Statistical models for analyzing human genetic variation

Sriram Sankararaman

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2010-51
http://www.eecs.berkeley.edu/Pubs/TechRpts/2010/EECS-2010-51.html

May 7, 2010

Statistical Models for Analyzing Human Genetic Variation

by

Sriram Sankararaman

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

and the Designated Emphasis

in
Computational and Genomic Biology

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Michael I. Jordan, Chair
Professor Kimmen Sjölander
Professor Richard Karp

Spring 2010

Statistical Models for Analyzing Human Genetic Variation

by

Sriram Sankararaman

Abstract

Statistical Models for Analyzing Human Genetic Variation

by

Sriram Sankararaman

Doctor of Philosophy in Computer Science

and the Designated Emphasis in Computational and Genomic Biology

University of California, Berkeley

Professor Michael I. Jordan, Chair

Advances in sequencing and genomic technologies are providing new opportunities to understand the genetic basis of phenotypes such as diseases. Translating the large volumes of heterogeneous, often noisy, data into biological insights presents challenging problems of statistical inference. In this thesis, we focus on three important statistical problems that arise in our efforts to understand the genetic basis of phenotypic variation in humans.

At the molecular level, we focus on the problem of identifying the amino acid residues in a protein that are important for its function. Identifying functional residues is essential to understanding the effect of genetic variation on protein function as well as to understanding protein function itself. We propose computational methods that predict functional residues using evolutionary information as well as from a combination of evolutionary and structural information. We demonstrate that these methods can accurately predict catalytic residues in enzymes. Case studies on well-studied enzymes show that these methods can be useful in guiding future experiments.

At the population level, discovering the link between genetic and phenotypic variation requires an understanding of the genetic structure of human populations. A common form of population structure is that found in admixed groups formed by the intermixing of several ancestral populations, such as African-Americans and Latinos. We describe a Bayesian hidden Markov model of admixture and propose efficient algorithms to infer the fine-scale structure of admixed populations. We show that the fine-scale structure of these populations can be inferred even when the ancestral populations are unknown or extinct. Further, the inference algorithm can run efficiently on genome-scale datasets. This model is well-suited to estimate other parameters of biological interest such as the allele frequencies of ancestral populations which can be used, in turn, to reconstruct extinct populations.

Finally, we address the problem of sharing genomic data while preserving the privacy of individual participants. We analyze the problem of detecting an individual genotype from

the summary statistics of single nucleotide polymorphisms (SNPs) released in a study. We derive upper bounds on the power of detection as a function of the study size, number of exposed SNPs and the false positive rate, thereby providing guidelines as to which set of SNPs can be safely exposed.

Acknowledgements

Grad school at Berkeley has been a wonderful learning experience and it is a joy to acknowledge the many individuals who have made it so.

I have been lucky to work with two great advisors, Kimmen Sjölander and Michael Jordan. Kimmen first introduced me to Bioinformatics and, in the years since, has given me advice on research, science and life in general. Writing papers with Kimmen has taught me how to communicate clearly and precisely to both biologists and to computer scientists. Mike fostered my understanding of statstics and machine learning. His emphasis on expertise building, breadth of interests and clarity of thought have been a constant source of inspiration for me. The intellectual freedom (and the countless espresso shots at Nefeli and Brewed Awakening!) at Berkeley has allowed me to become an independent thinker and to confidently explore new problem areas.

Dick Karp was kind enough to be on my qualifying exam and thesis committees. Dick was always generous with his time and I left every meeting with a clearer picture of the problem that I was working on. I would also like to thank Claire Tomlin for agreeing to be on my qualifying exam committee. Eran Halperin introduced me to the field of statistical genetics during a summer internship at ICSI – Eran taught me how to think about association studies and population structure and I have enjoyed the several hours-long discussions that we have had over the past few years.

I learnt a lot from Fei Sha while working on statistical models of catalytic residue prediction. I am grateful to Jack Kirsch for explaining to me all about enzyme catalysis and for his analysis of functional residue predictions in *Bovine* α-Chymotrypsin. I would like to thank Carolina Dallet, Ron Alterovitz and Aaron Arvey for discussions on functional residue prediction. Gad Kimmel, Srinath Sridhar and Bogdan Pasaniuc helped me to think about algorithmic and statistical aspects of the admixture and locus-specific ancestry problem.

The thread on genomic privacy began as a random discussion between Guillaume Obozinski and me in an attempt to understand a puzzling result on individual detection. Working on this problem helped me to understand the theory of local asymptotic normality, which I had learnt earlier in Mike's 210B, and for this I am indebted to Guillaume. I would also like to thank Alex Bouchard for discussions on varied topics that included sequential monte carlo, Bayesian inference of trees, and linguistic and genetic evolution. I would like to thank members of the Berkeley Phylogenomics Group (BPG) and SAIL for being ever-willing to provide a sounding board for, often half-baked, ideas.

Outside the lab, I have been fortunate to make some very close friends. I shared many common interests with Athulan, my housemate for five years here in Berkeley. Late night discussions with Athulan on topics ranging from Indian politcs to the nuances of Carnatic music never ceased to be stimulating. Music has grown to be an inseparable part of my life over these years and for this, I am grateful to Karthik who shared with me both his knowl-

edge and his meticulously-indexed collection of recordings. Subbu, Praveena, Danjo, Mary, Adarsh, Kranti, and Archana – a big thank you for making life in Berkeley so memorable.

My decision to pursue a Ph.D. owes itself, in no small measure, to the influence of inspirational teachers – the late David Chatterjee at St.Joseph's; C. Pandurangan, B. Yegnanarayana, V. Balakrishnan and the late Dilip Veeraraghavan at IIT Madras. Their constant curisoity and clarity of ideas brought the subject to life and left me convined of the value of a great teacher.

Finally, it is impossible to sufficiently thank my parents for everything – their constant support, their many sacrifices, and for the values that they have instilled in me. To my dear father and mother, I dedicate this thesis.

To my parents

Contents

Chapter 1

Introduction

The main theme of this thesis is the study of human genetic variation using statistical models. The central questions in genetics have centered around the link between our genes and our traits or phenotypes:

- Is a phenotype determined by the genetic code alone or is it also influenced by the environment? Diseases such as cystic fibrosis are often completely determined by the genetic code and are termed *Mendelian* because they follow Mendel's laws of inheritance. Mendelian phenotypes tend to be rare, however. A vast majority of phenotypes, including many common diseases such as hypertension or type-2 diabetes, are caused by a combination of genetic and environmental factors. Relative little is known about the genetic basis of these complex phenotypes. For instance, attempts to understand the genetic basis of type-2 diabetes in humans have lead to the discovery of at least 11 genetic variants that influence the risk of type-2 diabetes [Frayling, 2007]. Yet these discovered variants explain only a small fraction of the disease risk.

- How does variation in the genetic code produce the diversity of phenotypes across individuals? Is this diversity caused by direct changes to the protein sequences within each cell or does this variation affect other biological mechanisms ?

- How is this genetic variation shaped by population-level forces such as migration, mixture, and adaptation? Populations that are isolated from each other tend to become more differentiated by random genetic drift. The need to adapt to the local environment also leads to differentiation, as seen in the genes involved in skin pigmentation. On the other hand, migration and mixing tend to reduce the genetic differences amongst the mixing populations. The distribution of genetic variants across populations is a result of a combination of these forces.

Rapid advances in whole-genome sequencing and genotyping technologies are enabling us to answer each of these questions with increasing precision. The large volumes of data pro-

duced by these technologies coupled with our relative ignorance of the underlying biological processes have led naturally to the development of statistical models for studying genetic variation. To be applicable to large-scale datasets, these models need to permit efficient inference while accurately capturing the essential properties of the underlying biological processes.

1.1 Genetic and phenotypic variation

The human genetic code consists of a little over 3 billion base paired nucleotides organized in two sets of 23 chromosomes (22 pairs of autosomes and a pair of sex chromosomes). An individual inherits one set from each parent. Each of these sets is termed a *haploid* genome.

Genetic variation refers to the differences in this genetic code at a given position across individuals in the human population. Several kinds of genetic variant are known to occur. These include a change at a single nucleotide (known as a single nucleotide polymorphism or a SNP), an insertion of a nucleotide sequence in the genome or a change in the number of copies of a genomic sequence (known as copy number variants or CNVs).

It has been observed that two haploid genomes differ in about 0.05% of their bases leading to an estimated 1.5 million SNPs that differentiate two haploid genomes [Levy *et al.*, 2007] (The total number of SNPs discovered in the human genome is much higher with an estimate of at least 12 million as of June 2008 [http://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi]). More recent studies have estimated the fraction of non-SNP DNA variation between two haploid genomes to be around 0.4%. These non-SNP variants account for a larger fraction of genetic variation than was previously believed [Levy *et al.*, 2007] and their effect on variation in phenotypes remains to be studied.

We can understand the link between genetic and phenotypic variations at multiple levels. Phenotypic variation is, to a large degree, driven by changes in the level and timing of proteins that are expressed in the various cells of the human body. Variation in the genetic code that alters the amino acid sequence of a protein would be expected to influence some of these phenotypes. Thus, we can begin to understand the link between genotype and phenotype at the *molecular level*. At this level, we would like to understand the effect of genetic variation on protein function.

Focusing on the molecular level has its limitations. A large fraction of the genetic variation does not alter the protein sequence because protein-coding genes form only a small fraction of the human genome. Genetic variation that does not alter the protein sequence may still affect the phenotype through, as yet poorly understood, regulatory mechanisms such as transcriptional control, chromatin accessibility, and alternative splicing. How can we understand the impact of these variants without understanding the biological mechanisms that they influence? One strategy to work around this difficulty involves studying *populations* of individuals instead of a single individual. Such an approach would allow us

to analyze the population-level distributions of the variants and infer their impact on various phenotypes. While a number of population-level approaches have been proposed [Risch and Merikangas, 1996; Lander, 1996; Collins *et al.*, 1997; Chakraborty and Weiss, 1988; McKeigue, 2005], as we will explain shortly, their effectiveness depends on our ability to understand and exploit the genetic structure of human populations.

1.2 Genetic association: common and rare variants

One approach to identifying the variants that influence a phenotype, termed an association study, screens either a candidate gene or the entire genome for variants that are highly-correlated with the phenotypes in a sample of unrelated individuals [Risch and Merikangas, 1996; Lander, 1996; Collins *et al.*, 1997]. In the case of a binary trait where the phenotype is the presence of absence of a disease, such a study looks for variants that have differing frequencies in the two groups and is termed a case-control association study (cases refer to individuals diagnosed with the disease and controls refer to unaffected individuals). To perform association studies with little prior knowledge of the biological mechanisms involved in the disease requires an assay of SNP variation across the human genome. Such genomewide SNP genotyping assays have become feasible due to the development of high-density SNP microarrays [Wang *et al.*, 1998] that are capable of simultaneously genotyping hundreds of thousands of SNPs.

A major challenge in these association studies is to maximize the power to detect true associations while controlling the fraction of false positives (spurious associations). This problem is particularly severe in a genomewide setting due to the hundreds of thousands of variants that are tested. False positive associations often arise because of a failure to account for the underlying structure of the population being studied [Lander and Schork, 1994b]. Many of the early association studies have focused on relatively homogeneous European subpopulations and have thus avoided the need to account for population structure. To gain a comprehensive understanding of the genetic basis of complex phenotypes, it is essential that these studies be applied to non-European populations. However, performing association studies in populations such as African Americans or Latinos presents a major challenge. These populations have a complex genetic structure. To a first order, African Americans are a mixture of Africans and Europeans while Latinos are a mixture of Europeans, Africans and native Americans.

An association study that does not account for the structure of an African American population would produce associations simply because a SNP has different allele frequencies in the African and European groups [Pritchard and Rosenberg, 1999]. On the other hand, population structure can also be exploited to improve the power of detecting associations [Chakraborty and Weiss, 1988; McKeigue, 2005]. Thus, inferring the genetic structure of human populations is an important step to understanding human genetic variation.

Since they only exploit the population-level distribution of genetic variants, genomewide association studies are expected to detect associations with common SNPs (SNPs with a minor allele frequency of at least 1% in the population). It was reasoned that these common SNPs would explain a significant fraction of the variation in common diseases like type-2 diabetes and hypertension (this assumption is referred to as the Common Disease Common Variant (CDCV) hypothesis [Collins *et al.*, 1997; Lander, 1996; Risch and Merikangas, 1996]) and hence, the detected associations would be useful for predicting disease risk. However, most of the SNPs identified in the hundreds of genomewide association studies performed so far have explained only a tiny fraction of the disease variation [Goldstein, 2009; Hirschhorn, 2009; Kraft and Hunter, 2009].

A number of reasons have been proposed for this gap – one possible reason is that the causal variants are rare and current association study designs have little power to detect them [Goldstein, 2009]. One strategy to detect rare variants is to use other sources of functional data to direct the search for associations (such association studies would no longer be "unbiased"). For example, we could scan for variants that fall in protein-coding regions and rank these variants according to their functional impact. To do so, we would need to understand the functional effect of the variants within protein-coding genes. However, the protein-coding regions comprise a small fraction of the genome so that most variants would fall outside these regions. Such an approach would fail to detect these other variants. In the future, functional data such as proximity of a genomic region to transcription-factor binding sites, positioning of nucleosomes around a region etc. would need to be integrated to fully understand the impact of rare variants. In effect, these approaches would combine molecular-level and population-level analyses of genotype-phenotype relationships.

1.3 Predicting functional residues in proteins

The problem of understanding the functional effect of variation in a gene translates directly to one of identifying the functionally important amino acid residues in the corresponding protein product. While we have motivated this problem in the context of understanding genetic variation, identifying functional residues provides valuable clues about the function of proteins [George *et al.*, 2005] and is an important problem in its own right. The residues of a protein are involved in different roles: catalytic residues in an enzyme are responsible for the extraordinary efficiency of the reactions that an enzyme facilitates; ligand-binding and specificity-determining residues target the protein towards specific substrates while avoiding others; protein-protein interaction residues mediate the formation of complex assemblies; allosteric residues allow effector molecules to control the activity of the protein. Since experimental methods to determine the roles of individual residues are time-consuming and expensive, computational methods are essential for functional residue prediction; computational methods provide initial clues that can be followed up by experiments.

The biggest challenge in developing a computational method to predict each of these classes of residues is the availability of a benchmark dataset of proteins in which (some fraction of) the residues have been annotated with their functional role. We focus mainly on the task of predicting catalytic residues because of the availability of an extensive manually-curated dataset of catalytic residues [Bartlett *et al.*, 2002; Porter *et al.*, 2004].

Computational methods for identifying functional residues fall into two broad classes:

- Sequence-based methods: Sequence-based methods predict functional residues based on the primary sequence of the protein alone. These methods examine the patterns of conservation of each residue in the protein across a set of evolutionarily-related sequences (homologues). Residues that are more conserved are expected to be functionally important. Methods that only use sequence information are useful because 3D structural information is not available for a majority of proteins.

  The sequence-based methods gather sequences that are homologous to the protein of interest, build an alignment of these sequences and analyze the conservation patterns in each column of the alignment. Such methods operate on the assumption that all residues in a column are homologous; this assumption can be violated due to structural and functional variability across specific lineages (where a residue conserved in one subgroup is not conserved in another due to changes in function) and errors in alignments. While we can reduce variability and alignment errors by restricting the set of homologues to closely-related sequences, it turns out that analyzing a divergent family of proteins can improve the ability to detect truly functional residues. In chapter 2, we present a sequence-based method that uses phylogenomic information to predict functional residues in large, highly divergent, protein families.

- Structure-based methods: Structure-based methods use information from the protein 3D structure such as the solvent accessibility of the residue or presence of the residue in a cleft or pocket, sometimes in combination with sequence information. For example, it is well-known that catalytic residues tend to be located in one of the three largest pockets on the enzyme. Although methods that use structural information in combination with sequence information would be expected to better predict functional residues, the accuracies of these methods has remained low. In chapter 3, we present a statistical method that combines sequence and structural information and significantly improves over current catalytic residue prediction methods. This method combines the sequence-based method that we describe in chapter 2 with features computed at residues near each other in the protein 3D structure within a logistic regression model. The attempt to combine features from neighboring residues leads to a proliferation of features and we use a statistical regularization technique to control the complexity of the resulting model.

1.4 Locus-specific ancestries in admixed populations

Obtaining a comprehensive understanding of genotype-phenotype relationships requires genomewide association mapping in different population groups. This would enable the detection of variants that are common in some groups but rare in others. A major difficulty in studying genetic variation in populations such as Latinos, is the complex genetic structure of these populations. These populations are *admixed*, i.e., they are formed by the intermixing of several ancestral populations. Each individual genome in these admixed populations is a mosaic of chromosomal segments inherited from the ancestral populations. Inferring the ancestry of these admixed genomes is critical for discovering variants associated with diseases. One such technique, known as admixture mapping [Chakraborty and Weiss, 1988; McKeigue, 2005], scans the admixed population for regions which are preferentially inherited from one of the ancestral populations compared to the genomewide average. Admixture mapping has been successfully applied in African-Americans to identify regions linked to diseases such as hypertension [Zhu *et al.*, 2005b], prostate cancer [Freedman *et al.*, 2006], and multiple sclerosis [Reich *et al.*, 2005a]. On the other hand, studies that do not account for the underlying ancestries can produce spurious signals of association. Thus, we need to infer the locus-specific ancestries of these admixed populations and correct for these ancestries in tests of association.

Ancestry inference in admixed populations relies on two properties of admixed genomes: i) nearby locations on a chromosome tend to be inherited from the same ancestral population and ii) the patterns of variation at these nearby locations can be used to infer their ancestral origin. Ancestry inference is a challenging statistical problem due to several reasons: genomes from the ancestral populations may not be available, the ancestral populations may be very similar, or the admixture may be *ancient*, i.e., the ancestral populations may have been mixing for a long period of time. In chapters 4 and 5, we describe a method that can accurately infer locus-specific ancestries even when the ancestral genomes are not available. This method is based on a probabilistic model of the admixture process (Chapter 5). Inference in the model is intractable on genomewide datasets; hence, we propose a fast and accurate approximation algorithm that is used as initialization for the inference algorithm.

1.5 Genomic Privacy

The statistical power to detect associations in genomewide association studies can be enhanced by combining data across these studies in meta-analysis or replication studies. Such methods require data to flow freely in the scientific community, but this raises privacy concerns. Till recently, many studies pooled individuals together, making only the allele frequencies of each SNP in the pool publicly available. However, even this summary data does not preserve privacy. It was shown recently that, using the large number of SNPs genotyped

in these studies, the presence of an individual genotype in such a pool can be determined with high power [Homer *et al.*, 2008]. An immediate response to this result might be that detecting the presence of an individual in a pool will not provide any valuable information in addition to what can already be obtained from the genotype (which is needed to test for presence in the pool). However, detecting an individual in the pool of cases in a case-control study reveals the disease status of the individual. For most common diseases, it is difficult to obtain this disease status directly from the genotype alone. thus, such tests have the ability to reveal more information than what is already known based on the individual's genotype.

This result prompted organizations such as the NIH to restrict public access to summary data as a conservative means of protecting privacy [Gilbert, 2008]. In response, a number of solutions have since been proposed to deal with the twin issues of sharing genomic data and protecting individual privacy [Church *et al.*, 2009]. Solutions on extremes of the spectrum involve giving up the demand for sharing or the requirement of privacy. Amongst solutions that occupy a middle-ground, one approach proposes setting up a secure infrastructure to facilitate data sharing. Another approach would involve determining the privacy guarantees provided by various data-sharing mechanisms and using these to develop guidelines for data-sharing. To come up with privacy guarantees, we need to determine which set of SNPs can be safely exposed while preserving an acceptable level of privacy. In chapter 6, we address this issue by providing an upper bound on the power achievable by any detection method as a function of factors such as the number and the allele frequencies of exposed SNPs, the number of individuals in the pool, and the false positive rate of the method.

Chapter 2

Functional site prediction using phylogenomic information

2.1 Introduction

The problem of identifying the positions in a protein critical for its structure or function plays a significant role in biological discovery. These residues (such as the catalytic triad of serine, aspartate and histidine found in proteases) provide valuable clues about the functions of proteins. Since experimental methods to determine the roles of individual positions are time-consuming and expensive, computational methods are widely used for protein functional residue prediction; these provide initial clues that can be followed up by experiments. In these experiments, we use the definition of catalytic residues provided by the authors of the Catalytic Site Atlas and of the CATRES benchmark dataset [Bartlett *et al.*, 2002]. They defined catalytic residues as those residues in an enzyme active site that participate directly in catalysis as revealed by structural studies.

Casari, Sander, and Valencia developed one of the first computational approaches to identify positions conferring functional specificity [Casari *et al.*, 1995]. Another method for functional residue prediction is Evolutionary Trace (ET) [Lichtarge *et al.*, 1996]. The original ET method defines progressively more conservative cuts of a phylogeny. The level of the cut at which a column shows a specific pattern of conservation (either family-wide or subfamily-specific) is used to assign a score to each position in a protein. A more recent method, ConSurf [Landau *et al.*, 2005], computes the rate of evolution at each position based on phylogenetic analysis; residues with lower rates of evolution are considered more important. Variants of both ET, one of which uses an entropy-based score, [Aloy *et al.*, 2001; Mihalek *et al.*, 2004] and ConSurf [Mayrose *et al.*, 2004; Nimrod *et al.*, 2005; Glaser *et al.*, 2006] have also been developed. In general, predictive methods have relied on protein surface geometry [Peters *et al.*, 1996], energy considerations [Laurie and Jackson, 2005; Elcock, 2001], chemical properties [Ko *et al.*, 2005; Ondrechen *et al.*, 2001] and sequence conservation

[Lichtarge *et al.*, 1996; Casari *et al.*, 1995; Landgraf *et al.*, 2001; Landau *et al.*, 2005] or have attempted to combine different features [Gutteridge *et al.*, 2003; Petrova and Wu, 2006; Youn *et al.*, 2007].

A number of methods focusing exclusively on specificity-determining residues have also been developed [Del Sol Mesa *et al.*, 2003; Kalinina *et al.*, 2004; Mirny and Gelfand, 2002; Donald and Shakhnovich, 2005; Pei *et al.*, 2006; Hannenhalli and Russell, 2000]. [Capra and Singh, 2008] developed a method for scoring the positions in an alignment, termed GroupSim, which was found to be competitive with a number of previous methods. Some of the methods proposed for specificity determinant prediction require the subtypes to be specified [Mirny and Gelfand, 2002; Kalinina *et al.*, 2004; Pirovano *et al.*, 2006; Capra and Singh, 2008; Hannenhalli and Russell, 2000] while others [Pei *et al.*, 2006; Del Sol Mesa *et al.*, 2003; Donald and Shakhnovich, 2005] do not. In practice, subtypes are seldom known for a protein family. Thus, methods which can work without explicit knowledge of subtypes (*i.e.*, from a *tabula rasa*) are more suitable for general use.

Here we present a new method - INTREPID (INformation-theoretic TREe traversal for Protein functional site IDentification). INTREPID takes as input a target protein, a multiple sequence alignment (MSA) and a gene tree of the family containing the target protein; a protein structure can also be included to boost performance but is not required. We focus on methods that exploit only sequence information, since structural information is not available for a majority of proteins. Methods employing an MSA as input operate on the assumption that all residues in a column are homologous; this assumption can be violated due to structural and functional variability across specific lineages and errors in alignments. A number of enzyme families exhibit variability in the location of catalytic residues [Todd *et al.*, 2002], while other enzyme families exhibit variation at catalytic positions. The inteins have been known to exhibit variations in their catalytic residues that in turn affect the intein-mediated splicing mechanisms. For instance, functional inteins with an N-terminal alanine instead of the catalytic cysteine or serine have been observed [Johnson *et al.*, 2007; Southworth *et al.*, 2000]. INTREPID is designed to be robust to these issues.

The key idea in INTREPID is the use of phylogenetic information by examining the conservation patterns at each node of a phylogenetic tree on a path from the root to the leaf corresponding to the sequence of interest. For instance, catalytic residues tend to be conserved across distant homologs and thus will appear conserved at (or near) the root of a gene tree. By contrast, specificity determinants will not be conserved across all members of a family but are likely to be conserved within one or more subtypes. Thus, prediction of these two distinct types of positions requires a different approach for each task. Any suitable conservation score can be used within the tree traversal of INTREPID depending on the type of functional residue to be predicted. A number of functions have been developed for determining functional residues by scoring the columns of a MSA, including information-theoretic scores based on Shannon Entropy [Shenkin *et al.*, 1991; Sander and Schneider, 1991], Relative Entropy [Wang and Samudrala, 2006], and Jensen-Shannon divergence [Capra

and Singh, 2007]. INTREPID uses the Jensen-Shannon divergence as it has been found to be the most accurate conservation-based score for functional residue identification [Capra and Singh, 2007].

In the catalytic residue prediction problem, we apply INTREPID to large protein families for enzymes in the Catalytic Site Atlas (CSA) [Porter *et al.*, 2004]. We compare INTREPID to other sequence-based methods, such as ET, ConSurf, and baseline methods based on global conservation scores. We also compare INTREPID to the machine-learning methods reported in [Petrova and Wu, 2006] and in [Youn *et al.*, 2007]. We also analyze the effect of alignment diversity on the accuracy of catalytic residue prediction. Finally, we apply INTREPID-SPEC, a variant of INTREPID adapted to specificity determinant prediction, to the dataset of putative specificity-determining positions (SDPs) generated by Capra and Singh [Capra and Singh, 2008].

2.2 Catalytic residue prediction

The input to INTREPID comprises a target protein p whose functional residues are to be predicted, a multiple sequence alignment (MSA) of proteins homologous to p and an estimated evolutionary tree of these homologs *i.e.*, the gene tree.

Each residue in p is analyzed independently to derive its predicted importance, based on the conservation patterns at each node on a path from the root to the leaf corresponding to protein p. INTREPID uses a key observation that was first exploited in the context of functional residue identification by Casari, Sander and Valencia [Casari *et al.*, 1995] and reinforced since then by numerous studies: residues playing critical roles for protein structure or function are often under strong negative selection. This negative selection enables these residues to be detected due to their strong conservation across a family of related proteins. Catalytic residues in enzyme active sites are an example of such a class. In predicting catalytic residues based on sequence conservation, the evolutionary context is critical *i.e.*, the degree of sequence divergence across homologs included in the analysis will have a significant impact on the method performance. In a closely related set of proteins, even positions that are not critical for function may appear well-conserved. Thus, truly critical residues may only be revealed against a backdrop of evolutionary divergence.

Unfortunately, conservation patterns in an MSA can be affected by inadvertently included non-homologs, alignment and phylogeny errors, and functional divergence in specific lineages *e.g.*, where a residue conserved in one subtree is not conserved in another subtree due to changes in function. INTREPID is designed to detect catalytic residues exhibiting such behaviour by combining the conservation patterns observed at different nodes of the tree.

2.2.1 Computing the Positional Importance Score

INTREPID computes an importance score $IMP_p(x)$ for every position x in protein p using a traversal of the phylogenetic tree from the root to the leaf corresponding to p. The tree traversal enables us to exploit the information over the entire tree, instead of requiring us to select a particular cut of a tree into subtrees. It also helps to avoid the contribution of noise from subfamilies or entire lineages that may disagree on the importance of particular positions.

Every node encountered in this traversal corresponds to a subtree containing p and one or more homologs, and provides a different perspective on the potential importance of each position in p. For instance, at the leaf corresponding to p, no homologs are available to highlight which positions are conserved and which are variable, and it is impossible to predict which of the positions in p are likely to be critical for function. At the other extreme, residues that are perfectly conserved across the entire family will be evident when viewed from the root of the tree. As we traverse a path from the root to the leaf, positions formerly appearing to be variable will become fixed in specific lineages; at a leaf, all positions will be perfectly conserved. To enable us to compensate for subtrees with highly correlated or very few sequences, the score IMP_p accounts for the evolutionary distance spanned as estimated by the sequence divergence.

We denote by S the subtree corresponding to a node encountered in the tree traversal, $cons(S, x)$ is the conservation of position x within subtree S, and $\overline{cons(S)}$ is the average conservation across all columns in subtree S. The importance score at a position x is computed as

$$IMP_p(x) = max_s cons(S, x) - \overline{cons(S)} \tag{2.1}$$

Here we use the Jensen-Shannon (J-S) divergence [Lin and Wong, 1990] between the amino acid distribution and the background (with prior weight $= \frac{1}{2}$ as in [Capra and Singh, 2007]). The importance score thus assigns a high score to those residues that are conserved over a large subtree of divergent sequences. When subtrees with many highly similar sequences are considered, the average conservation will be high. In this case, even though the positional conservation is also high, the difference between these two numbers will be fairly low. The maximum observed positional conservation on the path from the root to the leaf at each position x is its importance. We finally normalize the score across all the positions in the protein p so that the reported score at position x is $Z - IMP_p(x) = \frac{IMP_p(x) - \overline{IMP_p}}{\sigma(IMP_p)}$ where $\overline{IMP_p}$ and $\sigma(IMP_p)$ are the average and standard deviations of the importance scores across all the columns in the MSA.

We illustrate INTREPID with an example.

Figure 2.1 shows six protein sequences of length four each. The target protein is marked with an arrow. The nodes traced by the tree traversal are $S1, S2, S3, S4$, and $S5$. We first compute the average Jensen-Shannon divergence in each of the subtrees. In subtree S_1,

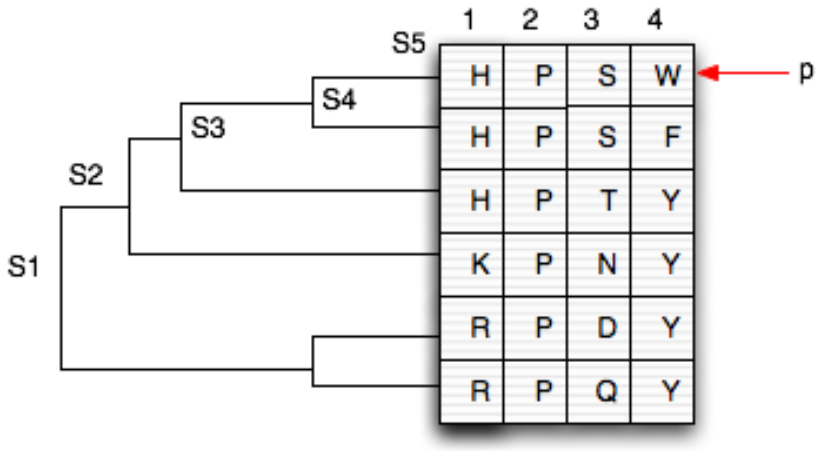

Figure 2.1: An example of the INTREPID algorithm. This example shows six protein sequences of length four each. The target protein p is marked with an arrow. The nodes visited by the tree traversal are $S1$, $S2$, $S3$, $S4$, and $S5$. As explained in the text, INTREPID ranks the positions in the order 2, 1, 4 and 3 while simple global conservation would rank position 4 above position 3.

the average Jensen-Shannon divergence is: $\overline{cons(S_1)} = \frac{0.87+0.73+0.56+0.79}{4} = 0.74$. Repeating this calculation for each of the other subtrees, we get $\overline{cons(S_2)} = 0.79$, $\overline{cons(S_3)} = 0.83$, $\overline{cons(S_4)} = 0.87$, and $\overline{cons(S_5)} = 0.89$.

Now let us look at column 1. In a tree traversal from the root (node $S1$) to the leaf corresponding to p, we compute the following importance scores: $cons(S_1, 1) = 0.73 - 0.74 = -0.1$, $cons(S_2, 1) = 0.82 - 0.78 = 0.03$, $cons(S_3, 1) = 0.91 - 0.82 = 0.09$, $cons(S_4, 1) = 0.91 - 0.87 = 0.04$, $cons(S_5, 1) = 0.91 - 0.89 = 0.02$.

The maximal importance score $IMP_p(1) = 0.09$, corresponds to the score at node S_3 where position 1 is completely conserved. Computing these scores for other positions: $IMP_p(2) = 0.13$, $IMP_p(3) = -0.03$, $IMP_p(4) = 0.05$. As expected, we see that position 2 has the highest importance score followed by position 1. If simple global conservation had been used (*i.e.*, each position had been ranked based on its conservation across the family), then position 4 would have a higher rank than position 1. INTREPID gives a higher score to position 1 than to position 4 because of the higher conservation in position 1 in the subtree containing p. In other words, position 4 appears to be important for a majority of the family but may have evolved a different role in the lineage corresponding to subtree $S4$. On the other hand, position 1 appears to be associated with a function that is preserved within the subtree $S3$ but is lost or modified outside.

Different measures of positional conservation can be used within the tree traversal protocol. We also considered using the log-odds of the frequency of the most frequent amino acid and the relative entropy between the amino acid distribution of position x within subtree

S and a background distribution [Wang and Samudrala, 2006]. Consistent with the results reported in [Capra and Singh, 2007], the score based on J-S divergence was found to be the most accurate. We use the distribution from the BLOSUM62 alignments [Henikoff and Henikoff, 1992] as the background distribution. See Section 3.E.3 for details and experimental results using the different positional conservation scores.

