bioinformatics*, 24(11):1410–1412, 2008.

[140] Rzhetsky A, Seringhaus M, and Gerstein M. Seeking a new biology through text mining. *Cell*, 134(1):9–13, 2008.

[141] Darwin C. *The origin of species*. Penguin books ltd, London, England, Penguin Classics (1985) edition, 1859.

[142] Poli R, angdon B, and McPhee N. *A field guide to genetic programming*. Springer, 2008.

[143] Holland JH. *Hidden Order - How adaption builds complexity*. Addison Wesley, MA, USA, 1st edition, 1995.

[144] Dorn GW II. The fuzzy logic of physiologic cardiac hypertrophy. *Hypertension*, 49:962 – 970, 2007.

[145] Vineis P. Methodological insights: Fuzzy sets in medicine. *J Epidemiol Community Health*, 62:273 – 278, 2008.

[146] Bichindaritz I and Marling C. Case-based reasoning in the health sciences: What's next? *Artif Intell Med*, 36(2):127–135, 2006.

[147] Larranaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA Armananzas R, Santafe G, Perez A, and Robles V. Machine learning in bioinformatics. *Brief Bioinform 7(1):86-112*, 7(1):86 – 112, 2006.

[148] Lin J. Pagerank without hyperlinks: reranking with pubmed related article networks for biomedical text retrieval. *BMC Bioinformatics*, 9:270, 2008.

[149] Ruttenberg A, Clark T, Bug W, Samwald M Bodenreider O, Chen H, Doherty D, Forsberg K, Gao Y, Kashyap V, Kinoshita J, Luciano J, Marshall MS, Ogbuji C, Rees J, Stephens S andWong GT, Wu E, Zaccagnini D, Hongsermeier T, Neumann E, Herman I, and Cheung KH and. Advancing translational research with the semantic web. *BMC Bioinformatics*, 8 Suppl 3:S2, 2007.

[150] Cannata N, Schroder M, Marangoni R, and Romano P. A semantic web for bioinformatics: goals, tools, systems, applications. *BMC Bioinformatics*, 9 Suppl 4:S1, 2008.

[151] Splendiani A. Rdfscape: Semantic web meets systems biology. *BMC Bioinformatics*, 9 Suppl 4:S6, 2008.

[152] Oivind Braaten. *Computational Methods in Biophysics, Biomaterials, Biotechnology and Medical Systems: Algorithm Development, Mathematical Analysis and Diagnostics, I - IV*, chapter Foreword, pages vii – ix. Kluwer Academic Publishers, MA, USA, 2003.

[153] Bartosch-Harlid A, Andersson B, Aho U, Nilsson J, and Andersson R. Artificial neural networks in pancreatic disease. *Br J Surg*, 95(7):817–826, 2008.

[154] Wang X, Zheng B, Li S, Mulvihill JJ, Wood MC, and Liu H. Automated classification of metaphase chromosomes: Optimization of an adaptive computerized scheme. *J Biomed Inform*, 2008.

[155] Ramesh AN, Kambhampati C, Monson JRT, and Drew PJ. Artificial intelligence in medicine. *Ann R Col Surg Engl*, 86:334 – 338, 2004.

[156] Werner Horn, Christian Popow, Silvia Miksch, Lieselotte Kirchner, and Andreas Seyfang. Development and evaluation of vie-pnn, a knowledge-based system for calculating the parenteral nutrition of newborn infants. *Artif Intell Med*, 24(3):217–228, Mar 2002.

[157] Crispin J Miller and Teresa K Attwood. Bioinformatics goes back to the future. *Nat Rev Mol Cell Biol*, 4(2):157–162, Feb 2003.

[158] Albert-Lszl Barabsi and Zoltn N Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113, Feb 2004.

[159] Shu-Qin Zhang, Wai-Ki Ching, Nam-Kiu Tsing, Ho-Yin Leung, and Dianjing Guo. A new multiple regression approach for the construction of genetic regulatory networks. *Artif Intell Med*, 48(2-3):153–160, 2010.

[160] Vimla L Patel, Edward H Shortliffe, Mario Stefanelli, Peter Szolovits, Michael R Berthold, Riccardo Bellazzi, and Ameen Abu-Hanna. The coming of age of artificial intelligence in medicine. *Artif Intell Med*, 46(1):5–17, May 2009.

[161] Riccardo Bellazzi and Ameen Abu-Hanna. Artificial intelligence in medicine aime'07. *Artif Intell Med*, 46(1):1–3, May 2009.

[162] Penco S, Buscema M, Patrosso MC, Marocchi A, and Grossi E. New application of intelligent agents in sporadic amyotrophic lateral sclerosis identifies unexpected specific genetic background. *BMC Bioinformatics*, 9:254, 2008.

[163] Feinstein A. *Clinical Epidemiology*. W.B. Saunders Comp, Phil, USA, 1985.

[164] Feinstein A. *Cincal Judgment*. Robert Krieger, FL, USA, 1985.

[165] Wulff HR. *Rationel klinik*. Munksgaard, Copenhagen, 2nd edition, 1981.

[166] Scadding JG. Essentialism and nominalism in medicine: logic of diagnosis in disease terminology. *Lancet*, 348:594 – 96, 1996.

[167] Pirsig R. *Zen and the art of motorcycle maintenance*. Corgi, London, 1974.

[168] Ø Braaten. Artificial intelligence in pediatrics: Important clinical signs in newborn syndromes. *Comput Biomed Res*, 29:153 – 61, 1996. Reprinted in Yearbook of medical informatics 1997, p 566 – 74.

[169] Polanyi M. *The tacit dimension*. Peter Smith, Gloucester, Massachusetts, USA, 1983.

