Department of Computer Science
Series of Publications A
Report A-2013-9

Computational methods for augmenting association-based gene mapping

Lauri Eronen

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public criticism in Auditorium XII (University Main Building, Unioninkatu 34) on 20 September 2013 at twelve o'clock noon.

University of Helsinki
Finland

Supervisor
Hannu Toivonen, University of Helsinki, Finland

Pre-examiners
Jaakko Hollmén, Aalto University School of Science, Finland
Sven Laur, University of Tartu, Estonia

Opponent
Joost Kok, Leiden University, The Netherlands

Custos
Hannu Toivonen, University of Helsinki, Finland

Contact information

Department of Computer Science
P.O. Box 68 (Gustaf Hällströmin katu 2b)
FI-00014 University of Helsinki
Finland

Email address: info@cs.helsinki.fi
URL: http://www.cs.helsinki.fi/
Telephone: +358 9 1911, telefax: +358 9 191 51120

ISSN 1238-8645
ISBN 978-952-10-9177-3 (paperback)
ISBN 978-952-10-9178-0 (PDF)
Computing Reviews (1998) Classification: F.4.2, G.2.2, G.3, H.2.5, H.2.8, J.3
Helsinki 2013
Unigrafia

Computational methods for augmenting association-based gene mapping

Lauri Eronen

Department of Computer Science
P.O. Box 68, FI-00014 University of Helsinki, Finland
Lauri.Eronen@iki.fi

PhD Thesis, Series of Publications A, Report A-2013-9
Helsinki, September 2013, 84+93 pages
ISSN 1238-8645
ISBN 978-952-10-9177-3 (paperback)
ISBN 978-952-10-9178-0 (PDF)

Abstract

The context and motivation for this thesis is gene mapping, the discovery of genetic variants that affect susceptibility to disease. The goals of gene mapping research include understanding of disease mechanisms, evaluating individual disease risks and ultimately developing new medicines and treatments.

Traditional genetic association mapping methods test each measured genetic variant independently for association with the disease. One way to improve the power of detecting disease-affecting variants is to base the tests on haplotypes, strings of adjacent variants that are inherited together, instead of individual variants. To enable haplotype analyses in large-scale association studies, this thesis introduces two novel statistical models and gives an efficient algorithm for haplotype reconstruction, jointly called HaploRec. HaploRec is based on modeling local regularities of variable length in the haplotypes of the studied population and using the obtained model to statistically reconstruct the most probable haplotypes for each studied individual. Our experiments demonstrate that HaploRec is especially well suited to data sets with a large number or markers and subjects, such as those typically used in currently popular genome-wide association studies.

Public biological databases contain large amounts of data that can help in determining the relevance of putative associations. In this thesis, we introduce Biomine, a database and search engine that integrates data from

several such databases under a uniform graph representation. The graph database is used to derive a general proximity measure for biological entities represented as graph nodes, based on a novel scheme of weighting individual graph edges based on their informativeness and type. The resulting proximity measure can be used as a basis for various data analysis tasks, such as ranking putative disease genes and visualization of gene relationships.

Our experiments show that relevant disease genes can be identified from among the putative ones with a reasonable accuracy using Biomine. Best accuracy is obtained when a pre-known reference set of disease genes is available, but experiments using a novel clustering-based method demonstrate that putative disease genes can also be ranked without a reference set under suitable conditions.

An important complementary use of Biomine is the search and visualization of indirect relationships between graph nodes, which can be used e.g. to characterize the relationship of putative disease genes to already known disease genes. We provide two methods for selecting subgraphs to be visualized: one based on weights of the edges on the paths connecting query nodes, and one based on using context free grammars to define the types of paths to be displayed. Both of these query interfaces to Biomine are available online.

Computing Reviews (1998) Categories and Subject Descriptors:
F.4.2 Grammars and Other Rewriting Systems
G.2.2 Graph algorithms
G.3 Markov processes, Probabilistic algorithms, Statistical computing
H.2.5 Data translation
H.2.8 Data mining, Scientific databases
J.3 Biology and genetics

General Terms:
Algorithms, Experimentation

Additional Key Words and Phrases:
Association mapping, Bioinformatics, Context-free grammars, Disease gene prioritization, EM algorithm, Graph mining, Haplotyping, SNPs, Weighted graphs

Acknowledgements

I am most grateful to my supervisor Hannu Toivonen for his persistent guidance and support during my studies and the writing of this thesis. It is clear that this daunting task would never have been completed without his insight and optimism.

Trivially, I thank my co-pundits Petteri Hintsanen, Atte Hinkka and Kimmo Kulovesi for hard work and cooperation during the BIOMINE project and for frequenting the *!biomine* IRC channel that has been the mainstay of my daily computational experience during the long years of this thesis project. In particular, I want to thank my long-time roommate Petteri for the plethora of scientific discussions and general banter, as well as support in a myriad of computational and non-computational tasks. A honorary mention goes to my colleague Petteri Sevón for fruitful collaboration encompassing a lot of the work in this thesis, as well as fruitless opposition in table hockey. Thanks also to my former colleagues Floris Geerts and Päivi Onkamo for scientific collaboration and introducing me to the peculiarities of academic research at the beginning of my scientific endeavours.

I thank the pre-examiners, Jaakko Hollmén and Sven Laur, as well as Petteri Hintsanen, for their enlightened comments that greatly helped to increase the quality of this thesis. Needless to say, I am indebted to the Department of Computer Science for providing a murky but dependable bastion for conducting this research. The teaching, research community and IT infrastructure provided by the department have naturally all been quintessential to this thesis.

This research has been supported by Helsinki Institute for Information Technology (HIIT); the Finnish Funding Agency for Technology and Innovation (Tekes); Jurilab Ltd; Biocomputing Platforms Ltd; Graduate school in Computational Biology, Bioinformatics and Biometry (ComBi); Helsinki Graduate School in Computer Science and Engineering (HeCSE); Algorithmic Data Analysis (Algodan) Centre of Excellence of the Academy of Finland and the Department of Computer Science of the University of Helsinki. I am grateful for this fragmented but jointly steady stream of in-

come that has enabled me to indulge in this thesis work while maintaining a satisfactory standard of living. Furthermore, thanks to the folks at my current employer, Biocomputing Platforms, for their understanding towards my rather extended thesis project that persisted for almost three years while I worked there, and also for providing me with a valuable practical link to the field of disease gene mapping.

I would also like to mention my former colleagues Patrik Hoyer, Jukka Perkiö, Juho Muhonen, Jussi Lindgren, Jukka-Pekka Kauppi (leaving out many other memorable characters) for their contributions to scientific and, more pleasantly, non-scientific discussions and activities. Almost finally, I congratulate the Academic Table Hockey Association (PÖLY) for providing a splendid venue for counteracting the pressures of academic struggles.

Finally but very importantly, I thank my family, relatives and friends for various wonders of life which I often regrettably had to neglect in favor of this thesis work.

Contents

Original Publications

This thesis is based on the following articles, which are referred to in the text by their Roman numerals.

Paper I

L. Eronen, F. Geerts, and H. Toivonen. HaploRec: Efficient and accurate large-scale reconstruction of haplotypes. *BMC Bioinformatics*, 7:542, 2006.

Paper II

P. Hintsanen, P. Sevon, P. Onkamo, L. Eronen, and H. Toivonen. An empirical comparison of case–control and trio-based study designs in high-throughput association mapping. *Journal of Medical Genetics*, 43:617–624, 2006.

Paper III

P. Sevon, L. Eronen, P. Hintsanen, K. Kulovesi, and H. Toivonen. Link discovery in graphs derived from biological databases. In *Proceedings of Data Integration in the Life Sciences, Third International Workshop*, pages 35–49, 2006.

Paper IV

L. Eronen and H. Toivonen. Biomine: Predicting links between biological entities using network models of heterogeneous databases. *BMC Bioinformatics*, 13:119, 2012.

Paper V

P. Sevon and L. Eronen. Subgraph queries by context-free grammars. *Journal of Integrative Bioinformatics*, 5:100, 2008.

These articles and their roles in this thesis are summarized in Section 1.2, and the author's contributions to each article are described in Section 1.3. More detailed descriptions of the contributions contained in each article are given in Chapter 6. Papers II and III also appear in the doctoral thesis of Petteri Hintsanen [40].

Chapter 1

Introduction

The context and motivation of this thesis is *disease gene mapping*, the discovery of genes affecting disease susceptibility based on comparing genetic variation between healthy and affected individuals. This discovery of genetic variants affecting many diseases is the first step in the research process aiming at understanding disease mechanisms on the molecular level. The role of gene mapping is to enable targeting more elaborate follow-up studies to a limited set of candidate genes. The most important practical goals of gene mapping research are developing gene tests for selecting treatments suitable to genetically different variants of the same disease, and ultimately developing new medicines and treatments.

This thesis introduces two kinds of computational approaches to improve the power of gene mapping studies. First, we describe a computationally efficient statistical *haplotype*[1] reconstruction method which allows for more accurate localization of genes affecting disease susceptibility. Haplotype reconstruction factors the genetic variation measured from each individual into two separate components (haplotypes): variants inherited from the father and from the mother. This makes it possible to analyze genetic variation by considering inheritance of segments of consecutive genetic variants together, instead of just single variants at a time. The utility of haplotypes reconstructed with the presented method in disease gene mapping is demonstrated by extensive simulation experiments.

Second, we present methods for utilizing readily available background data from public biological databases for refining and exploring putative gene mapping results obtained by analysis of genetic variation data. A major outcome of this work is Biomine[2], a database and search engine that

[1] We will explain what haplotypes are in Chapter 2, where we give an overview of the genetics needed for this thesis

[2] http://biomine.cs.helsinki.fi

integrates several such databases under a uniform graph representation. The graph database is used to derive a general gene proximity measure, which can be used to rank putative *disease susceptibility (DS)* genes obtained from a gene mapping study. The proximity measure also provides a basis for visualization and explorative analysis of gene relationships.

The body of this thesis consists of 5 peer-reviewed publications. To provide a framework for situating the publications within the field of gene mapping, we next present the workflow of a hypothetical gene mapping project. After that we give summaries of the original articles and the author's contributions to them. We conclude this chapter by outlining the remaining chapters of this introductory part.

1.1 Workflow of a gene mapping project

Our illustratory workflow (Figure 1.1) is divided into two main phases: *primary analysis* (left) and *refinement phase* (right). In the primary analysis phase, genetic variation data (*genotypes*) from the subjects under study is analyzed to find regions of the genome that are statistically associated with the disease. In the refinement phase, putative results from the primary analysis are prioritized and explored by putting them into the context of known biology, as represented by the Biomine database in this thesis. The goal of the refinement phase is to identify the most promising putative DS genes resulting from the primary analysis as targets for follow-up analyses by more elaborate laboratory experiments. Such follow-up analyses include obtaining data from additional subjects to validate observed associations, targeted DNA resequencing to uncover all genetic variants within the disease-associated regions and ultimately functional genomics experiments to isolate the precise molecular mechanisms leading to the disease.

The presented workflow is simplified and designed with the purpose of illustrating the methodology presented in this thesis. It does not address aspects that are not directly relevant to the publications of this thesis, such as effects of population structure, interactions between DS genes, effects of environmental factors on disease, or follow-up analyses by further laboratory experiments.

We next go through the steps of the example workflow, briefly outlining the role of the methods introduced in the original publications. More complete summaries of the articles are deferred to Section 1.2.

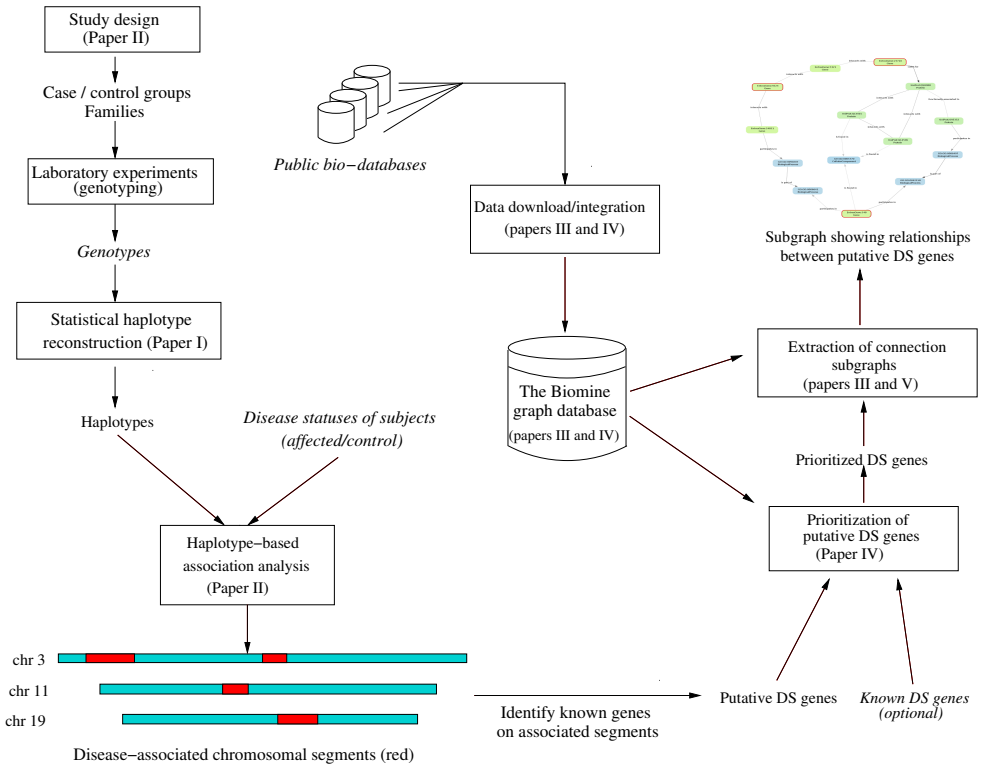

Figure 1.1: Overview of a hypothetical gene mapping study. The inputs to the analysis are depicted by italic text, the steps of the analysis are represented by boxes, and the results and intermediate results by normal text. The steps using methods presented in the papers forming this thesis are indicated in the corresponding boxes when applicable.

Primary analysis

The research process summarized by Figure 1.1 begins with a *study design* phase, where the researcher decides how subjects are ascertained for the study, what is the number of genotyped subjects (*sample size*), what genotyping method is used (this usually also determines the set of measured variants) and how the measured variants are to be analyzed. The study design step largely determines the *statistical power* of the study, meaning the probability of identifying the presence of genetic variants affecting disease susceptibility. Ideally, the study design phase should include analytic calculations or simulation studies to ensure sufficient statistical power. Paper II presents a simulation study addressing some of the issues arising in the design of association studies, most importantly the choice between family–based and case–control ascertainment strategies and an assessment of haplotype reconstruction strategies.

The first concrete step of the study is ascertaining the subjects and getting DNA samples from them. The DNA is then processed using laboratory methods (*genotyping*) to extract *genotypes* representing the genetic variants possessed by each subject, typically using a dedicated *genotyping chip* that can measure variants at hundreds of thousands of loci simultaneously.

The main step of the primary analysis is testing the obtained genotypes for association with the disease in question. In this thesis we focus on the simple case where the disease status is a binary variable (affected/unaffected). The genotypes can either be tested for association with the disease directly, or several consecutive variants can be tested jointly as haplotypes. The latter, more powerful option requires that the genotypes are first combined into haplotypes using e.g. *statistical haplotype reconstruction* methods. One such method, *HaploRec*, is introduced in Paper I. In Paper II, we experimentally compare association testing of haplotypes reconstructed with HaploRec to alternative strategies and find it to be a competitive option.

The result of the association analysis phase is a set of chromosomal regions associated with the disease. The genes contained in these regions are then assigned as putative *disease susceptibility genes (DS genes)*. Additionally, also nearby genes or genes otherwise known to be regulated by transcriptional DNA elements within the associated regions may be assigned as putative DS genes.

Refinement phase

In many cases the effects of individual mutations are too weak to be detected reliably by using the primary genetic variation data only, even if haplotypes are used to improve power. Due to the huge number of tested hypotheses (variants within the whole genome), the set of putative DS genes will almost certainly contain a large number of *false positives*: completely non-related genes which appear to have a high correlation with the disease, just by chance. As follow-up analyses are costly and time-consuming, it is important to focus these analyses only on the most promising candidates by filtering out the false positives as well as possible. This filtering process is addressed by the refinement phase of the example workflow (Figure 1.1, right).

Automatic prioritization of putative DS genes can be performed by considering pre-known biological and molecular relationships between putative genes and already known DS genes as additional evidence for the disease association. Even if pre-known DS genes are not available, it may be possible to do the prioritization based on the mutual relationships of the putative genes. Public biological databases contain a vast amount of readily accessible data relevant to the disease gene prediction problem, such as protein interactions [88, 47], genes' effects on diseases [38] and functional gene annotations [39, 79]. In this thesis, this knowledge is utilized through the Biomine database, introduced in Paper III, which represents the known network of public biological, medical and genetic knowledge derived from a number of heterogeneous source databases as a single weighted graph. In papers III and IV, we define a general similarity measure computed from the Biomine graph; in Paper IV this measure is applied to the disease gene prioritization task.

Although automatic DS gene prioritization methods can provide useful information, their results will rarely be used as such. Additionally, the researcher will typically want to inspect the putative DS genes manually (possibly concentrating on the top-ranked ones), by checking what is known about them in the original databases and literature. The Biomine database also facilitates such exploratory use, by finding and visualizing relationships between the putative genes and known DS genes, or relationships between the putative genes (the final step of the gene mapping workflow in Figure 1.1). These visualization functionalities can use the same proximity measures that are used for the prioritization, making them well suited for manual verification of the prioritization results. Alternatively, graph queries can be performed using *context-free grammars* as the query language (Paper V), allowing the querying of paths with specific semantics,

e.g., paths that suggest a causal relationship or paths that confer similarity.

1.2 Summary of original publications

The original publications of this thesis represent rather diverse fields of computer science and biology: algorithmic data analysis (Paper I), study design in genetic epidemiology (Paper II), bioinformatics (papers III and IV), and graph algorithms (papers III-V). Although they all share the common motivation of gene mapping, many of the contributions in the original publications are not restricted to gene mapping applications, but are potentially usable also in other bioinformatics and graph analysis tasks. We next give brief summaries of the original publications and their roles in the gene mapping workflow of Figure 1.1. More detailed descriptions of the contributions contained in each paper are given in Chapter 6.

Papers I and II address the issue of using haplotypes in association analysis, corresponding to the primary analysis phase of the gene mapping workflow in Figure 1.1. A problem with haplotype-based analyses is that commonly employed laboratory techniques do not directly reveal haplotype information. Paper I describes HaploRec, an accurate and efficient method suitable for genome-wide reconstruction of haplotypes needed in large-scale association studies based on haplotypes. (Our earlier conference paper [26] on the same topic was the first one to propose reconstructing haplotypes for longer chromosomal regions simultaneously.) HaploRec is based on statistically modeling local regularities in the genotype data.

Paper II contains a simulation-based evaluation of different study designs in gene mapping, comparing family-based and case–control association study designs under 3 different disease models. The experiments demonstrate the power of study designs based on haplotype-based association analysis, using HaploRec for the important haplotype reconstruction step. The main result is that statistically inferred haplotypes can be equally powerful to the true haplotypes for the purposes of association analysis. Moreover, the results suggest that the case–control study design (combined with statistical haplotype reconstruction) is a powerful alternative for the more laborious family-based ascertainment approach, especially for large data sets.

Paper III introduces the Biomine database that integrates data from several public biological databases under a common graph data model and repository, providing the foundation for the refinement phase of our gene mapping workflow (Figure 1.1). A fundamental component of the Biomine system is a proximity measure for assessing the strength of relationships

between graph nodes representing genes and other biological entities. This measure can be used e.g. to discover links between genes and diseases that consist of annotated relationships derived from different source databases. In Paper III, weights are assigned to individual graph edges based on the degrees of nodes adjacent to the edge, and overall proximity of nodes is measured using *two-terminal network reliability* [20] computed from a small subgraph connecting the query nodes. Experiments to discover two kinds of relationships are reported: ones between disease genes and the corresponding disease record in the OMIM database [38], and others to discover connections between interacting proteins. The article also includes an assessment of the statistical significance of discovered relationships by comparing them to a null distribution obtained by random sampling.