2.3 Experiments

We start by describing experiments to assess INTREPID on the prediction of catalytic residues, and examine the effect of protein family divergence on the accuracy of catalytic residue prediction.

2.3.1 Preliminaries

We compared INTREPID to two methods that use only sequence information to predict functionally important residues, Evolutionary Trace [Lichtarge *et al.*, 1996] and ConSurf [Pupko *et al.*, 2002]. We also included in our comparison a baseline method termed Global-JS which applies the JS-divergence score to each column of the alignment as performed by [Capra and Singh, 2007]. We used the results from servers implementing ET and ConSurf to ensure that each of these methods would be run with parameters for which it has been optimized: the Evolutionary Trace server from Baylor College of Medicine (`http://mammoth.bcm.tmc.edu/traceview/`) (BCMET), which implements the improved evolution-entropy hybrid version of Evolutionary Trace [Mihalek *et al.*, 2004], and the ConSurf web server at Tel Aviv University (`http://consurf.tau.ac.il`).

While evaluating these methods, the question of how the reported scores are typically handled by users needs to be addressed. We consider two ways of post-processing the scores reported. In the first case, we use the ranks of the residues instead of the scores. This treatment is more useful under the assumption that every protein should have some predicted residues (if, for instance, the protein is known to be an enzyme). In the second case, we normalize the scores of each method on each protein and then analyze all 100 proteins as a set, sorting the normalized scores for each position. In this approach, for some score cutoff, some proteins may have no predicted positions while others may have several. Normalizing the scores improved the accuracies of both BCMET and ConSurf compared to using unnormalized scores.

We computed the following metrics for comparison (note that although sensitivity and recall are synonymous terms, we follow convention and use each term according to the analysis):

$$Recall = Sensitivity = \frac{TP}{TP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$Specificity = \frac{TN}{TN+FP}$$

Here a true positive (TP) is a residue identified by the CSA as catalytic which is selected by a method, a false negative (FN) is a catalytic residue that is missed, a false positive (FP) is a residue erroneously selected by a method (*i.e.*, it is not listed in the CSA), and a true negative (TN) is a non-catalytic residue that is correctly not selected. Specificity measures how well a method rejects non-catalytic residues. Since the ratio of catalytic to non-catalytic residues is low, even apparently high values of specificity can correspond to a large number of false positives. Precision, which measures the fraction of predicted catalytic residues that are correct, is a more relevant measure of performance in this setting. We plot the ROC curve (Sensitivity *vs* 1-Specificity) and the Precision-Recall curve (Precision *vs* Recall) for each of these methods. The ROC curve has been truncated to the high-specificity region for clarity ($Specificity \geq 80\%$) although the trends shown are similar over the entire range of specificities.

2.3.2 Accuracy of INTREPID relative to other conservation-based methods

Figure 2.2 compares the performance of INTREPID, Global-JS, BCMET and ConSurf on the CSA-100 dataset (see Section 2.A for details). We see from the figure that INTREPID has the highest sensitivity over the entire range of specificities and is significantly more accurate than the other methods. Table 2.1 compares the different methods under various metrics. For example, at 90% specificity, INTREPID attains a sensitivity of 85.03% relative to sensitivities of 70.06% and 73.8% by BCMET and ConSurf respectively. The baseline method (Global-JS) performs quite well (a sensitivity of 78.66% at a specificity of 90%). At a precision of 10%, INTREPID attains a recall of 75.0% while Global-JS has a recall of 64.0%. ConSurf and BCMET never attain a precision of 10% resulting in 0% recall at this level. When the normalized scores are used in place of the ranks, we see from Table 2.1 that INTREPID has the highest sensitivity followed by Global-JS, BCMET, and ConSurf.

Since the ConSurf server selects a smaller, more closely related set of sequences as input to Rate4Site (the program that computes the site-specific evolutionary rates as part of the ConSurf protocol), we also tested the prediction power of Rate4Site on the CSA-100 dataset which contains a greater level of sequence divergence. Rate4Site failed to complete on 77 of the 100 alignments due to memory allocation problems. By removing sequences with greater than 80% identity, we obtained Rate4Site results on 71 out of the 100 inputs. We refer to these 71 families as the CSA-71 dataset.

We also ran INTREPID on these reduced alignments as well as the full alignments for these 71 families. Figure 2.4 compares the performance of INTREPID, run on alignments made non-redundant at 80% identity and on the original alignments for the CSA-71 dataset, with Rate4Site. INTREPID, when run on the reduced MSA, has a small but statistically

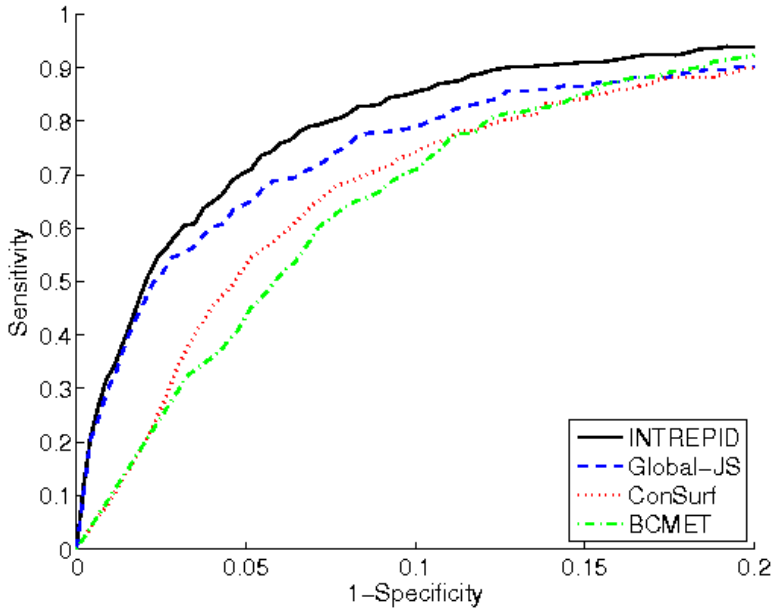

Figure 2.2: Results for catalytic residue prediction on a subset of 100 manually curated enzymes from the Catalytic Site Atlas (termed CSA-100) using rank-based scores: ROC curves comparing INTREPID with other conservation-based functional residue prediction methods: Global-JS, BCMET and ConSurf (refer Section 2.3.1 for details). The ROC curve shows INTREPID to have the highest sensitivity over the range of high specificity ($\geq 80\%$) followed by Global-JS. BCMET performs better as the specificity decreases. Refer to Table 2.1 for full AUC scores.

significant improvement over Rate4Site (Wilcoxon paired sign-rank test p-value of 1.3×10^{-5}). At 90% specificity, INTREPID attains sensitivities of 83.6% on the full MSA and 85.1% on the reduced MSA while Rate4Site attains a sensitivity of 84.6%. Similarly, at 10% precision, INTREPID on the full MSA, INTREPID on the reduced MSA and Rate4Site have 75%, 80%, and 75% recall. See Figure 2.4 for details. The figure also shows the considerable difference in accuracies between Rate4Site when run on these alignments and when run as part of ConSurf; this difference is likely a result of the different alignments used. Importantly, INTREPID has an average running time of 25.7 seconds on this dataset compared to Rate4Site which requires 2 hours and 52 minutes on average.

We also evaluated INTREPID on two other datasets consisting of the protein families used by [Petrova and Wu, 2006] and by [Youn *et al.*, 2007] respectively. On the Petrova dataset, INTREPID, with a sensitivity of 90.57% at a false positive rate of 13%, is as accurate as their method which attains a sensitivity of 90% at the same false positive rate (*i.e.*, the results are essentially indistinguishable). This is a very surprising result because

		INTREPID	Global-JS	ConSurf	BCMET
Residue ranks	Sensitivity_{95}	70.06	64.33	49.20	40.76
	Sensitivity_{90}	85.03	78.66	73.80	70.06
	Sensitivity_{80}	93.95	90.13	89.78	92.04
	Recall_{10}	75.0	64.0	0.00	0.00
	AUC	0.944	0.924	0.907	0.914
	AUC_{95}	0.024	0.022	0.011	0.010
	AUC_{90}	0.063	0.058	0.046	0.039
	AUC_{80}	0.154	0.145	0.127	0.124
	p-value	–	3.89×10^{-18}	1.64×10^{-17}	1.34×10^{-17}
Normalized scores	Sensitivity_{95}	67.83	58.28	36.74	54.46
	Sensitivity_{90}	85.03	75.48	59.42	74.84
	Sensitivity_{80}	92.99	89.81	87.86	91.72
	Recall_{10}	71.0	56.0	3.83	31.21
	AUC	0.935	0.910	0.884	0.919
	AUC_{95}	0.022	0.018	0.011	0.016
	AUC_{90}	0.060	0.053	0.036	0.048
	AUC_{80}	0.149	0.137	0.111	0.134
	p-value	–	3.89×10^{-18}	3.89×10^{-18}	5.27×10^{-18}

Table 2.1: Statistics comparing the different algorithms on a subset of 100 manually curated enzymes from the Catalytic Site Atlas (termed CSA-100). BCMET refers to the Evolutionary Trace server from Baylor College of Medicine. In the top panel, the ranks of the residues were used while in the bottom panel, the normalized scores were used. Sensitivity is measured at specificities of $95, 90$, and 85% respectively and recall at 10% precision. AUC_x $(x = 80, 90, 95)$ refers to the area under the ROC curve when specificity is at least $x\%$; AUC is the area under the entire curve. The p-value refers to the Wilcoxon signed rank p-values between the AUC of the INTREPID and each of the other methods. INTREPID improves significantly over the other methods on all metrics. BAsed on their ranks, ConSurf and BCMET do not reach a precision of 10% and hence have zero recall. The confidence intervals on these statistics are reported in Table 2.2.

INTREPID uses only sequence conservation while the method reported in [Petrova and Wu, 2006] uses a learning algorithm to combine sequence and structural features. [Youn *et al.*, 2007] present two variants of their method, one employing only sequence information while the second combines sequence and structural information. They present results for both variants on a dataset based on ASTRAL 40 v1.65 [Brenner *et al.*, 2000] selected to be non-redundant at the SCOP family level. On a similarly constructed dataset, INTREPID attains a recall of 28.13% at a precision of 15% and an AUC of 0.906. When restricted to sequence features alone, their method attains a sensitivity of about 16% at 15% precision and an AUC of 0.866. Thus, INTREPID improves over the method used in [Youn *et al.*, 2007] when restricted to sequence features alone. By contrast, their method that combines sequence

		INTREPID	Global-JS	ConSurf	BCMET
Residue ranks	Sensitivity_{95}	[65.47,74.31]	[59.67,68.46]	[44.95,54.28]	[36.30,46.01]
	Sensitivity_{90}	[82.31,87.86]	[74.92,82.17]	[69.44,77.92]	[66.35,74.39]
	Sensitivity_{80}	[91.82,95.83]	[87.42,92.75]	[86.79,92.24]	[89.35,94.65]
	Recall_{10}	[66.00,81.00]	[57.00,72.00]	[0.00,45.00]	[0.00,0.00]
	AUC	$[0.932, 0.955]$	$[0.901, 0.941]$	$[0.894, 0.916]$	$[0.898, 0.916]$
	AUC_{95}	$[0.016, 0.020]$	$[0.019, 0.024]$	$[0.010, 0.013]$	$[0.005, 0.007]$
	AUC_{90}	$[0.049, 0.057]$	$[0.054, 0.061]$	$[0.041, 0.048]$	$[0.032, 0.037]$
	AUC_{80}	$[0.132, 0.144]$	$[0.137, 0.150]$	$[0.122, 0.136]$	$[0.114, 0.124]$
Normalized scores	Sensitivity_{95}	[62.93,72.24]	[53.66,62.68]	[31.58,40.88]	[50.15,59.46]
	Sensitivity_{90}	[81.36,88.07]	[73.31,80.68]	[70.64,78.59]	[52.86,63.79]
	Sensitivity_{80}	[90.61,95.55]	[87.12,92.36]	[83.59,90.83]	[89.33,94.10]
	Recall_{10}	[60.00,76.00]	[38.00,64.00]	[0.00,24.00]	[16.00,54.00]
	AUC	$[0.924, 0.946]$	$[0.896, 0.923]$	$[0.869, 0.898]$	$[0.907, 0.928]$
	AUC_{95}	$[0.020, 0.024]$	$[0.016, 0.020]$	$[0.009, 0.012]$	$[0.014, 0.017]$
	AUC_{85}	$[0.057, 0.064]$	$[0.049, 0.056]$	$[0.031, 0, 038]$	$[0.045, 0.051]$
	AUC_{80}	$[0.145, 0.156]$	$[0.131, 0.142]$	$[0.103, 0.117]$	$[0.127, 0.138]$

Table 2.2: Confidence Intervals for statistics comparing the different algorithms on a subset of 100 manually curated enzymes from the Catalytic Site Atlas (termed CSA-100). BCMET refers to the Evolutionary Trace server from Baylor College of Medicine. In the top panel, the ranks of the residues were used while in the bottom panel, the normalized scores were used. Sensitivity is measured at specificities of $80, 90$, and 95% respectively and recall at 10% precision. $AUC_x, x = 80, 90, 95$ refers to the area under the ROC curve when specificity is at least $x\%$; AUC is the area under the entire curve. The 95% confidence interval are computed from 200 bootstrap replicates.

and structural information attains a much higher recall of about 65% at about the same precision. Reassuringly, the performance of INTREPID is approximately the same across these different datasets suggesting that these results would generalize well to new protein families.

2.3.3 Effect of evolutionary divergence on the accuracy of INTREPID

To measure the impact of evolutionary divergence on method performance, we controlled the sequence diversity of the alignment used. We created restricted alignments at the $x\%$-level, *i.e.*, sequences were discarded from each of these alignments so that the minimum percent identity from any sequence to the seed was at least $x\%$. We varied x over $10, 15, 20$, and 25% respectively. For comparison, we also included the original alignment which is labeled "Unrestricted". The effect of evolutionary divergence on INTREPID is shown in Figure 2.5. We see that as the divergence of the family increases, INTREPID accuracy increases. At 90% specificity, INTREPID has 42% sensitivity at 25% identity trimming. INTREPID

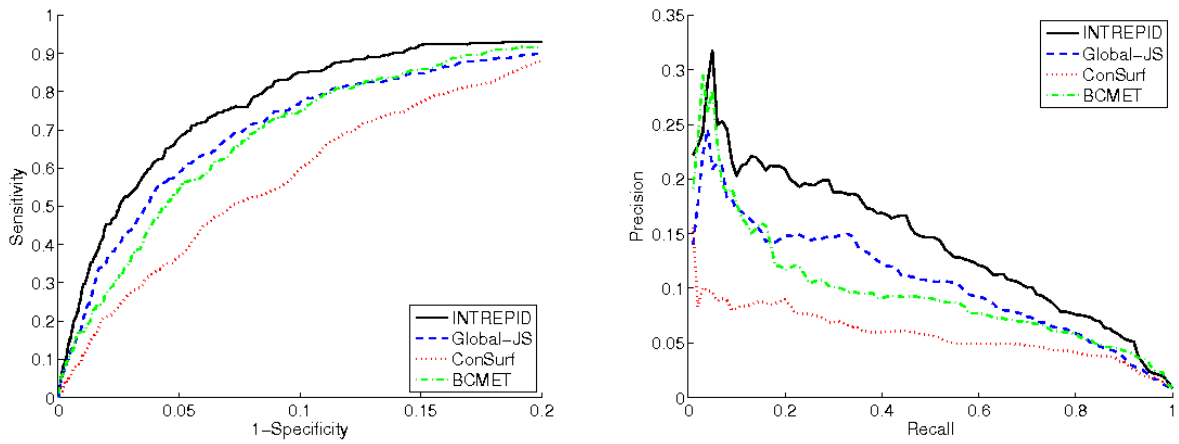

Figure 2.3: Results for catalytic residue prediction on a subset of 100 manually curated enzymes from the Catalytic Site Atlas (termed CSA-100) using normalized scores: ROC curves comparing INTREPID with other conservation-based functional residue prediction methods: Global-JS, BCMET and ConSurf (refer Section 2.3.1 for details). The methods have AUCs of $0.935, 0.910, 0.919$, and 0.884 respectively and AUC_{90} of $0.060, 0.053, 0.036$, and 0.048 respectively.

reaches 85% sensitivity when no sequences are removed. The trends shown here suggest that INTREPID is robust to divergence in protein families. All methods tested for the impact of sequence divergence on catalytic residue prediction – INTREPID, Global-JS and Rate4Site – benefit from increased sequence diversity (see Supplementary Materials).

We have shown in Section 2.3.3 of the main paper that greater evolutionary divergence improves the accuracy of INTREPID. Global-JS also improves with the inclusion of additional homologs but appears to be somewhat less robust to sequence divergence (Figure 2.6). For instance, at 90% specificity, the sensitivity of Global-JS prediction is 42% on the most restricted alignment (removing sequences with less than 25% identity to the seed) but increases to about 79% on the unrestricted alignment.

Our results are in agreement with previous studies [Panchenko *et al.*, 2004; Aloy *et al.*, 2001; Landgraf *et al.*, 2001]. In [Landgraf *et al.*, 2001], the recall of their scoring functions improved when the E-value cutoff for homolog inclusion was reduced from 10^{-50} to 10^{-20} while [Aloy *et al.*, 2001] observed a considerable improvement in accuracy of their method when their alignments had sequence identity less than 30%. Similar results were also reported by [Panchenko *et al.*, 2004], where the performance doubled on alignments with average sequence identity of 20% relative to those with average identity of around 45%. An important difference is that our alignments are highly divergent. In the experiments reported by [Panchenko *et al.*, 2004], the least divergent dataset had a minimum percent identity of 25% to the seed. For catalytic residue prediction, we observe that it is beneficial to use

highly divergent alignments with minimum percent identities extending as low as 10%.

2.3.4 Robustness of INTREPID to non-conserved catalytic residues

The advantage of INTREPID over global conservation analysis can be inferred from the level at which the maximum score is attained in the tree traversal. A little less than 50% of the catalytic residues have their maximum scores at the root. However, for 56 of the catalytic

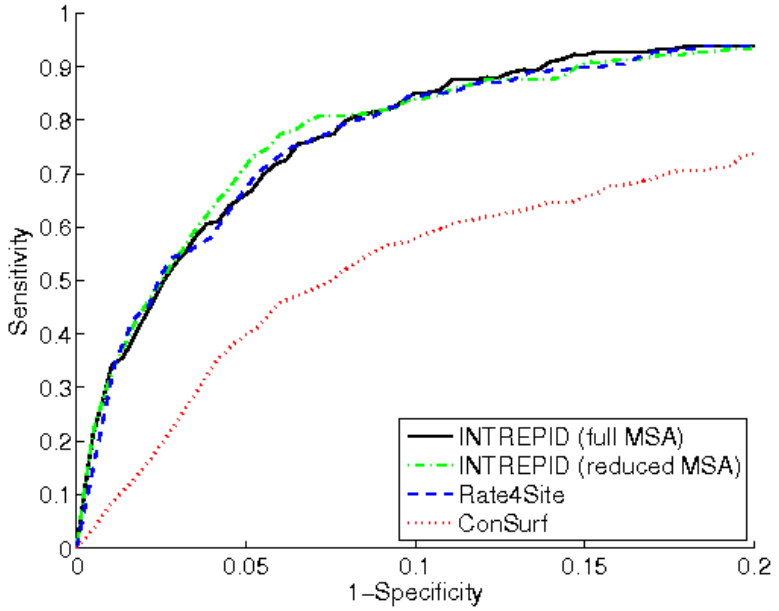

Figure 2.4: Results on a subset of 71 manually curated enzymes from the Catalytic Site Atlas (termed CSA-71) comparing INTREPID, Rate4Site and ConSurf using rank-based scores. Results were obtained on alignments derived from the original CSA-100 dataset by removing sequences with more than 80% sequence identity to one another; the 71 alignments used here were the alignments on which Rate4Site completed successfully. INTREPID was run on the reduced MSA as well as on the full MSAs for these 71 families. INTREPID, when run on both MSAs, and Rate4Site have similar accuracies though INTREPID is slightly more accurate (Area under the ROC curve for specificity $\geq 90\%$, AUC_{90}, for INTREPID, Rate4Site and ConSurf are $0.061, 0.061$, and 0.059 respectively; Area under the ROC curve, AUC, for these methods are $0.941, 0.938$, and 0.940 respectively; the difference in accuracy between INTREPID, run on the reduced MSA, and Rate4Site is statistically significant with a p-value of 1.3×10^{-5}). Rate4Site is considerably more accurate than the ConSurf webserver (which also uses the Rate4Site program) – this difference is likely a result of the different alignments used.

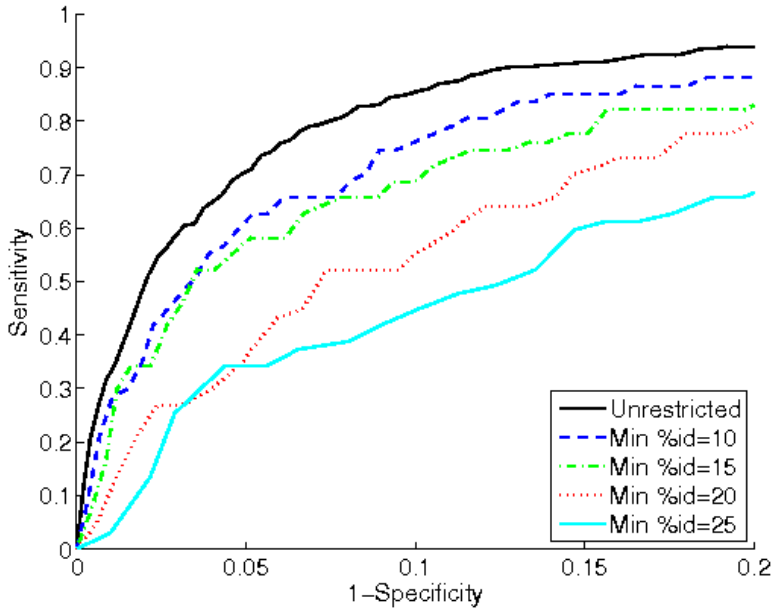

Figure 2.5: Effect of alignment diversity on catalytic residue prediction: ROC curve for INTREPID on alignments with varying degrees of evolutionary divergence, indicated by the minimum percent identity to the sequence used to gather homologs (seed sequence). The original alignment with no sequences removed is labelled "Unrestricted". INTREPID performs significantly better with increasing evolutionary divergence. For instance, INTREPID achieves 42% sensitivity at 90% specificity and 25% identity trimming but reaches 85% sensitivity when no sequences are removed.

residues ($\approx 18\%$ of all catalytic residues in the dataset), the maximum score is attained at least 5 levels away from the root. In 34 of the 56 residues, INTREPID assigns a better rank than Global-JS while Global-JS assigns a better rank on 15 (see Figure 2.7 in Supplementary Materials). Thus, INTREPID is more effective at identifying catalytic residues that are not conserved across the entire protein family. To illustrate this point, we consider two such families.

The first example is the enoyl-[acyl-carrier-protein] reductase from *Escherichia coli* (PDB id: 1mfp). CSA lists two catalytic residues: K163 and Y156. All methods give high ranks to K163 while Y156 is far more challenging. INTREPID given Y156 a rank of 18 (out of 258 positions), and BCMET, Global-JS and ConSurf give ranks of 31, 58, and 100 respectively. The homologs gathered for this protein family are found to include other short chain dehydrogenases (such as 3-oxoacyl-[acyl-carrier-protein] reductase). In these other families, this position generally contains a glutamine. The catalytic role of this glutamine has been observed in human 15-hydroxyprostaglandin dehydrogenase [Cho *et al.*, 2006]. The global

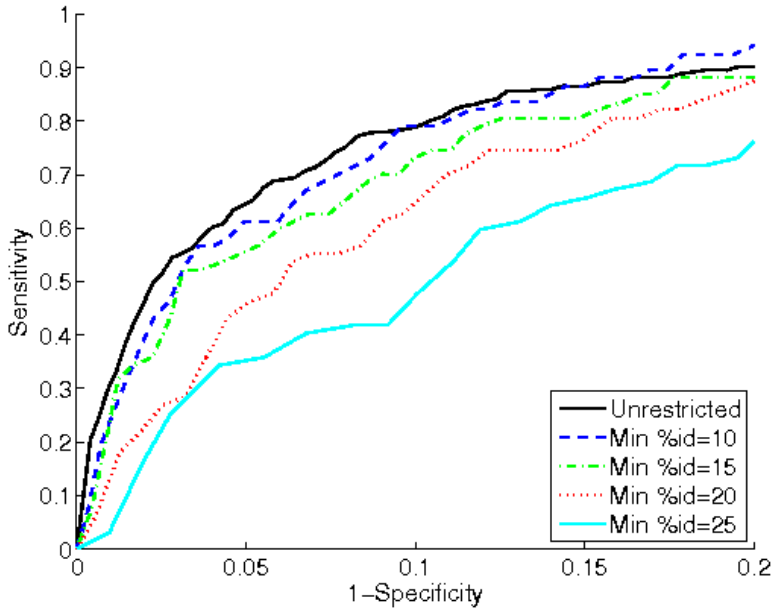

Figure 2.6: Effect of alignment diversity on the accuracy of a global conservation method (Global-JS) on a subset of 100 manually curated enzymes from the Catalytic Site Atlas (termed CSA-100): ROC curve for Global-JS on alignments with varying degrees of evolutionary divergence, indicated by the minimum percent identity to the seed. The original alignment with no sequences removed is labelled "Unrestricted". Global-JS performs significantly better with increasing evolutionary divergence - from 42% sensitivity at 90% specificity and 25% identity trimming to 79% sensitivity when no sequences are removed.

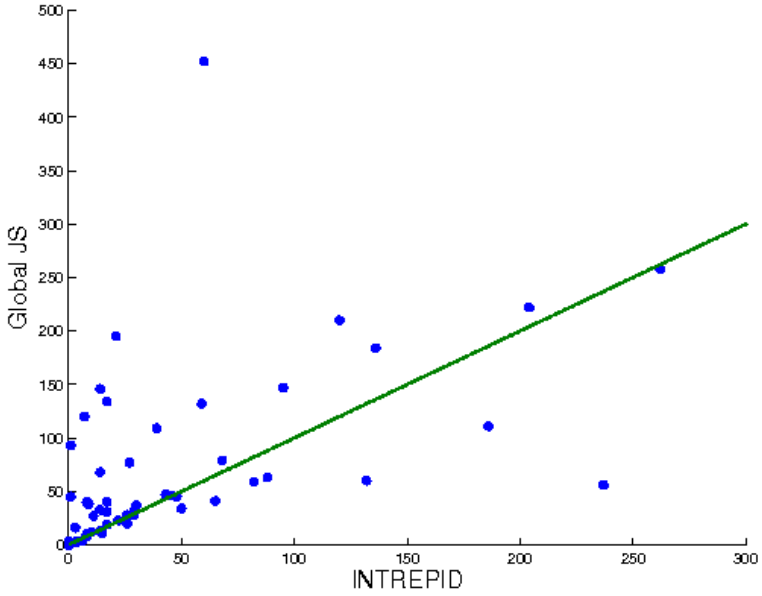

Figure 2.7: INTREPID more effectively identifies catalytic residues that are not conserved across the family because of its use of the phylogenetic tree: Scatter plot comparing the ranks assigned by INTREPID, a phylogenetic method, and Global-JS, a global conservation (non-phylogenetic) method, on catalytic residues that are not conserved across the family. Lower ranks correspond to residues that are easily identified as catalytic. The diagonal denotes residues on which both methods do equally well, residues above the diagonal are those for which INTREPID gives better ranks, and residues below the diagonal are those for which Global-JS gives better ranks. Here, INTREPID gives better ranks to 34 residues and Global-JS to 15.

frequency of tyrosine at position 156 is only about 25% though it is conserved within a subtree containing 199 sequences in a family with 833 sequences (see Supplementary Figures 2.8 and 2.9).

Another example is Flavocytochrome b2 from *Saccharomyces Cerevisiae* (PDB id:1fcB). This protein is part of the flavin mononucleotide (FMN)-dependent oxidoreductases. The poor conservation at the active site residues in this family has been observed by [Todd *et al.*, 2001]. This lack of conservation is most evident at the catalytic residues Y143, H373, and R376 (see Supplementary Figure 2.10). On H373, INTREPID, ConSurf and BCMET all give ranks of 1 while Global-JS gives a rank of 22. On R376, INTREPID, ConSurf, BCMET, and Global-JS give ranks of 7, 23, 4, and 9 respectively, while on Y143, the respective ranks are 23, 66, 56 and 20.

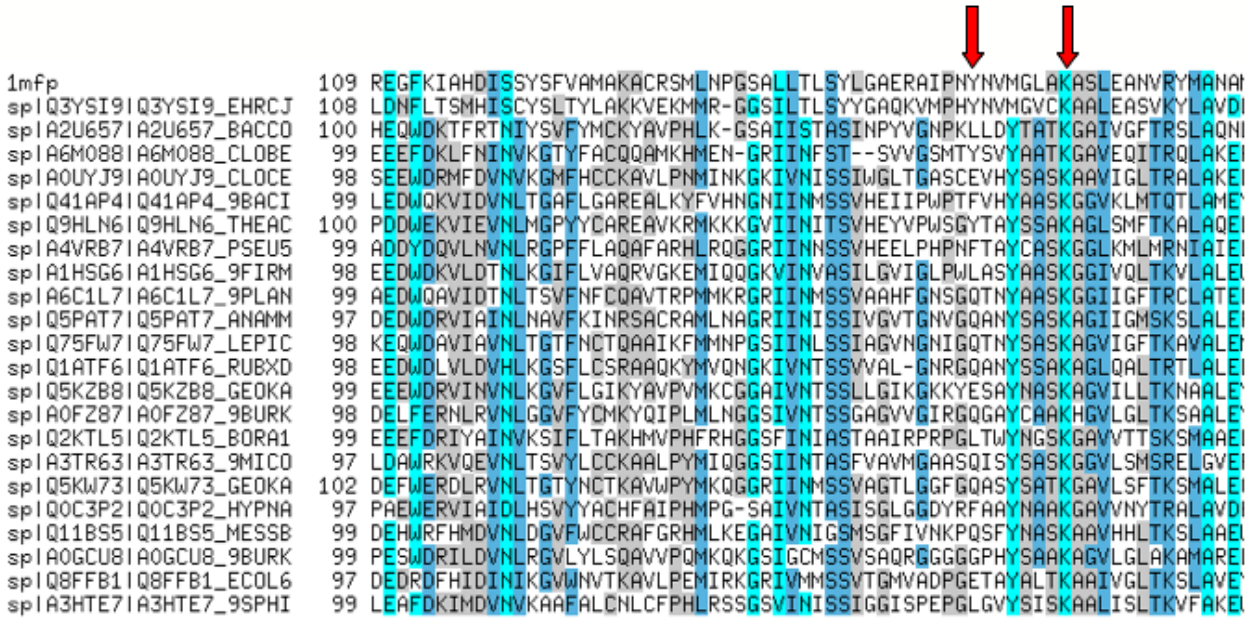

Figure 2.8: Alignment of the family containing enoyl-[acyl-carrier-protein] reductase from *Escherichia Coli* (PDB id: 1mfp) made non-redundant at 45% sequence identity. Positions marked in red correspond to catalytic residues (K163, Y156). Y156 is given a rank of 18 (out of 258 positions) by INTREPID, a rank of 58 by Global-JS, 100 by ConSurf and 31 by BCMET. See Figure 2.9 for an expanded view of a subtree containing the seed 1mfp and Section 2.3.4 of the main paper for details.