Index

accuracy, 134
AI references PubMed
    figure, 128
aims
    of the thesis, 3
algorithm
    genetic, 5, 29, 79, 116
analysis
    cluster, 18
    discriminant, 18
    factor, 19
    principal component, 19
ant hill, 44
anthropometrics, 16
application area
    syndromology, 12
article
    ID3, 7
article summaries, 8–9
articles
    list of, xi
    overview, 7
artificial intelligence, 5, 125
    medical, 4
    medicine, 125
artificially generated patients, 33
association, 13

Bayes' formula, 136
Berg, Kåre, vii
bias
    ascertainment, 138
    selection, 139
bioinformatics, 5
bioinformatics studies, introduction, 24
bioinformatics studies, results, 36
biological networks, 14–15
breadth first search, 115

case based reasoning, 22, 124
CBR, 22, 124
change with age, clinical sign(s), 16
classification
    syndrome, 14
clinical sign(s)
    change with age, 16
    diagnostic value, 10–12
    presence of, 16
closed universe, 124
cluster analysis, 18
code
    Lisp, 145
    script
        Perl, 161
        Python, 167
computation
    evolutionary, 116
consistency, 134
    diagnosis, 9
criterion standard of diagnosis, 5, 12
cut off point, 136

Darwin quotation
    evolution, 116
data mining, 113, 129
decision tree, 5, 11, 28, 113
definition

syndrome, 16
deformation, 12
depth first search, 115
diagnosis
consistency, 9
criterion standard, 5, 12
differential, 11
gold standard, 12
objective, 4, 57
philosophical background, 141
diagnostic methods
objective, 57
discriminant analysis, 18
disruption, 12
DNA pattern searching, 26
dysplasia, 12

essentialism, 142
evolution, 116
evolutionary computation, 116
evolutionary programming, 123
example output
genetic algorithm
hash table, 156
pack, 157
path, 154
population, 151
expert systems, 111

factor analysis, 19
feature vector, 4
figure
AI references PubMed, 128
fitness proportionate selection, 119
fuzzy logic, 123

GA pattern searching in DNA, 28
General introduction section, 3
general overview, 4
genetic algorithm, 5, 29, 79, 116
chromosomes, 116
crossing over, 116
diploid organisms, 117
example output
hash table, 156
pack, 157
path, 154
population, 151
first population, 119
fitness, 116, 119
fitness function, 121
greedy exploitation, 120
haploid organisms, 117
inversion, 117
listing
pack, 157
local maximum, 120
local optimum, 120
mating, 119
mutation, 117
parthenogenesis, 120
population, 116, 121
premature convergence, 120
recombination, 116, 117
reproduction, 116
selection, 119
genetic programming, 123

handle
in syndrome diagnosis, 10
hill climbing, 115

ID3, 4, 5, 113
identification tree, 5, 11, 28, 113
induction, 113
information theory formula, 114
intelligence
artificial, 125
medical artificial, 4
inter-observer variation, 135
intra-observer variation, 135
introduction section
General, 3

Jervell, Herman Ruge, vii

kappa, 135
knowledge
    tacit, 143

learning
    machine, 112
    supervised, 113
    unsupervised, 113
likelihood ratio, 132
Lisp code, 145
list of articles, xi
listing
    pack
        genetic algorithm, 157
logistic regression, 19

machine learning, 112
malformation, 12
Material and methods section, 28
Material section, 32
mathematical methods, 17
mating
    genetic algorithm, 119
medical artificial intelligence, 4, 125
medical philosophy of science, 141
MEME, 31
Methods section, 28
mining
    data, 129
    text, 129
multiple regression, 19
multiple tests, 136

natural language processing, 129
networks
    biological , 14–15
NLP, 129
nominalism, 142
non-negative matrix factorization, 19

objective diagnosis, 4
objective diagnostic methods , 57
overview
    articles, 7
    general, 4

pack, 30
parthenogenesis, 30
    genetic algorithm, 120
path, 30
patients
    artificially generated, 33
Perl
    script code, 161
philosophy of science, 4, 141
pivot sign, 10
polytopic field defect, 12
precision, 134
predictive value, 9–10, 132
presence of clinical sign(s), 16
principal component analysis, 19
project
    discussion, 37
    implications, 44
    introduction, 9
    material and methods, 28
    motivation for, 9
    results, 34
Python
    script code, 167

quality of data, 138

reasoning
    case based, 124
receiver operating curves, *see* ROC curves
regression
    logistic, 19
    multiple, 19
ROC curves, 134
roulette wheel selection, 119

screening, 10
script code
    Perl, 161
    Python, 167
search, 115
    breadth first, 115
    depth first, 115
    hill climbing, 115
search space, 115
searching
    DNA patterns, 26
selection
    fitness proportionate , 119
    genetic algorithm, 119
    roulette wheel, 119
    tournament, 119
semantic web, 129
sensitivity, 131
sequence, 13
sign(s)
    clinical, 10
        using more than one, 135
specificity, 131
statistical methods, 17
summaries
    of articles, 8
supervised learning, 113
supervisor, vii
support vector machines, 40
syndrome
    association, 13
    deformation, 12
    disruption, 12
    dysplasia, 12
    handle, 10
    malformation, 12
    polytopic field defect, 12
    sequence, 13
syndrome classification, 14
syndrome definition, 16
syndrome diagnoses
    AI approaches, 20–23
    statistical approaches, 17–20
syndromology, 5
    as application area, 12
syndromology studies, results, 34

tacit knowledge, 143
tests
    multiple, 136
text mining, 113, 129
theme
    the thesis', 3
thesis
    aims of, 3
    theme of, 3
tournament selection, 119
tree
    decision, 113
    identification, 113

universe
    closed, 124
unsupervised learning, 113