Paper IV extends the work done in Paper III in several directions, most importantly by two concrete applications: prediction of future links based on a previous version of the database and, more importantly for the topic of this thesis, prioritizing lists of putative DS genes based on proximities computed in the integrated graph. A new cluster-based method is introduced for the little-studied problem of gene prioritization in the case that there are no previously known "reference" genes for the disease under study. The paper also introduces weighting of different relationship types based on optimizing link prediction performance, as well as improvements to the edge weighting scheme based on node degrees. Finally, the database is compiled from a more comprehensive set of source databases than in Paper III.

Visualization of indirect relationships in the integrated graph may be useful in exploratory analysis of putative gene mapping results (the final step of the gene mapping pipeline in Figure 1.1). In **Paper V**, we describe a query system that enables finding *connection subgraphs* linking a given set of query nodes, providing a concise summary about their relationships. The result of such a query is the graph spanned by all paths matching a query, where the queries are based on the types of nodes and edges on the paths. More specifically, the set of accepted paths is specified by a context-free grammar (*CFG*) where the node and edge are used as terminal symbols of the grammar. An important contribution of the paper is a modified version of the well-known Earley algorithm [25], adapted to efficiently extract subgraphs matching a given CFG from a large graph.

Connection subgraph queries can also be performed based on the node proximity measures defined in papers III and IV; this usage of Biomine is briefly described in Paper III, and a public prototype of a proximity-based

graph query/visualization system is available at our web site[3]. A brief description of the query engine is also given in Chapter 4.

1.3 Contributions of the author

The author played a major part in formulating the probability models and algorithms in Paper I, implemented the methods, performed the experiments and wrote a major part of the text. In Paper II, the author provided the implementation for the haplotyping algorithm, implemented the automated test environment used for performing the study, and participated in performing the experiments. The author implemented a major part of Biomine, the data integration and query system used in papers III-V. In Paper III, the author participated in developing the original idea and methods, and performed the experiments. The author developed and implemented the methods of Paper IV, performed the experiments and wrote the majority of the article. In Paper V, the author participated in conceiving and implementing the algorithms, performed the experiments and participated in writing the article.

1.4 Structure of the introductory part

The rest of the thesis is organized as follows. In Chapter 2, we give an overview of genetic variation and gene mapping, including study design issues tackled in Paper II. In Chapter 3, we review the computational haplotype reconstruction problem and outline our statistical haplotype reconstruction approach, HaploRec, introduced in Paper I. We also discuss the relevance of HaploRec and haplotypes in general for gene mapping. In Chapter 4 we outline the Biomine database, and a general node proximity measure derived from it, based on Papers III and IV. We also describe experiments for optimizing the parameters of the proximity measure and validate it using link prediction as an example application. Furthermore, we outline the use of Biomine for discovering and visualizing relationships between biological entities, based on papers III and V. In Chapter 5 we describe the disease gene prioritization problem, and outline the use of Biomine for prioritization of putative disease genes, based on the contents of Paper IV. In Chapter 6, we summarize the contributions of the thesis. Finally, in Chapter 7 we discuss the relevance of the presented work in the field of gene mapping and give an outlook for future research.

[3]Biomine search engine: http://biomine.cs.helsinki.fi

Chapter 2

Principles of disease gene mapping

In this chapter, we will give a brief overview of genetics and the principles of disease gene mapping, to provide a basis for understanding the rest of the thesis. We also review the problem of study design in gene mapping and describe a particular type of gene mapping studies that is of interest in this thesis: genome-wide association studies (GWAS), where the complete genome is screened for disease-associated genetic variants without any prior hypotheses on the location of the DS genes. The contents of this chapter are based on standard textbook knowledge [7, 21, 44, 19] and previous theses on disease gene mapping [69, 82, 40].

2.1 The human genome

The *genome* is the collection of hereditary information of an organism, a copy of the whole genome being present in each cell of an organism. The genome consists of DNA (deoxyribonucleic acid), which is made up of simple units called nucleotides. There are four different nucleotides: adenine (A), thymine (T), cytosine (C) and guanine (G), and the whole human genome can be thought of as a very long sequence consisting of approximately 2.9 billion nucleotides. More specifically, a DNA molecule consists of two complementary strands where A is always paired with T and C is always paired with G, so that both strands actually contain the same information. These paired nucleotides are called *base pairs*.

The main purpose of DNA is to act as a template for the production of proteins, which are functional units participating in most cellular processes. The parts of genome that are used for producing proteins are called genes. Only a small fraction of the genome comprises of genes; of the remaining part, some regions have a role in regulating the activity of genes, but for

most part this *non-coding* DNA has no known function.

Human DNA is organized into 46 *chromosomes*, 44 of which are structurally pairwise similar, or *homologous*. These homologous chromosomes are referred to as *autosomes*. The autosomes form 22 *chromosome pairs*. In each pair, one of the chromosomes is inherited from the mother and the other from the father. In addition to the 22 pairs of autosomes, the normal human genome has two *sex chromosomes*; a female has two X chromosomes whereas a male has one X and one Y sex chromosome.

2.2 Genetic variation and genotyping

The majority of DNA is similar in all humans' genomes. However, due to mutations that have occurred during the population history, there is variability at some locations of the genome. There are several types of such variation, usually called *polymorphisms*. Of these, the most common are *single nucleotide polymorphisms (SNPs)*, variations of a single base pair. Other types of genetic variations include deletion of one or more base pairs from the sequence, *insertions* of a stretch of DNA into another place in the genome and variable number of repetitions of a short DNA sequence.

A certain location in the DNA sequence is called a *locus*, and different variants of the DNA at a polymorphic locus are called *alleles*. It is possible find out almost all genetic variants of an individual by using *genome sequencing* methods to read the complete DNA sequence of the individual. However, genome sequencing is quite expensive and slow, and for most studies it is more cost-efficient to only examine the DNA at a predefined set of *marker* loci that are known beforehand to contain relatively common variants. The list of specific alleles that an individual possesses at each marker locus is called a *genotype*, and the process of determining the genotype using laboratory methods is called *genotyping*.

SNPs are the most commonly used type of markers genotyping subjects for the purpose of gene mapping. Current SNP genotyping methods are able to determine hundreds of thousands or even millions of SNPs per subject in a single study, enabling the screening of the genome with a reasonable resolution and cost.

As there are two copies of each chromosome (except for the sex chromosomes), each individual also has two separate alleles for each studied polymorphic locus. A genotype is thus represented as a list of *allele pairs*, which in the case of SNP markers each represent the two nucleotides of an individual at a single locus. A part of an individuals' genotype containing allele pairs for a single chromosome can alternatively be viewed as a pair

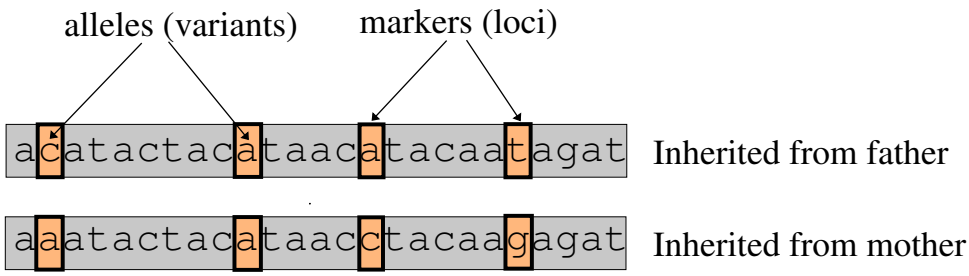

Figure 2.1: Example of a chromosomal segment containing 4 SNP markers

of two *haplotypes*: strings of consecutive alleles located on the same copy of the chromosome. Each haplotype represents the variants inherited from one of the parents.

Figure 2.1 illustrates the basic concepts of genotype data, showing the genotype and haplotypes of an individual on a short chromosomal segment containing four SNP markers. The genotype is $(\{a, c\}, \{a, a\}, \{a, c\}, \{g, t\})$. The haplotype inherited from the father is (c, a, a, t) and the haplotype inherited from the mother is (a, a, c, g). The haplotype representation thus contains more information than the genotype representation, as it also specifies the parental origin of each allele (the order of allele pairs is not relevant in the genotype representation).

Haplotypes of nearby markers are inherited together from generation to generation (with some exceptions), which means that considering haplotypes instead of independent genotypes makes it possible to more accurately estimate whether a particular segment of DNA in the genomes of two different individuals has been inherited from the same ancestor or not. This information is particularly useful in disease gene mapping, as will be explained in the following sections. Because genotyping methods do not directly provide haplotype information, *haplotype reconstruction* is an essential intermediate task in gene mapping methods using haplotypes. In Chapter 3 we will provide an overview of statistical haplotype reconstruction, and outline the method *HaploRec* introduced in Paper I.

2.3 Inheritance, recombination and genetic distance

An offspring inherits two copies of each autosome, one from the mother and one from the father. The DNA is transmitted from parent to offspring through germ line cells, *gametes*, which unlike other cells contain just one

copy of each chromosome (the other copy will be received from the other parent). *Meiosis* (Figure 2.2) is the process where gametes are formed, by combining DNA from both copies of each chromosomes of an individual.

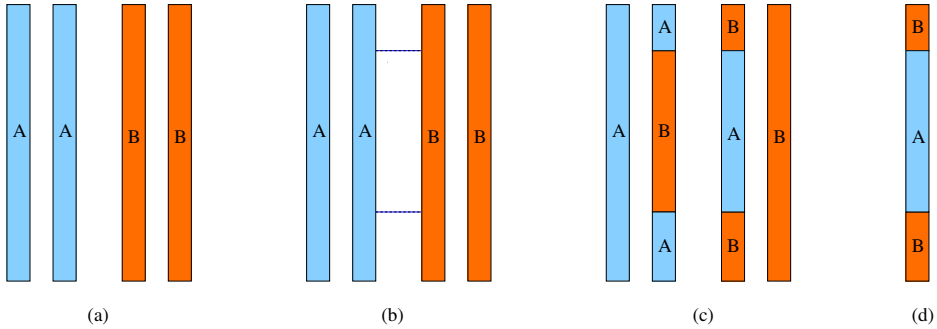

Figure 2.2: A schematic figure of a meiosis event for one chromosome. The individual has inherited two copies of the chromosome from its parents (denoted by A and B). (a) The DNA in the two chromosomes is duplicated to form 4 *chromatids*. (b) The chromatids exchange genetic material; in this example, two *crossovers* occur, denoted by the horizontal lines. (c) Two of the chromatids have exchanged their DNA in the region between the two crossover loci. (d) One of the four resulting chromatids is inherited to the offspring.

During meiosis, usually one or more *crossovers* occur between the original chromosomes of the parent, so that the chromosomes inherited by the offspring from a single parent are not usually exact copies of the either of the parent's chromosomes, but instead a combination of shorter segments from the two chromosome copies originating from the grandparents. In disease gene mapping, it is often of interest whether alleles at two markers on the same chromosome have been inherited together from the same parental chromosome, or instead from the two different copies of the chromosome (having been separated by a crossover). When an offspring inherits the DNA in two loci from different chromosomes of the parent (i.e. from different grandparents), it is said that a *recombination* has happened between the loci.

Two loci are inherited independently, if they reside in different chromosomes. Likewise, alleles in two loci far apart in the same chromosome are inherited nearly independently, due to the high probability of crossovers between the two loci. When the distance between two loci is small, there is an increased probability that no crossovers occur between the two loci during a meiosis, so that the alleles in the same chromatid are passed to

the offspring unchanged. In this case, the loci are said to be *genetically linked*. This *linkage* between the disease locus and genotyped marker loci is the key factor that enables using genetic markers for disease gene mapping.

2.4 Linkage disequilibrium and haplotypes

During the course of population history, the phenomenon of linkage leads to non-random associations, called *linkage disequilibrium (LD)*, between alleles of closely spaced markers. Two loci are said to be in linkage disequilibrium when certain combinations of alleles at these loci occur together more often than they would do just by change, assuming the markers were independent of each other.

Consider two alleles A and B in different loci with population frequencies P(A) and P(B). Assuming that the loci are mutually independent in the population, their probability of occurring together is a product of the individual allele frequencies: $P(AB) = P(A)P(B)$. However, if a combination of alleles occurs more or less often than would be expected assuming independence ($P(AB) \neq P(A)P(B)$), the markers are said to be in linkage disequilibrium. Several measures have been proposed to characterize the amount of LD between two loci [23]. One widely used measure is the absolute difference $\delta = P(AB) - P(A)P(B)$.

Although most measures in the literature consider LD between only two loci, the concept of LD is not limited to pairs of loci. Pairwise measures may fail to reveal even strong LD contained in a sample of chromosomes. As an example, consider the haplotype frequencies in Figure 2.3.

haplotype	frequency (%)
A B C	0.25
A b c	0.25
a B c	0.25
a b C	0.25

Figure 2.3: Example of LD over multiple markers.

There are three markers, where $\{A, a\}$, $\{B, b\}$ and $\{c, C\}$ are the pairs of alleles occurring at each marker, respectively. The frequency for all alleles is the same: $P(x) = 0.5$ for all $x \in \{A, a, B, b, C, c\}$. Looking at pairs of markers does not reveal any LD; haplotype frequencies for any of the three marker pairs are identical to ones obtained by assuming independence of the markers (for instance, $P(AC) = 0.25 = 0.5 \cdot 0.5 = P(A) \cdot P(C)$). However,

when inspecting all the three markers jointly, the allele at marker C can always be predicted based on the alleles at markers A and B.

2.5 Disease gene mapping

Disease gene mapping (see e.g. Altshuler et al. [3] for a more complete review) is the process of localizing mutations in DNA that increase (or decrease) the risk of obtaining the disease under study. In the most basic gene mapping setting, the *phenotype* of interest is a binary variable indicating whether an individual has the disease of interest or not, and the goal is to locate mutations affecting the probability of obtaining the disease by looking at the correlations between the *disease status* and genotypes of a sample containing both healthy and affected individuals. The locations of the genome containing mutations correlated with the disease are called *disease susceptibility (DS) loci*. Often, mutations affecting disease susceptibility are located within genes, in which case we speak of *disease susceptibility (DS) genes*. Also mutations outside of genes may affect disease susceptibility. This is not surprising, as comparative genome analysis has shown that 5% of the human genome is evolutionarily conserved and thus functional, while only less than one-third of this 5% consists of genes that encode proteins [100]. Disease-affecting mutations outside of genes typically affect disease through their role in regulating nearby genes [3]. Thus, also DS loci outside of genes should ideally be mapped to genes for subsequent analysis, based on their proximity on the DNA sequence or knowledge about regulatory relationships [80, 93].

If the complete DNA sequence were to be measured for all subjects, gene mapping would be relatively straightforward (ignoring the statistical problems from multiple testing and correlation between nearby markers): just test each variant independently for statistical association with the disease. However, genome sequencing is still currently slow and expensive compared to genotyping, and gene mapping studies are most often performed by genotyping a limited set of predefined marker loci. This is a feasible strategy owing to linkage between the DS locus and nearby marker loci. The idea is that although each actual DS variant is probably not among the set of genotyped variants, it is possible to approximately determine the DS locus by observing nearby markers that are in strong linkage with it, and therefore also correlated with the disease status.

Disease gene mapping methods can be roughly categorized into two categories: *linkage* and *association-based* methods. Family-based *linkage methods* [65, 57] examine the transmission of both the disease and marker

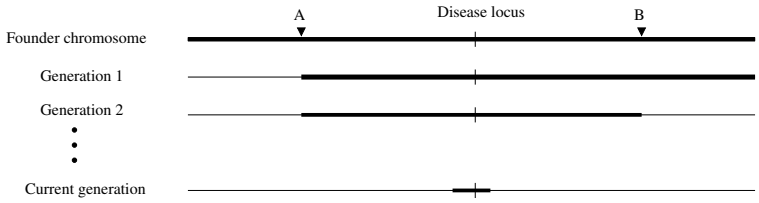

Figure 2.4: Evolution of a chromosomal region containing a disease mutation. The disease mutation has been introduced to the *founder chromosome*. In the first two generations, crossovers at locations A and B have replaced the ends of the chromosome by material from other chromosomes. In the current generation, a short fragment of the founder chromosome around the disease mutation still remains intact.

alleles within a known pedigree. The idea in linkage methods is that the distance between a genotyped marker and the disease locus can be inferred by observing how often the disease and the alleles at the marker locus are inherited together, i.e. without being separated by a recombination in between. The location of the DS variant can then be estimated by combining information from markers estimated to be located close to the DS variant. A problem with linkage studies is that as they only consider inheritance in a few generations, only a small number of recombinations occur within time period spanned by the pedigree. This means that also markers relatively far from the disease locus will be inherited together (without recombinations) in the observed data, making it impossible to pinpoint the DS locus very accurately.

In this thesis, we mainly concentrate on *association mapping* [17, 19, 6], which is an alternative way of locating disease genes based on genotype data. In association mapping, the disease locus is sought by finding markers that are in LD with the disease status, using genotypes from a set of *unrelated* (or more precisely, very distantly related) individuals. The key assumptions in association mapping are that each DS locus contains a single mutation that has been introduced to the population by a certain distant ancestor (*founder*), and that LD between the mutation and markers surrounding it is strong enough to be detected statistically. Such LD is expected to exist because the mutation locus has been inherited through the generations from the to the current carriers of the mutation as part of a preserved ancestral chromosomal segment (see Figure 2.4).

A chromosomal segment shared by members of the current population that has been inherited from a common ancestor without being split by recombinations during the population history is said to be inherited *identical*

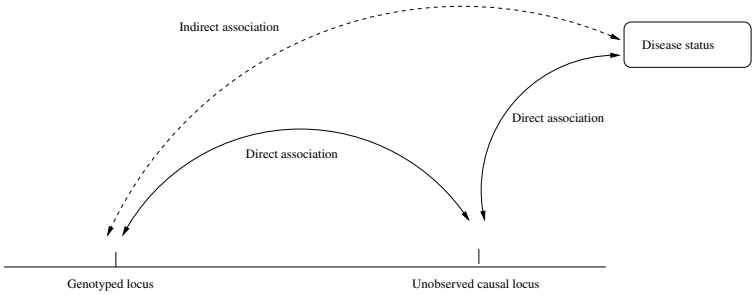

Figure 2.5: Indirect association between a genotyped locus and disease status

by descent (IBD). In the ideal case that two specific alleles only occur in a single IBD segment shared by some members of the current population, they will be in perfect linkage disequilibrium. Even if this is not the case, there will generally be at least some LD between variants located in the IBD region. Due to the combination of this LD between the unobserved DS locus and genotyped markers within the same IBD region, and the correlation between the DS locus and the disease status, the genotyped markers near the disease locus will also be statistically associated with the disease (Figure 2.5), although more weakly than either of the direct associations. Because the IBD segments where associations are observed will typically be much shorter than the ones inherited intact without recombination in the pedigrees considered in linkage studies, association mapping has the potential to pinpoint the DS locus with far greater accuracy than do linkage methods.

Tests of association and statistical significance An association study is performed by comparing the frequencies of alleles or genotypes at marker loci between groups of affected and unaffected individuals, in order to determine whether a statistical association exists between the trait and the marker. Perhaps the simplest method to test such associations is based on constructing a 2 by 2 *contingency table* (Figure 2.6) for each marker to be tested, where the presence of the disease and the presence of a particular allele (A/a) are the variables, and the table cells contain the counts of haplotypes that have the particular allele within each group of subjects (healthy/affected). This is of course only one of a large number of possibilities; instead of alleles, genotypes or haplotypes can be used as variables to be tested, and these can be grouped in various ways according to different assumptions about the inheritance of the disease

(dominant, recessive, etc.). Also, instead of directly comparing allele or genotype counts using a contingency table, the testing can be done using an explicit *disease model* that expresses the likelihood of the genotype data conditional on the disease status. In this case, the test of association is based on comparing the likelihood of the observed data under the disease model to its likelihood under a null model of no association. We will not consider these alternative methods for testing association any further in this thesis; see e.g. Clarke et al. [17] for a review.