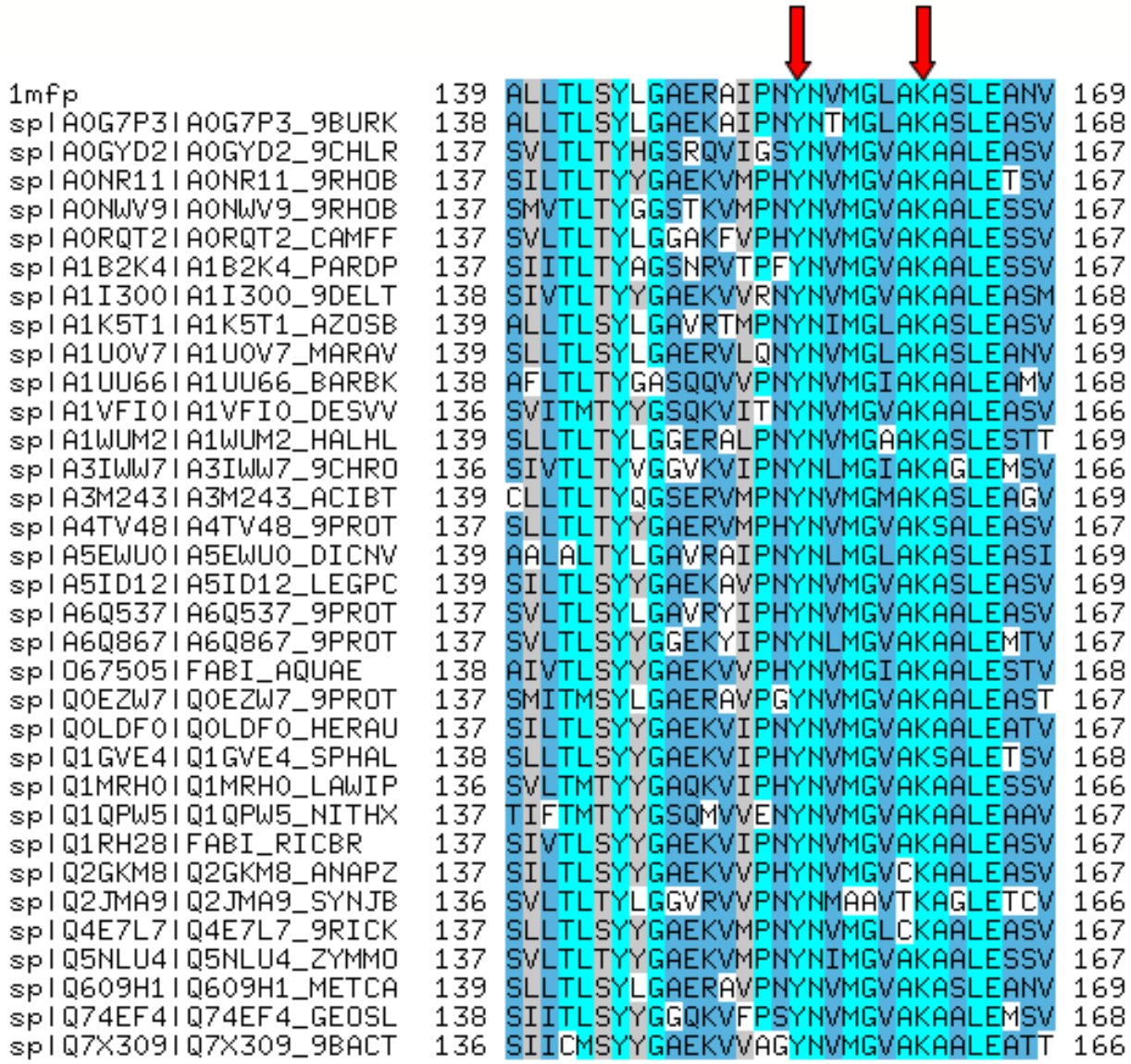

Figure 2.9: Alignment of sequences in the subtree containing enoyl-[acyl-carrier-protein] reductase from *Escherichia Coli* (PDB id: 1mfp) made non-redundant at 70% sequence identity. Positions marked in red correspond to catalytic residues (K163, Y156). This subtree contains 199 sequences out of the original 833 sequences. Notice that Y156 is conserved within this subtree while it has a frequency of only 25% in the entire family. See Section 2.3.4 of the main paper for details.

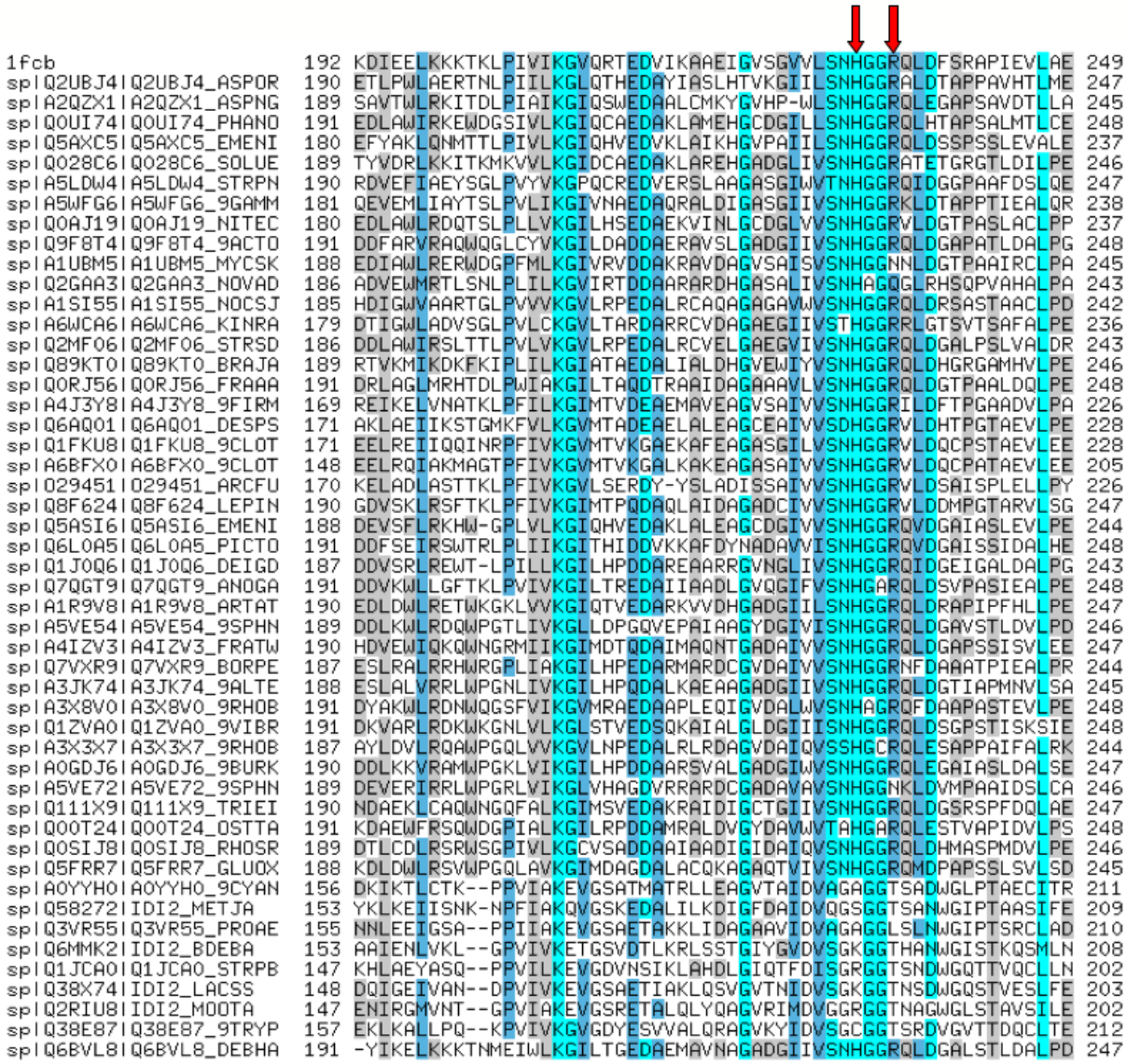

Figure 2.10: Alignment of the family containing Flavocytochrome b2 from *Saccharomyces Cerevisiae* (PDB id:1fcB) made non-redundant at 40% sequence identity. Positions marked in red correspond to catalytic residues (H373, R376).

PDB id	Percent id	Rank as non-seed	Rank as seed
1hti	44.4	57,72,2,29,30	54,67,2,11,29
1pxv	32.9	124,124,4,4	9,32,7,17
1azw	10.8	201,120,64	8,2,1

Table 2.3: INTREPID accuracy decreases with distance from the seed (sequence used to gather homologs): The table shows the ranks (Rank as non-seed) assigned to the CSA catalytic residues by INTREPID on a sequence which was not the seed. These ranks are compared to the ranks (Rank as seed) when the same sequence was used as the seed. As the sequence identity to the seed decreases, the accuracy decreases as seen from the numerically higher "Rank as non-seed" column.

2.3.5 Effect of distance of the target protein from the seed on INTREPID accuracy

To test the effectiveness of INTREPID prediction for proteins not used as seeds for selecting and aligning homologs, we took three families of enzymes from CSA containing at least two members each in the core (manually curated) dataset. One of these sequences for each family was used as a seed for clustering homologs, and we used INTREPID to predict critical residues for all members. We ensured that sequence identities with the seed were not so high as to make the experiment uninformative (i.e., homologous enzymes from CSA were selected with less than 50% identity). Table 2.3 compares the ranks of catalytic residues in a sequence that was not used as seed to the ranks when the same sequence was a seed. As sequence identity to the seed decreases, the accuracy of INTREPID also decreases. In the context of predicting catalytic positions in a single protein as opposed to the entire family, these results would apply to other sequence-based methods as well. Based on these limited results, we would recommend ensuring that the sequence of interest has sequence identity $> 50\%$ to the seed.

2.4 Examples of INTREPID predictions

In this section, we analyze INTREPID predictions on families found in the PhyloFacts resource (`http://phylogenomics.berkeley.edu/phylofacts`). For the families found in PhyloFacts, homologs were gathered from UniProt [Apweiler *et al.*, 2004] using Flowerpower [Krishnamurthy *et al.*, 2007] (with global-local settings and number of SHMM iterations set to 3) and re-aligned using MUSCLE [Edgar, 2004]. For this analysis, we built neighbor-joining trees using the PHYLIP software though other tree construction algorithms may be used in practice. PhyloFacts displays the top 5% of the INTREPID predictions though the cutoff may be varied by the user (we handle tied scores by selecting all residues at a given score).

2.4.1 Dihydroneopterin aldolase

Dihydroneopterin aldolase catalyzes the conversion of 7,8-dihydroneopterin (DHNP) to 6-hydroxymethyl-7,8-dihydropterin (HP) playing an essential role in the folate biosynthesis pathway. Mammals, unlike bacteria, plants, and yeast, lack a complete folate biosynthesis pathway and obtain folate from their diet [Lawrence *et al.*, 2005]. Hence, dihydroneopterin aldolase, along with other enzymes in the folate biosynthesis pathway, has served as a target for antimicrobial and antibacterial agents [Lawrence *et al.*, 2005].

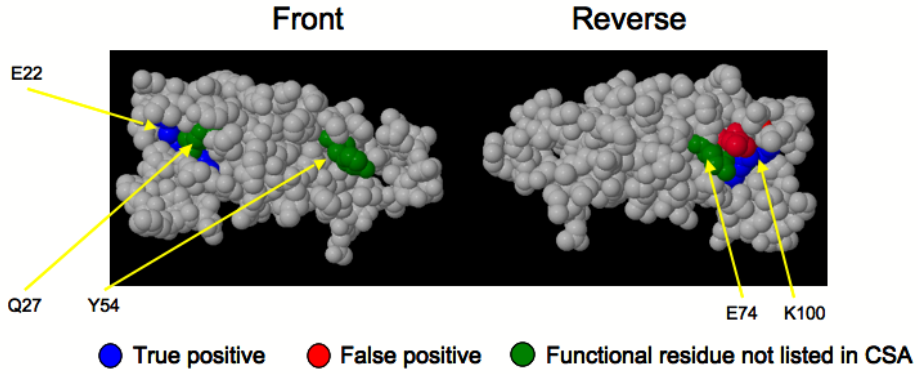

Figure 2.11: INTREPID predictions for dihydroneopterin aldolase from *Staphylococcus aureus* (PDB Id:2dhn, BPG accession:bpg020587). INTREPID correctly predicts the catalytic residues E22, K100, Q27, K74, and Y54. Of these, only E22 and K100 are listed in the CSA. The non-CSA functional residues refer to INTREPID predictions that are not listed in the CSA but have experimental evidence of being catalytic (Q27, K75, and Y54) (see text).

INTREPID predictions on dihydroneopterin aldolase from *Staphylococcus aureus* (PDB Id:1dhn) are shown in Figure 2.11. INTREPID correctly predicts the residues E22 and K100, which are listed as catalytic in the CSA. INTREPID also predicts residues Q27 which forms a hydrogen bond to the substrate in the crystal structure of neopterin(NP), an analog of 7,8-dihydroneopterin (DHNP) [Wang *et al.*, 2006]; Y54 which is known to coordinate catalysis with E22; K74 which influences the affinity of the enzyme for the substrate [Wang *et al.*, 2006; Hennig *et al.*, 1998]; and G17 and H16.

2.4.2 Src Homology 2 (SH2) domain

SH2 domains are found in multi-domain intracellular signaling proteins and play key roles in assembling signaling complexes by binding to phosphotyrosine moieties in target proteins. Key residues in SH2 binding pockets determine the specificity of interaction, and thereby the pathways in which these proteins participate. The protein used in this analysis, Src SH2 domain from Rous sarcoma virus (PDB accession 1SPS chain C), has been crystallized with an

11-residue polypeptide [Waksman *et al.*, 1993]. Three distinct ligand-binding pockets have been identified; a phosphotyrosine binding pocket, glutamate binding pocket and hydrophobic binding pocket. The key residues in these ligand-binding pockets are R12, R32, S34, E35, T37, C42, K57, H58, Y59, K60, I71, T72, Y87, D92, G93 and L94 [Waksman *et al.*, 1993; Songyang *et al.*, 1993]. Amongst the top 6 residue predicted by INTREPID are R32, H58 with the other predictions being W5, G27, F29 and F77.

2.5 Specificity determinant prediction

While the scoring functions discussed in the previous section are designed to detect family-defining positions (and catalytic positions in particular), this basic tree traversal protocol can be adapted to detect specificity-determining positions as well. Specificity-determining positions tend to be conserved within – but different across – subfamilies. For this problem, we compute the positional conservation score as the relative entropy of the amino acid distributions within and outside a subtree. This variant is termed INTREPID-SPEC. The importance score at position x is computed as

$$SP_p(x) = max_s RE(p_x^S, p_x^{S^c}) \tag{2.2}$$

where S is a node on the path from the root to the leaf corresponding to p, p_x^S denotes the probability distribution of amino acids at position x for the sequences within subtree S, and $p_x^{S^c}$ denotes the probability distribution of amino acids at position x over the other sequences. In computing the scores in Equation 2.2, S ranges over all the nodes in the tree traversal except the root. To avoid saturated probabilities (and handle subtrees with very few sequences), we use add-one pseudocounts [Durbin *et al.*, 1998]. Such a relative entropy score was used by [Hannenhalli and Russell, 2000] for specificity-residue prediction when the subtypes are known. Using the score within the tree traversal allows us to predict specificity determinants even when the subtypes are not known.

2.6 Experiments

2.6.1 Preliminaries

Methods for specificity determinant prediction can be classified as those that require the subtypes to be known *a priori* [Mirny and Gelfand, 2002; Kalinina *et al.*, 2004; Pirovano *et al.*, 2006; Hannenhalli and Russell, 2000; Capra and Singh, 2008] and those that do not [Pei *et al.*, 2006; Del Sol Mesa *et al.*, 2003; Donald and Shakhnovich, 2005]. INTREPID does not require knowledge of the subtypes. For specificity determinant prediction, we use INTREPID-SPEC (described in Section 2.5). We can implicitly provide subtype information

to INTREPID-SPEC by building a separate tree for each subtype which are then joined at the root to obtain a tree for the family. We compared INTREPID-SPEC to the GroupSim heuristic that was found to be competitive with other sequence-based methods in [Capra and Singh, 2008]. Note that all the methods that were benchmarked in [Capra and Singh, 2008], including GroupSim, use subtype information. We used the dataset generated by [Capra and Singh, 2008] for the evaluation. Following the definitions used in [Capra and Singh, 2008], residues that pass the SDP_O filter (low overlap of residues across subtypes and conserved in at least one subtype) are considered positives and those that do not pass the SDP_L filter (low overlap of residues across subtype) are considered negatives. We used the alignments from this original dataset.

We ran INTREPID-SPEC on this dataset by choosing each protein in turn as the target p, computing an importance score and then averaging this score across all the proteins. Since we are interested in SDPs, we ignore the conservation score at the root during the tree traversal.

2.6.2 Comparison of INTREPID-SPEC to other sequence-based methods for specificity determinant prediction

INTREPID-SPEC, when subtype information is used, has accuracies similar to GroupSim as seen in figure 2.12. ([Capra and Singh, 2008] have shown that GroupSim is competitive with other sequence-based methods suggesting that INTREPID-SPEC would have similar accuracies to these other methods as well). Although INTREPID-SPEC does a tree traversal even when subtype information is provided implicitly, our results show that the maximum scores for the specificity determinants are attained at the point in the tree that separates the known subtypes.

We also ran INTREPID-SPEC on trees constructed without knowledge of subtypes (Figure 2.12). INTREPID-SPEC with subtype information has 10% greater precision across the range of recall values than when no subtype information is available. This difference in performance can be attributed to the bias induced by the rooting of the tree on the process of averaging the INTREPID-SPEC scores across all the sequences in the family. In a family with multiple subtypes, this procedure gives higher ranks to those SDPs that differentiate a subtype that is joined to the rest of the family at the root. This bias explains why trees built using subtype information lead to improved accuracy. When the subtype information is not used, the top ranked residues often separate subtrees which do not correspond to the original subtypes. While such predictions are penalized in our present evaluation, these may be biologically interesting.

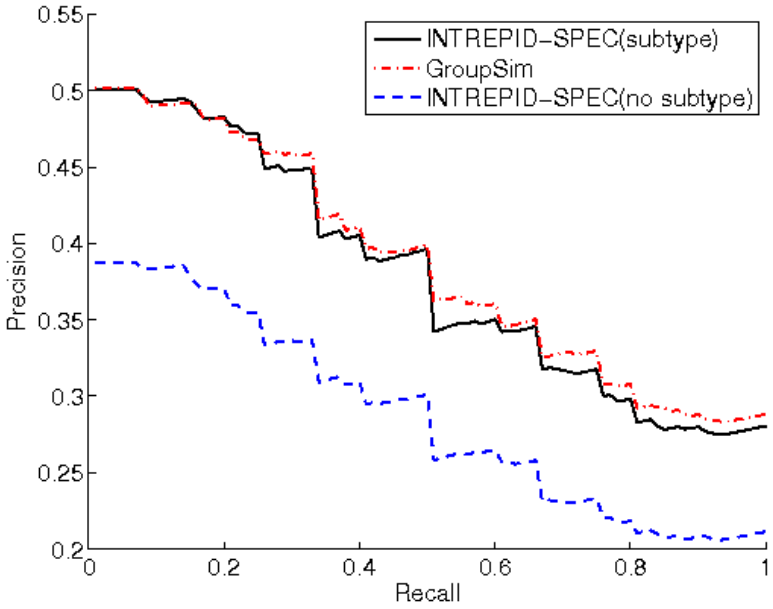

Figure 2.12: Comparison of methods for specificity determinant prediction: INTREPID-SPEC run on trees built using subtype information and INTREPID-SPEC run with no subtype information are compared to GroupSim. INTREPID-SPEC (with subtypes provided) attains accuracies competitive with GroupSim. Including subtype information improves INTREPID-SPEC recall by roughly 10% at all levels.

2.7 Conclusions

In this chapter, we have presented INTREPID, a novel method to predict functional residues from sequence information only. The primary innovation in INTREPID is its use of the phylogeny of the family to infer the evolutionary pressures on positions within different subgroups. INTREPID infers functionally important positions through a traversal of the phylogeny from the root to the target protein located at a leaf; at each point on this path and for each position independently INTREPID computes a positional conservation score based on Jensen-Shannon (J-S) divergence between the distribution of amino acids at that position and a background distribution. Positional scores are adjusted to take into consideration the scores of other positions within the same subtree; thus positional scores for a subtree containing highly similar sequences will be small, even though individual positions may be highly conserved. By contrast, a position that is highly conserved within a subtree that is otherwise highly variable will have a high JS divergence. Each position is then assigned the maximal JS score achieved over all nodes on the path. Positions that are conserved across the

entire family achieve their maximum score at the root, whereas other positions will achieve their maximum at some distance from the root. Since even catalytic residues are not always perfectly conserved across a family (if, for instance, sequences with divergent functions are included in the analysis, or due to alignment errors), this tree traversal enables INTREPID to exploit the information in highly divergent datasets. In fact, our analysis of CSA-defined catalytic residues shows that 18% of catalytic residues in the dataset have their maximum score at least 5 levels from the root of the tree.

We have presented results comparing INTREPID with two of the leading methods in functional residue prediction that make use of sequence information only – Evolutionary Trace (ET) and ConSurf and with a simple baseline method that computes the Jensen-Shannon divergence between the amino acid distribution at a position and a background distribution (Global-JS). We compared each method on a benchmark dataset of 100 manually curated sequence-divergent enzymes from the Catalytic Site Atlas. Our results show that INTREPID has significantly superior accuracy than each of these methods, attaining a sensitivity of 85% at 90% specificity (in contrast, ET and ConSurf attain sensitivities of 70% and 74% respectively at the same specificity) and attaining a recall of about 64% at 10% precision (in contrast neither ET nor ConSurf attain a precision greater than 10%). Since the ConSurf server selects a more conservative set of sequences than those we selected, we also did a separate experiment in which we submitted our larger alignments to the Rate4Site algorithm (the core algorithm within ConSurf). As Rate4Site failed to complete on the full alignments, we filtered the alignments to reduce highly similar sequences. The method performances are very close on the 71 alignments on which Rate4Site completed successfully (ROC analysis shows INTREPID has a small but statistically significant edge over Rate4Site on this dataset).

In addition to these comparisons with methods using sequence information only, we compared INTREPID to the machine learning algorithms reported by [Petrova and Wu, 2006] and by [Youn *et al.*, 2007] which make use of structural information. Surprisingly, on the Petrova dataset, INTREPID is as accurate as their SVM-based method, even though the latter uses both sequence and structure-based features. On the [Youn *et al.*, 2007] dataset, INTREPID is more accurate than the variant of their method that makes use of only sequence features. Reassuringly, the performance of INTREPID is approximately the same across these different datasets suggesting that it would generalize well to new protein families.

To analyze the effect of the evolutionary divergence on prediction accuracy, we created alignments in which the minimum pairwise identity to the seed was restricted. The sensitivity of INTREPID was found to increase as the alignments became more divergent. These results, while in agreement with several previous studies [Landgraf *et al.*, 2001; Aloy *et al.*, 2001; Panchenko *et al.*, 2004], suggest that highly divergent families (with minimum pairwise identity as low as 10%) can significantly improve catalytic residue prediction.

Prediction of active site residues based on sequence information alone is clearly affected

by the quality of the sequence data, in particular, on the effective coverage and extent of the sequence space around the protein of interest. To test the impact on this kind of sequence space coverage, we analyzed the accuracy of INTREPID in predicting catalytic residues for sequences not used as seeds in clustering homologs (i.e., which may be towards the periphery of the sequence space). As expected, accuracy decreases as evolutionary distance to the seed increases. Our limited results suggest that the sequence of interest should have sequence identity $> 50\%$ to the seed.

In summary, the utility of INTREPID in catalytic site prediction can be traced to the following features. First, INTREPID relies solely on sequence information, making it useful when no structural data are available. Secondly, INTREPID is computational efficient, making it useful in large-scale application, and allowing it to be used on large datasets. For instance, INTREPID is considerably faster than Rate4Site, with 400-fold lower average runtime. Third, INTREPID can be used on datasets including highly divergent sequences; in fact, its accuracy improves as more divergent sequences are included. While INTREPID is designed to make use of sequence information alone, it can be used as a component in a prediction protocol that attempts to combine sequence information with other types of information.

On the task of specificity determinant prediction, a variant of INTREPID, INTREPID-SPEC, was as accurate as the GroupSim method proposed by [Capra and Singh, 2008] when both methods were given subtype information. Unlike GroupSim however, INTREPID-SPEC does not require subtype information since the tree traversal provides an implicit grouping of sequences. We found that subtype information results in an improvement in precision of about 10% across the range of recall values.

In this work, we have focused on functional residue prediction in enzymes. In future work, we plan to assess the performance of these methods on non-enzymes as well as on other types of functional residues. Scoring functions that may be better suited to detect other types of conservation signals can be plugged into the INTREPID framework to obtain improved predictions. Finally, all the estimated accuracies of catalytic residue prediction methods depend critically on the characteristics of the dataset used to benchmark method performance. The poor performance of a method on a protein family may simply be the result of insufficient experimental data available for that family.

Appendix 2.A Datasets

For catalytic residue prediction, we identified a set of 100 enzymes from the manually curated section of the Catalytic Site Atlas [Porter *et al.*, 2004] selected to ensure that no pair had detectable homology (*i.e.*, we required a BLAST E-value > 1). We term this the CSA-100 dataset. A PSI-BLAST [Altschul *et al.*, 1997] search was performed with each of these 100 enzymes as a seed against the UniProt database [Apweiler *et al.*, 2004]. PSI-BLAST was run

for 4 iterations with an E-value inclusion threshold of 1×10^{-4}, from which a maximum of 1000 homologs were retrieved. The resulting homologs were realigned using MUSCLE [Edgar, 2004] with MAXITERS set to 2. Identical sequences were discarded. Columns in which the seed had a gap were removed. A neighbor-joining tree was built from this alignment using the PHYLIP package [Felsenstein, 1993]. The dataset has alignments with a minimum of 32 sequences, a maximum of 1033 sequences, and a median of 843 sequences. The average percent identity of the alignments varies from 6.4% to 31.14% with a median of 14.99%. The dataset contains a total of 314 catalytic residues out of a total of 36,229 residues with a median of 3 catalytic residues per enzyme.

For the comparison with the [Petrova and Wu, 2006] dataset, we generated alignments and trees by the protocol described above using the 79 enzymes reported in their paper [Petrova and Wu, 2006]. The resulting dataset contains 244 catalytic residues out of a total of 23,332 residues. For the comparison with the [Youn *et al.*, 2007] dataset, we picked a random domain from each SCOP family for which we generated alignments and trees as described above. This dataset contains 1,172 catalytic residues out of a total of 119,433 residues.

For specificity determinant prediction, we used the alignments from the dataset constructed by [Capra and Singh, 2008]. Neighbor-joining trees were built using the PHYLIP package [Felsenstein, 1993].

Appendix 2.B INTREPID variants

In this section, we define different INTREPID variants based on the positional conservation score $cons(S, x)$ which is used to compute the importance score in Equation 2.3.

$$IMP_p(x) = max_s cons(S, x) - \overline{cons(S)} \tag{2.3}$$

- INTREPID-JS is the method referred as INTREPID in the previous sections. In this variant, $cons(S, x)$ is the Jensen-Shannon divergence between the amino acid distribution and the background amino acid distribution derived from the Blocks database [Henikoff and Henikoff, 1992] with prior weight $= \frac{1}{2}$ as in [Capra and Singh, 2007].
- INTREPID-LO where $cons(S, x)$ is the log probability of the most frequent amino acid at position x within subtree S.
- INTREPID-RE where $cons(S, x)$ is computed as the relative entropy between the amino acid distribution within subtree S of position x and a background distribution derived from the Blocks database alignments [Henikoff and Henikoff, 1992].

We compare these INTREPID variants to their respective global conservation baselines which we term Global-JS [Capra and Singh, 2007], Global-LO, and Global-RE (the scoring function introduced by [Wang and Samudrala, 2006]) respectively. Consistent with the results

reported in [Capra and Singh, 2007], the score based on J-S divergence was found to be the most accurate.

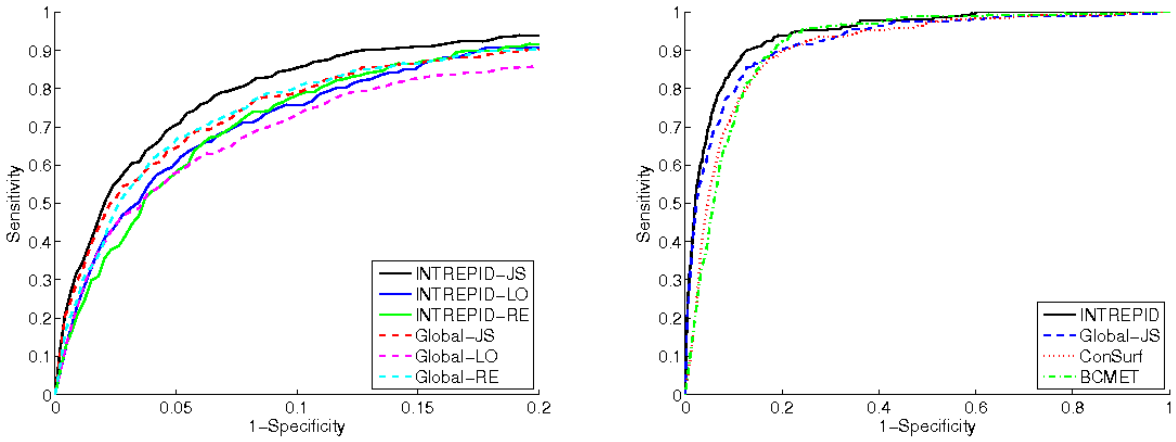

Figure 2.13: Results for catalytic residue prediction on a subset of 100 manually curated enzymes from the Catalytic Site Atlas (termed CSA-100) using rank-based scores. (Left) ROC curve comparing the variants of INTREPID (INTREPID-JS referred to as INTREPID in the text, INTREPID-RE and INTREPID-LO) and the respective variants of global conservation (Global-JS, Global-RE, and Global-LO). See Section 3.E.3 for a detailed description of these variants. INTREPID-JS and INTREPID-LO are significantly more accurate than Global-JS and Global-LO respectively while INTREPID-RE performs worse than Global-RE. (Right) ROC curves comparing INTREPID, Global-JS, BCMET and ConSurf. See Section 2.3.1 for a description of these methods. The ROC curve shows INTREPID-JS to have the highest sensitivity over the range of specificity followed by Global-JS. BCMET performs better as the specificity decreases.

		INTREPID-LO	INTREPID-JS	INTREPID-RE	Global-LO	Global-JS	Global-RE
Residue ranks	Sensitivity_{95}	59.55	70.06	57.32	57.64	64.33	64.97
	Sensitivity_{90}	75.80	85.03	77.39	72.29	78.66	79.94
	Sensitivity_{80}	90.76	93.95	91.72	85.67	90.13	90.45
	Recall_{10}	84.00	91.00	86.00	80.00	86.00	86.00
	AUC	0.923	0.944	0.926	0.907	0.924	0.926
	AUC_{95}	0.019	0.024	0.018	0.019	0.022	0.021
	AUC_{90}	0.054	0.063	0.051	0.051	0.058	0.058
	AUC_{80}	0.139	0.154	0.139	0.131	0.145	0.145
	p-value	–	–	–	3.89×10^{-18}	3.89×10^{-18}	0.17
Normalized scores	Sensitivity_{95}	51.91	67.83	55.10	47.77	58.28	61.15
	Sensitivity_{90}	73.89	85.03	75.48	69.43	76.75	80.25
	Sensitivity_{80}	90.13	92.99	92.68	86.31	89.81	89.49
	Recall_{10}	83.00	92.00	84.00	78.00	83.00	85.00
	AUC	0.911	0.935	0.918	0.891	0.910	0.916
	AUC_{95}	0.014	0.022	0.017	0.014	0.018	0.019
	AUC_{85}	0.046	0.060	0.050	0.045	0.053	0.055
	AUC_{80}	0.130	0.149	0.136	0.124	0.137	0.141
	p-value	–	–	–	3.89×10^{-18}	3.89×10^{-18}	0.33

Table 2.4: Statistics comparing the variants of INTREPID and the baseline global conservation on the CSA-100 dataset. In the top panel, the ranks of the residues were used; in the bottom panel, the normalized scores are shown. Sensitivity is measured at specificities of $80, 90$, and 95% respectively and recall at 10% precision. $AUC_x, x = 80, 90, 95$ refers to the area under the ROC curve when specificity is at least $x\%$; AUC is the area under the entire curve. The p-value refers to the Wilcoxon signed rank p-values between the AUC of the INTREPID variant and the respective global conservation baseline. INTREPID-JS improves over other methods on all metrics. The INTREPID variants, INTREPID-JS and INTREPID-LO improve over their respective baseline while INTREPID-RE performs worse. The confidence intervals on these statistics are reported in Table 2.5.

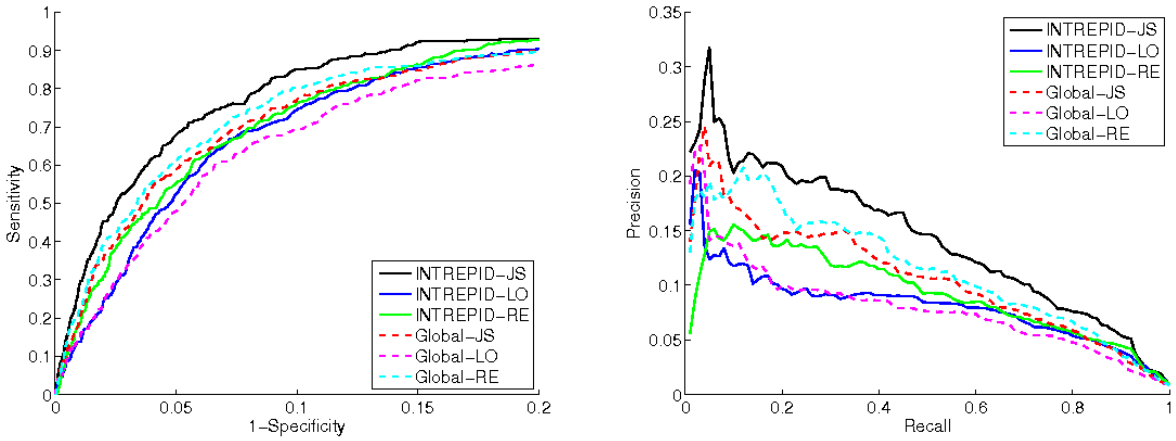

Figure 2.14: Comparison of the INTREPID variants and the respective global conservation baselines based on the normalized scores on a subset of 100 manually curated enzymes from the Catalytic Site Atlas (termed CSA-100).