Allele	a	A	Total
Healthy	f_{Ha}	f_{HA}	$f_{H\cdot}$
Case	f_{Ca}	f_{CA}	$f_{C\cdot}$
Total	$f_{\cdot a}$	$f_{\cdot A}$	

Figure 2.6: Contingency table for computing allelic association

Returning to the contingency table example (Figure 2.6), under the *null hypothesis* of no association between the tested marker and the disease, we expect the relative allele frequencies to be similar in the case and healthy groups ($\frac{f_{Ca}}{f_{C\cdot}} = \frac{f_{Ha}}{f_{H\cdot}}$ and $\frac{f_{CA}}{f_{C\cdot}} = \frac{f_{HA}}{f_{H\cdot}}$). The deviation between the observed allele counts and ones that would be expected under the null hypothesis can be measured e.g. by computing a χ^2 score [101, 17] that is defined as the sum of normalized squared differences between the observed and expected values of the table cells.

The *statistical significance* of an association it characterized by a *p-value* that tells how likely it is that, under the null hypothesis of no association, the value of the test score would be at least as high as the one computed from the observed frequencies. For the χ^2 score, the theoretical distribution is known and the p-value can be computed straightforwardly by comparing the score based on the observed frequencies to this known distribution.

If the p-value is under a predefined *significance threshold* (a value of 0.05 is typically used), then the association is said to be *statistically significant* (with the given threshold). Using 0.05 as the significance threshold thus also means that when testing a marker with no association to the disease, the probability of getting a *false positive* association is 0.05. We will discuss the statistical significance in the case of multiple tested markers in the section on genome-wide association studies below.

Tests using multiple markers In the relatively rare case where the disease locus is within the set of tested markers, or there exists a marker that is in perfect LD with the disease locus, testing of single markers is a sufficient strategy for detecting associations. However, often there is no single SNP that would correlate completely with the disease locus, even if all carriers of the disease-predisposing allele share a chromosomal segment around the disease locus that is inherited IBD. This is because also members of the current population that do not share the same IBD haplotype can have the same marker alleles that are present in the haplotype, having inherited them from other ancestral sources, thus diluting the LD between the markers and the disease locus. Consider for example the haplotype frequencies in Figure 2.7, where the (not genotyped) DS locus D is in perfect LD with the haplotype consisting of genotyped markers A and B. However, there is only an imperfect LD with alleles A and B individually, as they also occur in haplotypes without the disease allele (rows 2 and 3).

observed haplotype	DS locus	frequency (%)
A B	D	0.25
A b	d	0.25
a B	d	0.25
a b	d	0.25

Figure 2.7: Example of LD between marker haplotype and disease alleles. D denotes the causal disease allele and d denotes the normal allele at the disease locus.

Thus, the power to detect associations can be increased if several markers are considered simultaneously [1, 78]. Several association mapping methods that utilize LD across multiple markers have been published [89, 83, 60, 86, 52, 97]. Many of these methods require haplotypes as input, and cannot work directly on genotype data. In addition to describing the haplotyping method HaploRec introduced in Paper I and statistical haplotyping in general, the next Chapter will also discuss the use of haplotypes in gene mapping.

2.6 Genome-wide association studies

In this thesis, the main interest is in *genome-wide association studies* (*GWAS*), where typically hundreds of thousands or even millions of variants are screened across the complete genome, without any prior hypotheses on

the location of the DS variants [99, 64]. Before the recent advances in genotyping technology that allow for the screening of the whole genome, association analysis could only be used in studies when the set of candidate regions is already restricted to a small subset of the genome by some prior knowledge, such as linkage analysis results. Such studies include *candidate gene* and *fine-mapping studies*, which consider only variants within a single gene or within a candidate region containing approximately $5 - 50$ genes, respectively [6].

An inherent problem in GWA studies is the problem of *multiple testing* of a very large number of hypotheses (markers), which almost invariably leads to the discovery of a number of *false positive* associations. For instance, when testing 100.000 markers, using a standard significance threshold of 0.05 for each individual test would lead to approximately 5000 actually non-associated markers to be declared as associated (in addition to any genuinely associated ones). There are methods to correct the p-values to account for multiple testing [82, 66], but rigorous application of these methods leads to only the strongest associations being accepted as statistically significant. In current gene mapping research, the interest is usually in *complex diseases* [67] which are affected by a combination of multiple genes and environmental factors. For complex diseases, there are typically no "causal" mutations, but instead the contribution of each DS variant is only moderate or weak. Therefore, applying stringent correction for multiple testing may lead to most or all actual DS variants being missed, and in practice a larger set of putative variants (almost certainly containing many false positives) has to be included for further examination. Instead of using a pre-defined significance level to rule out false positive findings, an alternative possibility is to determine a *false discovery rate* [94], which allows the markerwise significance threshold to be set dynamically based on the set of observed p-values and the desired ratio between the number of true positive and false positive findings within the set of significant associations.

As the set of DS variants discovered by a GWA study necessarily contains many false positives, further studies of the putative variants are needed to validate the observed associations. Possible approaches include genotyping these variants in additional independent samples in order to identify the real associations, or sequencing the regions around the (genotyped) putative variants to discover the actual variants increasing susceptibility to disease. As these follow-up studies are expensive, validation and prioritization of putative disease genes based on existing biological data is required before proceeding with the study. These will be the topics of Chapters 4 and 5, respectively.

2.7 Study design in association analysis

A gene mapping project begins with a study design phase (Figure 1.1, upper-left corner), where the goal is to specify with a way of conducting the study that maximizes the power to detect genetic variants, constrained by the available resources. Issues constituting the design of a study include how the trait of interest is defined, how the subjects to be studied are ascertained, what genotyping method is used (this typically determines the set of studied variants), and how the relationship between measured genotypes and the trait of interest is analyzed. Expensive or otherwise constrained components of the study include the ascertainment of subjects, collecting phenotype data and performing the genotyping and analysis. We will only briefly address the concept of study design here; for reviews on study design in association analysis, see e.g. [67, 4, 64].

Statistical power

Ideally, the study design phase should include analytic calculations or simulation studies to ensure sufficient statistical power for detecting the presence of genetic variants affecting disease susceptibility, and also for assessing the accuracy of localizing these variants. In gene mapping, statistical power refers to the probability of detecting the presence of a genetic variant affecting disease susceptibility, using a statistical test of significance with a predefined significance threshold. Inadequate design may lead to insufficient power of detecting DS variants. In terms of the gene-mapping pipeline (Figure 1.1), this means that there will be less true positives within the set of putative DS genes passed as input to the latter part of the pipeline. As a consequence, either a larger fraction of true positives will be missed altogether, or the significance threshold has to be made less strict, increasing the number of considered putative genes and thus making it more difficult to identify the true positive ones later in the pipeline.

Sample ascertainment and haplotyping

Association studies are typically built around case–control groups or families. In case–control studies, both *cases* (affected individuals) and *controls* (healthy individuals) are sampled independently from a population. In family-based studies, instead, cases are sampled from the population, their parents are also recruited and genotyped, and control genotypes are then formed by deducing the non-transmitted parental haplotypes using trio haplotyping; these are called *pseudo-controls* (see Paper II for details).

The major difference between the family-based and population-based designs are as follows: (1) the formation of the control group, (2) haplotyping and (3) effective sample size. In case–control studies controls are separate individuals, while in family studies pseudo-controls are used. The use of pseudo-controls or siblings makes family-based studies robust against *population stratification*, systematic differences in allele frequencies between subpopulations that may cause false associations if the cases and controls of the study are ascertained from different subpopulations [28]. On the other hand, the use of pseudo-controls may cause imbalance between the number of cases and controls, as parents may also be affected with the disease. Furthermore, with late-onset diseases, the trio-based option may not be usable at all, as parents may not be available for genotyping.

In the family-based design, haplotypes can be straightforwardly determined by comparing the genotypes of the parents and offspring to see which alleles have been inherited from which parent, while in the population-based approach haplotypes must be determined using statistical inference. While the family-based option does not suffer from errors introduced by statistical inference, it leaves some alleles missing, as trio-based inference cannot be performed for markers where all genotypes are heterozygous for all the members of the trio. This may be a problem in downstream analysis.

Finally, the effective sample size differs between the two approaches: in the trio-based design, the transmitted haplotypes of the parents are duplicates of the haplotypes of the children, which means that the population-based design achieves a larger effective sample using the same genotyping costs, increasing the power of the study.

A method for statistical haplotyping suitable for large-scale association studies is presented in the next chapter. In Paper II, we present an empirical simulation study addressing the power and accuracy of gene mapping study designs as a function of sample ascertainment method, effective sample size, and haplotyping method. Results of Paper II are discussed at end of the next Chapter.

Chapter 3

Computational haplotyping and its use in gene mapping

As discussed in the previous chapter, haplotypes are the unit of inheritance, and it is beneficial to perform gene mapping using haplotypes instead of genotypes. However, only genotype data is readily obtainable by laboratory methods, while haplotypes are considerably more difficult and expensive to measure directly. When the genotyped subjects are related, it is relatively straightforward to infer most of the haplotypes by considering the inheritance of alleles within the pedigree, deducing for each allele which parent it has been inherited from. Fortunately, even when the genotyped subjects are unrelated, LD between nearby markers usually makes it possible to infer the haplotypes with a reasonably high accuracy. This process is called *computational haplotype reconstruction*, or more shortly *haplotyping*. In this chapter, we will review the computational haplotype reconstruction problem and outline our statistical haplotype reconstruction approach, *HaploRec*, introduced in Paper I. In the end of this chapter, we will also discuss the relevance of haplotypes in gene mapping.

3.1 Computational haplotype reconstruction

The computational haplotype reconstruction problem is stated as follows: given a set of genotypes of unrelated individuals from the same population, the task is to output the most likely pair of haplotypes for each individual. The number of all possible haplotype pairs for a given genotype is exponential in the number of markers, and assuming the markers were completely independent of each other, it would be impossible to say which of the possible solutions is the correct one. Fortunately, the haplotypes within a

population usually share some genetic history, which means that there usually is strong LD between nearby markers. This LD can be exploited by computational haplotyping approaches to greatly narrow the set of possible solutions.

Formal description of the problem We assume a sequence (map) M of ℓ markers $1, \ldots, \ell$ and denote the set of alleles of marker i by A_i. The set of possible (unordered) allele pairs for marker i is denoted as $\mathcal{A}_i = \big\{\{a_1, a_2\} : a_1, a_2 \in A_i\big\}$. A *haplotype* H over M is then a sequence of alleles: $H \in A_1 \times A_2 \times \ldots \times A_\ell$, and a (multi-marker) genotype G over M is a sequence of unordered allele pairs: $G \in \mathcal{A}_1 \times \mathcal{A}_2 \times \ldots \times \mathcal{A}_\ell$. For SNP markers, $|A_i| = 2$. Assuming alleles are labeled "1" and "2", SNP haplotypes are vectors in $\{1,2\}^\ell$ and SNP genotypes are vectors in $\big\{\{1,1\},\{1,2\},\{2,2\}\big\}^\ell$.

The allele of haplotype H at marker i is denoted by $H(i)$. Similarly, the allele pair of a genotype G at marker i is denoted by $G(i)$. An allele pair $\{a_1, a_2\}$ is said to *homozygous* if $a_1 = a_2$; otherwise it is said to be *heterozygous*. Given a pair of haplotypes $\{H_1, H_2\}$ and a genotype G such that $G(i) = \{H_1(i), H_2(i)\}$ for all i, we say that $\{H_1, H_2\}$ is *consistent* with G, or that $\{H_1, H_2\}$ is a (possible) *haplotype configuration* for genotype G. Two haplotypes determine a unique consistent genotype in the obvious way. A genotype, on the other hand, can have several haplotype configurations. For a genotype G with k heterozygous markers $\big(i.e., k = |\{G(i) = \{a_1, a_2\} \mid a_1 \neq a_2\}|\big)$, there are 2^{k-1} different haplotype configurations. The set of all possible haplotype configurations for a genotype G is denoted by C_G, where $|C_G| = 2^{k-1}$. A genotype that has only one or no heterozygous markers has only one possible haplotype configuration, and is called *unambiguous*. Analogously, a genotype with more than one heterozygous marker is called *ambiguous*.

The set of input genotypes is denoted by $\mathcal{G}$. The haplotype reconstruction problem is now defined as finding the most plausible haplotype configuration $\{H_1, H_2\} \in C_G$ for each genotype $G \in \mathcal{G}$. Here, the interpretation of "most plausible" is left open; it will depend on the assumptions made by each different haplotyping method.

3.1.1 Haplotyping short regions of strongly linked markers

Clark's algorithm When considering haplotypes consisting of only a small number of markers in strong LD, the set of different haplotypes occurring in the population is usually very small, perhaps only 4 or 5 different haplotypes. In this case, it is often possible to solve the problem by finding a small set of haplotypes that are consistent with the observed genotypes.

The first computational approach that uses this idea was published by Clark [16]. Clark's algorithm uses as starting point a set of haplotypes $\mathcal{H}$ trivially obtained from the unambiguous genotypes in the input. The ambiguous genotypes are then resolved one by one, by finding a compatible haplotype configuration $\{H_1, H_2\}$, where at least one of the haplotypes $\{H_1, H_2\}$ is found in the set $\mathcal{H}$, and adding the complementary haplotype to the set $\mathcal{H}$. This process is repeated until all genotypes have been resolved, or no more genotypes can be resolved. A drawback of Clark's algorithm is that the order in which the genotypes are resolved affects the result. It is also possible that the algorithm fails to resolve all the haplotypes. Clark proposes to run the algorithm multiple times with different random seeds, and then use results from the order of execution which maximizes the number of genotypes resolved.

Pure parsimony The problem can be formulated with a more explicit objective function as the combinatorial *pure parsimony* problem [37], where the goal is to find the smallest set of haplotypes that is consistent with the set of observed genotypes. The pure-parsimony formulation can be approximated efficiently for small data sets (30 markers, 50 individuals), by using linear programming. It is, however, impractical for larger data sets.

Perfect phylogeny A *perfect phylogeny* [36] approach to haplotyping attempts to capture the evolutionary process leading to the present-day haplotypes. This approach works by constructing an evolutionary tree (Figure 3.1) such that the observed genotypes are consistent with the haplotypes in the tree. The current haplotypes are assumed to be derived from a common ancestor using an *infinite sites* mutation model, where each marker locus is assumed to mutate exactly once during the history.

Multinomial haplotype model A major drawback of the approaches above is that they do not utilize information about the frequencies of different haplotypes. The statistical approach to haplotype inference is to jointly model the distribution of the haplotypes in the population and estimate the probability of compatible haplotype configurations based on the model. To make the modeling feasible, the simplifying assumption of random mating within the population is usually made, stating that the two haplotypes of each individual are independent of each other. In the basic statistical haplotyping approach [29, 68], haplotype probabilities are simply represented by a multinomial model, where all possible haplotypes are enumerated and

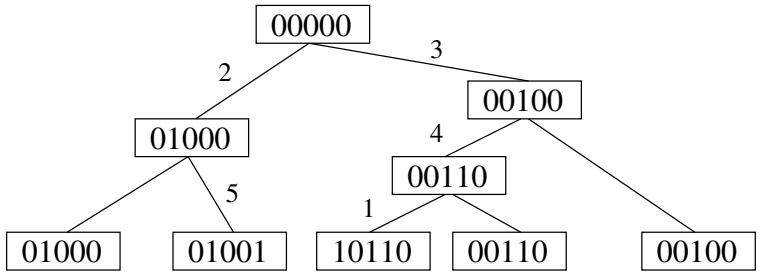

Figure 3.1: A phylogenetic tree for 5 haplotypes. The boxes on the last row represent the haplotypes underlying the observed genotypes, and the other boxes represent ancestral haplotypes. The numbers on the lines joining haplotypes indicate positions where a mutation has occurred in the lineage joining the two haplotypes. Note that some, but not all shown ancestral haplotypes are present in the current-day population (last row).

each haplotype is assigned an independent probability $P(H)$, corresponding to the estimated frequency of that haplotype in the population. Under this model, the likelihood of the observed genotype data $\mathcal{G}$ is defined as

$$L(\mathcal{G}) = \prod_{G \in \mathcal{G}} \sum_{\{H_1, H_2\} \in C_G} P(H_1)P(H_2). \tag{3.1}$$

Here, the genotype data is interpreted as a random sample of haplotype pairs drawn from the haplotype distribution specified by the model. In practice, neither the parameters (haplotype frequencies $P(H)$) nor the individual haplotypes are known in advance, and the model has to be estimated from the same genotype data that is to be haplotyped. The haplotype probabilities that maximize Equation 3.1 can be found using the *Expectation maximization (EM)* algorithm [29]. Once the model parameters are estimated, each genotype is resolved into the pair of haplotypes $\{H_1, H_2\}$ that has the maximal probability $P(H_1)P(H_2)$ among all haplotype pairs consistent with the genotype.

3.1.2 Haplotyping longer regions

All the methods described above are suitable only for very short genomic regions, where recombinations are very rare, and consequently only a few different haplotypes exist in the population. When analyzing a larger number of markers spread over longer genomic regions, it is no longer practical nor meaningful to model the haplotype distribution as a list of distinct

haplotypes. Instead, the haplotyping method has to account for the recombinations that have randomly fragmented the haplotypes during the evolutionary history. This means that only local regularities are observable in the present-day haplotypes, while completely similar haplotypes rarely occur in different individuals.

Multinomial model with partition ligation The first method to partially address this issue was the *partition ligation* algorithm by Niu et al. [68], which works in a bottom-up fashion so that haplotypes are first reconstructed on short regions, using the EM algorithm based on the multinomial model. For each subject, the B most probable haplotype pairs on each short region are combined into $2 * B^2$ longer haplotype pairs (each pair of haplotypes can be combined in two ways). EM algorithm is then again run on these longer regions, such that the set of possible haplotype configurations for each subject is limited to the $2 * B^2$ pairs formed above. This process is continued so that at each step, B most probable haplotype pairs from each pair of neighboring regions are combined using the EM algorithm, until haplotypes for the whole genotyped region have been obtained. While the partition ligation algorithm is computationally efficient and works relatively well for a moderate number of markers, it still models the complete haplotypes with a simple multinomial model, and cannot accurately model the complex haplotype distributions of longer genomic regions.

Haplotype block models A popular model for handling longer haplotypes is the concept of *haplotype blocks* [22]. The block model assumes that the genome is divided into relatively short regions where haplotype diversity is low, and that these blocks are separated by *recombination hotspots* where recombination occurs frequently, causing the linkage between adjacent blocks to be weak. Haplotype blocks are often estimated from haplotype data by finding a fixed set of block boundaries, and modeling the haplotype distribution within each block separately [22, 55]. Several haplotyping methods based on blocks have also been proposed [27, 51], where the haplotypes and block structure are estimated jointly.

The block model is conceptually simple and also convenient for association analysis, as the haplotypes within each block can be interpreted as multi-allelic markers and straightforwardly tested for association with any method designed for analyzing single markers independently. However, the block model fails to model longer-range LD which can extend over several consecutive blocks. Also, while in some genomic regions there is a clear

block structure, often this is not the case, and shared haplotype fragments do not necessarily follow any fixed block boundaries.

Haplotyping based on frequent fragments Motivated by the aim to more flexibly capture as much as possible of the local regularities present in the haplotypes, we introduced in Paper I the haplotyping method *HaploRec*, which models the haplotype distribution based on the set of *frequent haplotype fragments*, that is all partial haplotypes that occur in the study sample with at least a specified minimum frequency. The idea in HaploRec is to estimate the set of frequent fragments to capture local patterns of LD, and then combine information from these fragments into a probability model for the complete haplotypes. This model can then be used to evaluate the probabilities of alternative haplotype configurations, and to find the most probable haplotype configuration for each subject. HaploRec uses two alternative haplotype probability models: one based on a *variable-order Markov chain* and the other based on segmenting each haplotype into a mosaic of frequent haplotype fragments. We here briefly describe these models; full specifications of the models and the HaploRec algorithm are given in Paper I.

Let $H(i, j)$ denote the sequence, or haplotype fragment, from the ith to the jth marker in a given haplotype H. In the *variable-order Markov model* the conditional probabilities at each marker i are estimated from fragments $H(s_i, i-1)$ of varying length:

$$P(H) = P(H(1)) \prod_{i=2,\ldots,\ell} P(H(i) \mid H(s_i, i-1)). \tag{3.2}$$

The length of the context $H(s_i, i-1)$, and thus the order of the Markov chain, is individually adjusted for each position and each haplotype by choosing the longest matching context that exceeds a predetermined minimum frequency.

The *segmentation model* considers each haplotype as a sequence of independent, non-overlapping fragments, and defines the probability of a haplotype to be the product of fragment probabilities. Instead of only finding a single segmentation, the probability for the complete haplotype is obtained by averaging over the set $\mathcal{S}$ of all possible segmentations:

$$P(H) = C^{-1} \sum_{S \in \mathcal{S}} q^{|S|-1} \cdot \prod_{(s_i, e_i) \in S} P(H(s_i, e_i)), \tag{3.3}$$

where S is a segmentation of H into (non-overlapping) segments (s_i, e_i), $q \in [0, 1]$ is a parameter determining how much more weight is given to

segmentations with a smaller number of fragments, $|S|$ is the number of segments in segmentation S, and C is a normalization factor.

HaploRec is implemented using an EM algorithm that alternates steps for reconstructing the haplotypes of each subject and estimating the model parameters (set of frequent fragments). At each parameter estimation step, the algorithm uses data mining techniques [90] to efficiently discover and store the set of frequent haplotype fragments found in the current estimate of the haplotype configurations. At each haplotype reconstruction step, dynamic programming is used to efficiently compute the average over all possible segmentations (Equation 3.3) for each potential haplotype and find the most probable pairs of haplotypes, based on the current estimate of the model parameters.

In contrast to most previous approaches, long-range LD between markers is not required for HaploRec to work, but it can be utilized where it does exist. Our segmentation-based model bears some resemblance to previous methods which combine haplotype block finding and haplotyping [27, 51]. However, whereas these models place universal block boundaries across the whole population, our model averages over all possible segmentations for each haplotype separately, without any fixed block boundaries. This makes it possible to utilize LD also when a clear block structure does not exist, and also enables utilizing LD between blocks in the presence of block structure.

Concurrent and later work on haplotyping Concurrently with the development of HaploRec, several methods which model haplotypes using a fixed number founder haplotypes (or clusters) have been proposed [75, 81], where the cluster memberships are allowed to change continuously along the chromosome, according to a hidden Markov model. Similar to HaploRec, these models allow for both block-like patterns and gradual decline of linkage disequilibrium with distance. However, HaploRec can more flexibly represent regularities at various resolutions and various stages of the population history, as it is not limited to a fixed set of founder haplotypes, enabling it to more accurately model rare shared fragments.

After the publication of HaploRec, also several other block-free methods based on variable order Markov chains [12, 76] have been published. However, the author is not aware of any significant improvements on accuracy or efficiency in comparison to HaploRec; in contrast, in the more recent experiments reported in the PhD thesis of Pasi Rastas [74], and an article by Rosa et al.[77], HaploRec was found to be very competitive with state of the art methods (see also Section 3.2 below).

3.2 Summary and discussion of haplotyping results

The main motivation for developing the HaploRec algorithm was to obtain an efficient and scalable haplotyping method suitable for whole genome association studies. Particular design goals included (1) the ability to jointly model a large number of potentially sparsely spaced markers while capturing also long range LD in the data, and (2) computational efficiency, in practice time complexity roughly linear with respect to both the number of markers and subjects. The experiments of Paper I, as well as later work [74] indicate that these goals were reached to a large extent. In the following, we will briefly describe the experimental setting of Paper I, and discuss its results. In Paper II, we applied HaploRec in a simulation study assessing different gene mapping strategies, validating its applicability in an actual gene mapping application. The experimental results of Paper II will be discussed in the next section.

Experimental setting The experiments of Paper I consist of evaluating haplotyping accuracy and running times of HaploRec and state-of-the-art (at the time) methods, using simulated data sets with varying properties. Five publicly available haplotyping programs were used for benchmarking: fastPhase [81], Gerbil [51], Phase [85], PL-EM [73] and Snphap [18]. The varied quantities include the number of subjects and markers, the distance between markers, a well as the fraction of genotyping errors and missing genotypes. Some experiments were also performed using real data from the HapMap project [87], to validate and complement the results obtained from simulated data.

We used Hudson's coalescence simulator [43] to simulate genotype data sets of 25 to 1000 subjects. The data had between 5 and 500 markers, with average marker spacings between 6.6 and 166 kb, within chromosomal regions having lengths between 166 kb and 16.6 Mb. These marker distances correspond to genome-wide studies having 20,000 to 500,000 markers in the whole genome. For details on the data simulation procedure, see the Methods section of Paper I. As an accuracy measure, we used *switch accuracy*, which is defined as the fraction of neighboring phases (between each pair of consecutive heterozygous markers) reconstructed correctly. All experiments were run separately for 10 independently simulated data sets, over which the accuracies were averaged.

Summary of results In the rest of this section, we summarize the results of Paper I, referring to the figures of Paper I. In the experiments, HaploRec scales in a unique way to large data sets: its accuracy improves with both the number of markers (Figure 3) and the sample size (Figure 5) in an unparalleled way (Phase being an exception in some aspects), while being robust to increasing the distance between markers (Figure 4). This combination of properties makes HaploRec especially suitable for genome-wide association analysis, where large numbers of relatively sparsely located markers are analyzed for thousands of individuals. In such settings, HaploRec can outperform Phase in accuracy while being $2-3$ orders of magnitude faster. Although Phase is very accurate and can also benefit from large samples, it is computationally very intensive, making it unsuitable for genome-wide studies. In the experiments of Paper I, the fastPhase [81] method scales reasonably well computationally, but with large data sets it is clearly less accurate and also significantly slower than HaploRec.

An useful property of the HaploRec algorithm is that its accuracy increases both when increasing the number of jointly haplotyped markers and when increasing the number genotyped subjects. The first property means that it is not an optimal strategy to perform haplotyping independently in short regions; instead, as long regions as possible should be haplotyped jointly, to fully utilize information from longer shared haplotype fragments. These aspects are also linked: longer shared haplotypes are often too rare to be reliably detected from smaller samples, meaning that the utility of having a large number of subjects grows when longer regions are haplotyped jointly, and vice versa.

The ability to benefit from increasing the number of haplotyped subjects is getting more important, as current genome wide association studies require several thousands of subjects to obtain sufficient statistical power [64]. Also, such studies often aim at identifying rare disease haplotypes, which can be modeled by the flexible models of HaploRec. Based on the experiments of Paper I and Rastas [74], it appears that popular methods based on hidden Markov models [81, 75] are not able to benefit from additional information provided by increasing the number of subjects.

Our results also show that the running times of HaploRec scale linearly with the number of markers (Figure 7), and roughly linearly with the number of subjects (Figure 5). This is an important practical property, as both the number of genotyped markers and subjects in typical gene mapping studies have increased rapidly in the recent years. Currently, it is not uncommon to have studies with one million markers and several thousand subjects.

A theoretically optimal haplotyping method would model the population history of the haplotypes, like is done in the perfect phylogeny method described above for short haplotypes. A biologically plausible method to achieve this in the presence of recombinations is the *ancestral recombination graph* [35]. This method is, however, computationally intractable for all reasonably sized data sets. Phase [85], which is the most accurate of the compared methods, uses an approximation based on coalescence theory, but also it is computationally intractable for modern data set sizes. All practical haplotyping methods are thus trade-offs between computational efficiency and accuracy.

Based on the experiments in Paper I, it appears that already relatively simple statistical models, such as HaploRec, are sufficient to obtain an accuracy comparable with the most accurate method (Phase) based on a more complex model, while being significantly more efficient computationally. Simulation experiments of Paper II comparing gene mapping results obtained with haplotypes reconstructed using HaploRec to ones obtained using the true haplotypes show that probably not much is to be gained by more elaborate modeling in practice.

3.3 Haplotypes in gene mapping

The main benefit of using haplotypes in association-based gene mapping is that they provide more accurate estimates of IBD status between chromosomal segments of different individuals than do single markers, providing more power for detecting associations between marker alleles and the (usually not genotyped) DS allele. Haplotype analyses may also be useful for e.g. capturing epistatic interactions of closely located mutations [15, 34, 49, 24] or identifying rare DS variants that cannot be detected by GWAS analyses [50]. In thesis, we will not discuss these latter possibilities, however, and the rest of this chapter handles the use of haplotypes in detecting the effects of individual mutations in GWAS analyses.

There are two simple practical approaches to using haplotypes in association analysis: using a sliding marker window of haplotypes, or first splitting the chromosomes into haplotype blocks, and analyzing haplotypes within each block separately. In both cases, analysis methods using individual markers are straightforwardly applicable to haplotypes, by interpreting the different haplotypes within each window or block as different alleles.

As discussed in Section 2.7, the choice on whether to base analysis on haplotypes or genotypes is also closely related to other issues in study design, particularly the choice on whether to study trios of related subjects

or unrelated subjects from the same population. With a population-based sample, it is necessary to use statistical haplotyping, while in a trio-based setting the haplotypes can be inferred directly. This choice also has an indirect effect on the sample size that is obtained when the genotyping costs are kept constant, as using trio-based sampling, one third of the genotypes are redundant and cannot be utilized for the association analysis. In Paper II, we evaluate the use of haplotypes in gene mapping by performing a simulation study where we compare trio-based and case–control study designs having equivalent genotyping costs. The main questions addressed by the experiments are as follows:

1. How much do we gain from using haplotypes in the analysis (instead of single-marker genotypes)?

2. What is relative performance of population-based and trio-based sample ascertainment strategies?

3. Do errors introduced by the statistical haplotype inference have an impact on the gene mapping performance?

4. Can the good performance of HaploRec be validated in a gene mapping setting?

To study these questions, we simulated genotypes and occurrence of disease in a population of $100,000$ individuals using three alternative, challenging disease models. Each setting (combination of disease model, sample ascertainment strategy, analysis method and sample size) was studied using 100 independent simulations. Performance under each setting was evaluated by statistical power of detecting the presence of a DS variant in a GWA study, and by the accuracy of locating the DS variant.

Haplotyping in the case–control setting was performed using HaploRec. Gene mapping was performed by three alternative methods: simple allelic association, haplotype association with sliding windows between 1 and 10 markers, and EATDT [60]. Of these methods, EATDT is applicable to trio data only, while the association methods can be used both with trio and case–control data.

The following main conclusions can be drawn from the results of Paper II. First, maybe the most interesting result in the context of this thesis is that using haplotypes statistically inferred by HaploRec is equivalent to using the true haplotypes, in terms gene mapping performance (Table 3 and Figure 1 F). Second, among the considered designs, haplotype association using case–control samples is the most powerful way of conducting the study for all considered disease models, both in terms of statistical power

and localization accuracy. As expected, the effective sample size has a clear effect on both statistical power and mapping accuracy, which naturally favors the case–control setting enabled by statistical haplotype reconstruction. Third, the sample ascertainment method does not have much effect on mapping accuracy. The results suggest that the case–control design is a powerful alternative for the more laborious family-based ascertainment approach, especially for large data sets, assuming population stratification can be controlled.

Although it is convenient to perform haplotype-based association analysis based on small marker windows using either a sliding window or block-based approach, it is beneficial to perform the reconstruction for longer genomic regions jointly, as shown in the experiments of Paper I. Also, it is preferable to do this with a method that can utilize as much of the information present in the data as possible, such as HaploRec. Using HaploRec enabled us to perform haplotype reconstruction for the whole genomic regions considered in a single run, improving the accuracy and efficiency compared to the case of using a sliding window approach also for the haplotyping.

Also several more elaborate association mapping methods have been published in which haplotype reconstruction is an important intermediate step (e.g. [89, 83, 60, 86, 52, 97]). HaploRec can naturally be used in conjunction with any of these methods. However, the experiments we performed for Paper II show that already a very simple haplotype-based association analysis can perform as well as many of these more elaborate methods. Variable-order Markov models much like the ones originally proposed for haplotyping in Paper I have later been used as a component of a gene mapping method [11].

Although haplotypes provide a significant theoretical increase in power to detect associations compared to single-point analyses, haplotype-based analyses have been used relatively little in currently popular GWA studies. This may be due to several reasons, most notably the uncertainty introduced by the computational haplotyping and the more complex interpretation of results. The problem of multiple testing may become more difficult when using haplotype-based analysis, as correlations between overlapping haplotypes are not as straightforward to account for as correlations between individual markers. Moreover, some associations might be detected more readily using individual SNPs while others might only be discovered using haplotypes [84]. For example, if already a single genotyped marker is in perfect LD with the causal variant, haplotypes are not needed to detect the association, and using them may unnecessarily dilute the signal of association.

Haplotypes in genome imputation Maybe the most concrete benefit of haplotyping currently comes from *genome imputation* methods [63] that have become very popular recently. Genome imputation works such that a *reference panel* of genotypes is genotyped with a very high marker density for a smaller set of subjects, and a larger sample of study individuals is genotyped at a small subset of these markers. Then, untyped variants for the study individuals are inferred based on the corresponding markers in the haplotypes of the reference panel, by matching the haplotypes of the study samples to the reference haplotypes, and filling in the missing values from the corresponding markers in the reference haplotypes. This makes it possible to test a significantly larger number of markers for disease association with reasonable cost, while avoiding the additional complexity caused by haplotype-based association analysis. While single-point analysis using imputed genotypes provides some of the benefits of haplotype analyses (namely, the additional information from using the imputed markers that may contain the DS variant or be in stronger LD with it than any single genotyped marker), it cannot capture potential interactions between nearby DS variants, as do haplotype analyses [9]. Also, haplotype reconstruction is a crucial step in imputation studies: both the reference and study samples need to be haplotyped to enable the imputation. We hypothesize that HaploRec could also be suitable for performing the haplotype reconstruction step needed in imputation analyses.

Chapter 4

Biomine: an integrated graph database and search engine

The topic of this chapter is Biomine, an integrated database of biological relationships derived from public biological databases. In this thesis, Biomine forms the basis for the refinement phase of the gene mapping workflow described in Section 1.1. Biomine was introduced in Paper III, and further developed in papers IV and V.

We begin this chapter by describing the background and motivation for developing Biomine. We then describe in Section 4.2 the data model and types of data integrated by Biomine. In Section 4.3 we describe the graph-based proximity measure which is a core component in utilizing the database for practical applications. The use of the proximity measure for biological link prediction is demonstrated in Section 4.4, which summarizes the link prediction experiments from Paper IV. Finally, we outline the use of Biomine as an explorative graph query engine for the discovery and visualization of relationships between biological entities, such as genes. Use of Biomine for disease gene prioritization will be covered in the next chapter.

4.1 Background and overview

The traditional gene mapping paradigm ignores the existing wealth of knowledge about genes and their relationships to diseases and other genes. Such knowledge can be used to explore the network of biological relationships surrounding putative disease genes, and also to automatically prioritize the putative predictions.

A large amount of background data is readily available in public biological databases. The relevant data includes e.g. previously known gene-

phenotype associations, annotations of gene function, and protein interactions. This background data is typically scattered across multiple source databases, each accessed with its own set of query interfaces. Further, they typically only provide localized access to the data, that is, one can query the links and attributes for a single entity (e.g. gene). More complex queries, such as finding chains of relationships linking two or more entities are usually not supported. To enable more global analysis of such data, the data needs to be integrated and made accessible under a uniform query interface.

The idea in the Biomine project[1] was to develop a system that integrates data from diverse biological databases under a common graph data model and repository. Figure 4.1 illustrates the contents of Biomine by showing a subgraph containing the strongest (indirect) relationships between two genes related to gastric cancer (the query and visualization system will be described in Section 4.5). The goal of Biomine is to enable discovery and evaluation of connections spanning multiple types of relationships derived from different source databases. Such indirect relationships can act as hypotheses for potential, yet undiscovered links, or they can be used to describe and validate relationships obtained from experimental data. For instance, in Figure 4.1, suppose that the leftmost gene (PIK3CA) was already known to be related to the disease under study (gastric cancer), while the rightmost one (KLF6) was one of the putative candidates. The close relationship between these genes (depicted in the figure) could then act as evidence for the involvement of KLF6, increasing its priority in subsequent analyses of the putative genes. A central tool for this kind of link discovery is a general proximity measure derived from the integrated graph. Such a proximity measure can be used for various purposes, e.g. predicting new links, refining predictions based on some external measurements (e.g. association analysis results obtained from genotypes), and guiding visualization of relationships. In Section 4.3, we will describe several such measures.

4.2 Data model and database contents

The Biomine graph database essentially is an integrated index of several biological databases, each with different contents and format. Biomine is based on a relatively simple data model: a labeled graph with typed nodes and edges. Distinct entities of the source databases, such as genes, proteins and gene ontology (GO) concepts, are mapped to nodes in the Biomine, and cross-references between entities, such as GO annotations, gene-protein

[1] `http://www.cs.helsinki.fi/group/biomine`

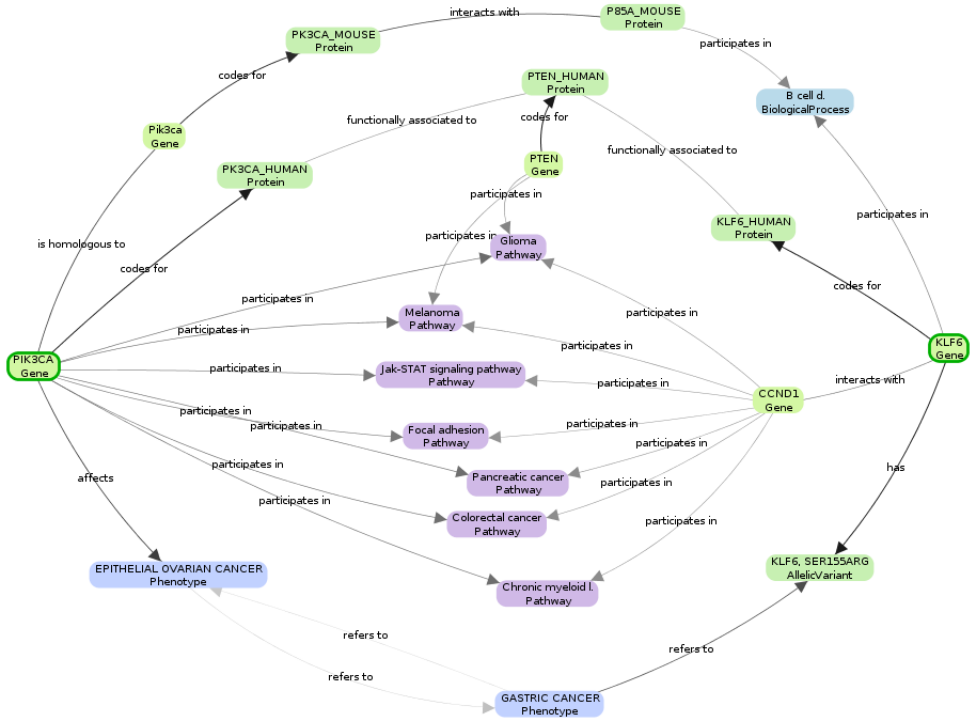

Figure 4.1: A subgraph summarizing the relationships between two genes related to gastric cancer. The query genes, PIK3CA and KLF6, are marked with a green border. The other nodes in the graph are the ones on the strongest paths linking the query genes, as defined by the edge-weighting scheme described later in this chapter.

relationships and protein interactions, are mapped to edges between nodes. Additionally, nodes and edges can have arbitrary attributes, such as names and reliabilities, to represent additional data from the source databases. The chosen data model is deliberately simple, and our aim has not been to comprehensively integrate all data from the source databases. Instead, we just store the identities and various aliases of the entities and the link structure, which is sufficient for performing the prediction and visualization tasks in this thesis.

Biomine is based on data from a small set of representative public biological databases. These databases have been chosen based on the needs of our primary application (disease gene mapping), as well as availability of data in easily accessible format. Next, we briefly summarize the source databases and types of data derived from them. See section “The Biomine

database" in Paper IV for a more complete description.