		INTREPID-LO	INTREPID-JS	INTREPID-RE	Global-LO	Global-JS	Global-RE
Residue ranks	Sensitivity_{95}	[55.39,64.22]	[65.47,74.31]	[51.74,62.24]	[52.41,62.16]	[59.67,68.46]	[60.87,69.87]
	Sensitivity_{90}	[72.02,79.62]	[82.31,87.86]	[73.89,81.38]	[68.22,76.60]	[74.92,82.17]	[76.85,83.67]
	Sensitivity_{80}	[88.29,93.26]	[91.82,95.83]	[89.32,94.23]	[83.00,89.28]	[87.42,92.75]	[87.72,92.69]
	Recall_{10}	[46.00,68.00]	[66.00,81.00]	[35.00,62.00]	[43.00,62.00]	[57.00,72.00]	[59.00,73.00]
	AUC	$[0.914, 0.934]$	$[0.935, 0.952]$	$[0.915, 0.935]$	$[0.893, 0.919]$	$[0.909, 0.936]$	$[0.912, 0.937]$
	AUC_{95}	$[0.016, 0.020]$	$[0.021, 0.026]$	$[0.015, 0.019]$	$[0.016, 0.020]$	$[0.019, 0.024]$	$[0.018, 0.022]$
	AUC_{90}	$[0.049, 0.057]$	$[0.059, 0.066]$	$[0.047, 0.055]$	$[0.047, 0.055]$	$[0.054, 0.061]$	$[0.054, 0.061]$
	AUC_{80}	$[0.132, 0.144]$	$[0.148, 0.159]$	$[0.132, 0.144]$	$[0.126, 0.138]$	$[0.137, 0.150]$	$[0.138, 0.149]$
Normalized scores	Sensitivity_{95}	[46.78,56.68]	[62.93,72.24]	[50.00,59.55]	[42.90,52.56]	[53.66,62.68]	[55.90,65.58]
	Sensitivity_{90}	[69.94,78.59]	[81.36,88.07]	[71.57,79.40]	[64.46,89.79]	[73.31,80.68]	[76.53,83.56]
	Sensitivity_{80}	[87.64,93.20]	[90.61,95.55]	[90.20,94.98]	[83.28,89.79]	[87.12,92.36]	[86.96,92.23]
	Recall_{10}	[5.00,50.00]	[60.00,76.00]	[35.00,57.00]	[9.00,40.00]	[38.00,64.00]	[46.00,66.00]
	AUC	$[0.898, 0.924]$	$[0.924, 0.946]$	$[0.907, 0.929]$	$[0.875, 0.906]$	$[0.896, 0.923]$	$[0.902, 0.929]$
	AUC_{95}	$[0.012, 0.016]$	$[0.020, 0.024]$	$[0.015, 0.019]$	$[0.012, 0.015]$	$[0.016, 0.020]$	$[0.017, 0.021]$
	AUC_{85}	$[0.042, 0.050]$	$[0.057, 0.064]$	$[0.046, 0.053]$	$[0.041, 0.048]$	$[0.049, 0.056]$	$[0.051, 0, 058]$
	AUC_{80}	$[0.123, 0.137]$	$[0.145, 0.156]$	$[0.129, 0.141]$	$[0.119, 0.130]$	$[0.131, 0.142]$	$[0.135, 0.146]$

Table 2.5: Confidence Intervals for statistics comparing the variants of INTREPID and the baseline global conservation on a subset of 100 manually curated enzymes from the Catalytic Site Atlas (termed CSA-100). In the top panel, the ranks of the residues were used while in the bottom panel, the normalized scores were used. Sensitivity is measured at specificities of $80, 90$, and 95% respectively and recall at 10% precision. $AUC_x, x = 80, 90, 95$ refers to the area under the ROC curve when specificity is at least $x\%$; AUC is the area under the entire curve. The 95% confidence intervals are computed from 200 bootstrap replicates.

Chapter 3

Functional site prediction using phylogenomic and structural information

3.1 Introduction

In this chapter, we focus on the task of predicting catalytic residues in enzymes using information from sequence and structure.

The earliest methods for catalytic residue prediction relied on sequence conservation patterns across a family [Casari *et al.*, 1995; Lichtarge *et al.*, 1996; Landau *et al.*, 2005], followed by increasingly powerful sequence-based scoring functions [Aloy *et al.*, 2001; Mihalek *et al.*, 2004; Mayrose *et al.*, 2004; Sankararaman and Sjölander, 2008]. Methods relying exclusively on information from solved 3D structures have been developed, analyzing features such as the geometric arrangements of residues [Fetrow and Skolnick, 1998], surface geometry [Peters *et al.*, 1996], electrostatics [Bate and Warwicker, 2004], energetics [Laurie and Jackson, 2005; Elcock, 2001], and chemical properties [Ondrechen *et al.*, 2001; Tong *et al.*, 2008]. Other methods combine features derived from sequence and structure [Aloy *et al.*, 2001; Landgraf *et al.*, 2001; Gutteridge *et al.*, 2003; Petrova and Wu, 2006; Youn *et al.*, 2007; Innis *et al.*, 2004; Pazos and Sternberg, 2004; Ota *et al.*, 2003].

We present here a new method for predicting catalytic residues, which we have named DISCERN. DISCERN is a statistical predictor that achieves a significant improvement in performance over published reports of catalytic residue prediction through: (i) the use of phylogenomic methods (i.e., methods that analyze the evolution of multi-gene families) to exploit the signal from sequence conservation [Sankararaman and Sjölander, 2008], (ii) the incorporation of features from structural neighbors (i.e., residues near each other in the protein 3D structure), and (iii) the use of regularization—specifically, L_1-regularization [Tibshirani, 1996]—to control model complexity. It is the combination of these three ingredients that underlies the effectiveness of DISCERN. While a predictor that includes both sequence conservation and features from structural neighbors would be expected to have improved

accuracy, including features from structural neighbors yields a proliferation of parameters to be estimated from the data, and, as we show, an unregularized predictor based on these features leads to a decrease in accuracy. Statistical regularization is essential to being able to make effective use of a rich description of the protein for the purposes of prediction.

We report results on cross-validation experiments on the CATRES benchmark dataset of experimentally characterized catalytic residues [Bartlett *et al.*, 2002]. CATRES defines a residue as catalytic if it has been shown to be involved in catalysis either directly or through other molecules, to stabilize an intermediate transition state, or to influence a cofactor or substrate that aids catalysis. Previously, the best *recall* (the fraction of true catalytic residues that are predicted to be catalytic) reported on homology-reduced datasets is 57% at a *precision* (the fraction of predicted catalytic residues that are indeed catalytic) of 18.5% [Youn *et al.*, 2007]. In comparison, DISCERN yields a recall of 69% at the same precision on a homology-reduced version of the CATRES dataset.

3.2 The DISCERN methodology for catalytic residue prediction

The statistical model underlying the DISCERN predictor is a binary logistic regression model [Hosmer and Lemeshow, 2000] that predicts whether a site is catalytic or not based on a list of features describing a site. Logistic regression takes a weighted linear combination of these features and then transforms the result to a probability scale. The parameters (i.e., the weights) in the weighted combination are estimated based on data from the CATRES dataset of experimentally characterized enzymes [Bartlett *et al.*, 2002].

While similar statistical models have been used previously for catalytic residue prediction [Gutteridge *et al.*, 2003; Petrova and Wu, 2006; Youn *et al.*, 2007], DISCERN brings together three ideas that differentiate it from existing predictors. The first is the use of a sequence conservation score based on phylogenomics as a component of the feature vector describing a site. In particular, DISCERN makes use of the INTREPID phylogenomic conservation score, described in the previous chapter. INTREPID is based on a tree traversal that enables it to be applied to highly divergent datasets (e.g., pairwise identities below 5%) and extract a conservation signal that may only appear at deeply nested subtrees in the superfamily phylogeny.

The second critical aspect of DISCERN is its use of information from structurally proximal amino acids. For instance, it is known that the active site is typically conserved structurally across homologs, even when sequence identity is low [Baker and Sali, 2001]. This structural conservation is reflected by high sequence conservation across a family of related proteins in the structural vicinity of the actual catalytic residue(s). The DISCERN predictor represents this fundamental characteristic of active sites by including information from residues that are proximal in the structure.

In addition to features based on the phylogenomic conservation score, our feature vec-

tor includes features that have been noted in large-scale studies as typical of catalytic sites [Bartlett *et al.*, 2002], including relative solvent accessibility, presence in a cleft, secondary structure, charge, and so on. Given that these features are used to describe not only a given site but also its structural neighbors, the resulting model has a large number of features, and overfitting is a concern, particularly given the redundant, noisy nature of many of these features [Hastie *et al.*, 2001]. The third differentiating aspect of DISCERN is thus to use an L_1-regularization procedure to estimate the parameters of the model. This procedure maximizes the likelihood of the logistic regression model under a constraint on the sum of the absolute values of the parameters in the model; such a constrained estimation procedure yields a sparse model in which many of these parameters are set to zero [Tibshirani, 1996]. There is a large literature in statistics justifying this overall approach to estimation, where it is shown that L_1-regularization can yield models that are better predictors than those based on unregularized estimates [Tibshirani, 1996; van de Geer, 2008; Hastie *et al.*, 2001; Greenshtein and Ritov, 2004; Zhao and Yu, 2006]. L_1-regularization has also been used in a number of bioinformatics applications including gene expression microarray analysis [Shevade and Keerthi, 2003; Segal *et al.*, 2003] and genome-wide association studies [Hoggart *et al.*, 2008].

3.2.1 L_1-regularized logistic regression

Given an enzyme i with n_i amino acid residues, we denote by $\boldsymbol{x}_j^{(i)}$ the d-dimensional vector of residue-specific features at site j, $j = 1, \ldots, n_i$, by $\boldsymbol{X}^{(i)}$ the $d \times n$ matrix of all such features, and by $z_j^{(i)} \in \{+1, -1\}$ the catalytic label of residue j (whether the residue is catalytic or not). We denote the set of structural neighborhood features by a $dN \times n$ matrix $\boldsymbol{Y}^{(i)}$. Here N refers to the number of structural neighbors of each residue.

We pick the ten residues closest to residue j to form the set of structural neighbors (the distance $d_{j,k}$ between two residues is defined as the minimum of the distance among all pairs of atoms).[1]

We model the conditional distribution of the random variable $Z_j^{(i)} \in \{+1, -1\}$ by a logistic regression

$$\Pr(Z_j^{(i)} = 1 | \boldsymbol{X}^{(i)}, \boldsymbol{Y}^{(i)}, b, \boldsymbol{w}_1, \boldsymbol{w}_2) = \frac{1}{1 + \exp\left(-\left(b + \boldsymbol{w}_1{}'\boldsymbol{x}_j^{(i)} + \boldsymbol{w}_2{}'\boldsymbol{y}_j^{(i)}\right)\right)}. \tag{3.1}$$

The model has parameters $(b, \boldsymbol{w}_1, \boldsymbol{w}_2)$; b is the intercept term which controls the tradeoff between false positives and false negatives, $\boldsymbol{w}_1$ controls the weights of the residue features while $\boldsymbol{w}_2$ controls the weights of the features from the structural neighbors. Given a training

[1] The choice of ten residues as neighbors is arbitrary. It is also possible to treat the size of the structural neighborhood as a parameter and estimate it.

set of enzymes and their catalytic residue annotations, we can estimate the parameters $(b, \boldsymbol{w}_1, \boldsymbol{w}_2)$. To encode a preference for a "sparse" parameter vector, we adopt a regularized maximum likelihood approach in which we maximize the sum of the likelihood and an L_1 penalty term:

$$\max_{\boldsymbol{w}} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log \Pr(z_j^{(i)} | \boldsymbol{X}^{(i)}, \boldsymbol{Y}^{(i)}, b, \boldsymbol{w}) - \lambda \|\boldsymbol{w}\|_1, \tag{3.2}$$

where $\boldsymbol{w} = (\boldsymbol{w}_1, \boldsymbol{w}_2)$ and where $\|\boldsymbol{w}\|_1 = \sum_k |w_k|$ is the L_1 norm. The non-negative regularization parameter λ controls the sparsity of the estimate of $\boldsymbol{w}$; larger values of λ lead to estimates with increasing numbers of zero components. We chose the value of λ by a cross-validation procedure. The optimization problem is solved using an interior point method as implemented in [Koh *et al.*, 2007].

Enforcing sparsity on the parameter vector using L_1-regularization not only leads to a more interpretable fitted parameter vector; it also helps to prevent *overfitting*. The problem of overfitting, which is well known in statistics [Hastie *et al.*, 2001], is as follows. When the model contains a large number of parameters relative to the size of the *training set*, the model tends to fit the noise in the training set leading to high accuracy on the training set but poor performance on the *test set*. Regularization imposes a constraint on the parameter space (e.g., by limiting the size of the parameters as measured by the L_1 norm) reducing the "effective degrees of freedom" of the model and forcing the model to generalize more effectively.

3.2.2 Features for catalytic residue prediction

The feature vector used in our logistic regression model consists of a total of 528 features—48 features at the residue of interest and at ten neighboring residues. We provide a brief description of these features in this section as well as some of the options we considered; further details are provided in Section 3.A.

3.2.2.1 Sequence conservation features

We made use of three sequence conservation scores. The first, termed Global-JS, is the Jensen-Shannon divergence [Lin and Wong, 1990] between the amino acid distribution at a column and a background distribution derived from the BLOCKS [Henikoff and Henikoff, 1992] database (with prior weight = 0.5 as in [Capra and Singh, 2007]). The other two sequence conservation scores make explicit use of the phylogenetic tree topology using the INTREPID algorithm. The two variants used the Jensen-Shannon divergence (INTREPID-JS) and the log frequency of the modal amino acid (INTREPID-LO).

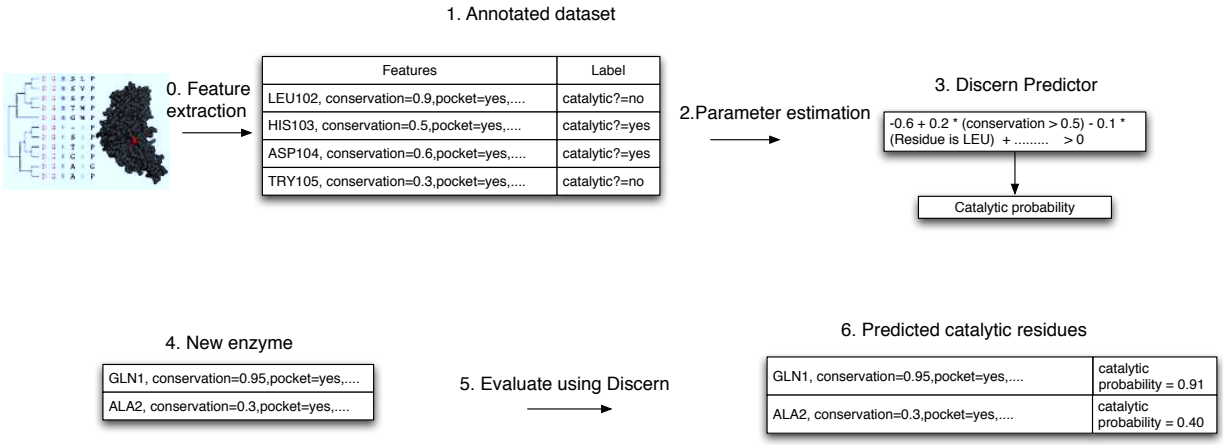

Figure 3.1: **Overview of the system for catalytic residue prediction:** (0) Features are derived from the sequence and 3D structure of an enzyme and from homologs identified using PSI-BLAST. Many features are considered, including the identity of the amino acid, evolutionary conservation scores, and presence in a pocket or cleft. (1) Annotated dataset (training data): A dataset of enzymes with catalytic and non-catalytic residues along with features derived for each residue. (2) We estimate the parameters of the logistic regression model from the training dataset (this is known as a *supervised learning* procedure) using L_1-regularized maximum likelihood. The parameters refer to the weights associated with the features. The L_1-regularization tends to set many of the parameters to zero, resulting in a sparse model. (3) The output of the training phase is a predictor. (4) To predict catalytic residues for a new enzyme, features are derived for the enzyme as in step 1 and the features are used by the logistic regression to classify each residue. (5) The predictor derived in step 3 is used to predict the probability that each residue is catalytic (step 6).

3.2.2.2 Amino acid properties

Amino acids have varying catalytic propensities as noted in [Bartlett *et al.*, 2002]. We use the amino acid types as features and also classify the amino acid into one of three categories—charged (D,E,H,K,R), polar (Q,T,S,N,C,Y) or hydrophobic (A,F,G,I,L,M,P,V,W). See Section 3.A.2 for a description of this classification.

3.2.2.3 Structure-based features

For each residue, we compute the residue centrality, the B-factor, solvent accessibility, presence in a cleft and secondary structure as follows. We compute the B-factor, a measure of thermal motion for each residue as the average of the B-factors of all its atoms. We compute a measure of centrality for each residue j as the inverse of the average distance from a residue to all other residues in the enzyme; i.e., $C_j = \frac{n-1}{\sum_{k \neq j} d(k,j)}$ where $d(k,j)$ is the distance from j to k along the contact map. A residue that is located in the center of the protein has smaller average distance to all other residues and hence a high centrality measure. We use the 7-state secondary structure representation output by DSSP [Kabsch and Sander, 1983]. The area of a residue accessible to solvent is obtained from NACCESS [Hubbard and Thornton, 1993]. We use LigSite^{csc} [Huang and Schroeder, 2006] to detect the presence of a residue in one of the three largest pockets in the enzyme.

3.3 Results

In this section we report results of large-scale experiments on manually curated enzymes from the Catalytic Site Atlas [Bartlett *et al.*, 2002]. We compare DISCERN to the best methods for catalytic residue prediction reported in the literature. Two of these methods make use of machine learning algorithms to combine information from sequence and structure: a neural network approach [Gutteridge *et al.*, 2003] (denoted *NN-Thornton*) and a support vector machine (SVM) method [Youn *et al.*, 2007] (denoted *SVM-Mooney*). We also compare DISCERN to methods that use sequence information only: ConSurf [Landau *et al.*, 2005], Evolutionary Trace (ET) [Mihalek *et al.*, 2004] and INTREPID [Sankararaman and Sjölander, 2008]. Webservers, software or pre-computed results were available for the sequence-based methods, making possible a head-to-head comparison with these methods. However, neither software nor webservers were available for the methods that also use information from 3D structure. We therefore compared DISCERN against NN-Thornton and SVM-Mooney based on precision and recall statistics reported by the authors.

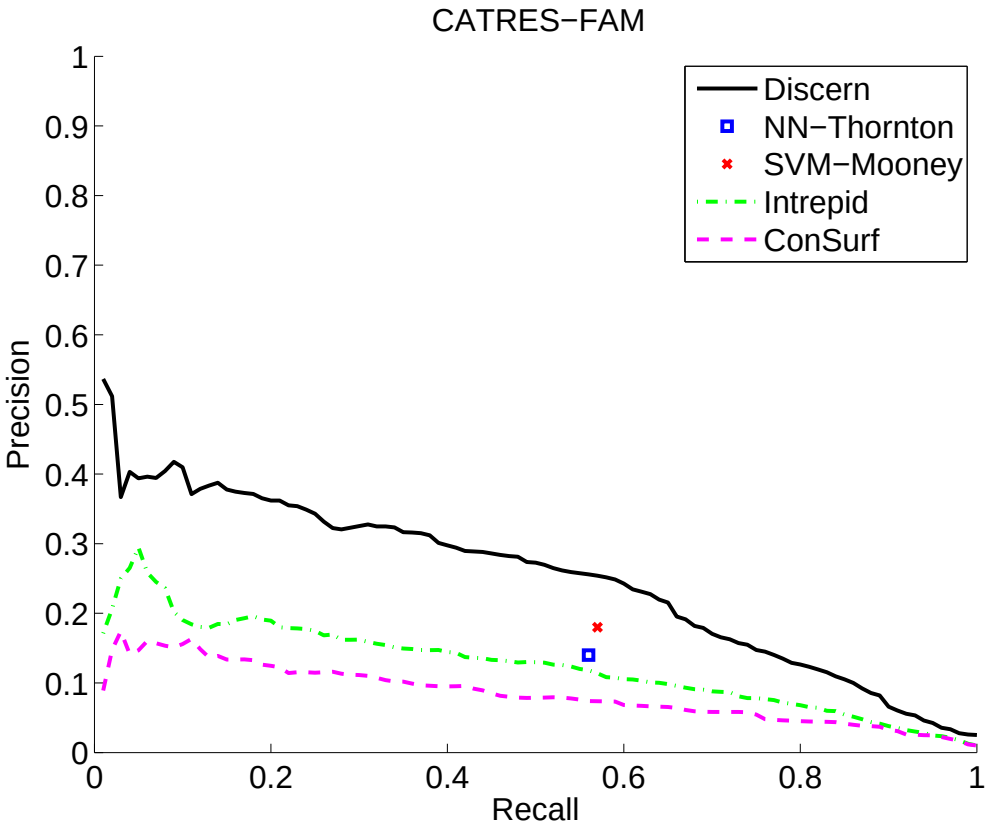

Figure 3.2: Results on the CATRES-FAM benchmark dataset comparing DISCERN against results on related datasets for four state-of-the-art methods for catalytic site prediction: a neural-network method [Gutteridge *et al.*, 2003] (denoted NN-Thornton), a Support Vector Machine approach [Youn *et al.*, 2007] (denoted SVM-Mooney), ConSurf [Landau *et al.*, 2005] and INTREPID [Sankararaman and Sjölander, 2008]. The overlap between CATRES-FAM and the dataset used by Gutteridge *et al.* [Gutteridge *et al.*, 2003] is $> 76\%$. Since the dataset used to validate SVM-Mooney is not available, we report their published performance on a dataset selected such that no pair belongs to the same SCOP family (i.e., a similar selection process as in CATRES-FAM). We include two additional methods for comparison: INTREPID [Sankararaman and Sjölander, 2008] and ConSurf [Landau *et al.*, 2005]. At 18% precision, DISCERN reaches 69% recall, corresponding to an increase of almost 50% over INTREPID which reaches only 19% recall at this precision. At the precision levels reported by methods on related datasets, DISCERN shows a gain in recall of 20% over NN-Thornton and 12% over SVM-Mooney. ConSurf does not reach 18% precision on this dataset.

3.3.1 DISCERN performance on CATRES-FAM

We used 140 enzymes from the CATRES dataset selected such that no pair were from the same SCOP (Structural Classification of Proteins) [Murzin *et al.*, 1995] family to produce a dataset similar to those used in NN-Thornton and SVM-Mooney. We call this dataset CATRES-FAM. This dataset is described in more detail in Section 3.B.

As shown in Figure 3.2, DISCERN recall is 12-20% higher on the CATRES-FAM dataset than that of NN-Thornton and SVM-Mooney at the levels of precision reported by these authors. As expected, the difference is greater in comparison with ConSurf and INTREPID (which makes use of sequence information only): at a precision of 18%, DISCERN has 69% recall while INTREPID reaches only 19% recall and ConSurf does not attain a precision of 18% over the entire range of recalls (see Figure 3.7 and Table 3.4). At a lower precision of 10%, DISCERN obtained a recall of 87% compared to a recall of 64% and 35% by INTREPID and ConSurf respectively.

We also included a control method in these experiments designed to evaluate the contributions of the different ingredients of the DISCERN predictor (i.e., it was trained identically to DISCERN but did not use features for structural neighbors or the INTREPID phylogenomic conservation scores, nor was any attempt made to enforce model sparsity). Notably, the performance of the control is very similar to the results reported in SVM-Mooney, suggesting that the improved performance of DISCERN relative to SVM-Mooney is unlikely to be an artifact of differences between the CATRES-FAM dataset and the datasets used by these authors.

As discussed in the following section, we also performed a detailed analysis of *Escherichia coli* Asparagine Synthetase (PDB id:12as), comparing predictions made by DISCERN, INTREPID, ET and ConSurf. Additional experiments on datasets filtered to remove members from the same SCOP superfamily are reported in Section 3.E.1.

3.3.2 Case Study of a Discern prediction: *Escherichia coli* Asparagine Synthetase (PDB id:12as)

L-Asparagine synthetase catalyzes the conversion of L-aspartic acid and ammonia to L-asparagine in the presence of a magnesium ion while hydrolyzing ATP to AMP and pyrophosphate [Meister, 1974]. L-Asparagine synthetase from *Escherichia coli* has three catalytic residues identified in the CATRES dataset—D46, R100 and Q116 [Nakatsu *et al.*, 1998].

Figure 3.3 compares the predictions of DISCERN with those made by INTREPID, Evolutionary Trace, and ConSurf. Predicted residues are shown at a recall of 100%; i.e., at the point at which all the catalytic residues listed in CATRES have been selected. The number of residues selected by each method is thus equal to the worst rank it gives to a catalytic residue. The figure shows that DISCERN predicts a total of 16 residues. By contrast, IN-

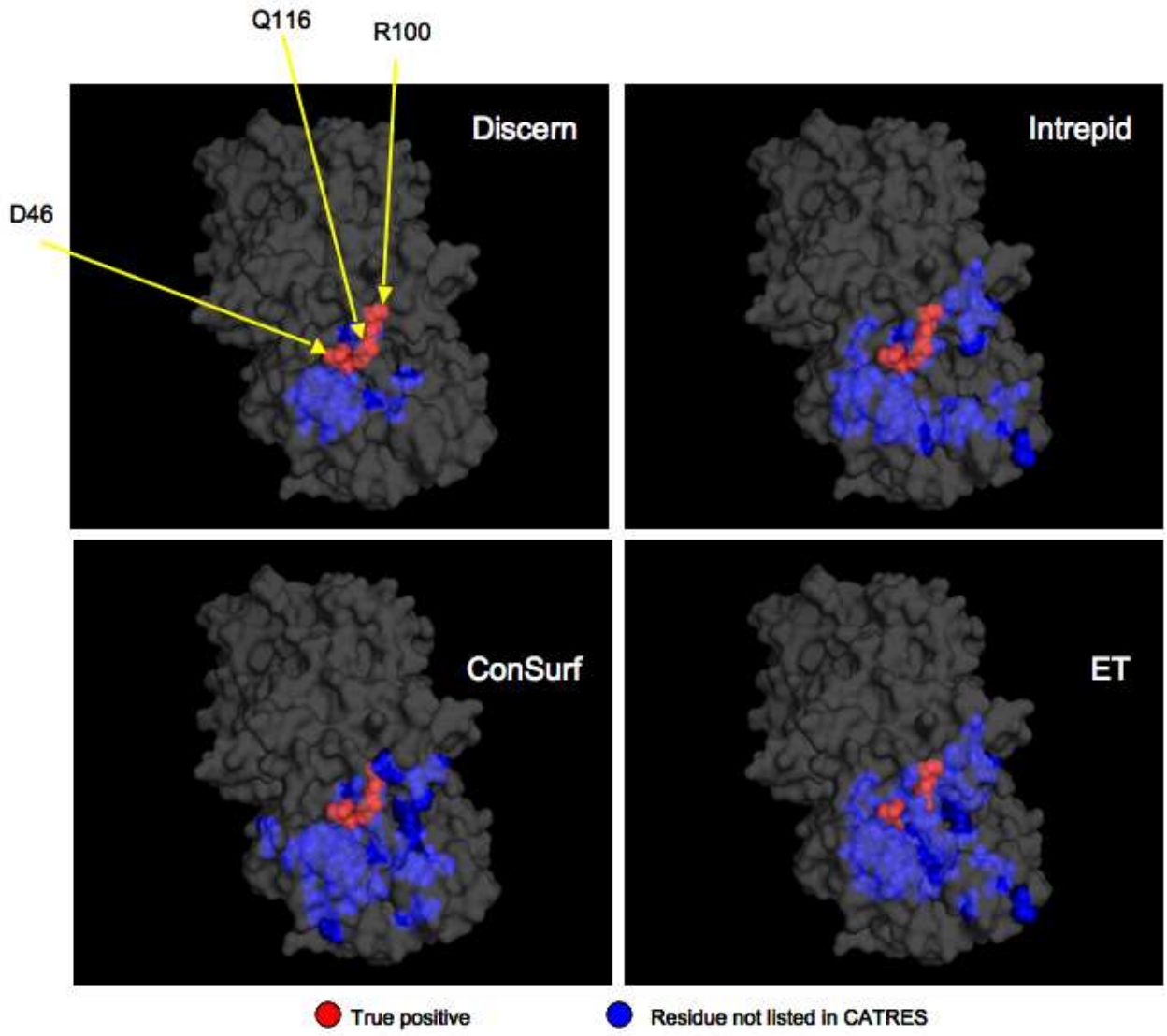

Figure 3.3: Comparison of DISCERN, INTREPID, ConSurf and ET predictions on *Escherichia coli* Asparagine Synthetase (PDB id:12as): The predictions from all methods are shown at a recall of 100%; i.e., when all the catalytic residues listed in CATRES have been selected. DISCERN predicts the three catalytic residues listed in CATRES (D46, R100, and Q116) and 13 additional residues (R214, D115, Y218, D219, D118, E120, H71, K75, K77, R78, D235, E248 and R255) of which seven have been proposed to play functional roles on the basis of structural studies [Nakatsu *et al.*, 1998]. In contrast, INTREPID, ConSurf and ET require a total of 33, 44, and 50 residues respectively to achieve perfect recall. Note that the catalytic residues predicted by the methods are sometimes visually obscured by the false positives. See Table 3.3 for a list of predicted residuesand Figure 3.10 for a view of the active site.

TREPID, ConSurf and ET require a total of 33, 44 and 50 residues respectively to reach perfect recall.

We separately examined the 20 top-ranked residues for DISCERN, INTREPID and ConSurf (see Table 3.3). (The Baylor College of Medicine ET server results were not included since it predicted 50 residues with equal scores.) DISCERN places all three catalytic residues in its top 20, INTREPID detects one, and ConSurf detects two of the three. Moreover, using CATRES to assess the performance of prediction methods underestimates the relative accuracy of DISCERN, because several of the residues that are not listed in CATRES have actually been shown or inferred to play functional roles in the literature. In particular, residues that are described in the literature as playing functional roles but are not listed in CATRES include P35, K77, E120, D219, D235, E248, S251, R255, and I295 [Nakatsu *et al.*, 1998]. Of these nine residues, seven are found among the top 20 for DISCERN, one is found by INTREPID and two are found by ConSurf. The two residues not included in the DISCERN top-20 are P35 (found only by INTREPID), and I295 (found only by ConSurf).

In addition, many of the residues predicted by DISCERN that have not been described in the literature as catalytic are actually found in clusters with residues that have been functionally characterized. These form three sequence motifs that are near each other in the 3D structure but separate in primary sequence (see Figures 3.10 and 3.12). Motif 1 includes H71, K75 and K77. Of these, K77 has been proposed, based on homology with the catalytic domain of yeast class II aspartyl-tRNA synthetase, to interact with the *beta*-carboxylate group of L-aspartic acid [Nakatsu *et al.*, 1998]. Motif 2 includes D115, Q116, D118, W119 and E120; all lie on a single beta strand that lines the active site cleft (referred to as β-6). Of these, Q116 is included in CATRES, and E120 has been proposed to interact with the β-carboxylate group of L-aspartic acid [Nakatsu *et al.*, 1998] (see Figure 3.11). Motif 3 includes R214, Y218, D219 and D220. Of these, the side chain carboxyl group of D219 has been observed to interact with the amino group of the L-asparagine through a water molecule [Nakatsu *et al.*, 1998].

3.3.3 Aspects of the DISCERN predictor

DISCERN combines three ingredients in making a prediction—the use of phylogenomic scores, information from structural neighbors, and a statistical regularization to control for overfitting. To investigate the relative importance of these three aspects of the predictor, we conducted a set of experiments in which subsets of these aspects were used. The results are shown in Table 3.1. We see that a performance gain is obtained by including phylogenomic scores and that—for the unregularized model—a decrease in performance is seen when structural neighborhood features are also included. This is presumably due to overfitting. Indeed, when the model is regularized, a significant performance gain is observed.