The core entities in Biomine are genes and proteins, which are derived from the NCBI's *Entrez Gene* [62, 79] database and EMBL's *UniProt* [88] database, respectively. Genes and proteins from other source databases are mapped to ones in Entrez and UniProt and are represented using a single node in Biomine where possible. Physical interactions between proteins are also derived from the UniProt and Entrez databases, which themselves integrate data from several protein interaction databases. In addition, a large number of predicted protein interactions are derived from the String database [47]. Homology relationships between genes of different organisms are derived from the *HomoloGene* database [79].

Phenotype nodes describing diseases and other inherited traits are derived from the OMIM database [38], as are nodes representing allelic variants of genes and cytogenetic gene locations. Genes are linked to related disease phenotype nodes based on annotations in Entrez gene and between-record cross-references derived from OMIM.

Gene Ontology (GO) [39] is a taxonomy of molecular functions, biological processes and cellular components used to annotate genes and proteins. Also the GO categories are represented in Biomine as nodes, and the relationships within the ontology are represented as edges. Biomine also incorporates protein structure classifications derived from the InterPro database [45] in the form of protein families and structural features of proteins, organized in a hierarchy in the same way as GO categories. The annotations linking genes and proteins to the above-described nodes representing GO and InterPro records are derived mainly from Entrez Gene and UniProt, and represented in Biomine as edges.

From the KEGG database [48], Biomine derives nodes corresponding to biological pathways, drugs and other compounds from the KEGG database, each represented by a dedicated node type in Biomine. Relationships of genes to these entities are derived from the KEGG and Entrez Gene databases and represented as edges in the Biomine database.

In addition to the specific biomedical relationships described above, most of the source databases also provide annotated cross-references from their records to articles (PubMed entries) where the corresponding biological entity is mentioned. These cross-references are incorporated in Biomine as article nodes linked to the corresponding biological entities. While this kind of information is generally less informative than more specific biological links, a large fraction of available information is only present in the form of generic article references. The link prediction experiments of Paper IV demonstrate the usefulness of incorporating article references in the

database.

In the current version of the Biomine system, data is extracted and stored for human and four model organisms: mouse, rat, fruit fly and nematode (*c. elegans*). The focus is on human biology, and the additional organisms are included to enable predictions based on potentially more comprehensively annotated homologous genes in these model organisms.

Related work Concurrently with the development of Biomine, several other data integration systems have been proposed in the literature. Of these, most similar to our approach are ONDEX [53] and BIOZON [10], which both collect the data from various sources under a single data store using a graph data schema. In both, the data model is a graph with typed nodes and edges, allowing for the incorporation of arbitrary data sources. In addition to curated data derived from the source databases, both ONDEX and Biozon include in-house data such as similarity links computed from sequence similarity of proteins and predicted links derived by text mining. Biozon provides several types of queries, most interestingly searching by graph topology and ranking of nodes by importance defined by the graph structure. In ONDEX, the integrated data is accessed by providing a pipeline, in which individual filtering and graph layout operations may be combined to process the graph in application-specific ways. BioWarehouse [58] aims to provide generic tools for enabling users to build their own combinations of biological data sources. Their data management approach is rather similar to ONDEX and Biozon, but the data is stored in a relational database with a dedicated table for each data type instead of a generic graph structure. This approach allows database access through standard SQL queries, and is not directly suitable for graph-oriented queries.

4.3 Node proximity in graphs

Many analysis and visualization tasks in biological networks are based on measuring node proximity. For instance, putative disease genes can be ranked by their proximity to another set of genes already known to be related to the disease [54]. Another application of node proximity is in visualization of relationships between nodes: given a set of query nodes, the task in *subgraph extraction* [30, 91, 41] is to select a subgraph of the original graph that best summarizes the links (paths) between the nodes of interest. A proximity measure can also be used to predict future links in the source databases [59]. While arbitrary links usually cannot be predicted with reasonable accuracy based on the graph data alone, the link

prediction problem using balanced sets of positive and negative examples nonetheless provides a convenient framework for systematically evaluating different proximity measures and optimizing their parameters, as we demonstrate in Paper IV.

In the following, we summarize our work on general node proximity measures derived from the graph structure and weights of individual links. This work has been originally presented in Papers III and IV.

4.3.1 Weighting of links

Often, analysis of biological networks is done using homogeneous graphs (such as a protein interaction network), where it is natural to use uniform weights for the edges of the network. However, with the heterogeneous Biomine network considered in this thesis, better results can be obtained when the edges of the network are weighted suitably (see e.g. Figure 4 in Paper IV). For example, an edge corresponding to an experimentally verified protein interaction should probably have a higher weight than an association between proteins predicted using only indirect evidence. Similarly, a manually curated annotation about a gene's effect on a disease should be more important than just knowing that the gene and phenotype are mentioned in the same article.

As an example of a second type of edge importances, consider two articles, where one refers to 2 genes and the other one to 20 genes. Since the former article is more specific, the corresponding edges are likely to be more informative. A third and most obvious case of different importances is when a source database specifies a weight or score for a relation such as the confidence of predicted interactions in the STRING database [47].

Following the weighting scheme introduced in Paper III, we formalize the above-mentioned factors as follows.

1. *Relevance.* Each edge type τ has a fixed *relevance coefficient* $q_\tau \geq 0$ representing the relative importance of that relationship type. We denote the relevance of an arbitrary edge e of type $\tau$by $q(e) = q_\tau$. The suitable choice of values for each q_τ is ultimately dependent on the specific application at hand. In Paper IV, we show how to choose the relevances such that they maximize link prediction accuracy. Alternatively, the relevances can also be set manually by the user.

2. *Informativeness.* The *informativeness* $i(u, v) \in [0, 1]$ of an edge (u, v) is measured based on the degrees of its incident nodes. As a simple method to penalize a node u with a high degree $deg(u)$, we take some negative power $deg(u)^{-\alpha}$ of it. Here $0 \leq \alpha \leq 1$ is a parameter

controlling how steeply the informativeness decreases with increasing node degree. The informativeness of an edge (u, v) is then defined by the degrees of its both endnodes:

$$i(u, v) = \sqrt{deg(u)^{-\alpha} \cdot deg(v)^{-\alpha}}.$$

Based on preliminary experiments with different values of α, we by default set $\alpha = 0.25$. (Where needed, this parameter like any other one can be optimized, e.g., by systematically testing different values, possibly in combinations with other parameters. A thorough optimization of all parameters is not within the scope of this thesis.)

3. *Reliability.* The *reliability* of an edge e, denoted by $r(e) \in [0, 1]$, measures how confident we are that the relation (and consequently the edge) really exists. From the STRING database, we obtain a reliability value for each predicted edge e, directly mapped to Biomine as $r(e)$. For edges derived from other databases, $r(e)$ is currently defined to be one, as they contain manually curated information which is expected to be reliable.

We combine these three factors into an overall edge weight $p(e)$ by simply taking their product:

$$p(e) = q(e) \cdot i(e) \cdot r(e).$$

In the next section, we will define general node proximity measures based on the edge weights. The above definition is directly applicable when using random walk as the node proximity measure. However, for probabilistic proximity measures, edge weights need to be in $[0, 1]$, and consequently the following modification is used:

$$p(e) = \min(q(e) \cdot i(e) \cdot r(e), 1).$$

In this case weight $p(e)$ can be interpreted as the probability that e represents an actually existing, relevant and informative relationship.

4.3.2 Node proximity measures

Several measures have been proposed for measuring the proximity of nodes in unweighted graphs. For an experimental comparison of these measures, see Liben-Nowell and Kleinberg [59]. For the weighted graphs considered in this thesis, much less has been published. In papers III and IV, we have used the four following, alternative proximity measures: probability of best

path [Paper III], network reliability [5], expected reliable distance [72] and rooted random walk [59].

Of these four measures, the first three are specifically defined for probabilistic graphs, while the rooted random walk is normally used for unweighted graphs, but can be straightforwardly modified to handle general weighted graphs. We will next review the definitions of the above-mentioned measures.

Probability of best path Each edge e has a *probability* $p(e) \in [0, 1]$ of being "true". Let *path* P consist of edges $e_1, \ldots, e_k$. The path is true only if all of its edges are true, and correspondingly the probability of P is the product of the probabilities of its edges: $Pr(P) = p(e_1) \cdots p(e_k)$.

The simplest possible proximity measure for two nodes $s, t \in V$ is the *probability of the best path*:

$$p_{bp}(s,t) = \max_{P \text{ is a path from } s \text{ to } t} Pr(P). \tag{4.1}$$

An obvious shortcoming of this measure is that it does not take into account other paths between s and t.

Network reliability To specify the next two, more complex proximity measures that are not restricted to considering the single best path, we first define a probabilistic graph model. Let $G = (V, E, p)$ be a probabilistic, or uncertain graph, where V and E are the sets of nodes and edges, and p is a function assigning a probability to each edge. A non-probabilistic graph $g = (V, E_g)$ is a random realization of G if its set of edges E_g is sampled from E according to the probabilities p, i.e., each edge $e \in E$ is selected to be an edge of g with probability $p(e)$, independently of other edges. The probability of a given random realization g is thus

$$Pr(g) = \prod_{e \in E_g} p(e) \prod_{e \in E - E_g} (1 - p(e)).$$

The *network reliability*, $p_r(s,t)$ between nodes s and t is defined as the probability that a randomly picked instance of G contains a path between s and t:

$$p_r(s,t) = \sum_{g | s \text{ and } t \text{ are connected in } g} Pr(g). \tag{4.2}$$

Expected reliable distance Given a graph g sampled from G, we denote the shortest-path distance (measured as the number of edges) between s and t by $d_g(s,t)$. The *expected reliable distance* [72] is now defined as the expected shortest-path distance, computed over all instances g in which a path exists between s and t:

$$d_{ER}(s,t) = \frac{1}{p_r(s,t)} \cdot \sum_{g|s \text{ and } t \text{ are connected in } g} Pr(g) \cdot d_g(s,t). \tag{4.3}$$

The expected reliable distance reflects the expected proximity of nodes s and t, but does this on the condition that they are connected.

Random walk with restart As the final proximity measure, we consider a symmetric, weighted version of a standard random walk stationary distribution score with restarts [59]. We first define a directed version of the score. A random walk starts at node s. It then iteratively moves to a random neighbor of the current node, such that the probability of traversing edge e is proportional to the edge weight $p(e)$. Additionally, there is a constant probability β of returning to the initial node s at each step, instead of traversing an edge. The directed version of the score, $d_{RW'}(s,t)$ is defined as the stationary distribution probability of the walker being at node t after indefinitely many iterations. The final, symmetric version of the score is defined as the average of the corresponding directed scores:

$$d_{RW}(s,t) = \frac{d_{RW'}(s,t) + d_{RW'}(t,s)}{2}. \tag{4.4}$$

The probability of best path and network reliability measures are used in the link discovery experiments of Paper III, and network reliability is found to give slightly more accurate results. In Paper IV, we evaluate the four measures in the task of link prediction. In these experiments, random walk is found to give most accurate results. Rest of the experiments in Paper IV are performed using the random walk measure. The public www-based visualization engine of Biomine (described in Section 4.5) uses probability of best path to select the displayed subgraph.

4.3.3 Statistical significance of links

While the proximity measures defined above can be used for ranking of putative relationships, their values may be difficult to put into perspective. The properties of a node, especially its degree, and more generally the local network topology around the node have a large effect on the expected

proximity to other nodes: nodes with more links are generally expected to be more proximal to randomly chosen other nodes than are ones with fewer links. This should ideally be taken into account when using proximity values in practical applications.

In Paper III, we measure the statistical significance of observed node proximity between a given node pair (s, t) by comparing it to the distribution of proximity values between randomly chosen node pairs in the graph. Probability of best path or network reliability are used as alternative test statistics, and the null distribution of proximity values is obtained by randomly sampling pairs of nodes having similar degrees with s and t.

We consider two alternative null hypotheses:

1. Nodes s and t of types τ_s and τ_t are not more strongly connected than randomly chosen vertices s' and t' of types τ_s and τ_t having similar degrees with s and t, respectively.

2. Vertex s of type τ is not more strongly connected to vertex t than a randomly chosen vertex s' of type τ having similar degree with s.

The choice between the two null hypotheses depends on what we are testing. In a symmetrical case, e.g. testing for significance of connection between two candidate genes, the first null hypothesis is appropriate. If the roles of the vertices are asymmetric, as in testing for the connection from a set of candidate genes to a single phenotype, the second null hypothesis should be used. In the experiments of Paper III, we have applied the first null hypothesis to assessment of gene–gene links, and the second one to assessment of gene–phenotype links.

For null hypotheses 1 and 2, the null distribution can be estimated by sampling pairs of vertices (s', t') (Null hypothesis 1) or single vertices s' (Null hypothesis 2), and computing the proximities $p(s', t')$ or $p(s', t)$, respectively, for all pairs in the sample (here, p is one of the four above-defined proximity measures). The p-value for the connection between s and t is then the proportion of samples having at least as high proximity as the one observed for (s, t). Because vertices of the same type may have greatly varying degrees, we only sample vertices s' and t' that have degrees close to (but not necessarily identical with) s and t, respectively. If several hypotheses are tested (several candidate genes, for example), the resulting p-values should be adjusted accordingly to account for multiple testing.

4.4 Link prediction

In Paper III, we performed experiments for discovering indirect links between nodes, using a preliminary version of the Biomine database. We performed experiments to find suitable values for the edge type-specific relevances q_τ and the informativeness parameter α that together define the weighting of edges. In these experiments, probability of best path and network reliability were used as alternative proximity measures. Promising results were obtained in discovering gene-phenotype links. However, the evaluation setting was somewhat problematic, as trivial links (ones directly reflecting the same information that is to be predicted) can bias the experiments. As there is no systematic way to avoid trivial links in the experimental setting of Paper III, we resorted to removing all paths with length below 3 as trivial.

To avoid this problem, in Paper IV we performed experiments using a *link prediction* setting where the goal is to predict pairs of nodes that will be connected by a edge in a future version of the graph, based on currently existing indirect links. In practice this was done by using two temporally separated versions of the Biomine database. This setting provides a systematic framework for comparing different proximity measures and adjusting their parameters. We next overview the link prediction results of Paper IV.

Experimental setting We performed experiments for predicting two types of links: protein interactions and phenotypic relationships of genes. A three-year old version of the Biomine database was used to predict the appearance of new links (ones that exist in the current database version, but not in the old version). The tests are based on sampling a set of *positive instances*, that is node pairs that are linked in the current database version, but not in the old one, and an equal number of randomly chosen *negative* instances obtained by randomly pairing the nodes appearing in the positive instances, excluding ones that are linked in the current database version. The positive and negative sets are combined, and the goal is to identify the positive node pairs. This is done by simply ranking the potential links (node pairs) according to their proximity in the old database version. The ability to discriminate positive instances from negative ones is then evaluated using *ROC analysis* [31], a generic framework for analyzing and comparing classifiers. The methods are compared by plotting ROC curves and computing the AUC (area under ROC curve) as a composite statistic to measure the performance of each tested prediction method. Additionally, the statistical significance of the differences between the AUC of different methods is evaluated by computing p-values using the web-based

ROC analysis tool StAR [96].

We first compared the four proximity measures defined in Section 4.3.2, and observed that the rooted random walk measure gave most accurate predictions in both settings. We then performed a simple parameter optimization procedure to approximately optimize weights of different edge types for maximizing link prediction accuracy in the protein interaction and disease gene prediction tasks. Finally, we performed two sets of experiments with the adjusted parameter values, with the goal of validating the proposed approach of combining data from heterogeneous data sources into a single node proximity measure.

The goal of the first experiment was to evaluate how much is gained by suitably adjusting the weights, instead of using uniform edge weights. This was done by comparing results obtained with adjusted parameter values to ones using uniform relevances for all edge types, and also to ones where degrees of nodes were not used for weighting links. The goal in the second experiment was to identify the relative importances of different types of links, and see how much is gained by combining all data, instead of using any of the link types alone. This was done by comparing the prediction accuracy obtained using all data in Biomine to that obtained by using each individual type of link separately (only the most important link types were considered here). Both of these experiments were performed with separately sampled validation sets of positive and negative instances, distinct from the corresponding sets used for optimizing the parameters.

Results Figure 4.2 shows the results from these experiments for the disease gene prediction setting. Roughly similar results were obtained also when predicting protein interactions; for more details on the results, see Paper IV.

The results for the parameter adjustment setting (Figure 4.2, left) demonstrate that adjusting relevances of link types suitably clearly improves the accuracy ($AUC = 0.792$ vs. $AUC = 0.814$, p-value 0.0002). Moreover, the experiment shows that having a degree-based informativeness component in the edge weights is clearly useful as well ($AUC = 0.758$ vs. $AUC = 0.792$, p-value < 0.0001).

For the combined data vs. individual data types setting (Figure 4.2, right), the results show that using all data in Biomine gives significantly better prediction accuracy than any single data type alone, validating the adopted integrative approach.

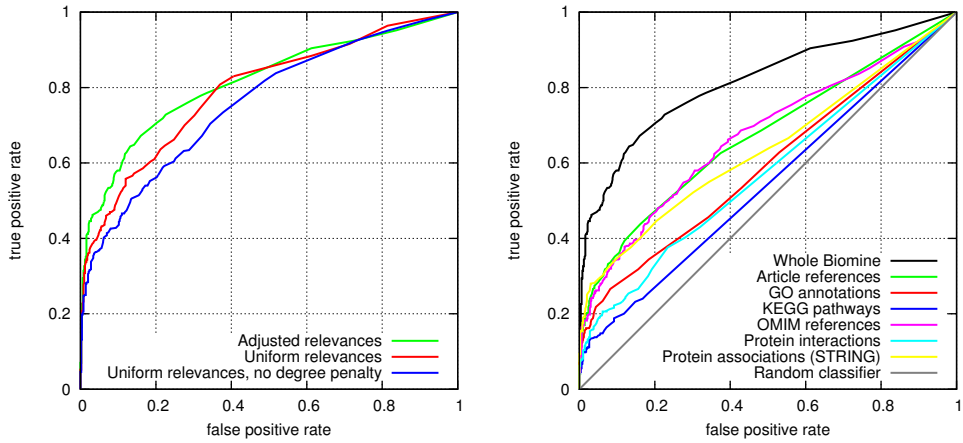

Figure 4.2: ROC curves for predicting gene pairs related to same disease. Left: Effect of weighting edges. Right: Combined data vs. individual data types.

4.5 Biomine as a query engine

Biomine can also be used as an exploratory tool to discover and visualize relationships between graph entities, corresponding to the last step in the gene mapping pipeline illustrated in Figure 1.1. In a gene-mapping project, the researcher will typically want to inspect what is known about the discovered putative disease genes in the original databases and literature. Of particular interest is how the genes are related to each other, and how they are related to already known disease genes. These questions can be answered to some extent by searching for and visualizing paths linking the genes of interest. The search functionality is not restricted to the gene mapping application; also any other type of node represented in Biomine can be used as query nodes.

Basic query functionality can be accessed through a web-based query interface at `http://biomine.cs.helsinki.fi`, which allows for searching for and visualizing connections between given biological entities. This public query interface has been designed to be as simple to use as possible. The user just gives a list of genes (or other entities) of interest as input. The system then produces a subgraph of suitable size, containing paths that aim to best summarize the relationship between the query nodes (see next subsection for details on how the subgraph is selected). All nodes having

the same set of neighbor nodes are collapsed into a single "group node" to be displayed, to increase the number of nodes that can be conveniently visualized as a single subgraph.

The resulting graph (see Figure 4.1 for an example) can be downloaded to the user's computer, or viewed in the browser using a graph visualization program *BMVis*[2]. *BMVis* provides an interactive view of the subgraph, where the user can zoom in and out and focus on different parts of the graph. The user can also move graph nodes by dragging them to new locations with the mouse. BMVis also provides interactive links to the records of the original source databases.

The main query interface described above selects the subgraph to be displayed based on the probability of best path measure (Equation 4.1). There is also another interface that allows the user to explicitly specify the types of queried paths using *context-free grammars*. In the next two subsections, we will describe how the graph to be displayed is selected in both of these query interfaces, respectively. The graph search functionality of the main query interface is not covered by the original publications of this thesis. We nevertheless describe its principles here due to its practical importance in making the Biomine database available and applicable.

4.5.1 Queries based on probability of best path

In this query interface, the input given by the user is a set of *query nodes*, and the result is a subgraph consisting of the most probable paths between any pair of the query nodes. We will first describe the case when there are exactly two query nodes, s and t, and then describe a straightforward extension to the case of more than two query nodes. The abstract problem is to find a *connection subgraph* $H(s,t,k) \subset G$ linking the query nodes that contains both s and t, fulfills some quality criterion and has (at most) a specified number of nodes k. Although more elaborate criteria for selecting connection subgraphs have been studied (see e.g. [41, 91]), we have chosen to use a simple quality criterion based on probability of best paths in the web interface, to obtain response times suitable for interactive use. Loosely speaking, $H(s,t,k)$ is constructed by adding s–t-paths in the order or decreasing probability until the given size limit of k nodes has been reached. As it is simpler to find a subgraph containing all paths with a fixed minimum probability *minp*, instead of directly enumerating the most probable paths, in practice the graph is constructed by finding successively larger graphs restricted by *minp* as follows. We define a *candidate graph* $H(s,t,minp)$,

[2]https://github.com/DiscoveryGroup/bmvis

to contain all nodes on any s–t-path P with probability $Pr(P) > minp$, and all edges which have both their endpoints among such nodes. First, $minp$ is set to a large value. Then graphs $H(s, t, minp)$ are searched by iteratively decreasing $minp$ until the candidate graph contains at least k nodes. At the end each iteration, all groups of nodes sharing the same set of neighbors are collapsed into a single group node each, and each of the group nodes is then counted as one node to be displayed.

Finding candidate graphs A concise way to describe the candidate graph $H(s, t, minp)$ is to define a betweenness measure $bw(v, s, t)$ for all nodes $v \in G$ as the probability of the best s–t-path going through v:

$$bw(v, s, t) = p_{bp}(s, v) \cdot p_{bp}(v, t). \tag{4.5}$$

Here, the probability of the best such a path is simply the product of probabilities of the best paths from v to both s and t. Each candidate graph $H(s, t, minp)$ then consists of all nodes v for which betweenness is at least $minp$:

$$H(s, t, minp) = \{v \in V | bw(v, s, t) \geq minp\}.$$

Fortunately, to find all nodes v with betweenness above a given threshold $minp$, it is not required to compute Equation 4.5 for all nodes of the complete graph G; instead, it is sufficient to consider the *neighborhood* of both s and t up to the path length $\sqrt{minp}$. A *q-neighborhood* $N(s, q)$ of a node s is defined as $N(s, q) = \{v \in V : p_{pb}(s, v) \geq q\}$. All nodes of $H(s, t, minp)$ are contained in the union of the $\sqrt{minp}$-neighborhoods of the query nodes:

$$H(s, t, minp) \subseteq N(s, \sqrt{minp}) \cup N(t, \sqrt{minp}). \tag{4.6}$$

This is because for any node v' not in the union, $bw(v', s, t) = p_{bp}(s, v') \cdot p_{bp}(v, t') < \sqrt{minp} \cdot \sqrt{minp} = minp$, and thus v' is not in $H(s, t, minp)$ either. According to Equation 4.6, the betweenness values only need to be computed for nodes in the union and from these the ones with $bw(v, s, t) \geq minp$ make up the desired candidate graph $H(s, t, minp)$.

Performing neighborhood searches Depth-first search restricted by the current $minp$ threshold is used to find the neighborhoods $N(s, \sqrt{minp})$ and $N(t, \sqrt{minp})$ at each iteration of the search algorithm. Performing these neighborhood searches is the most time-consuming part of the search process. Therefore the search engine has a two-tier architecture: the graph is stored on a separate *cache server* using a concise bit-optimized graph

data structure, which provides efficient batch retrieval for the neighbors of a node, and the actual search logic (including the weighting of edges) is implemented by a Java program that is run separately for each user query. The Java program calls the cache server to perform efficient batch neighborhood queries required for implementing the neighborhood searches. Results of the neighborhood searches are cached, so that performing identical queries in various stages of the iterative deepening process does not cause significant overhead. Overall, the optimized search process allows the search engine to answer typical queries in a few seconds.

Extension to more than two query nodes The algorithm described above can easily be extended to handle a set S with an arbitrary number of query nodes, by simply changing the definition of betweenness measure to consider the maximum of all pairs within the query set:

$$bw(v, S) = \max_{s,t \in S} p_{bp}(s, v) \cdot p_{bp}(v, t) \tag{4.7}$$

There exists also a more flexible query interface that is available only for registered users. This interface allows the tuning the query parameters and storing of queries and query results on the server. Adjustable parameters include specifying weights of different edge types and controlling the amount of degree penalty (see Section 4.3.1), and choosing the size of the result graph. This query interface supports two types of queries: all pairwise connections within a single set of query nodes (similarly as the public query interface, Equation 4.7), and all pairwise connections between two sets of query nodes, excluding any connections within either of the sets (Equation 4.8 below).

$$bw(v, S, T) = \max_{s \in S, t \in T} p_{bp}(s, v) \cdot p_{bp}(v, t) \tag{4.8}$$

4.5.2 Queries based on Context-free grammars

In the previous section, connection subgraphs were defined quantitatively, using the probability of best path for selecting which nodes are to be displayed. An alternative, qualitative approach is to specify the set of interesting paths based on the types of nodes and edges. Instead of all paths connecting two vertices, an investigator is often interested in paths with specific semantics, e.g, paths that suggest a causal relationship or paths that confer similarity.

With labeled graphs, it is natural to base queries on the *path type*—the string of vertex and edge types on a path. A *path class* is a set of path

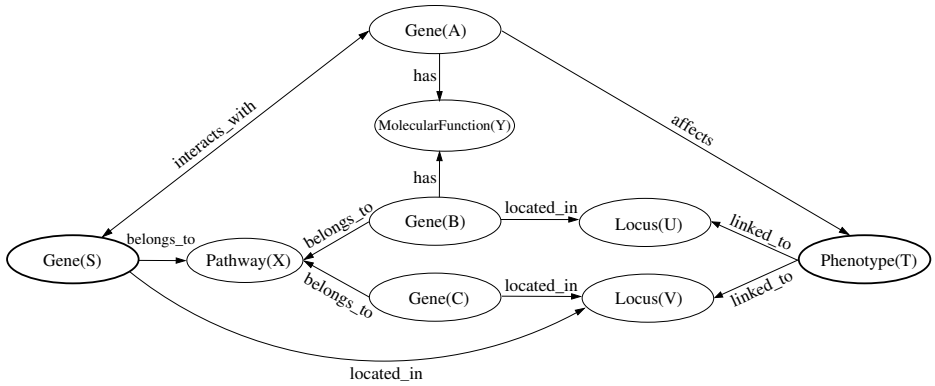

Figure 4.3: A fictional connection subgraph summarizing the link between gene S and phenotype T. The user has specified the source and target vertices S and T, and the class of path types of interest (path types suggesting causal relationship between a gene and a phenotype).

types with shared semantics. Figure 4.3 illustrates the research problem of Paper V: extract a subgraph between two sets of query nodes that consists of paths belonging to a given path class. In this example, the aim is to find paths that suggest a possible causal link between a given gene (S) to a given phenotype (`T`). This is done by defining a path class capturing relationships through intermediate genes (`A`,`B` and `C` in the figure) that are 1) known to be related to gene `S`, and also 2) to be linked to the phenotype T. The first type of relationship is represented by path types `interacts_with` and `belongs_to Pathway -belongs_to`, while the latter type is captured by paths of types `affects` and `located_in Locus linked_to` (here, a '-' is added as a prefix to an edge type to indicate the corresponding inverse edge type).

In Paper V, we introduce the use of *context-free grammars* (CFGs) as a path query language for extracting subgraphs from a large graph. Node and edge types are used as terminal symbols of the grammar, and non-terminal symbols of the grammar correspond to path classes. A CFG consists of a number of *production rules* that specify how paths in a given class are constructed from sequences of node and edge types and other path classes. For example, a rule `A -> c B d` states that concatenating terminal symbol `c`, any path in class `B`, and terminal symbol `d` produces a path in class `A`. The set of strings accepted by a CFG is defined by the productions of a special query path class `QUERY`.

Returning to the query of the example figure, we define three path classes: one (`SIM`) for capturing similarity of the genes, one (`AFF`) for cap-

turing a gene's (potential) effect on phenotype and the query class (`QUERY`) as a combination of the first two. The productions of the grammar are as follows:

(1) `SIM -> interacts_with`
(2) `SIM -> belongs_to Pathway -belongs_to`
(3) `AFF -> affects`
(4) `AFF -> is_located_in Locus linked_to`
(5) `QUERY -> SIM Gene AFF`

Given this grammar, the example query is defined simply by specifying the source and target nodes `S` and `T`, and specifying `QUERY` as the query class. The query system described in Paper V enables both the source and target to be sets consisting of multiple nodes, in which case the result will contain all matching paths between any pair of source and target nodes.

In Paper V, we introduced a modified version of the well-known Earley algorithm [25] to find the subgraph containing all paths between the query nodes that match a given CFG. The basic Earley algorithm performs parsing of strings instead of paths in a graph, so the first task was to extend the algorithm to handle graph data. A relatively simple way to do this is to perform a depth-first search in the graph starting from the source node, and perform the parsing for each possible path separately while doing the search, such that the state of the CFG parsing is maintained separately for each traversed path. The search for each traversed path is terminated whenever the path type is found to conflict with the given grammar, or the target node is reached using a path in the query class. This is not an optimal strategy, as the number of paths to be parsed is potentially exponential in the search depth (length of longest path compatible with the grammar). To make querying large graphs (such as Biomine) feasible, the basic algorithm described above was extended in two respects to reduce the number of states required: (1) bi-directional search is used, which halves the required search depth; and (2) all paths between a pair of nodes having the same path class are collapsed into a single state, instead of using a separate state for each path. According to the experiments of Paper V, these modifications greatly speed up the algorithm, making it applicable to graphs within the size range of Biomine.

A public query interface to Biomine based on context-free grammars is available at `http://biomine.cs.helsinki.fi/cfg/`. This interface allows specifying the grammar, and setting the source and target nodes and maximum search depth. As in the query interface based on path probabilities, the resulting graph can be downloaded to the user's computer, or visualized directly in the web browser using the graph visualization software *BMVis*.

Chapter 5

Graph-based disease gene prioritization

The topic of this chapter is the gene prioritization phase of the gene mapping workflow (right-hand side of Figure 1.1): how to automatically prioritize putative disease susceptibility genes so that further efforts can be focused on the most promising candidates? Genome-wide association studies typically produce a large set of putative genes that appear statistically associated with the disease, while actually only a fraction of these are true positive findings. Before moving on with follow-up analyses of these genes, it is useful to prioritize and filter the list based on what is already known about them in the public databases.

In this chapter we describe how to automatically prioritize putative disease genes, based on the Biomine database and the graph-based proximity measures described in the previous chapter. This chapter is mainly based on Paper IV, where we adapted two existing graph-based disease gene prioritization methods to be used with Biomine. One of these [54] measures proximities of putative genes to a distinct *reference set* of genes (or other suitable graph nodes) already be known to be associated with the disease, while the other [32] uses only the mutual proximities or the putative genes. We also introduced a novel method for performing disease gene prioritization that does not require a reference set. In the rest of this chapter, we first give a short overview of related work on using graphs for disease gene prioritization, and then summarize the gene prioritization methods and results of Paper IV.

5.1 Related work

Although putative disease genes are often prioritized by manually going through the list of genes and using several databases and literature to check what is known about each of them, several methods have also been published for automatic disease gene prioritization. These methods are typically based on the principle that genes similar to ones already known to be related to the disease of interest are to be considered as the most plausible candidates.

The definition of similarity and types of utilized data vary a lot depending on each particular method, the former naturally being dependent on the latter. The most common source of data is protein interactions, but also shared functional annotations and pathways, as well as text mining and gene expression similarity are commonly used.

In the following, we will review some of these methods considered most relevant to the work in this thesis. For a more complete review of such methods, we refer the interested reader to a practical overview covering freely available web tools for prioritizing candidate genes, written by Tranchevent et al. [92]. A review on prioritizing GWAS results is provided by Cantor et al. [13], and a review on utilizing gene networks for research of human diseases is provided by Barabasi et al. [8].

We first list several gene prioritization methods that use only protein interaction data. A trivial approach for predicting potential disease genes is to just assign interaction partners of already known disease-related proteins as candidates [70]. Krauthammer et al. [56] use a slightly more elaborate method, where evidence from known disease genes is propagated to nearby putative genes according to a score based on shortest paths distance, while Kohler et al. [54] use a random-walk-based network proximity measure instead of considering only a direct neighborhood or shortest paths.

In contrast to the previous methods, which only use protein interaction data, Franke et al. [32] and Linghu et al. [61] both construct a network of functional associations ("functional linkage network") using multiple types of integrated source data. They construct the network of functional associations using machine learning techniques to combine evidence from different data sources, using a fixed cutoff value to remove unreliable associations. Franke et al. [32] evaluate each candidate gene based on the shortest path distance to other candidate genes, while Linghu et al. [61] only use information from neighboring genes. Notably, the method of Franke et al. does not require a pre-defined set of known disease genes, but instead prioritizes candidate genes based on their distance to other candidates.

Similarly to Kohler et al. [54], Vanunu et al. [95] also use a random walk

proximity measure, but they also expand the query from the given disease to include other diseases based on phenotypic similarity. Evidence is then propagated in a protein interaction network from all proteins known to be related to any of these diseases.

Hwang and Kuang [46] consider both multiple types of associations (edges) and multiple types of nodes. Instead of integrating them all together into a homogeneous network, they propose methods to propagate information in the network while taking its heterogeneity into account.

An alternative to the network-based gene prioritization approaches outlined above is to directly utilize pre-known sets of functionally related genes [98, 14] as the unit to be tested for disease association. The idea in these methods is that each known pathway (or other pre-defined set of functionally related genes) is tested for relative excess of disease-associated genes, potentially accumulating information from several weaker associations into a single, stronger signal. The results of Chasman [14] indicate that while strong associations remain best identified by conventional association mapping methods, the gene set approach provides a useful complementary mode of analysis for revealing modestly associated genes for complex diseases.

5.2 Using Biomine for disease gene prioritization

5.2.1 Problem definition

We formulate the disease gene prioritization task as a binary classification problem as follows. The set of statistically disease-associated genes from an association study is denoted by S, and the subset of genes that actually increase susceptibility to the disease (*true positives*) is denoted by $S_P \subset S$. The rest, $S_N = S - S_P$, are *negatives*, or false positives of the association study. The task now is to predict S_P (and S_N) by outputting an estimate $\widehat{S}_P \subset S$, using the information contained in the Biomine graph database described in Chapter 4. The key assumption here is that genes affecting the same disease tend to be more proximal in the graph of known biological associations than randomly chosen genes (false positives of the association study).

We consider two alternative formulations of the classification problem:

- **Supervised classification** using only positive instances (see, e.g., George et al. [33] and Kohler et al. [54]). In this easier formulation we are given, in addition to S, a separate reference set S_R of genes already known to increase susceptibility to the disease.

- **Unsupervised classification** (see, e.g. Franke et al. [32]). In this "de Novo" version of the problem, we do not assume information about known disease genes, and only S is given as input.

The idea in the supervised version of the problem is that among the statistically associated genes, those proximal to already known disease genes will be identified as the most promising candidates. In the unsupervised version such existing information is not assumed. Instead, associated genes that are close to other associated genes will be considered as the most likely candidates. In the unsupervised problem, the fraction of associated genes within S must naturally be sufficiently high to perform reliable predictions.

5.2.2 Classifiers for disease gene prioritization

Disease gene prioritization with Biomine can use any the proximity measures defined in Section 4.3.2. We denote the chosen proximity measure by $p(s,t)$, and use it to define three alternative classifiers for the disease gene prioritization task. The first of these is applicable in the supervised classification setting, while the other two are applicable in the unsupervised setting.

Supervised classifier ranks each gene $s \in S$ based on its average proximity to elements of the reference set S_R:

$$score^{A}_{prox}(s) = \frac{1}{|S_R|} \cdot \sum_{t \in S_R} p(s,t). \tag{5.1}$$

This definition is closely related, although not identical to the one used by Kohler at al. [54]. A binary classifier is obtained by setting a threshold q: $\widehat{S}_P = \{s \in S : score(s) \geq q\}$; $\widehat{S}_N = \{s \in S : score(s) < q\}$.

KNN classifier ranks each gene in S by its average proximity to the k nearest other elements of S:

$$score^{B}_{knn}(s) = \frac{1}{k} \cdot \max_{\substack{S' \subset S-s, \\ |S'|=k}} \sum_{t \in S'} p(s,t). \tag{5.2}$$

This definition is motivated by the assumption that random genes (false positives) are not likely to have many close neighbors in S; on the other hand, genes actually related to the disease are expected to be proximal to each other, and thus likely to be found in the set of k nearest neighbors. This definition can be seen as a generalization of the scoring scheme used by Franke et al. [32]. Again, a classifier can be obtained by simply thresholding, as for the Supervised classifier above.

Cluster-based classifier In Paper IV we propose the following new method applicable to the unsupervised version of the problem: find a single cluster $\widehat{S}_P \subset S$ of genes that maximizes

$$score^B_{clus}(\widehat{S}_P) = \sum_{\substack{s,t \in \widehat{S}_P \\ s \neq t}} (p(s,t) - q) = \sum_{\substack{s,t \in \widehat{S}_P \\ s \neq t}} p(s,t) - q \cdot \frac{|\widehat{S}_P|(|\widehat{S}_P| - 1)}{2}. \quad (5.3)$$

Here, q is a parameter governing how proximal a gene should be on average to the other members of the cluster to be considered positive. This definition may be best explained by considering the decision of whether to add a new gene s to some current estimate of $\widehat{S}_P$. Adding a gene s increases the score if the average proximity of s to the genes already assigned to $\widehat{S}_P$ is larger than the constant q. The definition is similar to the *maximum edge-weighted clique* problem [2], and also related to the outlier detection problem in clustering [42], where the goal is to identify objects that are not part of any cluster. The differences to the latter one are that here only one cluster is sought, searching the cluster and handling of outliers is done in a single integrated step, and most genes are expected to be "outliers".

While the gene prioritization task is here presented as a binary classification problem, the first two classifiers are based on computing a score that can also be used for ranking as such. The Cluster-based classifier does not directly provide such a score suitable for ranking, but instead performs a binary classification, given a fixed value for the sensitivity parameter q. A practical way of using the Cluster-based classifier for gene prioritization is to vary the parameter q in order to obtain a number of predicted sets of different sizes. Since the predicted sets are not monotone, that is, a smaller predicted set is not necessarily a subset of a larger prediction, there is not necessarily an immediately implied ranking of genes. However, as a rule, genes that appear in smaller predictions and more often can be given a higher rank.

5.2.3 Experimental setting

In this section, we summarize the gene prioritization results from Paper IV. In these experiments, we used random walk with restart (Equation 4.4) as the proximity measure $p(s,t)$, as it outperformed the other tested measures in the link prediction experiments, and also performed consistently well in comparison to the other methods in preliminary tests with the disease gene prediction problem. To implement the Cluster-based classifier, we used a simple greedy algorithm for finding a cluster of positive genes that approximately maximizes Equation 5.3 (see Paper IV for details).

We evaluated the gene prioritization performance using artificial gene lists simulating lists of top-ranking genes from an association study, where the positive instances come from 110 already known disease gene families compiled by Kohler et al. [54] for the purpose of similar evaluations. Each test case is a gene set S containing a fixed number of "positive" genes S_P from one of the 110 disease gene families, and a control set S_N of "negative" genes chosen at random from the other disease gene families. As a baseline setting, we considered prioritizing gene sets with $|S_P| = 5$ and $|S_N| = 15$, and thus $|S| = 20$ genes in total. We also included more challenging settings by varying both the number of positive and negative instances, with $|S_P| \in \{2, 3, 4, 5\}$ and $|S_N| \in \{5, 15, 25, 35, 45\}$. For each combination $(|S_P|, |S_N|)$, 100 test cases were generated, each containing $|S_P|$ genes sampled from a single disease gene family and $|S_N|$ genes sampled from among the other 109 disease gene families.