DISCERN is not the only method to use information from structural neighbors for catalytic residue prediction, but there are a few differences between DISCERN and approaches used

Method	Structural neighbors	Phylogenomic conservation scores	L_1-regularization	CATRES-FAM	
				Precision_{50}	Recall_{18}
Method 0 (Control)	-	-	-	17.00%	48%
Method 1	-	Y	-	20.45%	55%
Method 2	Y	Y	-	16.13%	41%
DISCERN	Y	Y	Y	27.30%	69%

Table 3.1: *Comparison of* DISCERN *to simplified models.* We compare DISCERN to simplified models that do not include one or more of (1) structural neighborhood features, (2) phylogenomic conservation scores and (3) L_1-regularization. Note that the "Control" only uses a non-phylogenomic sequence conservation score (Global-JS). Precision_{50} reports the precision at 50% recall, and Recall_{18} reports the recall at 18% precision (these precision and recall points were selected to allow direct comparison to the SVM-Mooney method). DISCERN provides an improvement over the control of 10.3% precision at 50% recall and an increase in recall of 21% at 18% precision. See Figure 3.8 for full precision-recall curves.

by others that may contribute to the improved performance. In particular, several methods use spatial clustering [Landgraf *et al.*, 2001; Aloy *et al.*, 2001; Panchenko *et al.*, 2004] as a post-processing step [Gutteridge *et al.*, 2003] based on classification of individual positions independently in an initial stage. In contrast, DISCERN uses features from structurally neighboring residues as an integral part of the model. Closer in spirit to DISCERN is SVM-Mooney [Youn *et al.*, 2007], which uses atom-level features [Bagley and Altman, 1995] in concentric shells (weighted equally within each shell) around the C_β atom of the residue of interest [Mooney *et al.*, 2005]. As in DISCERN, this yields a rich set of features describing the neighborhood. Crucially, however, Youn *et al.* do not use an explicit regularization penalty in fitting their model, and the poorer performance of [Youn *et al.*, 2007] relative to DISCERN may reflect the kind of overfitting that we observe in Table 3.1.

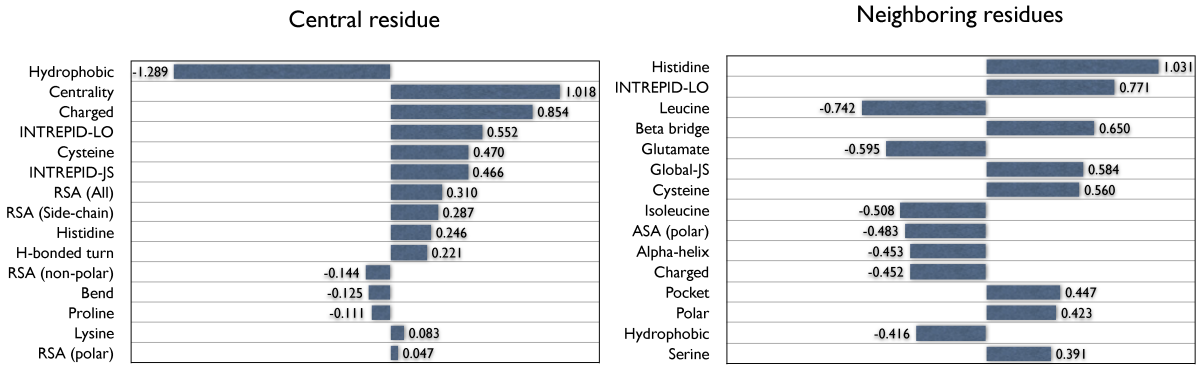

Figure 3.4: Features selected by DISCERN and the weights assigned to each based on fitting the logistic regression to the entire CATRES-FAM dataset and displaying the features with the 15 largest weights. Positive weights indicate positive correlation with putative catalytic residues; negative weights imply negative correlation. The magnitude of the weight is indicative of a feature's relative importance. **Left: Features computed at the residue of interest.** The feature with the largest weight (-1.289) is *hydrophobicity*; the negative weight is consistent with the observation that hydrophobic residues are rarely catalytic. The next highest-ranked feature is *residue centrality* with a weight of 1.018; high values for this feature indicate that a residue is located in the core of the enzyme 3D structure. *INTREPID-LO*, and *INTREPID-JS*, information-theoretic measures of the evolutionary conservation of a residue, jointly have a weight of 1.018, the same weight as centrality. Residue charge comes next with a weight of 0.854, followed by presence of a cysteine (0.470). *Relative solvent accessibility*, a measure of the fraction of a residue exposed to solvent averaged over all the atoms (RSA(All)) and over the side-chain atoms (RSA(Side-chain)) respectively, comes next with weights of 0.310 and 0.246 respectively. **Right: Features summed over residues that are nearby in the 3D structure.** For the purpose of visualization, feature weights are summed over all the structural neighbors; features that have large non-zero weights would tend to be predictive of a structurally neighboring catalytic residue. The feature with consistently large weights are the evolutionary conservation scores (*INTREPID-LO* and *Global-JS*). *INTREPID-LO* and *GLOBAL-JS* (a measure of sequence conservation across the family that does not use the phylogenetic tree) have a combined total weight of 1.255. The feature with the next largest weight (1.031) is the presence of a neighboring *histidine*. Two features with significant weights for residues in the structural neighborhood were *negatively* correlated with catalycity: presence of leucine (-0.742), glutamate (-0.595), and isoleucine (-0.508) and polar *absolute solvent accessibility* (ASA(polar)) (-0.483), i.e., solvent accessibility computed over all oxygens and nitrogens in the sidechain. ASA has large values for amino acids with large absolute surface areas, whereas RSA is normalized by the total surface area in the sidechain. Thus glycine could presumably have a large RSA under some circumstances, but will not have large ASA. The negative correlation of ASA at neighboring positions was unexpected; we hypothesize that this may be due to the function of a catalytic residue being inhibited by the presence of a nearby sidechain protruding into the cleft. The presence of a beta-bridge in the vicinity is indicative of a catalytic residue while an alpha-helix is negatively correlated. Note that the feature weights are summed over the structural neighbors. See Figure 3.9 for additional details.

It is also of interest to investigate quantitative aspects of the full DISCERN predictor after it has been fit to the CATRES dataset (see Figure 3.4). Among the 528 candidate features considered, 157 had non-zero weights in the final model. Examining these weights provides insight into the ability of DISCERN to discriminate between catalytic and non-catalytic residues. The highest weights are associated with features identified by others as highly correlated with catalytic sites (e.g., high degrees of sequence conservation across homologs, centrality in 3D structure and relative solvent accessibility), and the largest negative weights are those shown previously as anti-correlated (e.g., hydrophobicity) [Bartlett *et al.*, 2002].

A more subtle point is the fact that the DISCERN prediction is based on a combination of weighted features. For a residue to achieve a high rank (relative to other residues), a combination of features must be present (or absent, in the case of a feature with negative weights). For instance, while cysteine has a strong positive weight, this alone will be insufficient to rank a cysteine above other residues unless it is also highly conserved and has some level of relative solvent accessibility. This gives DISCERN the ability to differentiate between cysteine residues involved in disulfide bridges from those playing catalytic roles.

Note also that some features may be redundant with other features, or with other feature combinations, and the L_1-regularization may give them a zero weight; in such cases it is not correct to infer that the biological property encoded by the feature is not informative. Presence of a residue in a cleft or pocket is a case in point. We found that the explicit feature of presence in a cleft or pocket is given a weight of zero in our model, which is surprising given that presence in a cleft is known to be one of the hallmarks of catalytic residues [Bartlett *et al.*, 2002]. However, residue centrality and relative solvent accessibility jointly encode for presence in a cleft—if a residue is both near the center of the molecule and exposed, it must be in a deep cleft—and indeed these these two features were present in the final model.

In summary, the features selected by the regularized logistic regression jointly describe highly conserved, charged, solvent-accessible residues that are found in clefts or pockets, and whose neighbors in the 3D structure are also highly conserved.

3.4 Conditional Random Field for catalytic residue prediction

The logistic regression model in DISCERN exploits the structural context by combining features from the structural neighbors but still makes independent predictions of the catalytic label at each residue. In this section, we describe an alternate model based on the framework of Conditional Random Fields (CRFs) [Lafferty *et al.*, 2001]. CRFs allow us to capture contextual information by coupling the labels of the structural neighbors and making a joint prediction across all the residues. In principle CRFs can capture more complex dependencies than a model that treats each residue independently. A dependency of the form *structurally proximal residues X and Y tend to be in the same cleft if they are both catalytic* is one example

since it is a function of the features and the residue labels (which need to be inferred).

We define a CRF for the catalytic residue prediction problem as follows:

$$\begin{aligned}
&\log \Pr(\boldsymbol{z}^{(i)}|\boldsymbol{X}^{(i)}, b, \boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3) = \boldsymbol{w}'\phi(\boldsymbol{z}, \boldsymbol{X}^{(i)}) - Z^{(i)}(b, \boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3) \\
&= b + \sum_{j=1}^{n_i} \left(z_j^{(i)} \boldsymbol{w}_1{}' \boldsymbol{x}_j^{(i)} + z_j^{(i)} \boldsymbol{w}_2{}' \boldsymbol{y}_j^{(i)} + \boldsymbol{w}_3{}' \sum_{k \in N^{(i)}(j)} \psi(z_j^{(i)}, z_k^{(i)}, \boldsymbol{X}^{(i)}) \right) \\
&- Z^{(i)}(b, \boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3),
\end{aligned} \tag{3.3}$$

where $\boldsymbol{w} = (b, \boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3)$ and $Z^{(i)}(b, \boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3) = \log\left(\sum_{\boldsymbol{z}} \exp\left(\boldsymbol{w}'\phi(\boldsymbol{z}, \boldsymbol{X}^{(i)})\right)\right)$ is the log normalizer. Here, in addition to the features used in the logistic regression model, we have extra interaction features ψ to capture dependencies between the labels of two neighboring catalytic residues z_j, z_k. Setting $\boldsymbol{w}_3$ to zero in Equation 3.3 results in the logistic regression model discussed earlier.

To predict the labels of all the residues jointly, we would like to obtain the labeling $\boldsymbol{z}^{(i)*}$ with highest posterior probability.

$$\boldsymbol{z}^{(i)*} = \arg\max_{\boldsymbol{z}} \log \Pr(\boldsymbol{z}|\boldsymbol{X}^{(i)}, b, \boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3). \tag{3.4}$$

The configuration $\boldsymbol{z}^{(i)*}$ can be computed efficiently provided the interaction features ψ are chosen carefully. We use a maximum margin approach to estimate the parameters $\boldsymbol{w}$.

3.4.1 Maximum Margin Parameter Estimation for the CRF

For general interaction features ψ, the problem of computing the maximum *a posteriori* (MAP) configuration z^* of the CRF described in Equation 3.2 is NP-hard [Boykov *et al.*, 2001]. Efficient algorithms based on graph cuts exist for computing z^* when the interaction features are sub-modular; i.e., $\psi(0,0,x) + \psi(1,1,x) \geq \psi(0,1,x) + \psi(1,0,x)$ [Boykov *et al.*, 2001; Kolmogorov and Zabih, 2002; Boykov and Kolmogorov, 2004]. We therefore restrict the model to sub-modular interaction features ψ which take values in $\{0,1\}$—this restriction allows us to estimate the parameters $\boldsymbol{w}$ that respect the sub-modularity constraint for all inputs.

We use a maximum margin approach to estimate the parameters $\boldsymbol{w}$ of the CRF. The

maximum margin framework leads to the following optimization problem

$$\begin{aligned}
&\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i \qquad \text{such that} \\
&\boldsymbol{w}'\left(\phi(\boldsymbol{z}^{(i)}, \boldsymbol{X}^{(i)}) - \phi(\boldsymbol{z}, \boldsymbol{X}^{(i)})\right) \geq L(\boldsymbol{z}^{(i)}, \boldsymbol{z}) - \xi_i, \qquad \forall i = 1, \ldots, m, \forall \boldsymbol{z} \in \{+1, -1\}^{n_i} \\
&\xi_i \geq 0, \qquad \forall i = 1, \ldots, m \\
&\boldsymbol{w}_3\left(\psi(0,0,x) + \psi(1,1,x) - \psi(1,0,x) - \psi(0,1,x)\right) \geq 0 \qquad \forall x.
\end{aligned}$$

The first constraint requires the model to give the highest score to the true labeling $\boldsymbol{z}^{(i)}$. All other labelings are assigned scores lower than the score for the true labeling; the difference in the scores depends on a cost function $L(\boldsymbol{z}^{(i)}, \boldsymbol{z})$. We use the Hamming distance as the cost function—a labeling that is very different from the true labeling should be assigned a lower score than one that is more similar. To handle nonlinearly separable data, we introduce the non-negative slack variables $\xi_i, i = 1 \ldots, m$. The final constraint ensures that the fitted model has no non-sub-modular interaction features so that z^* can be efficiently computed.

We can replace the first constraint with the equivalent

$$\boldsymbol{w}'\phi(\boldsymbol{z}^{(i)}, \boldsymbol{X}^{(i)}) \geq \boldsymbol{w}'(\phi(\hat{\boldsymbol{z}}^{(i)}, \boldsymbol{X}^{(i)})) + L(\boldsymbol{z}^{(i)}, \hat{\boldsymbol{z}}^{(i)})) - \xi_i, \forall i = 1, \ldots, m,$$

where $\hat{\boldsymbol{z}}^{(i)} = \arg\max_{\boldsymbol{z}} \boldsymbol{w}'(\phi(\boldsymbol{z}, \boldsymbol{X}^{(i)})) + L(\boldsymbol{z}^{(i)}, \boldsymbol{z}))$. The Hamming distance loss does not affect any of the interaction features so that $\hat{\boldsymbol{z}}^{(i)}$ can be computed efficiently. The original optimization problem now reduces to

$$\begin{aligned}
&\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{m}\boldsymbol{w}'\left(\phi(\hat{\boldsymbol{z}}^{(i)}, \boldsymbol{X}^{(i)}) + L(\boldsymbol{z}^{(i)}, \hat{\boldsymbol{z}}^{(i)}) - \phi(\boldsymbol{z}^{(i)}, \boldsymbol{X}^{(i)})\right) \\
&\boldsymbol{w}_3(\psi(0,0,x) + \psi(1,1,x) - \psi(1,0,x) - \psi(0,1,x)) \geq 0 \qquad \forall x.
\end{aligned}$$

This is a convex program with a non-differentiable objective function which we solve using a subgradient method. In practice, we use the L_1-regularized logistic regression to estimate the parameters $(b, \boldsymbol{w}_1, \boldsymbol{w}_2)$, discard the zero weights and only estimate the interaction parameter vectors $(\boldsymbol{w}_2, \boldsymbol{w}_3)$.

3.4.2 Features used in the CRF

In addition to the features used in the logistic regression, we compute three additional feature functions for the CRF (described by the ψ terms in Equation 3.2). Each of these feature functions operates on pairs of neighboring residues; i.e., a pair is predicted as catalytic or not catalytic if they share one of these features: charged, polar or conserved. (Recall that $z_j = 1$ is residue j is predicted catalytic). The first two feature functions couple two neighboring

residues if they are both polar or both charged. The last feature function couples two neighboring residues that are both highly conserved (the INTREPID scores are normalized to have zero mean and unit variance for each enzyme).

$$\psi_1(z_j, z_k, x) = \begin{cases} 1 & \text{if } z_j = z_k = 1 \ \& \ j,k \text{ are polar} \\ 0 & \text{otherwise} \end{cases}$$

$$\psi_2(z_j, z_k, x) = \begin{cases} 1 & \text{if } z_j = z_k = 1 \ \& \ j,k \text{ are charged} \\ 0 & \text{otherwise} \end{cases}$$

$$\psi_3(z_j, z_k, x) = \begin{cases} 1 & \text{if } z_j = z_k = 1 \ \& \text{ INTREPID scores for j,k} > 1 \\ 0 & \text{otherwise} \end{cases}$$

3.4.3 Comparison of CRF to the L_1-regularized logistic regression

Table 3.2: **Comparison of DISCERN and the CRF.** Precision_{50} reports the precision at 50% recall, and Recall_{18} reports the recall at 18% precision (these precision and recall points were selected to allow direct comparison to the results reported in [Youn *et al.*, 2007]). The results are indistinguishable.

Method	CATRES-FAM	
	Precision_{50}	Recall_{18}
DISCERN	27.3%	69%
CRF	26.9%	69%

We see from Table 3.2 that the CRF has very similar accuracies to DISCERN with no change in recall on the CATRES-FAM dataset. The extra structural features used in the CRF attained low weights with the highest weight (0.122) being assigned to the feature that enforces agreement between two structural neighbors if each appears conserved. This is likely a result of the small number of catalytic sites observed in the dataset so that the new features introduced by the CRF do not capture any dependencies in addition to those captured at the feature level by the logistic regression model.

3.5 Discussion

In this chapter, we have described a new approach to the prediction of functional sites in proteins. DISCERN is a statistical predictor that brings together three important ideas, the combination of which are needed in order to obtain the striking improvements in accuracy that we obtain. First, DISCERN uses an evolutionary modeling approach (specifically, the

INTREPID phylogenomic method) to infer the degree to which residues are under selective pressure. Second, we incorporate information from the structural neighborhood of a residue including features (such as sequence conservation, charge, solvent accessibility, etc.) computed for structurally proximal residues. Third, and critically, we use statistical sparsification methods (specifically, L_1 regularization) to cope with the fact that our statistical model is based on a large number of redundant, noisy features. Without such regularization, we find that our method overfits—in particular the inclusion of information from structural neighbors leads to a decrease in accuracy. With regularization, we obtain a significant increase in accuracy. Regularization allows us to find a signal within the large set of candidate features that can be used to describe the structural and evolutionary neighborhood of an amino acid.

The parameters of the statistical model underlying DISCERN are the weights of various features that capture the evolutionary and structural context, computed both for the residue of interest and for its structural neighbors. The largest weights tend to be associated with features identified by others as highly correlated with catalytic sites (e.g., high degrees of sequence conservation across homologs, centrality in 3D structure and relative solvent accessibility), and the largest negative weights are those shown previously as anti-correlated (e.g., hydrophobicity). But the model is not restricted to such known features; it can create new features as linear combinations of the given features. Moreover, the model parameters act in concert: for a residue to achieve a high rank, a single feature is generally insufficient; multiple features must be present. This gives DISCERN the ability to differentiate between highly conserved residues playing functional roles from those that may be conserved for structural reasons.

While most catalytic site prediction methods exploit residue conservation across homologs as a primary source of signal [Gutteridge *et al.*, 2003; Youn *et al.*, 2007], most methods restrict homologs to closely related (or only moderately divergent) sequences, limiting the effective use of this signal [Landgraf *et al.*, 2001; Aloy *et al.*, 2001; Panchenko *et al.*, 2004; Sankararaman and Sjölander, 2008]. By contrast, DISCERN makes use of the INTREPID phylogenomic conservation score, which is able to exploit the conservation information in highly divergent sequence homologs.

We also considered an extension to logistic regression, based on the framework of Conditional Random Fields (CRF). CRF methods go beyond a simple logistic regression to allow the coupling of catalytic labels for different residues, enabling us to capture more complex dependencies and to make a joint prediction of the residue labels. In practice, we find that the accuracy of the CRF is virtually indistinguishable from DISCERN.

We have evaluated DISCERN on a homology-reduced subset of manually curated enzymes from the Catalytic Site Atlas [Bartlett *et al.*, 2002; Porter *et al.*, 2004]. While the CSA provides an essential benchmark for the prediction of catalytic sites, as the results in our case study show, not all functionally important or catalytic residues are listed in the CSA. Thus, some residues that are predicted as functional by a method may be labeled as false

positives based on not being present in the CSA even if they are, in fact, catalytic. Finite resources (e.g., a small number of biological curators entering data into the CSA and the inevitable lag between publication and data entry) make the development and maintenance of such a critical resource challenging. Used carefully, automated predictors such as DISCERN can help in surmounting this challenge.

Finally, our case study indicates that DISCERN identifies residues that are involved in other functions such as ligand-binding. In fact, the general approach underlying DISCERN is extensible and general, and can be applied to model other types of functional residues such as binding pocket specificity determinants and interaction interfaces. Each of these application areas depends only on the availability of high-quality training data, such as that provided in the Catalytic Site Atlas.

Appendix 3.A Features evaluated for catalytic residue prediction

The DISCERN logistic regression predictor is based on a feature vector having 528 component features. See Table 3.5.

3.A.1 Sequence conservation features

Sequence conservation has been observed to be the most important feature for catalytic residue prediction [Gutteridge *et al.*, 2003; Youn *et al.*, 2007]. We tested three sequence conservation scores. The first, GLOBAL-JS, is the Jensen-Shannon divergence [Lin and Wong, 1990] between the amino acid distribution at a column and a background distribution (with prior weight = 0.5 as in [Capra and Singh, 2007]). The other two sequence conservation scores tested make explicit use of the phylogenetic tree topology using the INTREPID algorithm. INTREPID has been shown to be sensitive for catalytic residue prediction in general and in particular is able to exploit the information in large divergent families. The two variants used the Jensen-Shannon divergence (INTREPID-JS) and the log frequency of the modal amino acid (INTREPID-LO).

3.A.1.1 Homolog selection and alignment

PSI-BLAST [Altschul *et al.*, 1997] was run for four iterations against the UniProt database [Apweiler *et al.*, 2004] with an E-value inclusion threshold of 1×10^{-4} from which a maximum of 1000 homologs were retrieved. A multiple sequence alignment (MSA) was estimated using MUSCLE [Edgar, 2004] with MAXITERS set to 2, followed by removing identical sequences and deleting columns in which the seed had a gap. The set of alignments built contain a minimum of 32 sequences, a maximum of 1033 sequences, and a median of 839 sequences.

The average percent identity between the seed sequence and homologs in the alignments varies from 6.42% to 31.14% with a median of 15.22%. Percent identity was computed as the fraction of the alignment columns that have identical characters in the sequence and the seed (i.e., the number of identical columns divided by the number of amino acids in the seed). The low percent identity is partly attributed to the inclusion of many sequences with local alignments in the MSA.

3.A.1.2 Tree construction

A neighbor-joining tree was built from this alignment using the PROTDIST and NEIGHBOR programs in the PHYLIP package [Felsenstein, 1993]. The programs were run with default parameters. We used midpoint rooting (placing the root at the midpoint of the longest span in the tree).

3.A.2 Amino acid properties

Amino acids have varying catalytic propensities. We use the amino acid types as features and also classify the amino acid into one of three categories—charged (D,E,H,K,R), polar (Q,T,S,N,C,Y) or hydrophobic (A,F,G,I,L,M,P,V,W). We used the classification described in [Bartlett *et al.*, 2002] with one modification. Tryptophan is included among the class of polar residues in [Bartlett *et al.*, 2002] but among hydrophobic residues by others [Eisenberg *et al.*, 1982]; we use the latter classification.

3.A.3 Structure-based features

For each residue, we compute the residue centrality, the B-factor, solvent accessibility, presence in a cleft and secondary structure as follows. We compute the B-factor, a measure of thermal motion for each residue as the average of the B-factors of all its atoms (derived directly from its PDB file). We compute a measure of centrality for each residue j as the inverse of the average distance from a residue to all other residues in the enzyme; i.e., $C_j = \frac{n-1}{\sum_{k \neq j} d(k,j)}$ where $d(k,j)$ is the distance from j to k along the contact map. A residue that is located in the center of the protein has smaller average distance to all other residues and hence a high centrality measure. We use the 7-state secondary structure representation output by DSSP [Kabsch and Sander, 1983]. The area of a residue accessible to solvent is obtained from NACCESS [Hubbard and Thornton, 1993]. We use LigSite^{csc} [Huang and Schroeder, 2006] to detect the presence of a residue in one of the three largest pockets in the enzyme.

Appendix 3.B Benchmark datasets

We used three datasets in these experiments, all derived from the CSA.

Our primary benchmark dataset, termed CATRES-FAM, consists of 140 enzymes from the CATRES [Bartlett *et al.*, 2002] dataset. The CATRES dataset consists of enzymes with PDB structures with catalytic site information assigned from the literature. Subsets of this dataset have been used by previous methods for catalytic residue prediction [Gutteridge *et al.*, 2003; Tong *et al.*, 2008]. The original CATRES dataset contains 178 enzymes. We discarded 26 enzymes as unusable in these experiments for various reasons: 21 enzymes presented problems for one or more of our feature extraction programs (18 had catalytic sites spanning multiple sub-units, and three enzymes had non-numeric PDB residue identifiers), one of the enzymes had no annotated catalytic residues, one had only one detectable homolog using PSI-BLAST, MUSCLE crashed on another, and two NMR structures were also discarded as unusable by the structure-based methods. The resulting set of enzymes was made non-redundant at the SCOP (Structural Classification of Proteins) [Murzin *et al.*, 1995] family level by removing an additional 12 enzymes. SCOP is a hierarchical classification of protein domains based on their structural, functional and sequence similarities. Domains in different SCOP folds are unrelated; domains in the same fold but different superfamilies have an uncertain relationship (i.e., although their topologies are similar, there is no other evidence to support homology); domains in the same superfamily are deemed homologous; domains in the same family have very similar functions and structures. The resulting dataset contains a total of 472 catalytic residues out of a total of 49,180 residues with a median of three catalytic residues per enzyme.

CATRES-SF is a second dataset of 121 enzymes that was created ensuring that no pair of enzymes belongs to the same SCOP superfamily. This dataset is thus filtered at a more stringent level, presenting a greater challenge to statistical models using this dataset in cross-validation.

The third dataset, CSA-FAM, consists of a set of 94 enzymes chosen from the manually curated section of the Catalytic Site Atlas [Porter *et al.*, 2004] such that no pair contained domains in the same SCOP family and no pair had detectable sequence homology (enforced by a BLAST E-value >1). We also required each of the sequences in this dataset to have pre-computed results in the Baylor College of Medicine Evolutionary Trace server to enable a direct comparison with Evolutionary Trace without putting undue load on their servers.

Appendix 3.C Performance measurements

We measure the precision and the recall on the test set where: Precision $= \frac{TP}{TP+FP}$, Recall $= \frac{TP}{TP+FN}$, a true positive (TP) is a residue included in the benchmark dataset that is predicted as catalytic, a false positive (FP) is a residue not listed in the benchmark that is predicted as catalytic, and a false negative (FN) is a catalytic residue in the benchmark which has been missed by a method. The precision-recall curves were averaged over all the cross-validation folds using the code from [Davis and Goadrich, 2006].

3.C.1 A note on cross-validation

To assess the performance of our method, we performed 10-fold cross validation over the enzymes in the benchmark dataset. k-fold cross-validation is a procedure to evaluate the accuracy of a predictor. The data is partitioned into k equal-sized subsets. In each fold, one partition is chosen as the test data and the rest of the data forms the training data; e.g., in 10-fold cross-validation, $9/10^{th}$ of the data would be used to estimate the model parameters, and then tested on the reserved $1/10^{th}$ of the data. In the next fold, a different $1/10^{th}$ is used to test. The accuracy of the predictor, as measured on the test dataset, is averaged over the folds to obtain a final estimate of the accuracy. Note that in cross-validation, the characteristics of the dataset can have a major impact on the performance. In particular, the presence of homologs in the dataset can lead to an increase in the *apparent* accuracy (i.e., an overestimate of the expected accuracy of the method when applied to novel data) when these homologs occur in both the training and the test set [Youn *et al.*, 2007]. We also observe a similar decrease in accuracy on the CATRES-SF dataset (non-redundant at the SCOP superfamily level) relative to CATRES-FAM (non-redundant at the SCOP family level). The L_1-regularization parameter was estimated by a similar cross-validation within the training set in each fold of the cross-validation.

Appendix 3.D Note on methods compared against

3.D.1 ConSurf and Evolutionary Trace results

The ConSurf-DB database of pre-computed results (http://consurfdb.tau.ac.il) was used to obtain results on the CATRES sequences while the ConSurf web server at Tel Aviv University (`http://consurf.tau.ac.il`) was used to obtain the results on CSA-FAM. Evolutionary Trace results were obtained from the pre-computed results of the Evolutionary Trace server at the Baylor College of Medicine (`http://mammoth.bcm.tmc.edu/report_maker`).

3.D.2 SVM-Mooney

SVM-Mooney [Youn *et al.*, 2007] refers to an implementation of a Support Vector Machine (SVM) for catalytic residue prediction. The features used in SVM-Mooney include amino acid residue type, sequence conservation, the structural environment of each residue represented by 4 shells of thickness $1.875\mathring{A}$, each consisting of 264 atom-based descriptors [Bagley and Altman, 1995], and a structural conservation obtained by comparing the structural environment at each residue. SVM-Mooney was evaluated by 10-fold cross-validation on three datasets derived from the set of 987 protein domains, classified into 396 families, 236 superfamilies and 189 folds, in ASTRAL 40v1.65 [Chandonia *et al.*, 2004]. Each of these datasets

was chosen to be non-redundant at the SCOP fold, superfamily and family levels respectively. SVM-Mooney attained a recall of 57.02% at a precision of 18.51% on the family-level dataset, a recall of 53.93% at a precision of 16.90% on the superfamily level dataset, and a recall of 51.11% at a precision of 17.13% on the fold-level dataset.

3.D.3 NN-Thornton

NN-Thornton [Gutteridge *et al.*, 2003] refers to an implementation of a neural network for catalytic residue prediction. The features used in NN-Thornton include amino acid residue type, sequence conservation features and structural features such as presence in a pocket, B-factor and solvent accessibility. Each residue was classified using the above features computed at the residue alone; i.e., features computed at the structural neighbors were not considered for prediction. The neural network was evaluated by 10-fold cross-validation on subset of 159 enzymes from the CATRES dataset. NN-Thornton attained a recall of 56% at a precision of 14%.

Appendix 3.E Results on additional datasets

Results on CATRES-FAM are reported in Section 3.3. In this section, we report results on the two other datasets.

3.E.1 DISCERN performance on CATRES-SF

CATRES-SF was selected so that no pair belongs to the same SCOP superfamily, making the dataset quite challenging. In contrast, the CATRES-FAM contains enzymes from distinct SCOP families (allowing sequences from the same superfamily to be included). Figure 3.5 shows that, as with other methods, the accuracy of DISCERN decreases on CATRES-SF. At a precision of 17%, DISCERN attains a recall of 65% on CATRES-SF compared to a recall of 70% on CATRES-FAM. Relative to the performance reported by SVM-Mooney on a dataset of enzymes made non-redundant at the SCOP superfamily level, on which they report a recall of 53.9% at 16.9% precision, DISCERN attains an improvement of 11% at the same level of precision.

3.E.2 DISCERN performance on CSA-FAM

The CSA-FAM dataset was designed to enable a direct comparison with Evolutionary Trace (ET) using pre-calculated results from the Baylor College of Medicine ET server [Mihalek *et al.*, 2004]. CSA-FAM consists of a set of 94 enzymes chosen from the manually curated section of the Catalytic Site Atlas [Porter *et al.*, 2004] selected such that pre-calculated

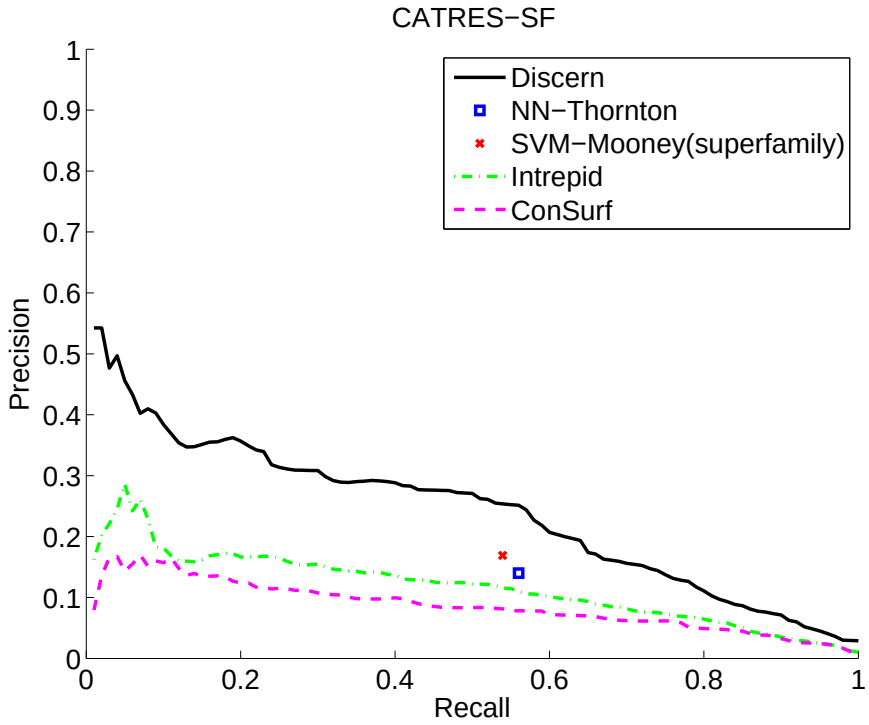

Figure 3.5: **Results on the CATRES-SF benchmark dataset comparing DISCERN against the SVM-Mooney method, the NN-Thornton method, INTREPID and ConSurf**. SVM-Mooney results shown are from their reported performance on a dataset containing single representatives from SCOP superfamilies (i.e., a similar dataset as CATRES-SF) on which they report a recall of 53.93% (the fraction of catalytic residues identified) at a precision of 16.90%. NN-Thornton results are from their reported performance on the CATRES dataset, which includes sequences from the same SCOP family (i.e., an easier dataset), on which they report 56% recall at 14% precision. These results show that DISCERN attains an improvement in recall of 11% over the SVM-Mooney superfamily-level results (achieving a recall of 65% at 17% precision relative to a recall of 53.93% reported by SVM-Mooney at the same precision), an improvement in recall of 16% over the NN-Thornton results at 14% precision, and an improvement of 34% over INTREPID at 18% precision. ConSurf does not reach 18% precision on this dataset.