Some of the test cases are unrealistically easy to solve using the complete Biomine dataset, since it may contain edges that directly reflect knowledge about the disease gene families. For instance, direct textual references between genes belonging to the same disease family could have been derived from the OMIM database. To include more realistic and challenging test cases, we carried out all experiments also using a version of the database where the most obvious sources of phenotype-related data were excluded (see Paper IV for details).

Results from the two problem settings, the supervised and unsupervised one, are not directly comparable since they are not really practical alternatives: the supervised method should be used whenever a reference set of known disease genes is available, since this helps in ranking; and when such a reference set is not available, there is no other option but to use an unsupervised method. Nevertheless, a comparison between the methods can provide insight on how crucial having a pre-known reference set is for the prediction task.

We compared the classifiers in settings where each (positive) gene is scored using information from the same number of other positive genes by all methods. For example, when $|S_P| = 5$, each gene within S_P is ranked using the 4 other positive genes by the Supervised classifier. For the KNN classifier, each positive gene is ranked using the 4 nearest neighbors (in the optimal case the 4 other positive genes), while in the Cluster-based method, each positive gene can potentially cluster with the 4 other positive genes.

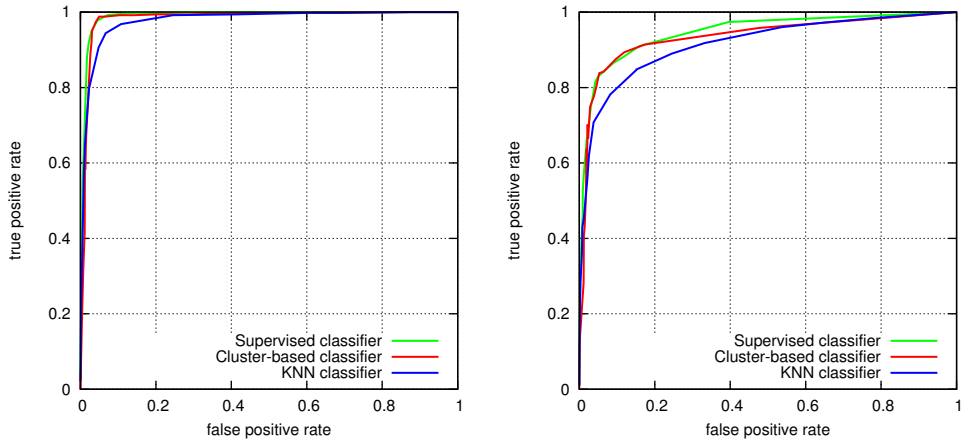

Figure 5.1: Comparison of classifiers using ROC curves. Left: all data. Right: obvious data sources removed.

5.2.4 Summary of experimental results

As a baseline setting, we tested the three proposed classifiers on the problem of identifying the 5 of true disease genes from among a set of 15 unrelated genes, using 100 independent test cases of 20 genes in total. The questions addressed by this setting were (1) how well disease genes can be prioritized using Biomine; (2) how much more difficult it is to prioritize disease genes without a reference set of pre-known disease genes; and (3) without a known reference set, how well the Cluster-based classifier works in comparison with the simpler KNN classifier.

Figure 5.1 reports the results from these experiments as ROC curves averaged over the 100 independent test cases, using either all data in Biomine (left) or the version of Biomine with obvious links removed (right). (See Figure 7 of Paper IV for a scaled version of the figure showing only the beginning of the curves.) There are several observations from this experiment. First, using all data in Biomine (left), the true disease genes can be predicted with a rather high accuracy. Also in the more challenging case of reduced data (right), predictions can be made with reasonable accuracy. In both settings, the Cluster-based classifier obtains practically identical accuracy with the Supervised classifier for most of the ROC space, although it uses less prior information. It is also clearly superior to the KNN classifier. However, in the very beginning of the ROC curve (see

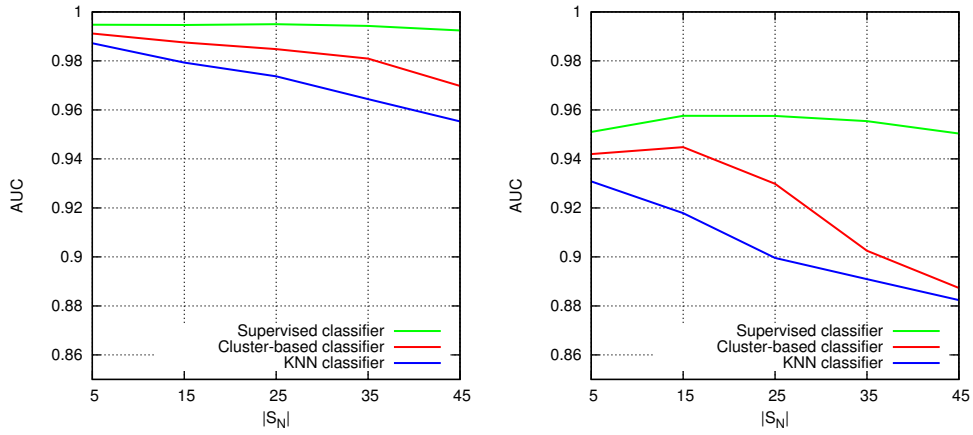

Figure 5.2: Effect of increasing number of false positives on prediction accuracy. $|S_P| = 5, |S_N| \in \{5, 15, 25, 35, 45\}$. Left: all data. Right: obvious data sources removed.

Figure 7 in Paper IV) the Cluster-based method does not perform so well. This is most likely because the beginning of the curve corresponds to stringent (large) values of q, where only a part of the true positive genes are included in the cluster; here, the Cluster-based method is not yet able to utilize information

Varying the ratio of true and false positives We also performed experiments to evaluate how increasing the number of false positives or decreasing the number of true positives affects prediction accuracy. First, we performed a similar experiment as the baseline setting reported above, but varied the number of false positives between 5 and 45 while keeping the number of positives fixed to 5, giving total number of genes to be ranked between 10 and 50 (Figure 5.2). Secondly, we performed an experiment where the number of positives $|S_P|$ varied between 2 and 5, with the number of negatives fixed to 15 (Figure 5.3). Again, we tested the three proposed classifiers, with complete and reduced data separately. Each point in the figures is an average AUC over the 100 independent test cases with a specific $|S_N|$ or $|S_P|$.

As expected, for the Supervised classifier AUC is not affected by increasing the number of negatives, as the ranking of each gene always occurs using a fixed reference set, irrespective of the number of negative genes

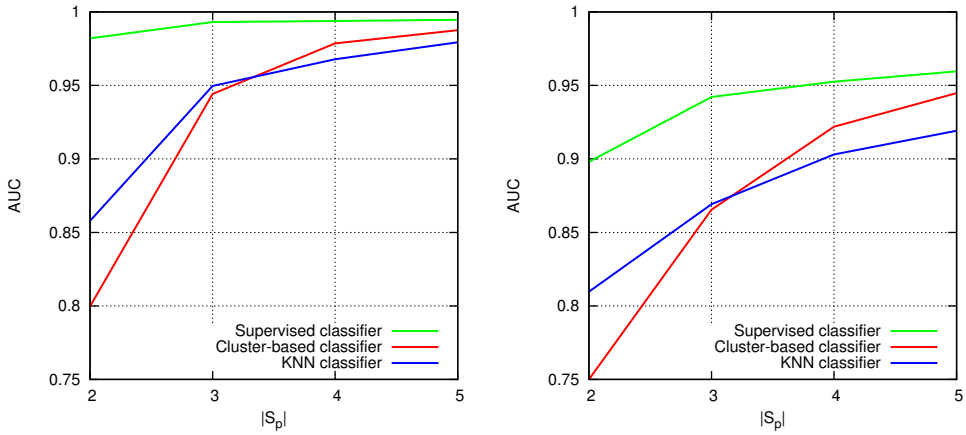

Figure 5.3: Effect of decreasing number of true positives on prediction accuracy. $|S_P| \in \{2, 3, 4, 5\}, |S_N| = 15$. Left: all data. Right: obvious data sources removed.

within the set of genes to be prioritized. On the other hand, the more challenging unsupervised problem becomes more difficult when the amount of negatives is increased, and the accuracy of the other methods decreases quite steeply as the number of negatives increases. However, the Cluster-based classifier is consistently superior over the KNN method, with a clear margin. On the other hand, decreasing the number of positives has a dramatic effect on accuracy, especially in the unsupervised version of the problem, but also in the supervised version.

These results indicate that a reference set S_R is obviously useful, but if one is not available, relatively good predictions can still be obtained with the Cluster-based method if the fraction of positive instances within the set of genes to be prioritized is sufficiently high.

Notably, the Cluster-based method does not work well with the smallest values of $|S_P|$, and is outperformed by the KNN method in these settings. Based on this experiment, it appears that at least 4 positive genes are required in the set to be prioritized in order for the Cluster-based method to be useful.

Chapter 6

Contributions of the thesis

The contributions of this thesis are summarized below.

Paper I

In Paper I, we define two novel statistical models of haplotypes and give an efficient algorithm for haplotype reconstruction using these models, jointly called HaploRec. HaploRec, originally published as a conference paper [26], was the first method that reconstructs haplotypes for longer chromosomal regions simultaneously, such that all handled markers do not need to be in strong LD with each other.

The power to handle genotypes from longer chromosomal regions comes from a novel, flexible use of short patterns of haplotypes to model local LD, while being able to also utilize longer-range LD where it exists. An important benefit of this model is that the accuracy of HaploRec increases both with increasing the number of jointly haplotyped markers and increasing the number genotyped subjects. Computational efficiency of HaploRec is demonstrated by the experiments of Paper I: it is roughly linear with respect to both the number of markers and subjects. To achieve this, we apply algorithmic techniques from the field of data mining in a novel way to efficiently search and store the local patterns of haplotypes used in implementing the models.

The above-mentioned properties make HaploRec suitable for handling data from genome-wide association studies where genotypes of complete chromosomes are haplotyped in a single run, and the number of subjects is very large. Of previous methods, only fastPhase [81] can handle such data, but with lower accuracy. The extensive systematic evaluations in Paper I are also a significant contribution: we are not aware of other studies where

the effect of marker spacing, number of jointly haplotyped subjects and markers, effects of missing data effects and genotyping errors are evaluated systematically.

Paper II

Paper II consists of a simulation-based evaluation of different study designs for haplotype-based association analysis. The main contribution of the paper is a systematic analysis of the gene mapping power as a function of three separate factors: sample ascertainment method, effective sample size, and haplotyping method. The main conclusion is that the case–control design is a powerful alternative for the more laborious family-based ascertainment approach, especially for large data sets. Another important result is that statistically inferred haplotypes reconstructed with HaploRec can be as powerful as the true haplotypes for the purposes of association mapping. An additional result is that the choice of sample ascertainment method in trio-based study designs does not have much effect on mapping accuracy.

Papers III & IV

The topic of Papers III and IV is Biomine, an integrated database of biological relationships, and its applications. Biomine was introduced in Paper III and further developed in Paper IV. The main contributions of these papers are as follows: (1) the Biomine database itself; (2) a novel proximity measure for assessing the relatedness of graph nodes; (3) application of the proximity measure to the tasks of disease gene prioritization and link prediction. An important further contribution of Biomine is the public query engine described in Section 4.5 which is, however, not covered by the original publications.

Data needed in the refinement of initial gene mapping results is typically scattered across multiple source databases. To enable joint analysis of such data from multiple biological databases, the data needs to be integrated and made accessible under an uniform data model and query interface. Biomine provides such a resource.

A fundamental component in the application of Biomine is a graph-based proximity measure for estimating the relatedness of biological objects represented in the graph. Biomine introduces a novel schema of weighting individual graph edges, based on three factors: *relevance* based on edge type, an *informativeness* measure based on node degrees, and *reliability* values extracted from the source databases. The resulting edge weights

are used for constructing a general node proximity measure, by adapting three existing proximity measures: one based on probability of best path (Paper III), another based on network reliability (Paper III) and a third one based on random walks (Paper IV).

In Paper III, the proposed node proximity measures are applied to link discovery in biological databases, i.e., for prediction and evaluation of implicit or previously unknown links between biological entities. Paper III also presents a method for assessing the statistical significance of discovered relationships.

In Paper IV, the node proximity measures are used for the task of link prediction, using two temporally separated versions of the database as a test setting. This setting provides a solid testbed for comparing and evaluating different proximity measures. Results from the link prediction experiments demonstrate that suitably weighting different link types clearly improves prediction accuracy compared to the case of using uniform relevance for all link types. The results also show that using all data in Biomine gives significantly better prediction accuracy than any single data type alone, validating the adopted integrative approach and edge weighting scheme.

Finally, in Paper IV the node proximity measure derived from Biomine is applied to the disease gene prioritization problem. Two existing disease gene prioritization methods are adapted to be used with Biomine, and also a novel method is introduced for the unsupervised version of the gene prioritization problem, based on finding a single cluster of proximal genes from the set of putative disease genes. The gene prioritization experiments show that putative disease genes can be ranked with a reasonable accuracy using Biomine. Best prediction accuracy is obtained when an already known reference set of disease genes is available, but experiments using the novel clustering-based method demonstrate that putative disease genes can also be ranked without an already established reference set, if the number and density of true disease genes in the candidate set is sufficient. In this unsupervised setting, the cluster-based formulation proved more accurate than a simpler approach based on k nearest neighbors score already used by Franke et al. [32].

Paper V

The main contribution of Paper V is the introduction of a novel problem: finding the connection subgraph between two sets of nodes that is induced by the set of paths matching a given context free grammar, where the node and edge types of the graph are interpreted as terminal symbols of the

grammar. Using a CFG enables the investigator to qualitatively define the paths of interest based on the path type, which is an alternative for the qualitative graph proximity measures used in papers III and IV. Such flexibility is needed for the heterogeneous data in Biomine, where the relevance of different path types may be highly dependent on the types of nodes and edges on the path.

We propose a modified version of the well-known Earley algorithm [25] to find the subgraph containing all acceptable paths, by adapting the basic algorithm to handle graph data instead of strings as input. To make querying large graphs feasible, the basic algorithm was improved in two respects to reduce the computational cost of the algorithm: (1) bi-directional search is used, which halves the required search depth; and (2) all paths between a pair of nodes having the same path class are collapsed into a single state, instead of using a separate state for each path. According to the experiments of Paper V, these modifications greatly speed up the algorithm, making it applicable to graphs within the size range of Biomine.

Chapter 7

Conclusions

The overall motivation and framework for this thesis is disease gene mapping, as illustrated in Figure 1.1. Current gene mapping projects are typically based on a genome-wide scan where hundreds of thousands of markers are genotyped from both affected and healthy individuals. This thesis introduces two kinds of computational approaches to aid in such studies. The first main contribution is HaploRec, a computational haplotyping method, and evaluation of its use in the primary analysis of genotype data (left side of Figure 1.1). The second main contribution is Biomine, a database and set of methods for prioritizing and exploring the set of putative disease genes obtained from the primary analysis (right side of Figure 1.1).

Although the topic of this thesis is disease gene mapping, many of the methods included in thesis are not restricted to this particular application. The developed graph query and analysis tools are applicable for any type of labeled graphs, and the graph search and visualization functionalities of Biomine can be used for other biological applications where the integrated databases are relevant. In the rest of this chapter, we will discuss the application of the presented methods in gene mapping and provide topics for further research.

Haplotypes in gene mapping While theory and experiments indicate that haplotype-based analyses have the potential to significantly increase power, most published results to date from genome-wide association studies are still based on testing of individual markers. This is at least partly attributable to the added complexity of statistical analysis and uncertainty introduced by the statistical haplotype reconstruction. Also, genotyping technology has developed at an enormous pace during the writing of this thesis, enabling the study of significantly larger number of markers and

subjects. When the number of tested markers grows, the relative benefit of haplotype-based association mapping methods may decrease, as the actual disease-affecting variants are more likely to be among the tested markers. As discussed in Chapter 3, a recent trend in GWA studies is the imputation of genotypes based on a reference set of completely sequenced genomes, which allows the statistical inference of a much larger number of genetic variants than is contained in the set of physically genotyped markers. While the use of imputation methods decreases the need of directly using haplotypes for association testing, haplotype reconstruction is still an crucial intermediate step in such methods. Also, haplotype analyses are expected to be beneficial due to their ability to capture LD with variants that cannot be detected by current sequencing methods, and due to capturing potential interactions between closely located DS alleles.

The Biomine database The Biomine database provides a foundation for the refinement of phase of the disease gene prioritization and exploration methods included in this thesis. The data model adapted in Biomine is deliberately simple and not tailored to any single application. The benefits of the chosen model are generality, and also the uniform representation of the graph for the proximity measurements used in the prediction tasks on one hand and for visualization of subgraphs on the other hand. While using a more specialized representation tailored for the target application (e.g. disease gene prioritization) could allow for more accurate predictions, promising results have been achieved already with the present version of Biomine. However, to increase its applicability, the database could be improved in some respects. The set of data currently integrated by Biomine is by no means complete, and Biomine in its current form is more a proof of concept than an exhaustive resource of biological data. In particular, Biomine does not have a method for mapping markers to genes, a crucial task in the analysis pipeline of Figure 1 which currently has to be performed as a separate preprocessing step. Such a mapping could be based on e.g. the genomic proximity of markers and genes, and markers being located within known regulatory regions of genes. Other examples of information that could potentially be added to Biomine is the internal structure of metabolic and signaling pathways and known gene regulation relationships. To enable wider applicability, Biomine would also benefit from the possibility to integrate user's own data.

Link prediction and evaluating proximity measures While the primary target application of Biomine in this thesis work has been the prioritization and exploration of putative disease genes, an important intermediate goal was to test how well new relationships can be predicted based on a previous version of the database, and to use this link prediction setting for adjusting the weights of different data types. The main motivation for this experimental setting was to provide an as unbiased as possible testbed for the proximity measures derived from Biomine, rather than to provide any directly applicable methodology for a particular problem. As discussed in the previous chapter, these experiments serve to demonstrate the validity of the chosen data integration approach and edge weighting scheme.

While relatively good prediction accuracy was obtained in the experiments, it has to be noted that the prediction performance observed using the two historical versions of the database does not necessarily imply that a similar prediction performance would be obtained by future predictions; it may be that research is biased by the current scientific knowledge represented in the source databases, and thus indirect links implied by this knowledge are more likely to turn up in future research. Also, it may be that research is likely to discover "low-hanging fruits" first, thus making it gradually harder to obtain new knowledge in the future.

Finally, the reported link prediction accuracy is clearly not sufficient for predicting links without any prior knowledge. While the settings consider cases where the input is already chosen so that there is an equal number of linked and unlinked gene pairs, the number of gene pairs that remain unlinked in the complete database is several orders of magnitude higher. However, when combined with additional data, this kind of predictions can still be useful.

Disease gene prioritization Use of Biomine in disease gene ranking enables identifying, from among a number of putative candidate genes, the ones that appear most plausible based on the data contained in the source databases. This approach is expected to work best in cases where several functionally related genes contribute to the disease, and knowledge about the functions of these genes is already present in the source databases. Obviously, less studied genes with little or no functional annotations cannot be identified in this way. The best results are naturally obtained when a reference set of already known disease genes is available. However, with a sufficient true positive density in the candidate list, the proposed cluster-based classifier without a reference set performed almost as well as a supervised classifier using a reference set.

While the disease gene prioritization results in the unsupervised setting are promising, more work is needed to make the methods practical. Accuracy of the predictions is poor when the density of true positives is not sufficient. An interesting future research topic in this area is combining the supervised and cluster-based methods into a semi-supervised method where information both from a reference set of disease genes and from mutual proximities of the candidate genes would be used. This might be useful especially in cases where only a small number of reference genes is available for the disease under study.

Disease gene prioritization using Biomine is based on measuring node proximities. This approach may be biased by the greatly varying node degrees, since nodes with high degrees have more closer neighbors. Therefore, considering the statistical significance of observed proximity values instead of measuring node proximity directly is expected to be beneficial. While estimation of statistical significance was done in the link prediction experiments of Paper III, it was not used in the experiments of Paper IV due to its computational cost. In practical applications, the set of potential relationships that needs to be evaluated is significantly smaller than in these exhaustive experiments, in which case statistical significance should be incorporated to the analysis.