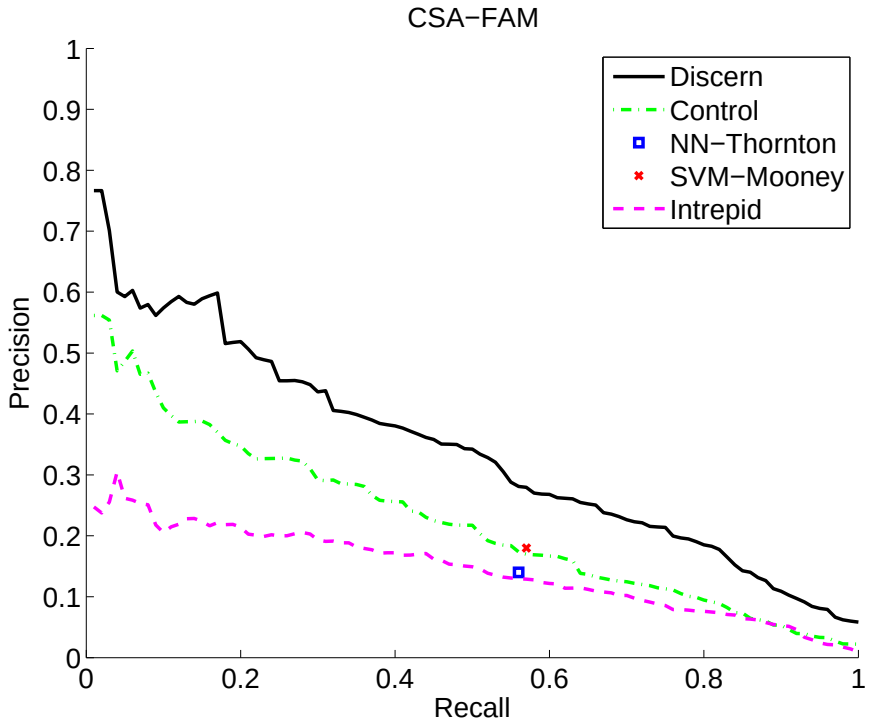

Figure 3.6: **Results on the CSA-FAM benchmark dataset comparing DISCERN against the SVM-Mooney method, the NN-Thornton method, INTREPID and a control**. SVM-Mooney results shown are from their reported performance on a dataset containing single representatives from SCOP families. SVM-Mooney reports 57.02% recall (the fraction of catalytic residues identified) at 18.5% precision (the fraction of predicted catalytic residues that are catalytic) on their family dataset. NN-Thornton results are from their reported performance on the CATRES dataset, which includes sequences from the same SCOP family, on which they report 56% recall at 14% precision. The control was trained identically to DISCERN but did not make use of INTREPID scoring functions or structural neighbors, and did not use L_1-regularization to enforce model sparsity. These results show that DISCERN attains an improvement in recall of 23% over the SVM-Mooney family-level results (achieving a recall of 75% at 18.5% precision relative to a recall of 57.02% reported by SVM-Mooney at the same precision), an improvement in recall of 26% over the NN-Thornton results at 14% precision, and an improvement of 39% over INTREPID at 18% precision. DISCERN also shows an improvement of 21% over the control at a precision of 18%.

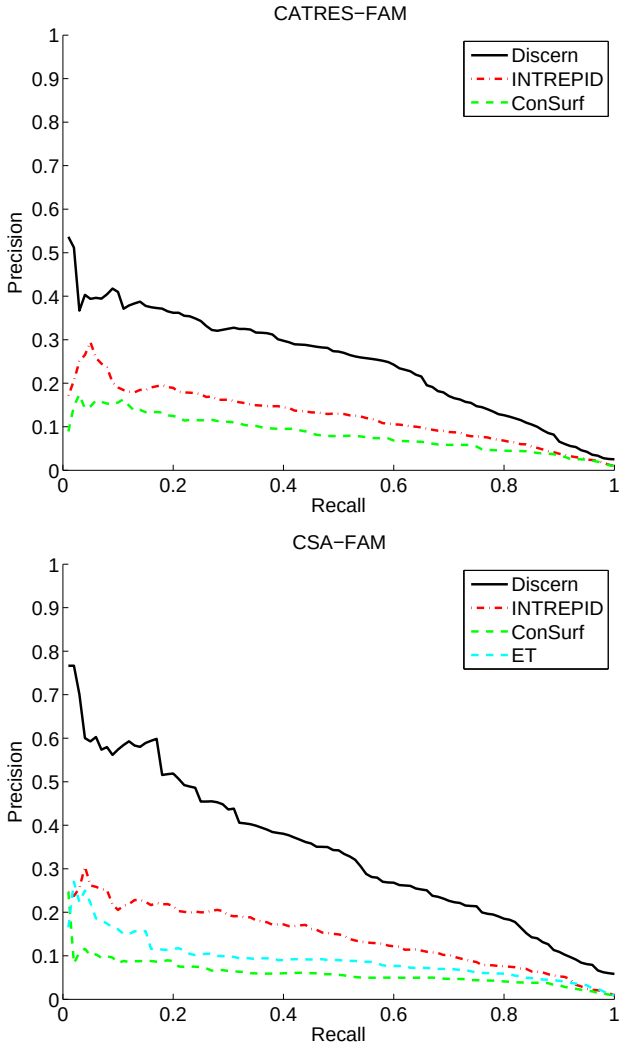

Figure 3.7: **Comparison of DISCERN to methods that rely only on protein sequence information.** Only DISCERN makes use of structural information giving it a significant advantage in these experiments. **Left:** On the CATRES-FAM dataset, at 18% precision, DISCERN has 69% recall and INTREPID has 19% recall while ConSurf does not attain a precision of 18%. At a lower precision of 10%, DISCERN obtained a recall of 87% compared to a recall of 64% and 35% by INTREPID and ConSurf respectively. **Right:** On the CSA-FAM dataset, at a precision of 10%, DISCERN has 90% recall while INTREPID, ConSurf and Evolutionary Trace (ET) have 71%, 3% and 31% recall respectively. ET results were obtained form the Baylor College of Medicine Evolutionary Trace server. ConSurf results were obtained from the ConSurf server DataBase (http://consurfdb.tau.ac.il).

results were available for ET, no pair had detectable sequence homology (enforced by a BLAST E-value > 1), and no pair came from the same SCOP family. We retrieved results for each enzyme from the ConSurf server [Landau *et al.*, 2005]. INTREPID scores were produced based on homologs gathered using PSI-BLAST.

We included the reported performance of the support vector machine (SVM) method from the Mooney lab [Youn *et al.*, 2007] (denoted SVM-Mooney) on a dataset derived from the CSA such that only a single representative from each SCOP family was included. SVM-Mooney reports 57.02% recall at 18.5% precision on their SCOP family dataset. We also included the neural network method from the Thornton lab (denoted NN-Thornton) which achieves a recall of 56% at 14% precision on the CATRES dataset.

Figure 3.6 shows that DISCERN attains an improvement in recall of 23% over the family-level results reported by SVM-Mooney (DISCERN achieves a recall of 75% at 18.5% precision compared to a recall of 57.02% at 18.5% precision achieved by SVM-Mooney). At a precision of 10%, Discern attains a recall of 90% while INTREPID, ConSurf and ET, all of which use only sequence information, attain 71%, 3% and 31% recall respectively (see Figure 3.7).

We include a control method, trained identically to DISCERN but not making use of INTREPID scoring functions or structural neighbors and without the use of L_1-regularization to enforce model sparsity (see Section 3.E.3 for additional details). DISCERN also shows an improvement of 21% over the control at a precision of 18%.

3.E.3 Controlled experiments to test the effect of including phylogenomic conservation score, features computed for structural neighbors, and L_1- regularization

The accuracy of the DISCERN predictor depends critically on the inclusion of discriminative features while avoiding model overfitting. To assess the relative contribution of different features we tested the predictive power of statistical models trained identically to DISCERN but withholding certain features. Performance was assessed on the CATRES-FAM dataset using 10-fold cross validation. Table 3.1 gives details on individual models and Figure 3.8 shows full precision-recall curves on the CATRES-FAM dataset.

Method 0, our control, is an unregularized logistic regression with no features from structural neighbors and no phylogenomic conservation scores (i.e., it uses only GLOBAL-JS, a measure of the family-wide conservation), similar to methods that exploit information from both sequence and structure but do not use features computed at structural neighbors, do not exploit the phylogenetic information and do not use L_1-regularization to enforce sparsity. The control attains a recall of 48% at 18% precision on the CATRES-FAM dataset.

Method 1 is identical to the control but includes INTREPID phylogenomic conservation scores. Including INTREPID provides an increase to recall of 7% over the control (recall=55%) at 18% precision.

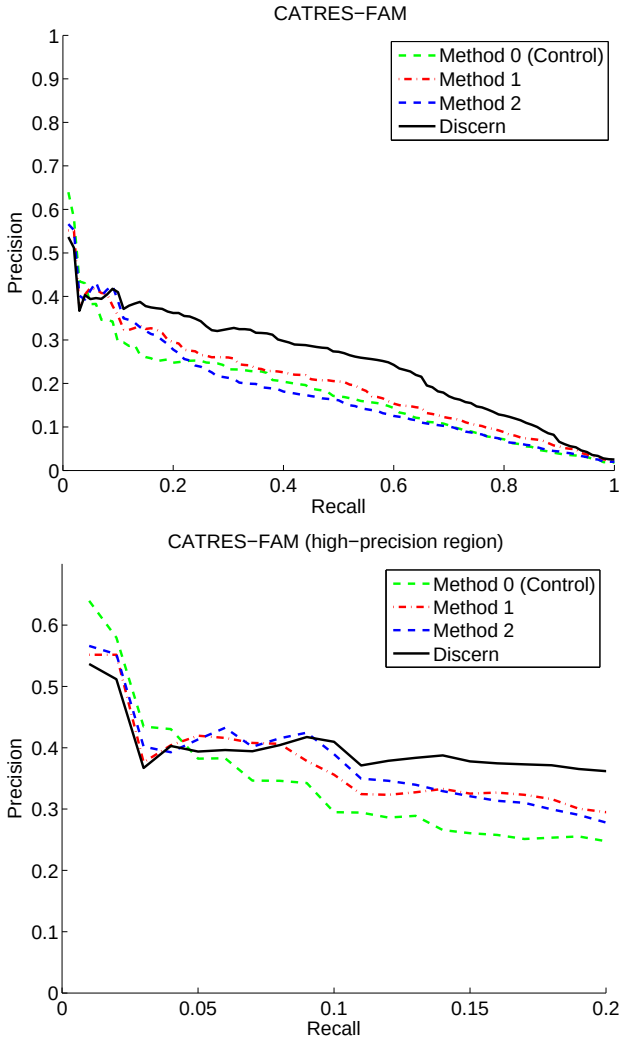

Figure 3.8: **Precision-Recall curves comparing the different predictors on the CATRES-FAM dataset**. **Left:** Full Precision-recall curves. **Right:** Precision-recall curves for the high precision region (Note that the axes are drawn to different scales). We varied the inclusion of structural neighbors, the use of L_1-regularization, and the inclusion of phylogenetic conservation scores from INTREPID. The control uses non-phylogenetic conservation scores, while other methods use INTREPID. DISCERN is more accurate than the other variants over the range of recalls, except between a recall of 0.05 and 0.1 where Method 2 is most accurate. Further, since the control has very similar accuracies to SVM-Mooney and NN-Thornton (as shown in Section 3.3.1), the improvement of DISCERN over these methods is significant and is unlikely to be an artifact of the dataset. See Table 3.1 for details on each variant and a comparison at fixed points of precision and recall.

Method 2 is identical to Method 1, but includes features computed at structural neighbors with no L_1-regularization. We see that naively including features from structural neighbors leads to a decrease in performance. Method 2 attains a recall of 41% at a precision of 18%.

DISCERN is identical to Method 2 (using phylogenomic conservation scores and features computed at structural neighbors), but also includes L_1-regularization to enforce sparsity. Relative to Method 1, DISCERN has 14% greater recall (recall=69%) at 18% precision. Relative to the control, DISCERN has 21% greater recall at the same level of precision.

Proceeding from the control to DISCERN also shows a dramatic reduction in false positive predictions (residues predicted as catalytic which are not listed in the CATRES dataset). Measuring precision (the fraction of predicted residues that are actually catalytic) at the point where half of the catalytic residues have been detected (i.e., a recall of 50%) shows that the control has precision of 17.0% while DISCERN has 27.3% precision. In other words, DISCERN effectively reduces the ratio of false positives to true positives from 4.1 to 2.8.

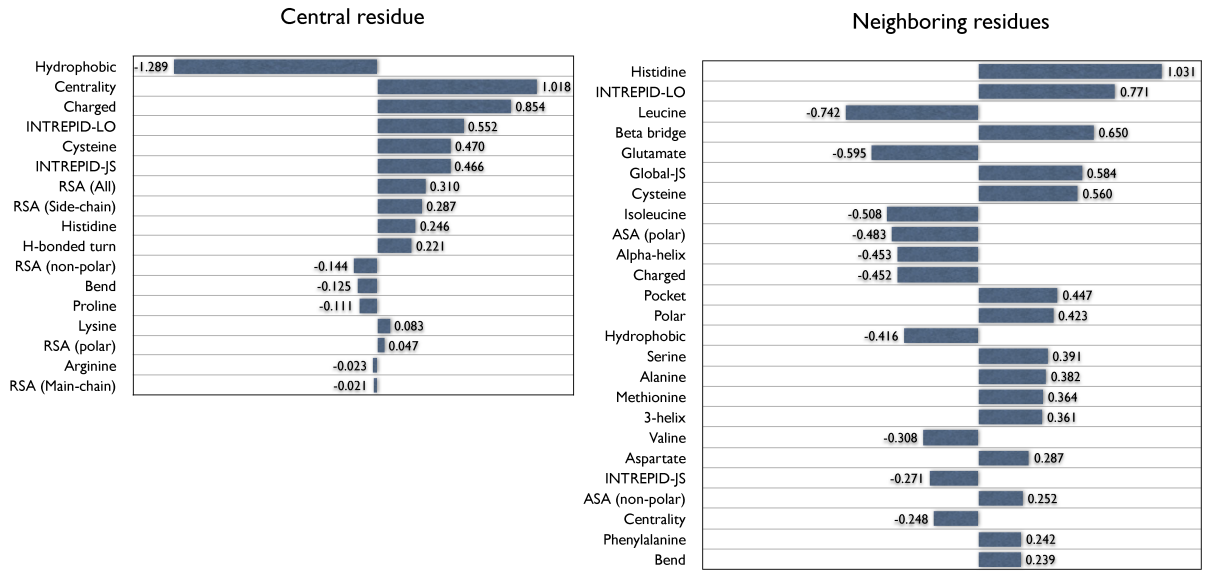

Figure 3.9: Features selected by DISCERN and the weights assigned to each feature based on fitting the logistic regression to the entire CATRES-FAM dataset and displaying the features with the largest weights. Positive weights indicate positive correlation with putative catalytic residues; negative weights imply negative correlation. The magnitude of the weight is indicative of a feature's relative importance. **Left: Features computed at the residue of interest.** For features at the central residue, only 17 had non-zero weights. The feature with the largest weight (-1.289) is *hydrophobicity*; the negative weight is consistent with the observation that hydrophobic residues are rarely catalytic. The next highest-ranked feature is *residue centrality* with a weight of 1.018; high values for this feature indicate that a residue is located in the core of the enzyme 3D structure. *INTREPID-LO*, and *INTREPID-JS*, information-theoretic measures of the evolutionary conservation of a residue, jointly have a weight of 1.018, the same weight as centrality. Residue charge comes next with a weight of 0.854, followed by presence of a cysteine (0.470). *Relative solvent accessibility*, a measure of the fraction of a residue exposed to solvent averaged over all the atoms (RSA(All)) and over the side-chain atoms (RSA(Side-chain)) respectively, comes next with weights of 0.310 and 0.246 respectively. **Right: Features summed over residues that are nearby in the 3D structure.** For the purpose of visualization, feature weights are summed over all the structural neighbors; features that have large non-zero weights would tend to be predictive of a structurally neighboring catalytic residue. The top 25 features with largest absolute weights are displayed. The feature with consistently large weights are the evolutionary conservation scores (*INTREPID-LO* and *GLOBAL-JS*). *INTREPID-LO* and *GLOBAL-JS* (a measure of sequence conservation across the family that does not use the phylogenetic tree) have a combined total weight of 1.255.

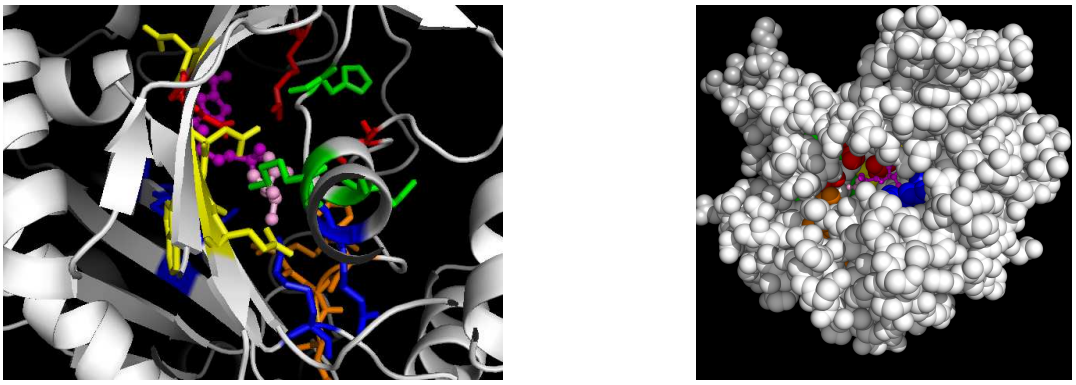

Figure 3.10: **Top 20 residues predicted by DISCERN on *Escherichia coli* Asparagine Synthetase (PDB id:12as).** (Left) Detailed view of the active site. Red indicates residues listed in CATRES (D46, R100, Q116). Green, yellow and orange indicate residues in motifs 1 (H71, K75 and K77), 2 (D115, Q116, D118, W119 and E120), and 3 (R214, Y218, D219, and D220) respectively. Other predicted residues are shown in blue. Also shown are the AMP and L-asparagine molecules. (Right) The predictions shown in space-fill representation. See Table 3.3 for a list of these residues.

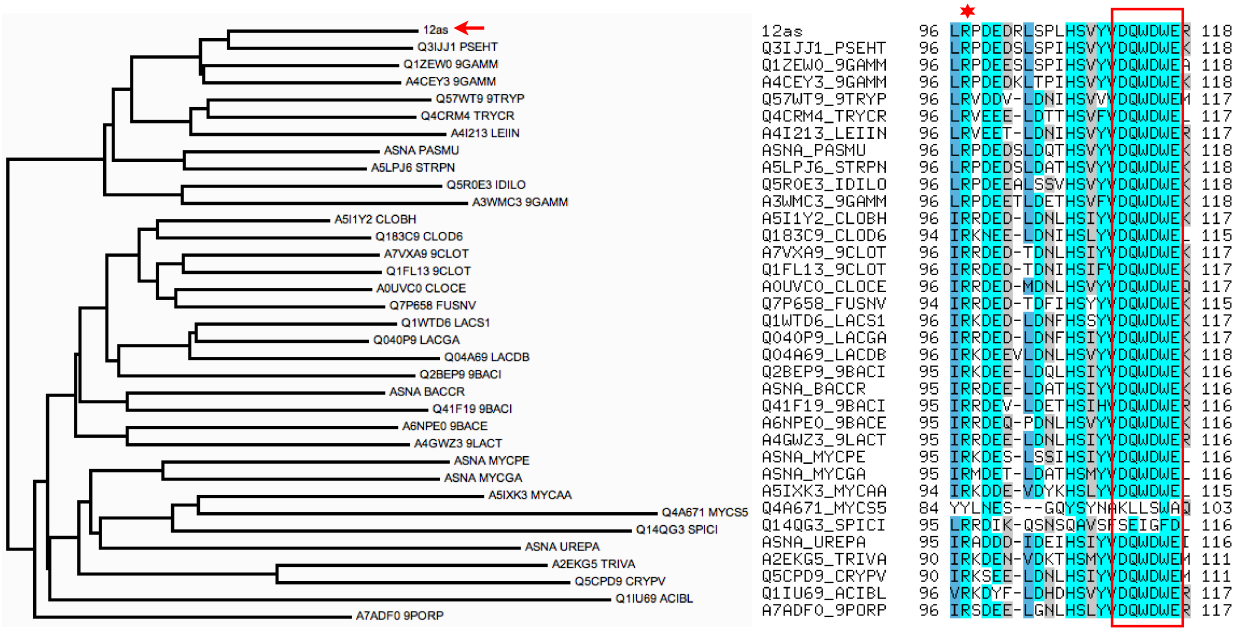

Figure 3.11: **Tree and alignment of the homologs for *Escherichia coli* Asparagine Synthetase (PDB id:12as).** Neighbor-joining tree and alignment derived by making the original alignment non-redundant at 70% identity relative to the seed. The positions in the seed sequence correspond to the residue number in PDB minus 3, e.g., the arginine at position 97 corresponds to R100 in the PDB record. R100 is marked with a star because it is listed as catalytic in CATRES. Note that not all sequences contain an arginine at this position. Positions in motif 2 (D115, Q116, D118, W119 and E120) have been boxed. The branch lengths of Q4A671_MYCS5 and Q14QG3_SPICI have been truncated from their original lengths of 1.089 and 1.181 respectively to a value of 0.5 for better visualization.

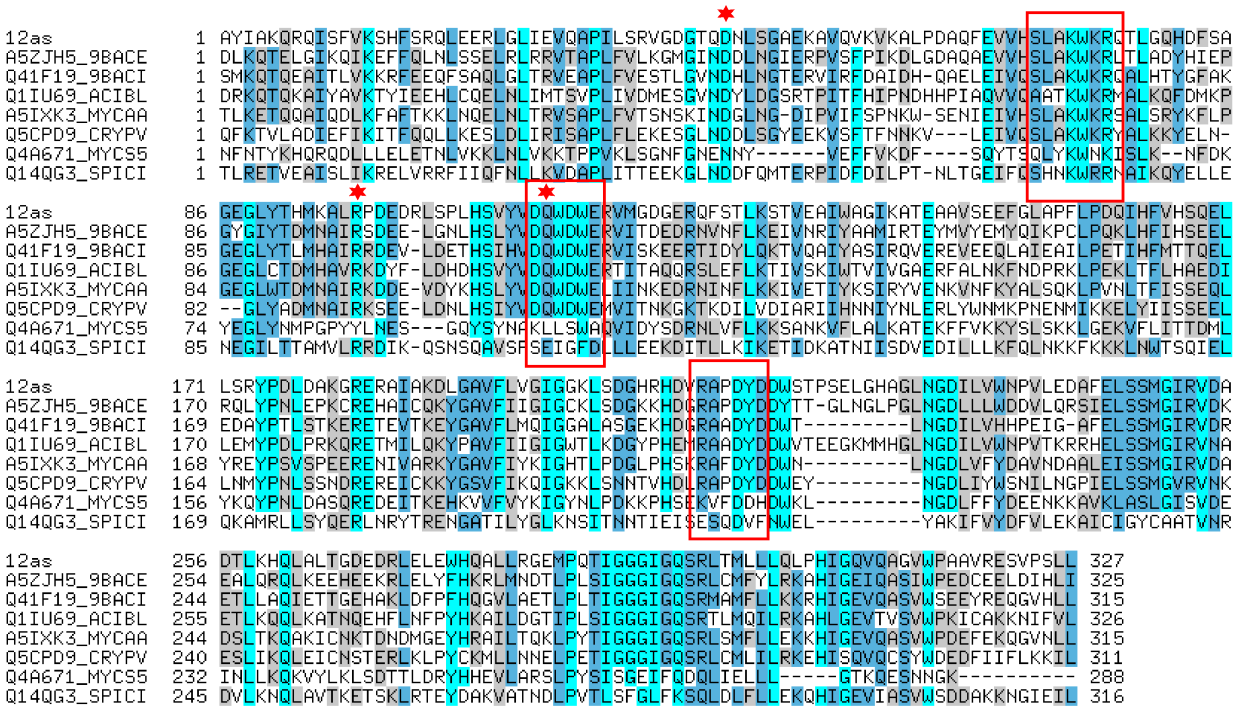

Figure 3.12: **Alignment of the homologs for *Escherichia coli* Asparagine Synthetase (PDB id:12as).** The displayed alignment was derived by making the original alignment non-redundant at 50% identity. Residues listed as catalytic in CATRES (D46, R100 and Q116) are marked with a star while positions that form motifs based on their DISCERN scores have been boxed. See Table 3.3 for the list of predicted residues. Note that even catalytic residues are not perfectly conserved across the family.

Table 3.3: **Top 20 residues predicted by different methods on *Escherichia coli* Asparagine Synthetase (PDB id:12as).** The three catalytic residues listed in CATRES (D46, R100 and Q116) are marked with *. Residues with a proposed functional role that are not listed in CATRES are marked with †. DISCERN detects all three catalytic residues in these top 20, INTREPID detects one, and ConSurf detects two of the three. Residues among these top 20 that are also described as functional in the literature but are not listed in CATRES include P35, K77, E120, D219, D235, E248, S251, R255, and I295. Of these 10 residues, seven are found among the top 20 for DISCERN, one is found by INTREPID and two are found by ConSurf. See Figure 3.10 for the DISCERN predictions plotted onto the structure of asparagine synthetase. Figure 3.12 shows an MSA for 12as and homologs. Refer to Section 3.3.2 for a detailed analysis of these predictions.

DISCERN	Intrepid	Consurf
R214	W76	S72
D219†	W119	S111
D115	W318	S250
D235†	W117	S251†
K77†	H309	S298
D46 *	W221	I201
R100 *	H211	N233
E248†	M252	I291
E120†	M96	I295†
R255†	Q264	A74
Y218	M302	A98
H71	H110	V32
D118	Y218	V55
R78	Q297	V70
K75	N233	V114
Q116 *	P35†	V137
S251†	Q116 *	V256
S250	F197	I12
W119	H279	M96
D220	P288	M252

Table 3.4: **Comparison of DISCERN, INTREPID and ConSurf.** Precision_{50} reports the precision at 50% recall, and Recall_{10} reports the recall at 10% precision (ConSurf does not achieve a precision of 18% on CATRES-FAM).

Method	CATRES-FAM		CSA-FAM	
	Precision_{50}	Recall_{10}	Precision_{50}	Recall_{10}
DISCERN	27.3%	86%	28.3%	90%
INTREPID	13.0%	64%	14.9%	70%
ConSurf	7.9%	35%	5.6%	6%

Table 3.5: **Features evaluated for catalytic residue prediction:** This set of features are evaluated at a residue and each of its ten structural neighbors resulting in $48 \times 11 = 528$ features. RSA and ASA refer to the relative and absolute solvent accessibility respectively. Refer to Section 3.A for detailed descriptions.

Type of feature	Description
Sequence conservation features	INTREPID-JS, INTREPID-LO, GLOBAL-JS
Amino acid properties	{Charged, Polar, Hydrophobic}, {20 amino acid sidechains}
Structure-based features	B-factor, Centrality, Secondary structure element (Alpha helix, Beta bridge, Strand, 3-helix, pi-helix, H-bonded turn, Bend) RSA (All atoms, Side chain, Main chain, Non polar, Polar), ASA (All atoms, Side chain, Main chain, Non polar, Polar), Presence in each of three largest pockets

Chapter 4

Estimating local ancestry in admixed populations

4.1 Introduction

Recent advances in genotyping technologies have opened up unprecedented opportunities to improve our understanding of complex diseases through disease association studies. In these studies, a population of cases and controls are genotyped across the genome, and the allele frequencies are compared across these two groups. Currently, in a typical study, hundreds of thousands of single nucleotide polymorphisms (SNPs) are genotyped for thousands of individuals [Bonnen *et al.*, 2006]. These numbers are expected to grow in the coming years due to the constant improvements in genotyping technologies [Bonnen *et al.*, 2006].

A significant discrepancy between the allele frequencies in the cases and the controls gives evidence for an association between the SNP and the phenotype, and therefore links the SNP to the disease. However, a growing concern is that many of the associations found are due to confounding effects. In particular, if the cases and the controls are not sampled from the same population, many spurious associations will be discovered, since the two populations may have different allele frequencies at a SNP regardless of the disease status [Price *et al.*, 2006; Freedman *et al.*, 2004; Clayton *et al.*, 2005; Lander and Schork, 1994a; Helgason *et al.*, 2005; Campbell *et al.*, 2005; Hirschhorn and Daly, 2005; Lohmueller *et al.*, 2003; Marchini *et al.*, 2004]. This bias can be observed in diseases that are more prevalent in one population than in another. In such cases, the collection of the cases is a biased sample of the population.

Various methods have been proposed to deal with population sub-structure in association studies [Price *et al.*, 2006; Devlin and Roeder, 1999]. One of the most intuitive approaches is to first find the population sub-structure within the cases and the controls using a clustering algorithm such as `STRUCTURE`[Pritchard *et al.*, 2000], and then to correct it using regression or other methods that take the sub-population variable into account [Setakis *et al.*, 2006]. The clustering algorithms need to be accurate enough, so that the signal obtained from

the difference in population sub-structure will be weaker than the signal obtained from the difference in the disease status.

The problem of inferring the population sub-structure is especially challenging when recently admixed populations are involved. In these populations (e.g., African Americans and Latinos), two or more ancestral populations have been mixing for a relatively small number of generations, resulting in a new population in which the ancestry of every individual can be explained by different proportions of the original populations. Due to recombination events, even within the DNA of a single individual, different regions of the genome may originate from different ancestral populations. This adds to the complexity of the problem of finding the ancestral information of an individual, since in non-admixed populations the whole genome can be used as an evidence for the population membership of an individual, while in the admixed case the genome of each individual is fragmented into shorter regions of different ancestry. It is therefore challenging to find the ancestral information of these individuals, and in particular, to find the locus-specific ancestries.

An accurate inference of locus-specific ancestry in admixed populations may lead to improved analysis of studies based on admixture mapping. In these studies, a set of cases from a recently admixed population is genotyped, and the genome is scanned for regions in which the proportions of ancestral populations are significantly different than the rest of the genome [Reich *et al.*, 2005b; Zhu *et al.*, 2005a]. Unfortunately, most of the current methods for inference of locus-specific ancestral information [Pritchard *et al.*, 2000; Patterson *et al.*, 2004; Falush *et al.*, 2003; Hoggart *et al.*, 2004] do not scale to large data sets. The only existing method that copes with large data sets is `SABER` [Tang *et al.*, 2006], which is based on an extension of a Hidden Markov Model [Rabiner, 1989] to deal with local haplotype blocks.

Here, we propose a new method, `LAMP` (Local Ancestry in adMixed Populations), for *de-novo* estimation of the locus-specific ancestry in recently admixed populations (see Figure 4.1). Our method is based on the observation that previous methods that use a Hidden Markov Model or extensions of it, are set to infer a very large set of parameters, including the exact position of the recombination events, which makes the search over the parameter space infeasible. Instead, our method operates on sliding windows of contiguous SNPs. We first calculate an optimal window length. Next, we use a clustering algorithm that operates on these windows and estimates each individual's ancestry. We then use a majority vote for each SNP, over all windows that overlap with the SNP, in order to decide the most likely ancestral populations at the SNP. This simple approach has two advantages over previous ones. First, we show analytically that the estimates of the algorithm are asymptotically correct across the entire genome. Second, it optimizes fewer parameters than previous methods and hence, the optimization is much faster and more robust than previous methods.

We tested `LAMP` extensively on various data sets of admixed populations generated from the HapMap resource [`http://www.hapmap.org`]. Our simulations show that `LAMP` is significantly more accurate than the state of the art methods such as `SABER` and `STRUCTURE`.

In addition, LAMP is highly efficient with a running time that is about 200 times faster than SABER and about 10^4 times faster than STRUCTURE. The efficiency of LAMP allows us to estimate ancestries across the genome in several hours on a single computer.

An additional advantage of LAMP is that unlike previous methods such as SABER, it does not require the ancestral genotypes to infer the locus-specific ancestries (though it can take advantage of these if available). This may be crucial when the ancestral genotypes cannot be typed or are unknown. For instance, if one studies the population genetics of populations in remote geographic locations where historical admixing has not been recorded, a method such as LAMP could be used to reveal such recent admixing. Furthermore, even in cases where the history of admixing is known, it is not always possible to genotype all the ancestral populations, since some of the subpopulations have become extinct and some have entirely mixed with other populations. On the other hand, as genotypes of major population groups become available, it would be beneficial to use LAMP-ANC which can take advantage of the pure genotypes.

Surprisingly, we find that in many cases where LAMP does not receive the genotypes of the ancestral populations as input, it performs considerably better than SABER. In particular, on a simulated dataset of African-Americans, when measuring the percentage of individuals that are predicted with an accuracy of at least 90%, LAMP achieves high accuracies on 90% of the individuals while SABER and STRUCTURE achieve less than 10%.

Finally, we used LAMP to estimate the individual admixture, and showed empirically that this results in much more accurate estimates than methods such as STRUCTURE[Pritchard *et al.*, 2000] or EIGENSTRAT[Price *et al.*, 2006]. This reduction in errors may be used to considerably reduce the rate of spurious association results in disease association studies.

4.2 Estimating local ancestry

The inference of locus specific ancestry depends on the mathematical model representing the mixing process of the populations. We will first describe the model assumptions, and then describe the inference algorithm under the model.

4.2.1 Model assumptions

We assume that there are K ancestral populations $A_1, \ldots, A_K$ that have been mixing for g generations. If the populations have mixed at different times, then g is taken to be an upper bound on the number of generations since the beginning of admixture. The fraction of population A_i in the ancestral population which we call the *admixture fraction* is α_i, where $\sum_i \alpha_i = 1$. We assume for convenience that $\alpha_1 \geq \alpha_2 \ldots \geq \alpha_K$. In each generation, we assume random mating within the combined pool of the k populations. We denote the recombination rate at position j by r_j. Note that r_j is the recombination rate at position j

at a specific meiosis (one generation), and not through history. We model the transmission of a chromosome from a parent to a child by walking along the chromosome from the 5' end to the 3' end with crossovers between chromosomes occurring as a Poisson process with rate r_j [Haldane, 1919]. For simplicity of the presentation, we will assume a uniform recombination rate, i.e., that $r = r_j$ for every position j. The algorithm and analysis remain qualitatively the same when applied to non-uniform recombination rates.

We denote the genotype data of individual i at position j as g_{ij}, where $g_{ij} \in \{0, 1, 2\}$ is the minor allele count at that position. At position j, the two alleles of individual i have descended from one or two of the K ancestral populations. We denote by $a_{ij}^p \in \{0, 0.5, 1\}$ the fraction of alleles descended from population p at position j in individual i. The quantities a_{ij}^p are unknown; our objective is to derive a method `LAMP` that accurately estimates these quantities.

4.2.2 The `LAMP` framework

We consider a recently admixed population in which the number of generations g since the beginning of the mixing is small. Therefore, we expect the total number of recombinations in these g generations to be small as well. The resulting chromosomes are mosaics of the k populations, where the *ancestral breakpoints* in which the chromosome ancestry changes from one population to the other are determined by the recombination events.

We assume that the quantities g, α_i, and r are known for the admixed population. The basic idea in `LAMP` is to estimate the ancestries of the individuals in a sliding window that spans l sites. We term l the length of the window. The choice of the length l will be discussed later. Intuitively, if l is small enough, and the number of generations g is not too large, a typical window of length l will have almost no recombination events throughout history, and therefore almost no breakpoints. Therefore, within each window, it is reasonable to use an inference algorithm that assigns the sequence of genotypes in the window to one or two of the populations under the assumption that there are no breakpoints in any of the chromosomes. The latter is a simple clustering problem, although the accuracy of the inference in a given window improves when the number of SNPs l in the window increases. We therefore search for a window length l, which is short enough so that most individuals have no breakpoints and large enough so that there is enough information to correctly cluster the individuals within the window. This procedure is repeated by sliding the window to cover all the SNPs on the genome. The windows that overlap a SNP are then combined into a single solution using a majority vote for the ancestry assignment. We note that unlike previous methods (e.g. `SABER` [Tang *et al.*, 2006], or `STRUCTURE` [Pritchard *et al.*, 2000]), we are not attempting to estimate the exact positions of the breakpoints, but instead we are trying to minimize the errors in the locus-specific ancestry prediction across the genome.

The `LAMP` algorithm works as follows. We first find the optimal window length based on the parameters g, α_i, and r. Then, we use a clustering algorithm that operates on a

window and estimates for each individual i, and for each ancestral populations A_j, A_k, the probability p^i_{jk} for individual i to have one chromosome descended from population A_j at this window and another descended from population A_k. We then use a majority vote for each SNP, over all windows that overlap with the SNP, in order to decide the most likely ancestral populations at the SNP. As we argue below, even though this scheme optimizes less parameters than previous methods, such as `SABER`, or a regular HMM, we show analytically and empirically that the estimates of the algorithm are asymptotically correct across the entire genome.

4.2.3 Estimating the ancestry in a single window

We assume that none of the individuals have a breakpoint within a window and estimate a single ancestral origin for each individual across the length of the window. This assumption is largely true if the length of the window is determined correctly (see Section 4.2.4 and Section 4.C in the supplementary materials). We further assume that the values $\alpha_1, \ldots, \alpha_K$ are known. These values are the admixture fractions of each of the populations across the whole genome, and they can be estimated using existing tools such as `STRUCTURE`[Pritchard *et al.*, 2000]. In the results section we show that our method is robust to reasonable inaccuracies in the estimates of $\alpha_1, \ldots, \alpha_K$.

4.2.3.1 Clustering Algorithm

We assume that sub-population A_i has minor allele frequencies $\vec{f_i} = f_{i1}, \ldots, f_{in}$ for n SNPs in a given window of length l, and that the different SNPs in the window are independent. The latter assumption can be achieved in practice by greedily removing SNPs having a high correlation coefficient ($r^2 > 0.1$) from the window. We look for a classification function $\theta : I \rightarrow \{1, \ldots, K\}^2$, where I is the set of individuals, and the range corresponds to the possible pairs of sub-populations. In particular, we write $\theta(i) = (\theta_1(i), \theta_2(i))$ to denote the ancestries of the two chromosomes of individual i in the current window. We use a clustering algorithm known as Iterated Conditional Modes (ICM) [Besag, 1986] to find an optimal classification of each individual in terms of the likelihood. For increased efficiency in the running time, we seed the algorithm with an initial classification as described in Section 4.2.3.2.

The updates in the ICM algorithm differ from those in a traditional EM method only in the E-step. In the latter, the E-step consists of obtaining the expected classification θ, given the values $\vec{f_i}$. This would provide fractional class membership for each individual i. However, since we assume that the initial classification provides a reasonable solution, we find the maximum aposteriori estimate of θ as shown below. For brevity we use $\mathcal{G}_i$ to refer

to the genotype $(g_{i1}, \ldots, g_{in})$ of the individual i.

$$\begin{aligned}\hat{\theta}(i) &= \text{argmax}_{A_sA_t \in \{1,\ldots,K\}^2} \Pr[\theta(i) = A_sA_t \mid \vec{f_1}, \ldots, \vec{f_K}, \mathcal{G}_i] \quad (4.1)\\ &= \text{argmax}_{A_sA_t \in \{1,\ldots,K\}^2} \Pr[\mathcal{G}_i \mid \vec{f_1}, \ldots, \vec{f_K}, \theta(i) = A_sA_t] \cdot \Pr[\theta(i) = A_sA_t \mid \vec{f_1}, \ldots, \vec{f_K}]\end{aligned}$$

Since $\alpha_1, \ldots, \alpha_K$ are known, under the assumption of random mating, we can estimate the first term $\Pr[\theta(i) = A_sA_t \mid \vec{f_1}, \ldots, \vec{f_K}]$ as $\Pr[\theta(i) = A_sA_t] = 2^{1-\delta(s,t)}\alpha_s\alpha_t$ where $\delta(x, y)$ is 1 iff $x = y$ and 0 otherwise.

The other term can be estimated as:

$$\begin{aligned}&\Pr[\mathcal{G}_i|\vec{f_1}, \ldots, \vec{f_K}, \theta(i) = A_sA_t]\\ &= \textstyle\prod_{g_{ij}\in\mathcal{G}_i|g_{ij}=2} f_{sj}f_{tj} \cdot \prod_{g_{ij}\in\mathcal{G}_i|g_{ij}=0} [(1 - f_{sj})(1 - f_{tj})] \cdot \prod_{g_{ij}\in\mathcal{G}_i|g_{ij}=1} [f_{sj}(1 - f_{tj}) + f_{tj}(1 - f_{sj})]\end{aligned}$$

In the M-step, we obtain the maximum likelihood estimate of $\vec{f_1}, \ldots, \vec{f_K}$ by finding

$$\text{argmax}_{\vec{f_1},\ldots,\vec{f_K}} \Pr[(\mathcal{G}_i)_{i=1}^m \mid \vec{f_1}, \ldots, \vec{f_K}, \theta] = \prod_{i=1}^{m} \Pr[\mathcal{G}_i \mid \vec{f_1}, \ldots, \vec{f_K}, \theta(i)] \quad (4.2)$$

If the phase of the individuals is known, the maximum likelihood estimate of $\vec{f_1}, \ldots, \vec{f_K}$ could have been computed by simply counting the number of alleles in each of the sub-populations at every position. However, when the phase is not known, the problem becomes more complicated. Consider for instance a heterozygous site j in an individual i, with $\theta_1(i) \neq \theta_2(i)$. In this case, it is not clear whether the minor allele count should be added to $f_{\theta_1(i)j}$ or to $f_{\theta_2(i)j}$. To solve this problem, we introduce another classification function per site, $\vec{\lambda_j} : I \to \{0, 1\}^K$. This function is defined on the set of SNPs for which the assignment of counts is ambiguous *i.e.*, heterozygous SNPs in individuals i with classification $\theta_1(i) \neq \theta_2(i)$. We denote this set of heterozygous SNPs $\mathcal{H}_i$. The function $\vec{\lambda}_j$ is defined as

$$\vec{\lambda}_j(i) = \begin{cases} \vec{e}_s, \text{if } j \in \mathcal{H}_i, \text{ one of } (\theta_1(i), \theta_2(i)) = s, \text{ and the minor allele is counted to } f_{sj} \\ \text{not defined for } j \notin \mathcal{H}_i \end{cases}$$

Here $\vec{e}_s$ is the vector with 1 in coordinate s and 0 elsewhere.

For a heterozygous site j in individual i such that $j \in \mathcal{H}_i$, we can now define

$$\begin{aligned}\Pr[\vec{\lambda}_j(i) = \vec{e}_{\theta_1(i)} \mid f_{1j}, \ldots, f_{Kj}, \theta(i)] &= f_{\theta_1(i)j}(1 - f_{\theta_2(i)j})\\ \Pr[\vec{\lambda}_j(i) = \vec{e}_{\theta_2(i)} \mid f_{1j}, \ldots, f_{Kj}, \theta(i)] &= f_{\theta_2(i)j}(1 - f_{\theta_1(i)j})\\ \Pr[\vec{\lambda}_j(i) = \vec{e}_{s\notin\{\theta_1(i),\theta_2(i)\}} \mid f_{1j}, \ldots, f_{Kj}, \theta(i)] &= 0\end{aligned}$$

Using the assumption of independence of the SNPs and the $\vec{\lambda}_j$ just defined, we can rewrite Equation 4.2 as follows. The usefulness of this will be apparent later.

$$(\hat{f}_{1j},\ldots,\hat{f}_{Kj}) = \text{argmax}_{f_{1j},\ldots,f_{Kj}} \prod_{i=1}^{m}\left(\prod_{j\in\mathcal{H}_i}\sum_{u=1}^{K}\Pr[\vec{\lambda}_j(i)=\vec{e}_u \mid f_{1j},\ldots,f_{Kj},\theta(i)]\right) \times\left(\prod_{j\in\{1,\ldots,n\}\ \mathcal{H}_i}\Pr[g_{ij}\mid f_{1j},\ldots,f_{Kj},\theta(i)]\right) \tag{4.3}$$

The MLE for $\hat{f}_{ij},\ldots,\hat{f_{Kj}}$ can be found using an EM algorithm where

$$\text{E-step} : \overline{\lambda_{j,s}}(i) = E[\lambda_{j,s}(i)\mid f_{\theta_1(i)j}, f_{\theta_2(i)j},\theta(i),g_{ij}=1] = \begin{cases} \frac{f_{\theta_1(i)j}(1-f_{\theta_2(i)j})}{f_{\theta_1(i)j}(1-f_{\theta_2(i)j})+(1-f_{\theta_1(i)j})f_{\theta_2(i)j}}, \text{for } s=\theta_1(i) \\ \frac{f_{\theta_2(i)j}(1-f_{\theta_1(i)j})}{f_{\theta_1(i)j}(1-f_{\theta_2(i)j})+(1-f_{\theta_1(i)j})f_{\theta_2(i)j}}, \text{for } s=\theta_2(i) \\ 0, \text{for } s\notin\{\theta_1(i),\theta_2(i)\}\end{cases} \tag{4.4}$$

$$\text{M-step} : \hat{f_{sj}} = \frac{2n_{2,2}^{sj}+n_{2,1}^{sj}+n_{1,2}^{sj}+\sum_{j\in\mathcal{H}_i}\overline{\lambda_{j,s}}(i)}{2n_{2,2}^{sj}+2n_{2,1}^{sj}+2n_{2,0}^{sj}+n_{1,2}^{sj}+n_{1,1}^{sj}+n_{1,0}^{sj}} \tag{4.5}$$

Here $\lambda_{j,s}(i)$ refers to coordinate s of the vector $\lambda_j(i)$. $n_{k,u}^{sj}$ refers to the number of individuals who have $u\in\{0,1,2\}$ minor alleles and $k\in\{1,2\}$ copies of alleles from population A_s at site j. The counts of these individuals can be obtained based on the classification $\theta(i)$. Notice that the term corresponding to the heterozygous sites which have a single allele from population A_s has its contribution modified by $\overline{\lambda_{j,s}}(i)$. We can now perform expectation-maximization iterations using equations 4.5 and 4.4. The convergence of these iterations provides us a maximum likelihood estimate of $\vec{f_1},\ldots,\vec{f_K}$. These estimates can then be used in the next iteration to estimate θ using Equation 4.1.

4.2.3.2 Initializing the clusters

We now describe how we obtain an initial setting of the parameters *i.e.*, the classification function θ or the allele frequencies $\vec{f_1},\ldots,\vec{f_K}$, which are used as starting points by the EM algorithm. We focus here on two specific scenarios. The first scenario is the case where there are two ancestral populations *i*.e., $K=2$ and unknown allele frequencies $\vec{f_1},\ldots,\vec{f_K}$. In this instance, we use an algorithm called `MAXVAR` to provide an initial solution to the EM algorithm. The main motivation behind `MAXVAR` is to quickly produce a reasonable classification. The algorithm runs in time linear in the number of SNPs and can take advantage of results computed from adjacent windows. We have also considered using spectral clustering but in practice we found that the final classification accuracy is nearly the same as `MAXVAR` though

the running time is increased. The result from MAXVAR is a classification of the individuals which is then used in Equation 4.2 of the EM.

The second scenario is the case where $K \geq 2$ and estimates of the allele frequencies $\hat{f}_1, \ldots, \hat{f}_K$ in the ancestral populations are known. In this case, these allele frequencies are used as a starting solution in Equation 4.1 of the EM algorithm.

4.2.3.2.1 The (MAXVAR) algorithm When we have two populations, we estimate a window length l such that most of the individuals have no breakpoints within a window. Thus the ancestries of these individuals are A_1A_1, A_1A_2 or A_2A_2. We define $\alpha = \alpha_2$ as the admixture fraction of the smaller of the two populations. We now describe a method to find the ancestry of each individual in this window. We call this the MAXVAR algorithm.

We first define a similarity score $\mathcal{S}$ between a pair of individuals. For each SNP j, let $\mu_j = \frac{\sum_i g_{ij}}{n}$, where n is the number of individuals, and let $\sigma_j = \sqrt{\frac{\sum_i (g_{ij} - \mu_j)^2}{n}}$. For two individuals i_1, i_2, we define

$$\mathcal{S}(i_1, i_2) = \sum_{j=1}^{l} \frac{(g_{i_1 j} - \mu_j)(g_{i_2 j} - \mu_j)}{\sigma_j^2}.$$

For each $i \leq n$, let $Var(i) = \sum_{i' : i' \neq i} \mathcal{S}(i, i')^2$ denote the similarity of all other individuals to individual i, and let $i^* = \text{argmax}_i\{Var(i)\}$. The MAXVAR algorithm simply finds i^*, and clusters the individuals according to the values $\mathcal{S}(i^*, i)$. In particular, we order the individuals according to these values, and the smallest $(1-\alpha)^2 n$ individuals are assigned as the ancestry of A_1A_1, the largest $\alpha^2 n$ individuals are assigned as the ancestry of A_2A_2, and the rest are assigned A_1A_2. We provide a formal proof of correctness of the MAXVAR method in the supplementary materials (Section 4.A).

4.2.3.2.2 Known ancestries. The problem of estimating the ancestry is considerably simpler if we are provided estimates of the ancestral allele frequencies. In this case, as before, we first estimate the window length l. Within each window, we then estimate the ancestries using the likelihood function given by Equation 4.1 with the given ancestries $\hat{f}_1, \ldots, \hat{f}_K$ used as the starting solution. The ancestries predicted at each SNP are combined using a majority vote.

4.2.4 Choosing the window length

In order for the local predictions to achieve reasonable accuracy, the length of the window l should be short enough so that most individuals do not have a breakpoint in the window and

long enough so that the SNPs provide sufficient information to observe a difference between the populations. Note that we use the term *breakpoint* to refer to a recombination event that results in a change in ancestry of the adjacent SNPs. The power of our method stems from the fact that long windows provide much more information than any local behavior, provided that there are not too many individuals with breakpoints in the window. We are looking for the maximum window length l so that the errors in the classification due to breakpoints in the window are bounded. We present empirical results that validate the window length estimates in Section 4.C of the Supplementary Material.

In each window, the errors in the classification depend on the length of the window, the number of individuals, and the distance between the populations. Evidently, it is hard to predict these errors as the distance between the populations is unknown, and the performance of the EM algorithm is unpredictable for a finite sample. To obtain a bound on the errors, we consider the most accurate classification of the individuals in a window. Such a classification is allowed to assign ancestries to the individuals in a window with knowledge of their true ancestral states a_{ij}^p for $p = 1, \ldots, K$. Thus an individual whose ancestry is A_sA_t over the length of the window is always classified correctly. The only errors made by such a classification are due to the locations of the breakpoints. In the presence of a breakpoint, an individual would be assigned an ancestry so that the number of errors is minimized. For instance, an individual with a breakpoint at position j and ancestries A_{s_1} and A_{s_2} on either side of the breakpoint gets assigned the majority ancestry over the length of the window *i.e.*, the individual gets classified as A_{s_1} if $j > \lceil \frac{l}{2} \rceil$ and A_{s_2} otherwise. It is easy to see that the larger the window size l, the more likely it is for an individual to have a breakpoint and hence, the more the errors in the optimal classification.

The number of recombination events throughout time along a specific window is assumed to be Poisson distributed with parameter $(g-1)lr$. Therefore, as long as $(g-1)lr << 1$, it can be verified that the probability to have a breakpoint in the window is upper-bounded by $2(g-1)lr\sum_{i<j}\alpha_i\alpha_j$ under the assumption of random-mating and that the admixture fractions of the population right before recombination are α_i. Therefore, the probability for a breakpoint on either chromosome is bounded by $\gamma = 4(g-1)rl\sum_{i<j}\alpha_i\alpha_j$.

For a given window, the above analysis shows that the expected fraction of individuals with no breakpoints is $1-\gamma$. We can now use this to obtain a bound on the fraction of errors in a window. Let X be the fraction of errors in a window of an algorithm that makes the optimal classification. Let I be the number of breakpoints in the window. We compute

$$E[X] = E[E[X|I]] = \sum_i Pr[I=i]E[X|I=i]$$

Note that $E[E[X|I=0]] = 0$ since the optimal classification in this case makes no errors. When there is a single breakpoint $I = 1$, the breakpoint is distributed uniformly over the length of the window. We denote the position of the breakpoint $J \sim Unif(1, l)$. The fraction

of errors in the presence of a single breakpoint at position J is

$$(X|I=1,J=j) = \begin{cases} 1-\frac{j}{l} & j > \lceil \frac{l}{2} \rceil \\ \frac{j}{l} & \text{otherwise} \end{cases} \tag{4.6}$$

We now have

$$E[X|I=1] = 2\sum_{j=1}^{\lceil \frac{l}{2} \rceil - 1} \frac{j}{l}\frac{1}{l} \leq \frac{1}{4}$$

If $glr << 1$,we can ignore $\Pr[I>1]$ so that

$$\begin{aligned} E[X] &\leq 0 \cdot \Pr[I=0] + \frac{1}{4} \cdot \Pr[I=1] + 1 \cdot \Pr[I>1] \\ &\approx \gamma\frac{1}{4} \end{aligned} \tag{4.7}$$

For a bound ϵ on the expected fraction of errors, we get $\gamma < 4\epsilon$. Rewriting the window length l in terms of ϵ, we get

$$l \leq \frac{\epsilon}{(g-1)r\sum_{i<j}\alpha_i\alpha_j} \tag{4.8}$$

While these arguments bound the errors in a single window, it is also possible to bound the errors due to overlapping windows at a SNP. In this case, the use of a majority vote can be shown to further improve the accuracy of the predictions. The details of this analysis can be found in the supplementary materials (Section 4.B).

The analysis presented here is specific to the model of admixture described at the start of Section 4.2.1. However, it is easy to see that the analysis can be extended to the case of non-uniform recombination rate, where the probability for a recombination in position i is r_i. In that case, the term $(g-1)lr$ should be replaced by $(g-1)\sum_{i=0}^{l} r_i$.

The model considered so far does not take into account the distance between the ancestral populations while choosing the window length. When the ancestral genotypes are known, the window length can be chosen to tradeoff the accuracy in separating the ancestral genotypes with an increase in the errors due to breakpoints. A binary search over the window lengths can then pick the optimal window length as discussed in the supplementary materials (Section 4.D).

4.3 Results

We empirically evaluated `LAMP` on various data-sets and compared its performance with other tools that infer ancestry in admixed populations. When comparing to previous methods, it is important to note that the inputs needed for the different methods are different. In

particular, in SABER [Tang *et al.*, 2006], the genotypes from the pure ancestral populations are assumed to be known, while in LAMP, we do not need this extra information. On the other hand, similar to SABER, LAMP assumes that the recombination rates across the genome and the admixture fraction $(\alpha_1, \ldots, \alpha_k)$ are known; the latter can be found with reasonable accuracy using existing methods such as STRUCTURE or EIGENSTRAT, while the former can be obtained from the previous estimates of recombination rates based on the HapMap data [Myers *et al.*, 2005]. We also provide LAMP with an estimate of the number of generations g of admixture which can be approximated if the history of the admixed populations is known. We show below that our method is robust to deviations in the estimate of g. For SABER, we set the parameter τ, which roughly corresponds to the number of generations since admixing, to g. We found that allowing SABER to estimate the values of τ yielded much poorer estimates of ancestry.

4.3.1 Simulated Datasets

We simulated admixed populations from the HapMap data in the following manner. We used the SNPs of chromosome 1 from the 500K Affymetrix GeneChip assay ® from each of the four HapMap populations: Yoruba people from Ibadan, Nigeria (YRI); Japanese from the Tokyo area (JPT); Han Chinese from Beijing (CHB); and Utah residents with ancestry from northern and western Europe (CEU). Overall, these span 38,864 SNPs for 60 unrelated individuals from CHB and YRI and 45 unrelated individuals from CHB and JPT.

For each pair of HapMap populations, we simulated admixed populations by random mating of individuals from the two populations across several generations. We started by joining a random set of αn individuals from the first population, and $(1-\alpha)n$ individuals from the second population. For the next generation, we repeatedly picked a random pair of individuals from the combined set of individuals, and generated a child for this pair by transmitting one chromosome from each individual. We repeated this process for g generations. We set the recombination rate to be 10^{-8} per base pair per generation consistent with previous studies [Nachman and Crowell, 2000]. We note that this model is a worst case scenario in the sense that in practice the populations are expected to mix in a slower rate, since individuals tend to mate with individuals from a similar ancestral background. We simulated admixture for various values of g and α; in the rest of this manuscript, the values of g and α are 7 and 0.2, unless stated otherwise. These parameters roughly match the nature of admixing in African-American populations [Patterson *et al.*, 2004; Falush *et al.*, 2003; Tian *et al.*, 2006; Parra *et al.*, 1998; Collins-Schramm *et al.*, 2003].

4.3.2 LAMP's performance and accuracy

We evaluated the accuracy of the ancestry estimates inferred by LAMP. We consider the two versions of LAMP, i.e., the de-novo inference of the local ancestries, as well as the inference of

the local ancestries based on genotype data of the original ancestral populations. We refer to the latter method as LAMP-ANC. For each individual i and SNP j, LAMP finds an estimate $\hat{a}_{ij}^p \in \{0, 0.5, 1\}$ for the true ancestry a_{ij}^p by a majority vote across the windows overlapping with position j. We measure the accuracy of LAMP as the fraction of triplets (i, j, p) for which $a_{ij}^p = \hat{a}_{ij}^p$.

We compared LAMP to two state of the art methods: STRUCTURE [Pritchard *et al.*, 2000] and SABER [Tang *et al.*, 2006]. SABER requires the input genotypes, admixture fraction α, physical location of the SNPs and the ancestral sequences that were used in the simulation (i.e., the original HapMap populations) and was also provided the number of generations g. For STRUCTURE, we only needed to provide the genotypes. We did not compare LAMP to methods such as AdmixMap [Hoggart *et al.*, 2004] and AncestryMap [Patterson *et al.*, 2004], since the high density of markers made these methods infeasible.

Table 4.1 summarizes the prediction accuracies of LAMP, LAMP-ANC, SABER, and STRUCTURE. LAMP and LAMP-ANC were run on the set of 38864 SNPs of chromosome 1. SABER and STRUCTURE were run on non-overlapping windows of 4000 SNPs that included 36000 of the original 38864 SNPs. This was done because STRUCTURE got into numerical instabilities when a large number of SNPs were used, and SABER crashed for an unknown reason when run on all the SNPs over the set of 500 individuals. For STRUCTURE the linkage model was used with 10,000 burn-in and 50,000 MCMC iterations. SABER was also seen to crash on some of the 4000 SNP blocks and these were excluded from the analysis. The accuracy of the ancestry estimates were obtained on the SNPs for which all methods returned a result. From Table 4.1, it is clear that LAMP achieves considerable improvement over the YRI-CEU and the CEU-JPT datasets, when compared to SABER or STRUCTURE. For the JPT-CHB dataset, LAMP is worse than SABER, but LAMP-ANC achieves a higher accuracy than SABER.

The accuracy of each of the methods varies across the population. We therefore measured the average accuracy in predicting the ancestries for each of the individuals. Figure 4.2 shows the cumulative distribution function of the accuracies achieved by each of the methods across the set of 500 individuals. As can be seen from the figure, the improvement of LAMP compared to STRUCTURE and SABER is quite significant. For the YRI-CEU dataset, when measuring the percentage of individuals that are predicted with an accuracy of at least 90%, LAMP achieves 90% while SABER and STRUCTURE achieve less than 10%. In general, the accuracy in the predictions that STRUCTURE makes has a higher variance than the predictions made by SABER and LAMP. On the CEU-JPT dataset, LAMP is more accurate than SABER. On the JPT-CHB dataset, SABER performs considerably better than LAMP; this is probably due to the fact that the ancestral populations, which are given to SABER but not to LAMP, are too similar to distinguish within a window; since LAMP-ANC uses the allele frequencies of the ancestral individuals as input while still inferring ancestries over entire windows, it is more accurate than SABER.

Table 4.1 also shows that LAMP achieves a gain in running time of at least two orders of magnitude. We found that, on a single computer, LAMP and LAMP-ANC take less than 30

seconds to run on a 4000 SNP block and less than 7 minutes to run on the entire set of 38864 SNPs.

These experiments suggest that LAMP is especially useful when the ancestral populations are sufficiently different from each other (e.g., CEU and YRI). In those cases, it is actually not essential to genotype the ancestral individuals, as we observe that LAMP-ANC and LAMP achieve similar accuracy levels. When the populations are closer (e.g., CHB-JPT), even for a modest number of generations of mixing (in our case, 7 generations), none of the methods performs well even when the ancestral populations are given.

4.3.3 Inferring individual admixture

Current studies often use the individual admixture of each individual across the genome to correct for population stratification [Falush *et al.*, 2003; Shriver *et al.*, 1997; Ziv and Burchard, 2003; Hanis *et al.*, 1986]. The individual admixture of an individual is defined by the proportion of ancestors of the individual from each of the ancestral populations. For instance, for an individual with a mother from CEU and a father from YRI, the individual admixture would be 50% YRI, and 50% CEU.

Even though LAMP is designed to estimate the locus-specific ancestry, we can use it to find the individual admixture. We compare the estimates of the individual admixture obtained from LAMP with those from STRUCTURE. We used the YRI-CEU dataset with $g = 7$ and $\alpha = 0.20$. We picked 4318 equally spaced SNPs from chromosome 1. This roughly matches the number of SNPs required to distinguish non-admixed individuals from even very closely related subpopulations [Sridhar *et al.*, 2007]. We ran STRUCTURE on this set of SNPs with 10000 burn-in iterations and 50000 iterations using the NOLINKAGE model and the NOADMIX mode option set to 0. We ran LAMP on the entire chromosome and then calculated the locus-specific ancestry of each individual by averaging the ancestries predicted across the same set of 4318 SNPs given to STRUCTURE. As shown by Figure 4.3, LAMP consistently achieves considerably better estimates for the individual admixture. In particular, the average error rate for LAMP is 2.1%, while the average error rate for STRUCTURE is 5.4%. The difference in the performance between the methods is statistically significant (Wilcoxon signed rank test p-value of 9.9×10^{-51}). This experiment suggests that since LAMP is more than 600 times faster than STRUCTURE (see Table 4.1), it would be better to use LAMP across the entire genome to infer the individual admixture, than to use STRUCTURE across a smaller set of AIMs. We also inferred the individual admixture using the LINKAGE model in STRUCTURE but found that this gave a significantly higher average error rate of 9.0%.

Another method to infer the individual admixture is EIGENSTRAT. We ran EIGENSTRAT on the SNPs used above and chose the largest eigenvector. The ancestries of the individuals were obtained by scaling the entries of the eigenvector to the interval $[0, 1]$. We found this procedure to result in individual admixtures with an average error rate of 13.4%. When we included 10 ancestral individuals each from the Hapmap YRI and CEU populations reduced

the average error to 4.1% (Wilcoxon signed rank test p-value of 1.3×10^{-83}). Using all 38864 SNPs decreased the average error to 11.1% and 3.6% respectively.

4.3.4 LAMP's performance across three admixed populations

When more than two populations are mixed, de-novo inference of the locus-specific ancestry is a more challenging task. We therefore compare LAMP-ANC, which uses the genotypes from the ancestral populations, to SABER, on a dataset generated by the mixing of three populations (YRI, CEU and JPT). We mixed these populations in the ratio 0.4 : 0.4 : 0.2 for seven generations. Figure 4.4 shows the ancestry estimates of LAMP-ANC for one of the individuals. LAMP-ANC accurately infers the ancestry over most of the chromosome, and it is clear that qualitatively the estimates are very close to the true ancestry. To give a more quantitative measure for the accuracy of LAMP-ANC, we calculated the cumulative distribution function of the accuracies for each individual of LAMP-ANC and of SABER (see Figure 4.5). Evidently, LAMP-ANC achieves a significantly better accuracy than SABER across the population (average accuracies of 92% and 74% respectively).

4.3.5 Empirical Robustness of LAMP

The performance of LAMP clearly depends on the nature of the data, on the number of generations g, and on α. We varied g for a simulated YRI-CEU admixed population with the fraction of CEU $\alpha = 0.20$. As shown in Figure 4.6, even when g is as large as 20, LAMP reaches an accuracy of 88%, and LAMP-ANC reaches an accuracy of 93%. For more realistic values of g, (i.e. $g < 10$) the accuracy of LAMP is above 93%.

To measure the effect of α on the performance of LAMP, we measured the performance for simulated data with $g = 7$ for different values of α (see Figure 4.6). We observe that LAMP performs well for values of α of up to 0.40 with it accuracy remaining above 90% and its performance drops sharply to a little above 50% accuracy for $\alpha = 0.5$.

Finally, we measured the effect of the distance between the ancestral populations, by comparing the accuracy of LAMP across the YRI-CEU, CEU-JPT and the JPT-CHB datasets. As shown in Table 4.1 (see also Figure 4.6), LAMP is quite accurate on the CEU-JPT and the YRI-CEU datasets, but its performance is quite poor on the JPT-CHB dataset. In such a situation, the availability of allele frequencies is essential for accurate inference, as we observe that LAMP-ANC maintains an accuracy of around 70%.

4.3.6 Robustness to Parameter Settings

Since LAMP requires as an input the values of α and g, it is important to verify that inaccurate estimates of these parameters do not affect the results significantly. We tested LAMP by benchmarking it over the simulated YRI-CEU dataset, with true values of $g = 7$ and $\alpha = 0.2$.

We ran LAMP on this dataset with different erroneous input values of g and α. In Figure 4.7 we observe that if the number of generations g is mistakenly given to LAMP as 4 or larger, than the accuracy of LAMP is kept reasonably high, and in particular it is at least 90%. On the other hand, it seems that if the input α is very different from the true α, LAMP can perform quite poorly. For instance, when the input α is set to 0.4 instead of 0.2, the accuracy level is about 85%. However, since α is a single parameter across all individuals, standard methods such as STRUCTURE [Pritchard *et al.*, 2000] give reasonable accuracy for α (e.g. the estimates for the YRI-CEU dataset are between 0.17 and 0.24 across 10 runs) , we can safely assume that the error in the prior estimate of α is within a factor of 0.5, in which case LAMP maintains a very good performance.