Subgraph extraction and visualization Although the results from gene prioritization experiments are promising, the current methods and data are far from perfect and the results from automatic prioritization methods need to be validated by manually by using biological databases and literature. The Biomine query system facilitates this by searching and visualizing relationships between the putative genes and known disease genes, or relationships between the putative disease genes, and providing links to the original databases. The search and visualization methods use the same edge weighting scheme that is used for the prioritization itself, making them well suited for manual exploration of the prioritization results.

The public search interface to Biomine has been designed to be as simple to use as possible: basically the user just gives a list of genes (or other entities) of interest. While this approach is convenient for simple exploratory use, in many cases there would be a need to access and further process the data in customized ways, which is not enabled by the current interface. In communications with potential users it has turned out that customization of the Biomine interface to meet the needs of specific applications would be needed to make it more useful in practice and to perform more customized queries and analyses.

Qualitative queries based on context free grammars provide an alternative way of extracting and visualizing relationships between graph nodes. While being applicable in their present form, CFG-based queries could be made more useful by augmenting them with weighting of paths, by attaching weights to the productions of the context-free grammar. Actually, a major motivation for developing the CFG-based query system in the first place was to provide a more flexible way of weighting different relationship types based on the context of the edges, instead of only having edge-specific weights. Due to time constraints, this extension to weighted productions was never completed, however.

Practical application of Biomine The gene mapping workflow presented in this thesis (Figure 1.1) is somewhat idealized, and tailored for providing a framework for the particular methods included in this thesis. In practice each gene mapping project has its own structure, and the methods introduced in this thesis are applied on a case-by-case basis. At the time of writing, HaploRec, the haplotyping method developed in this thesis has 140 registered users, and the web-based interface to Biomine is being used as an exploratory tool. Biomine has also been used as a component of the semantic micro-array analysis pipeline SegMine [71], to discover and visualize relationships between enriched gene sets.

The disease gene and link prediction functionalities of Biomine are not currently publicly available, and can only be accessed using command line tools within the network of the department of computer science where the Biomine server is located. Partly for this reason, the applicability of these methods in actual gene mapping research remains to be studied. They are currently being applied to the prioritization of gene lists associated with elevated lipid levels, but no results from this work are available yet.

There are two reasons for not having public access to disease gene prioritization and link prediction features of Biomine. Firstly, providing such access would require significant computational resources. Secondly, they are currently based on prototype implementations targeted for method development use, and it would require further efforts to build e.g. a web-based interface for accessing these functionalities. Another option would be to make the complete database and the related tools installable by any interested party; also this would be a substantial task due to the complexity of the Biomine system.

Outlook Overall, the methods of this thesis aim to help in identifying novel disease genes and understanding their function in the context of existing biological knowledge, with the ultimate goal of improving the prevention and treatment of diseases with a genetic component. While already a huge amount of disease genes have been identified to date, the area is still currently under active research, and the genetic basis of many important diseases is still poorly understood. Also the methods of producing data are continuously improving, giving rise to larger and larger data sets. Efficient and accurate haplotyping is an important component of genome imputation methods, which are a standard tool in current gene mapping research. Especially for finding yet uncovered genes with relatively weak effects, the importance of being able to explore existing biological knowledge and use it for automatic prioritization of putative genes is evident. The continuously increasing amount of information in public databases will naturally also increase the accuracy and applicability of the automatic gene prioritization methods based on such data.

References

[1] J. Akey, L. Jin, and M. Xiong. Haplotypes *vs* single marker linkage disequilibrium tests: what do we gain? *European Journal of Human Genetics*, 9:291–300, 2001.

[2] B. Alidaee, F. Glover, G. Kochenberger, and H. Wang. Solving the maximum edge weight clique problem via unconstrained quadratic programming. *European Journal of Operational Research*, 181:592–597, 2006.

[3] D. Altshuler, M. J. Daly, and E. S. Lander. Genetic mapping in human disease. *Science*, 322(5903):881–888, 2008.

[4] C. I. Amos. Successful design and conduct of genome-wide association studies. *Human Molecular Genetics*, 16(R2):R220–R225, 2007.

[5] S. Asthana, O. D. King, F. D. Gibbons, and F. P. Roth. Predicting protein complex memebership using probabilistic network reliability. *Genome Research*, 14:1170–1175, 2004.

[6] D. J. Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, pages 781–791, 2006.

[7] D. J. Balding, M. Bishop, and C. Cannings, editors. *Handbook of Statistical Genetics*, volume 2. John Wiley and Sons, Chichester, UK, second edition, 2003.

[8] A.-L. Barabasi, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 22:56–68, 2011.

[9] W. Barendse. Haplotype analysis improved evidence for candidate genes for intramuscular fat percentage from a genome wide association study of cattle. *PloS one*, 6(12):e29601, 2011.

[10] A. Birkland and G. Yona. BIOZON: a system for unification, management and analysis of heterogeneous biological data. *BMC Bioinformatics*, 7(1):70, 2006.

[11] B. L. Browning and S. R. Browning. Efficient multilocus association mapping for whole genome association studies using localized haplotype clustering. *Genetic Epidemiology*, 31:365–375, 2007.

[12] S. R. Browning and B. L. Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, 81:1084–1097, 2007.

[13] R. M. Cantor, K. Lange, and J. S. Sinsheimer. Prioritizing GWAS results: a review of statistical methods and recommendations for their application . *American Journal of Human Genetics*, 86:6–22, 2010.

[14] D. I. Chasman. On the utility of gene set methods in genomewide association studies. *Genetic Epidemiology*, 32:658–668, 2008.

[15] A. Clark. The role of haplotypes in candidate gene studies. *Genetic Epidemiology*, 27(4):321–333, 2004.

[16] A. G. Clark. Inference of haplotypes from PCR-amplified samples of diploid populations. *Molecular Biological Evolution*, 7:111–122, 1990.

[17] G. M. Clarke, C. A. Anderson, F. H. Pettersson, L. R. Cardon, A. P. Morris, and K. T. Zondervan. Basic statistical analysis in genetic case-control studies. *Nature protocols*, 6(2):121–133, 2011.

[18] D. Clayton. SNPHAP: a program for estimating frequencies of large haplotypes of SNPs. `http://www-gene.cimr.cam.ac.uk/clayton/software/snphap.txt`. Accessed: 2013-08-16.

[19] D. Clayton. *Population association*, chapter 19, pages 939–960. Volume 2 of Balding et al. [7], second edition, 2003.

[20] C. J. Colbourn. *The Combinatiorics of Network Reliability*. Oxford University Press, 1987.

[21] A. R. Collins. *Linkage disequilibrium and association mapping*. Springer, 2007.

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

[23] B. Devlin and N. Risch. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, 29:311–322, 1995.

[24] C. M. Drysdale, D. W. McGraw, C. B. Stack, J. C. Stephens, R. S. Judson, K. Nandabalan, K. Arnold, G. Ruano, and S. B. Liggett. Complex promoter and coding region β2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. *Proceedings of the National Academy of Sciences*, 97(19):10483–10488, 2000.

[25] J. Earley. An efficient context-free parsing algorithm. *Communications of the ACM*, 13:94–102, 1970.

[26] L. Eronen, F. Geerts, and H. Toivonen. A Markov chain approach to reconstruction of long haplotypes. In *Proceedings of the 9th Pacific Symposium on Biocomputing 2004 (PSB 2004)*, pages 104–115, Hawaii, USA, 2004. World Scientific.

[27] E. Eskin, E. Halperin, and R. M. Karp. Large scale reconstruction of haplotypes from genotype data. In *Proceedings of the seventh annual international conference on Computational biology*, pages 104–113. ACM Press, 2003.

[28] E. Evangelou, T. A. Trikalinos, G. Salanti, and J. P. A. Ioannadis. Family-based versus unrelated case–control designs for genetic associations. *PLoS Genetics*, 2:e123, 2006.

[29] L. Excoffier and M. Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biological Evolution*, 12(5):921–927, 1995.

[30] C. Faloutsos, K. S. McCurley, and A. Tomkins. Fast discovery of connection subgraphs. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 118–127, 2004.

[31] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters 27*, 27:861–874, 2005.

[32] L. Franke, H. van Bakel, L. Fokkens, E. D. de Jong, M. Egmont-Petersen, and C. Wijmenga. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *American Journal of Human Genetics*, 78:1011–1025, 2006.

[33] R. A. George, J. Y. Liu, L. L. Feng, R. J. Bryson-Richardson, D. Fatkin, and M. A. Wouters. Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Research*, 34(19):e130, 2006.

[34] J. W. Gregersen, K. R. Kranc, X. Ke, P. Svendsen, L. S. Madsen, A. R. Thomsen, L. R. Cardon, J. I. Bell, and L. Fugger. Functional epistasis on a common MHC haplotype associated with multiple sclerosis. *Nature*, 443(7111):574–577, 2006.

[35] R. Griffiths and P. Marjoram. Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology*, 3(4):479–502, 1996.

[36] D. Gusfield. Haplotyping as perfect phylogeny: conceptual framework and efficient solutions. In *Proceedings of the sixth annual international conference on Computational biology*, pages 166–175. ACM Press, 2002.

[37] D. Gusfield. Haplotype inference by pure parsimony. In *Proceedings of the 14:th Annual Symposium on Combinatorial Pattern Matching*, 2003.

[38] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(Database Issue):D514–D517, 2005.

[39] M. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32(Database issue):D258, 2004.

[40] P. Hintsanen. *Simulation and graph mining tools for improving gene mapping efficiency*. PhD thesis, University of Helsinki, Department of Computer Science, Report A-2011-3, 2011.

[41] P. Hintsanen and H. Toivonen. Finding reliable subgraphs from large probabilistic graphs. *Data Mining and Knowledge Discovery*, 17(1):3–23, 2008.

[42] V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22:85–126, 2004.

[43] R. R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18:337–338, 2002.

[44] R. R. Hudson. *Linkage Disequilibrium and Recombination*, chapter 19, pages 662–680. Volume 2 of Balding et al. [7], second edition, 2003.

[45] S. Hunter, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, et al. InterPro: the integrative protein signature database. *Nucleic Acids Research*, 37:D211–D215, 2009.

[46] T. Hwang and R. Kuang. A heterogeneous label propagation algorithm for disease gene discovery. In *Proceedings of the SIAM International Conference on Data Mining, SDM 2010*, pages 583–594, Columbus, Ohio, USA, 2010. SIAM.

[47] L. J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, and C. von Mering. STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, 37:D412–D416, Jan. 2009.

[48] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, 38:D355–D360, Jan. 2010.

[49] J. Kang, S. Kugathasan, M. Georges, H. Zhao, and J. H. Cho. Improved risk prediction for crohn's disease with a multi-locus approach. *Human molecular genetics*, 20(12):2435–2442, 2011.

[50] S. Karinen, S. Saarinen, R. Lehtonen, P. Rastas, P. Vahteristo, L. A. Aaltonen, S. Hautaniemi, et al. Rule-based induction method for haplotype comparison and identification of candidate disease loci. *Genome medicine*, 4(3):1–18, 2012.

[51] G. Kimmel and R. Shamir. GERBIL: Genotype resolution and block identification using likelihood. *Proceedings of the National Academy of Sciences*, 102(1):158–162, 2005.

[52] J. Knight, D. Curtis, and P. C. Sham. CLUMPHAP: a simple tool for performing haplotype-based association analysis. *Genetic Epidemiology*, 32(6):539–545, 2008.

[53] J. Köhler, J. Baumbach, J. Taubert, M. Specht, A. Skusa, A. Rüegg, C. Rawlings, P. Verrier, and S. Philippi. Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*, 22(11):1383–1390, 2006.

[54] S. Köhler, S. Bauer, D. Horn, and P. Robinson. Walking the interactome for prioritization of candidate disease genes. *American Journal of Human Genetics*, 82(4):949–958, April 2008.

[55] M. Koivisto, M. Perola, T. Varilo, W. Hennah, J. Ekelund, M. Lukk, L. Peltonen, E. Ukkonen, and H. Mannila. An MDL method for finding haplotype blocks and for estimating the strength of haplotype block boundaries. *Proceedings of the Pacific Symposium on Biocomputing 2003*, 2003.

[56] M. Krauthammer, C. A. Kaufmann, T. C. Gilliam, and A. Rzhetskya. Molecular triangulation: Bridging linkage and molecular-network information for identifying candidate genes in alzheimer's disease. *Proceedings of the National Academy of Sciences*, 101(42):15148–15153, 2004.

[57] L. Kruglyak, M. J. Daly, M. P. Reeve-Daly, and E. S. Lander. Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics*, 58(6):1347, 1996.

[58] T. Lee, Y. Pouliot, V. Wagner, P. Gupta, D. Stringer-Calvert, J. Tenenbaum, and P. Karp. BioWarehouse: a bioinformatics database warehouse toolkit. *BMC Bioinformatics*, 7(1):170, 2006.

[59] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, May 2007.

[60] S. Lin, A. Chakravarti, and D. J. Cutler. Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nature Genetics*, 36:1181–1188, 2004.

[61] B. Linghu, E. Snitkin, Z. Hu, Y. Xia, and C. DeLisi. Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. *Genome Biology*, 10(9):R91, 2009.

[62] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 35:D26–D31, Jan. 2007.

[63] J. Marchini and B. Howie. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11:499–511, 2010.

[64] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. A. Ioannidis, and J. N. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9:356–369, 2008.

[65] N. E. Morton. Sequential tests for the detection of linkage. *American Journal of Human Genetics*, 7(3):277, 1955.

[66] V. Moskvina and K. Schmidt. On multiple-testing correction in genome-wide association studies. *Genetic Epidemiology*, 32(6):567–573, 2008.

[67] C. Newton-Cheh and J. N. Hirschhorn. Genetic association studies of complex traits: design and analysis issues. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 573(1–2):54–69, 2005.

[68] T. Niu, Z. S. Qin, and J. S. Liu. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *American Journal of Human Genetics*, 70:17–169, 2002.

[69] V. Ollikainen. *Simulation techniques for disease gene localization in isolated populations*. PhD thesis, University of Helsinki, Department of Computer Science, Report A-2002-2, 2002.

[70] M. Oti1, B. Snel, M. A. Huynen1, and H. G. Brunner. Predicting disease genes using protein–protein interactions. *Journal of Medical Genetics*, 43:691–698, 2006.

[71] V. Podpečan, N. Lavrač, I. Mozetič, P. Novak, I. Trajkovski, L. Langohr, K. Kulovesi, H. Toivonen, M. Petek, H. Motaln, and K. Gruden. SegMine workflows for semantic microarray data analysis in Orange4WS. *BMC Bioinformatics*, 12(1):416, 2011.

[72] M. Potamias, F. Bonchi, A. Gionis, and G. Kollios. K-nearest neighbors in uncertain graphs. *Proceedings of the VLDB Endowment*, 3(1-2):997–1008, 2010.

[73] Z. S. Qin, T. Niu, and J. S. Liu. Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *American Journal of Human Genetics*, 71:1242–1247, 2002.

[74] P. Rastas. *Computational techniques for haplotype inference and for local alignment significance.* PhD thesis, University of Helsinki, Department of Computer Science, Report A-2009-9, 2009.

[75] P. Rastas, M. Koivisto, H. Mannila, and E. Ukkonen. A hidden Markov technique for haplotype reconstruction. In *Proceedings of Algorithms in Bioinformatics, 5th international workshop (WABI-2005)*, pages 140–151. Springer-Verlag, 2005.

[76] P. Rastas, J. Kollin, and M. Koivisto. Fast Bayesian haplotype inference via context tree weighting. In *Proceedings of Algorithms in Bioinformatics, 8th international workshop (WABI-2008)*, pages 259–270. Springer-Verlag, 2008.

[77] R. S. Rosa and K. S. Guimarães. Insights on haplotype inference on large genotype datasets. In *Advances in Bioinformatics and Computational Biology*, pages 47–58. Springer, 2010.

[78] R. Salem, J. Wessel, and N. Schork. A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. *Human Genomics*, 2(1):39–66, 2005.

[79] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, S. Federhen, et al. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 38:D5–D16, Jan. 2010.

[80] M. A. Schaub, A. P. Boyle, A. Kundaje, S. Batzoglou, and M. Snyder. Linking disease associations with regulatory information in the human genome. *Genome research*, 22(9):1748–1759, 2012.

[81] P. Scheet and M. Stephens. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, 78(4):629–644, 2006.

[82] P. Sevon. *Algorithms for association-based gene mapping.* PhD thesis, University of Helsinki, Department of Computer Science, Report A-2004-4, 2004.

[83] P. Sevon, H. T. Toivonen, and V. Ollikainen. TreeDT: Gene mapping by tree disequilibrium test. In *Proceedings of the 7:th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001)*, pages 365–370, San Francisco, California, 2001. World Scientific.

[84] H. Shim, H. Chun, C. Engelman, and B. Payseur. Genome-wide association studies using single-nucleotide polymorphisms versus haplotypes: an empirical comparison with data from the north american rheumatoid arthritis consortium. *BMC Proceedings*, 3(Suppl 7):S35, 2009.

[85] M. Stephens and P. Scheet. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American Journal of Human Genetics*, 76(3):449–462, 2005.

[86] A. R. Templeton, T. Maxwell, D. Posada, J. H. Stengård, E. Boerwinkle, and C. F. Sing. Tree scanning a method for using haplotype trees in phenotype/genotype association studies. *Genetics*, 169(1):441–453, 2005.

[87] The International HapMap Consortium. The international HapMap project. *Nature*, 426:789–796, 2003.

[88] The Uniprot Consortium. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Research*, 38:D142–D148, 2010.

[89] H. Toivonen, P. Onkamo, K. Vasko, V. Ollikainen, P. Sevon, H. Mannila, M. Herr, and J. Kere. Data mining applied to linkage disequilibrium mapping. *American Journal of Human Genetics*, 67:133–145, 2000.

[90] H. T. Toivonen, P. Onkamo, K. Vasko, V. Ollikainen, P. Sevon, H. Mannila, and J. Kere. Gene mapping by haplotype pattern mining. In *Proceedings of the IEEE International Symposium on Bio-Informatics & Biomedical Engineering (BIBE 2000)*, pages 99–108, Arlington, Virginia, Nov. 2000.

[91] H. Tong and C. Faloutsos. Center-piece subgraphs: problem definition and fast solutions. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 404–413, 2006.

[92] L.-C. Tranchevent, F. B. Capdevila, D. Nitsch, B. De Moor, P. De Causmaecker, and Y. Moreau. A guide to web tools to prioritize candidate genes. *Briefings in Bioinformatics*, 12(1):22–32, 2011.

[93] A. Z. D. Ullah, N. R. Lemoine, and C. Chelala. SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). *Nucleic acids research*, 40(W1):W65–W70, 2012.

[94] E. J. van den Oord. Controlling false discoveries in genetic studies. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 147(5):637–644, 2008.

[95] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan. Associating genes and protein complexes with disease via network propagation. *PLoS Computational Bioloogy*, 6:e1000641, Jan. 2010.

[96] I. A. Vergara, T. Norambuena, E. Ferrada, A. W. Slatẽr, and F. Melo. StAR: a simple tool for the statistical comparison of ROC curves. *BMC Bioinformatics*, 9:265, 2008.

[97] X. Wan, C. Yang, Q. Yang, H. Zhao, and W. Yu. HapBoost: A fast approach to boosting haplotype association analyses in genome-wide association studies. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 10(1):207–212, 2013.

[98] K. Wang, M. Li, and M. Bucan. Pathway-based approaches for analysis of genomewide association studies. *American Journal of Human Genetics*, 81(6):1278–1283, 2007.

[99] W. Y. Wang, B. J. Barratt, D. G. Clayton, and J. A. Todd. Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics*, 6:109–118, 2005.

[100] R. H. Waterston, A. T. Chinwalla, L. L. Cook, K. D. Delehaunty, G. A. Fewell, L. A. Fulton, R. S. Fulton, T. A. Graves, L. W. Hillier, E. R. Mardis, J. D. McPherson, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002.

[101] K. M. Weiss. *Genetic variation and human disease — Principles and evolutionary approaches*. Cambridge University Press, 1993.