The model used in LAMP requires the SNPs to be independent. To ensure this, we discard SNPs with r^2 above a threshold. We empirically chose a threshold of 0.10 for r^2 so that we retain a majority of the SNPs. However, as shown in Figure 4.8 the accuracy of LAMP does not change much even when this threshold is raised so that the SNPs retained are no longer independent. The accuracy begins to decrease only at stringent thresholds below 0.005 due to a tendency to discard informative SNPs. We also examined the impact of the sample size on the ancestry estimates. While an increase in sample size might lead to SNPs being significantly linked even when the mutual r^2 is small, for practical purposes, such SNPs are essentially uncorrelated. Thus, LAMP is also robust to the sample size as shown in Figure 4.8.

Finally, we measured the effect of the method used to simulate the data on the different algorithms. To achieve this, we amplified the Hapmap haplotypes for YRI and CEU populations using the model of Li and Stephens [Li and Stephens, 2003]. Briefly, the Li and Stephens model generates additional haplotypes based on the ones already observed. The newly generated haplotypes are composed from previous ones, assuming mutation and recombination. The recombination rate in this model depends on the number of observed haplotypes, such that the rate is higher when less haplotypes are observed. This reduces the recurrent sampling of haplotypes, and as was shown by Li and Stephnes, mimics more accurately the generation of haplotypes. This resulted in a set of 10000 ancestral individuals which then underwent admixture with $g = 7$ and $\alpha = 0.20$ as described earlier. On this new dataset, the accuracies obtained by LAMP, LAMP-ANC and SABER were 94.72%, 94.70%, and 89.09% respectively. The accuracies are close to the accuracies obtained on the YRI-CEU dataset described in Table 4.1.

4.4 Discussion

We have presented a new method, LAMP, for *de-novo* estimation of locus-specific ancestry in recently admixed populations. Unlike previous methods for locus-specific ancestry (e.g., SABER), LAMP does not use any information about the ancestral populations (i.e., it estimates the ancestries de-novo). We show that LAMP is analytically justified and that it achieves

significant improvements over existing methods both in terms of accuracy of prediction and speed. In particular, LAMP can easily be applied to whole genome datasets, and the resulting locus-specific ancestries can be estimated within a few hours.

De-novo estimation of the locus-specific ancestries is sometimes infeasible, especially when the ancestral populations are very close to each other (e.g., CHB and JPT). We therefore extended LAMP to a method called LAMP-ANC, which uses additional genotypes from the ancestral populations as priors. This approach has been shown to be useful before by methods such as SABER.

When compared to previous methods, LAMP is shown to achieve significantly better accuracy than other methods (SABER and STRUCTURE). The increase in accuracy may be crucial when trying to correct for population stratification in studies that involve recently admixed populations, as well as in studies that are based on admixed mapping. Furthermore, improved accuracy in the locus-specific ancestry estimation has potential applications in finding better signals for selection and other events across the genome.

While LAMP relies on a knowledge of the parameters g and α, we have shown the robustness of the ancestry estimates to inaccuracies in these parameters. These parameters control the window size. As the window size is decreased, each window may contain fewer informative SNPs. On the other hand, errors in classifying individuals who have breakpoints within a window are reduced. This tradeoff is illustrated in Figure 4.7 where we see that the ancestry estimates are robust when g is overestimated. In practice, we would therefore recommend using an upper bound on g when g cannot be estimated accurately. Furthermore, g may actually be a more complex parameter. For example, if some portions of the admixed population have admixed for g_1 generations and other portions have been admixed for only g_2 generations, where g_2 is smaller than g_1. In this case, g is set to be g_1, and more accurate results are expected than if the whole population has admixed for exactly g1 generation.

The fact that the LAMP algorithm performs better on the unbalanced case ($\alpha << 0.5$) than on the balanced case, seems counterintuitive at first. The reason for this phenomenon is the fact that a small α helps to break the symmetry. Even if all windows were perfectly clustered, combining the solutions of the different windows into one integrated solution is not a simple task when $\alpha = 0.5$ due to symmetry. That is, after clustering the individuals in every given window, we are still left with the problem of deciding which cluster is population 1, and which one is population 2. If $\alpha < 0.5$, then this decision is easier since the smaller cluster could be labeled as population 1, and the larger cluster as population 2.

Further, it is interesting to note that even though SABER models the LD structure while LAMP does not, it appears that LAMP performs better than SABER. This could be attributed to several possible reasons. First, it may be that the LD structure only adds slightly to the information captured by the independent SNPs. Second, it may be that optimizing the model in SABER is a harder task than optimizing the model in LAMP due to the larger number of parameters, and thus SABER may potentially not converge to the global optimum of its parameter space.

A simple extension to LAMP can be used to infer the individual admixture. As we show here, the resulting estimates of the individual admixture are considerably better than the estimates achieved by STRUCTURE or EIGENSTRAT. A number of recent studies have produced panels of Ancestry Informative Markers (AIMs) in admixed populations [Tian *et al.*, 2007; Mao *et al.*, 2007; Price *et al.*, 2007; Smith *et al.*, 2004], which are SNPs that have differing frequencies in the ancestral populations. It is possible that the AIMs may be used to improve the accuracy of individual admixture prediction done by STRUCTURE or other methods including LAMP. However, the AIMs have disadvantages since there is a risk of over-fitting, and the studied population may be somewhat different than the population for which the AIMs were found. As we show here, in an era where the genotyping technology is getting cheaper, it is useful to use the entire set of genotyped SNPs in the analysis of population stratification.

Dataset	Distance	LAMP	LAMP-ANC	SABER	STRUCTURE
YRI-CEU	0.055	0.94	0.95	0.87	0.84
CEU-JPT	0.036	0.87	0.93	0.82	0.47
JPT-CHB	0.0045	0.48	0.72	0.68	0.40
Time (sec)		394	246	7681	2.57×10^5
		(38864 SNPs)	(38864 SNPs)	(4000 SNPs)	(4000 SNPs)

Table 4.1: A summary of the comparison between LAMP, LAMP-ANC, SABER, and STRUCTURE (LAMP-ANC is LAMP run with knowledge of the ancestral genotypes). The accuracy across all positions on chromosome 1 is shown for the three admixed populations. The distance between the admixing population (measured by the mean squared distance between the allele frequency vectors) is also shown indicating the difficulty in separating alleles from the populations. The time taken to run each of the methods is shown. LAMP and LAMP-ANC were run on the entire set of 38864 SNPs while SABER and STRUCTURE were run on non-overlapping blocks of 4000 SNPs due to issues with scaling them to the entire dataset. For SABER and STRUCTURE the accuracies reported here are obtained by averaging the accuracies across the blocks while the running time is the time to run a single block. Each of these methods was run on a single computer.

Appendix 4.A Correctness of MAXVAR

In this section, we analyze the correctness of the MAXVAR algorithm. We have two populations A_1 and A_2. We denote $\alpha = \alpha_2$ as the admixture fraction of A_2 - the smaller of the two populations. The MAXVAR algorithm classifies the individuals into three types of ancestries, i.e., A_1A_1, A_1A_2, and A_2A_2. The algorithm works by first picking a specific individual termed a *pivot*, and then clustering individuals based on their similarity to the pivot. We show that

when the the populations are significantly different from each other, the pivot will have an ancestry A_2A_2 with high probability. In this case, we show that one can define a similarity score $\mathcal{S}$ (as defined in the Methods Section), such that the individuals who are also of ancestry A_2A_2 have positive similarity score to the pivot while those with ancestry A_1A_1 have negative similarity scores in expectation. Thus, the individuals with the smallest $(1-\alpha)^2 n$ values of the similarity score are assigned an ancestry of A_1A_1, the largest $\alpha^2 n$ values are assigned an ancestry of A_2A_2 and the rest are assigned A_1A_2.

Let $p_{A_1A_1}, p_{A_1A_2}, p_{A_2A_2}$ be the frequencies of individuals of the three types in the population. We assume that $p_{A_1A_1} = (1-\alpha)^2$, $p_{A_2A_2} = \alpha^2$, and that $p_{A_1A_2} = 2\alpha(1-\alpha)$. Let p_k, q_k be the minor allele frequencies of population A_1 and A_2 respectively in position k. Furthermore, we assume that the values of μ_k and σ_k (as defined in the Methods Section) are constants, and that $\mu_k = \alpha p_k + (1-\alpha)q_k$, $\sigma_k^2 = 2\alpha^2 p_k(1-p_k) + 2\alpha(1-\alpha)[p_k(1-p_k) + q_k(1-q_k)] + 2(1-\alpha)^2 q_k(1-q_k)$. If the number of individuals is large enough, the variance is quite low, and therefore this is not a restrictive assumption. We define the *distance* between the two populations as $W = \sum_k (p_k - q_k)^2 \sigma_k^2$. Under these assumptions, it is easy to see that if aa, ab, bb are three given individuals with ancestry A_1A_1, A_1A_2, A_2A_2 respectively in the window, then the expected similarity score $\mathcal{S}$ between pairs of individuals is:

$$\begin{aligned} E[\mathcal{S}(aa, aa')] &= 4\alpha^2 W, & E[\mathcal{S}(aa, ab)] &= -2(1-2\alpha)\alpha W, \\ E[\mathcal{S}(aa, bb)] &= -4(1-\alpha)\alpha W, & E[\mathcal{S}(ab, ab')] &= (1-2\alpha)^2 W, \\ E[\mathcal{S}(ab, bb)] &= 2(1-2\alpha)(1-\alpha)W, & E[\mathcal{S}(bb, bb')] &= 4(1-\alpha)^2 W, \end{aligned} \tag{4.9}$$

where aa', ab', bb' are individuals with ancestries A_1A_1, A_1A_2, A_2A_2, but they are different individuals than aa, ab, and bb. From this, it is easy to verify that the expected sum of squares over all individuals that are different from aa, ab, and bb can be approximated as:

$$\begin{aligned} \sum_{i:i\neq bb} E[\mathcal{S}(i, bb)]^2 &\approx 8(1-\alpha)^3 \alpha W^2 n \\ \sum_{i:i\neq ab} E[\mathcal{S}(i, ab)]^2 &\approx 2\alpha(1-\alpha)(1-2\alpha)^2 W^2 n \\ \sum_{i:i\neq aa} E[\mathcal{S}(i, aa)]^2 &\approx 8\alpha^3(1-\alpha) W^2 n \end{aligned}$$

The only reason for the approximation is that the number of individuals with ancestry A_2A_2 that are different from bb is $(1-\alpha)^2 n - 1$, while we consider it as $(1-\alpha)^2 n$. This approximation is not restrictive when the number of individuals is reasonably large.

Defining $P_k = p_k(1-p_k), Q_k = q_k(1-q_k)$, we can write the variance of these scores as

$$V[\mathcal{S}(aa,aa)] = 4\sum_k \tfrac{Q_k^2}{\sigma_k^4}, \qquad V[\mathcal{S}(aa,ab)] = 2\sum_k \frac{Q_k^2+P_kQ_k}{\sigma_k^4}$$

$$V[\mathcal{S}(aa,bb)] = 4\sum_k \tfrac{P_kQ_k}{\sigma_k^4}, \qquad V[\mathcal{S}(ab,ab)] = \sum_k \frac{P_k^2+Q_k^2+2P_kQ_k}{\sigma_k^4}$$

$$V[\mathcal{S}(ab,bb)] = 2\sum_k \tfrac{P_k^2+P_kQ_k}{\sigma_k^4}, \quad V[\mathcal{S}(bb,bb)] = 4\sum_k \frac{P_k^2}{\sigma_k^4}$$

We can conclude that the variance of the similarity scores from the rest of the individuals to an individual with one of the 3 ancestries is

$$\begin{aligned}
V[\sum_{i:i\neq bb}\mathcal{S}(i,bb)] &\approx 4\alpha^2 n\sum_k \frac{P_k^2}{\sigma_k^4} + 4\alpha(1-\alpha)n\sum_k \frac{P_k^2+P_kQ_k}{\sigma_k^4} + 4(1-\alpha)^2 n\sum_k \frac{P_kQ_k}{\sigma_k^4} \\
&= 4n\sum_k P_k \frac{\alpha^2P_k + \alpha(1-\alpha)(P_k+Q_k) + (1-\alpha)^2Q_k}{\sigma_k^4} = 2n\sum_k \frac{P_k}{\sigma_k^2} \\
V[\sum_{i:i\neq aa}\mathcal{S}(i,aa)] &\approx 4\alpha^2 n\sum_k \frac{P_kQ_k}{\sigma_k^4} + 4\alpha(1-\alpha)n\sum_k \frac{Q_k^2+P_kQ_k}{\sigma_k^4} + 4(1-\alpha)^2 n\sum_k \frac{Q_k^2}{\sigma_k^4} \\
&= 4n\sum_k Q_k \frac{\alpha^2P_k + \alpha(1-\alpha)(P_k+Q_k) + (1-\alpha)^2Q_k}{\sigma_k^4} = 2n\sum_k \frac{P_k}{\sigma_k^2} \\
V[\sum_{i:i\neq ab}\mathcal{S}(i,ab)] &\approx n\sum_k \frac{2\alpha^2(P_k^2+P_kQ_k) + 2\alpha(1-\alpha)(Q_k^2+P_k^2+2P_kQ_k) + 2(1-\alpha)^2(Q_k^2+P_kQ_k)}{\sigma_k^4} \\
&= n\sum_k \frac{Q_k+P_k}{\sigma_k^2}
\end{aligned}$$

Hence, the squared distances are given by

$$\begin{aligned}
E[\sum_{i:i\neq bb}\mathcal{S}(i,bb)^2] &= V[\sum_{i:i\neq bb}\mathcal{S}(i,bb)] + \sum_{i:i\neq bb}E[\mathcal{S}(i,bb)]^2 \\
&\approx n\left(2\sum_k \frac{P_k}{\sigma_k^2} + 8\alpha(1-\alpha)^3W^2\right) \\
E[\sum_{i:i\neq ab}\mathcal{S}(i,ab)^2] &= V[\sum_{i:i\neq ab}\mathcal{S}(i,ab)] + \sum_{i:i\neq ab}E[\mathcal{S}(i,ab)]^2 \\
&\approx n\left(\sum_k \frac{P_k+Q_k}{\sigma_k^2} + 2\alpha(1-\alpha)(1-2\alpha)^2W^2\right) \\
E[\sum_{i:i\neq aa}\mathcal{S}(i,aa)^2] &= V[\sum_{i:i\neq aa}\mathcal{S}(i,aa)] + \sum_{i:i\neq aa}E[\mathcal{S}(i,aa)]^2 \\
&\approx n\left(2\sum_k \frac{Q_k}{\sigma_k^2} + 8\alpha^3(1-\alpha)W^2\right)
\end{aligned}$$

Asymptotically, when n is large enough, the pivot i^* will be from A_2A_2 if $E[\sum_{i:i\neq bb}\mathcal{S}(i,bb)^2] > max(E[\sum_{i:i\neq ab}\mathcal{S}(i,ab)^2], E[\sum_{i:i\neq aa}\mathcal{S}(i,aa)^2])$. After simplifying the above expressions, we get that the requirement is that

$$\sum_k \frac{P_k - Q_k}{\sigma_k^2} + 4\alpha(1-\alpha)(1-2\alpha)W^2 \;>\; 0$$

The last inequality holds if the distance between the populations (W) is large enough. In that case, by Equation 4.9, the ordering of the individuals according to their similarity to the pivot should give the correct clustering asymptotically.

Appendix 4.B Accuracy of the window length and the majority vote

For a given window, the analysis in Section 4.2.4 shows that the expected fraction of individuals with no breakpoints is $1-\gamma$. Here, we strengthen this analysis under the assumption that the errors in the predictions of the different windows are independent.

It is easy to see that the expected fraction of individuals with two or more breakpoints in a window is smaller than γ^2. For a given individual with a breakpoint in position i, we denote the ancestry by $(A_{s_1}, A_{s_2}, i, A_{s_3})$, where A_{s_1} is the ancestry of the non-recombinant chromosome and A_{s_2} and A_{s_3} are the ancestries of the recombinant chromosome. We assume

that the probability to classify such an individual as $A_{s_1}A_{s_2}$ is $\frac{i}{l}$, and the probability to classify it as $A_{s_1}A_{s_3}$ is $1-\frac{i}{l}$. There are l windows that overlap with any SNP. Consider a SNP which is a distance d away from a breakpoint. Let X be the number of times that the SNP is incorrectly classified as $A_{s_1}A_{s_3}$. Clearly,

$$E[X] = \sum_{i=1}^{l-d} \frac{i}{l} \approx \frac{(l-d)^2}{2l}.$$

Using a Chernoff bound [Chernoff, 1952], the probability to incorrectly classify this SNP after the majority vote is

$$\Pr(X > \frac{l}{2}) = \Pr(X > (1+\frac{d}{l-d})^2 E[X]) < e^{-\frac{(\frac{d}{l-d}(2+\frac{d}{l-d}))^2 E[X]}{2}} = e^{-\frac{(d(2+\frac{d}{l-d}))^2}{4l}} < e^{-\frac{d^2}{l}}.$$

In the case that there are no other breakpoints within distance l from the breakpoint considered, the expected number of errors around the breakpoint for the individual is

$$\int_0^l e^{\frac{-x^2}{l}} dx = \int_0^{\sqrt{2l}} e^{\frac{-x^2}{2}} \sqrt{l/2} dx \approx \sqrt{l/2}\sqrt{2\pi} = \sqrt{l/\pi}$$

If there are breakpoints within distance l of each other, we take the worst-case assumption that all windows containing the two breakpoints make erroneous predictions over their entire length l. Since the expected fraction of breakpoints in an individual is $\frac{\gamma}{l}$, and the expected fraction of pairs of breakpoints that are of distance smaller than l is at most γ^2, we can bound the expected number of errors as $\frac{\gamma}{\sqrt{\pi l}} + \gamma^2 = 4\sum_{i<j}\alpha_i\alpha_j(g-1)r(\sqrt{l/\pi} + 4(\sum_{i<j}\alpha_i\alpha_j)(g-1)rl^3)$. Based on this analysis, a sufficient condition to achieve a desired error rate of ϵ is to have $\frac{1}{\pi}(\frac{8(g-1)r(\sum_{i<j}\alpha_i\alpha_j)}{\epsilon})^2 < l < \frac{1}{4(g-1)r\sum_{i<j}\alpha_i\alpha_j}\sqrt{\frac{\epsilon}{2}}$. For typical values of g and r, the lower bound on l is small enough to be negligible.

Appendix 4.C Estimate of window length

The window length derived in Equation 4.8 bounds the classification errors within each window to a desired error rate ϵ. Since all SNPs within a window are assigned the same ancestry, *any* algorithm that is used within this window will incur some errors in the presence of breakpoints. Hence the window length was calculated under the assumption that the classification algorithm within the window was the most accurate possible *i*.e., any errors in the classification were only a result of breakpoints within a window. Here we empirically show that, for the window lengths computed using Equation 4.8, the average classification error for a most accurate classification is bounded by the error rate ϵ which is set to 0.10.

Within each window, the most accurate ancestry assignment is inferred assuming that the true ancestries are known. The most accurate assignment consists of assigning to an individual the ancestry found in a majority of the SNPs in that window. Thus, an individual who has no breakpoints is always correctly classified while an individual with a breakpoint at position $i < \lfloor \frac{l}{2} \rfloor$ in a window of length l and ancestries A_{s_1} and A_{s_2} on either side of the breakpoint will have errors in positions $\{1, \ldots, i\}$. The error rate for a window is the fraction of positions that are incorrectly classified in the window.

We computed the average of these errors in overlapping windows that span chromosome 1 of the YRI-CEU dataset for different values of g and α. We see in Figure 4.9 that the average error is below ϵ. However, the variance of the estimates (indicated by the minimum and the maximum fraction of errors) increases with larger g or with $\alpha \to 0.5$. The window size estimates seem to provide a good bound on the average fraction of errors due to breakpoints.

Appendix 4.D Practical issues in implementing LAMP

In this section, we describe some of the issues that we faced while implemeting LAMP. One of the issues that we needed to address was how to determine the degree of overlap between adjacent windows. An extreme degree of overlap would require adjacent windows to differ in a single SNP. In practice, we found that a smaller degree of overlap, where consecutive windows overlapped in a fraction $c = 80\%$ of their length, did not significantly change the accuracy while resulting in faster running times. The overlap between adjacent windows can be exploited to further improve the running time. While using the MAXVAR algorithm to obtain an initial classification, each window requires a computation of the similarity score between all pairs of individuals. The similarity score is computed using an inner product of the normalized genotypes as described in Section 4.2.3.2. Instead of computing these similarity scores over entire windows of length l, we can compute these scores over chromosomes of length $(1-c)l$. The similarity score in a new window can then be computed from that of the previous window by adjusting for the non-overlapping regions.

As we mentioned at the end of Section 4.2.4, the window length calculation should take into account the distance between the two populations. This is feasible when the ancestral genotypes are known. In this scenario, the accuracy of the classification for a given window length can be obtained by running LAMP-ANC on the ancestral genotypes. With an increase in the window length, this accuracy is exptected to increase. On the other hand, the errors due to breakpoints, as given in Equation 4.8, increases with window length. We can then search for the window length that maximizes the product of the fraction of individuals who do not have breakpoints and the fraction of these individuals who are accurately classified. For populations that are well-separated such as YRI-CEU and CEU-JPT, we find that the number of SNPs needed to accurately classify a non-admixed individual is much smaller than the length of the window obtained from Equation 4.8, so that it suffices to simply set the

window length to the latter estimate.

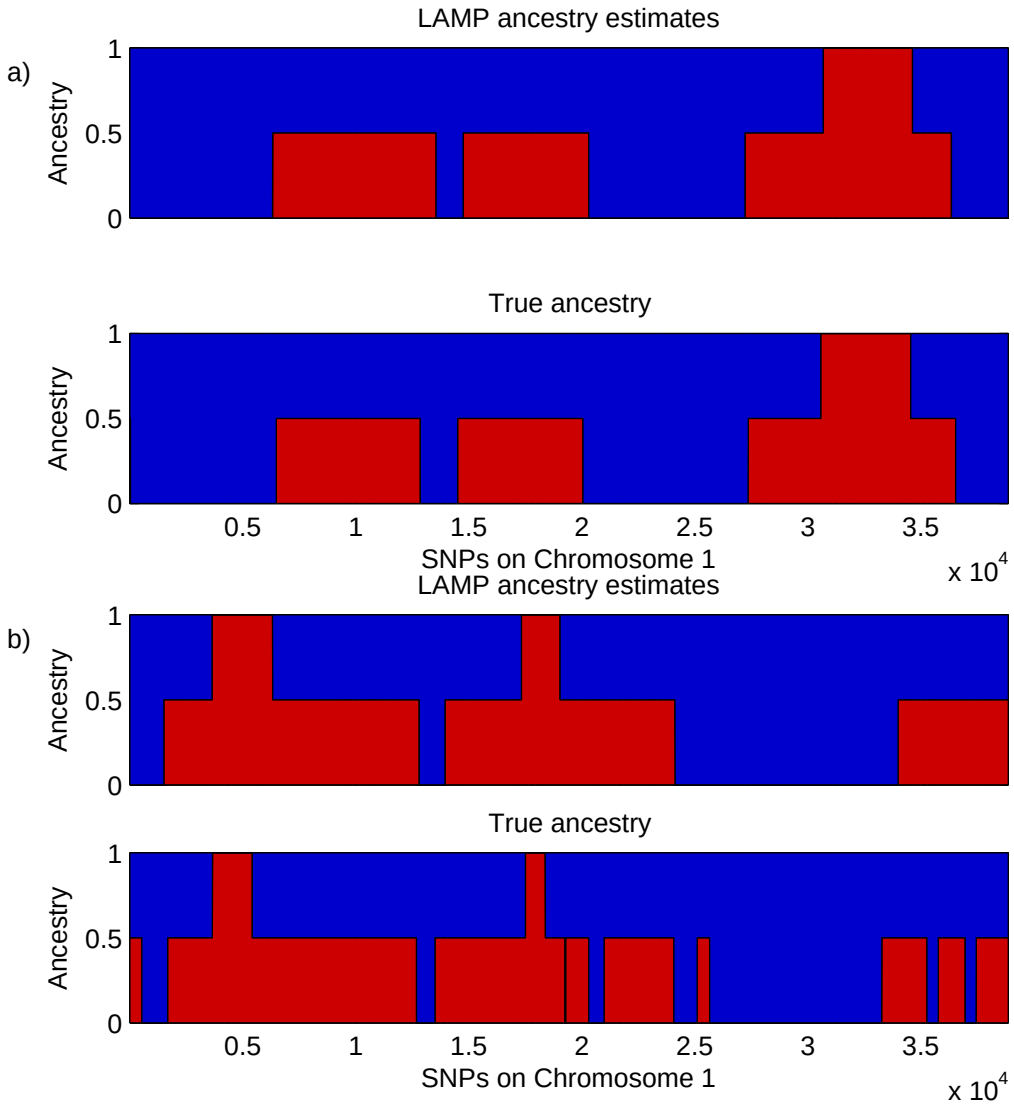

Figure 4.1: Two individuals in an admixed population. Ancestries predicted by LAMP(top panel) and true ancestries (bottom panel) are shown for each individual. As shown in the Figure, the ancestries (represented by red and blue) vary across the genome, and LAMP performs well in inferring the ancestry at each location.

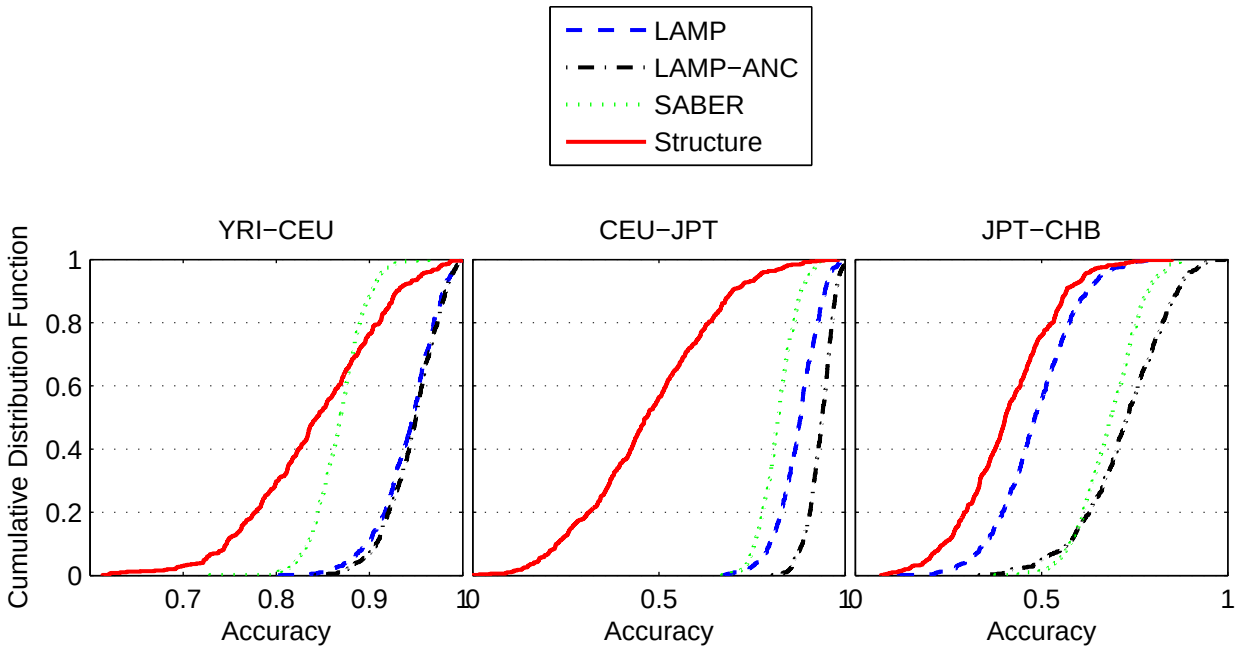

Figure 4.2: Comparison of the accuracies of LAMP, LAMP−ANC, SABER, and STRUCTURE on 3 admixed populations - YRI-CEU (left), CEU-JPT (middle), and JPT-CHB (right). The cumulative distribution function (CDF) is obtained from the accuracy of ancestry predictions for each individual. Note that the scales differ across the plots. CDFs that are to the right side correspond to higher accuracy. The graph on the left, for instance, shows that LAMP achieves an accuracy of at least 92% on 90% of the individuals. LAMP achieves an improved accuracy over SABER and STRUCTURE in the YRI-CEU and CEU-JPT populations while performing worse on the JPT-CHB population. LAMP−ANC performs consistently well on all three populations. Also notice the decrease in accuracy across all methods as we move from left to right as the ancestral populations become more similar.

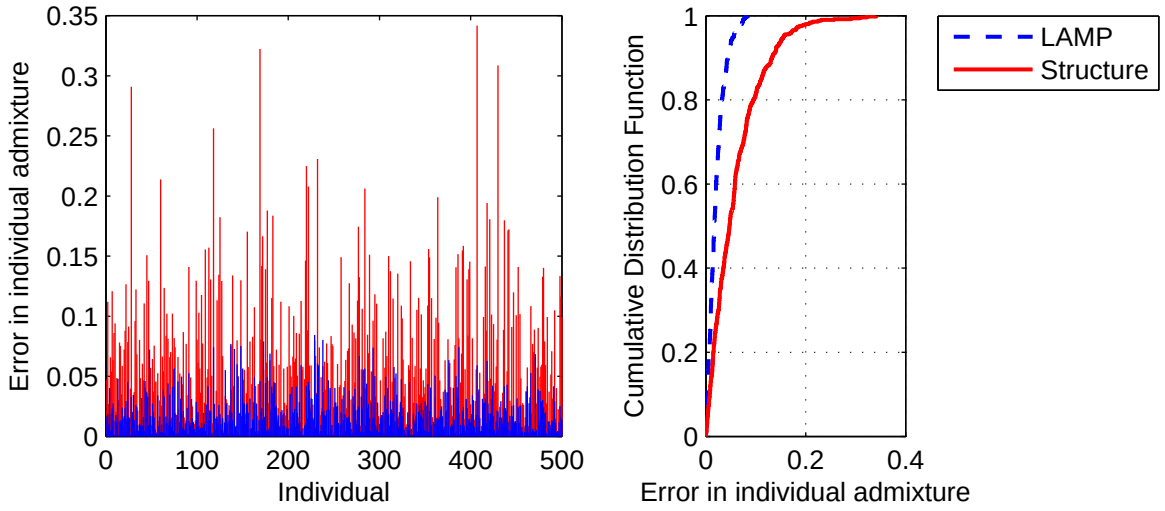

Figure 4.3: Comparison of the accuracy of methods for predicting individual admixture. a) The errors in the individual ancestries for each of the 500 YRI-CEU individuals. b) Errors in a) plotted as a Cumulative Distribution Function (CDF). The top-left region of the curve corresponds to higher accuracy. LAMP predicts the individual admixture with an error of less than 3% in 80% of the cases.

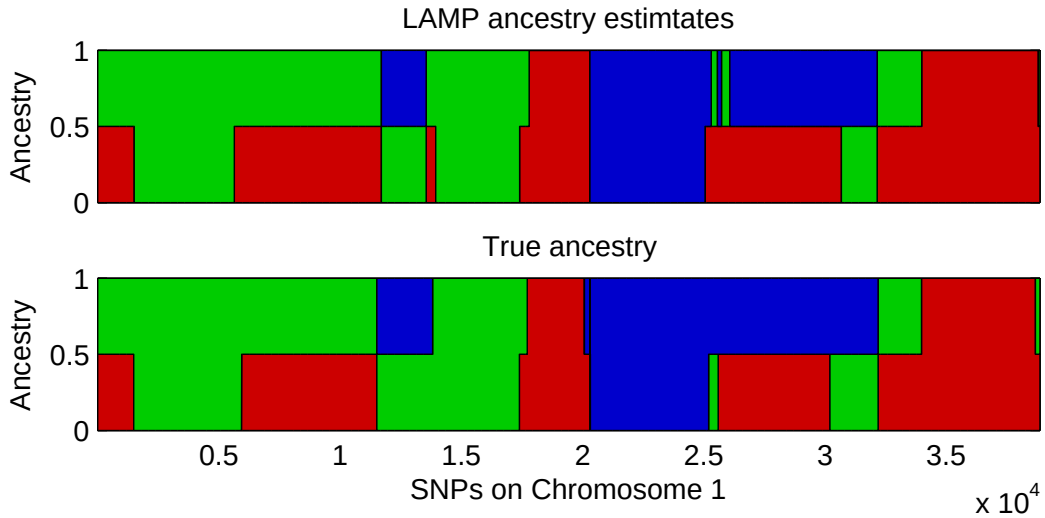

Figure 4.4: Ancestry estimates for an individual in an admixture of YRI-CEU-JPT in the ratio $0.4, 0.4, 0.2$. The top panel shows the LAMP ancestry estimates and the bottom panel the true ancestries. Red, green and blue represent YRI, CEU and JPT respectively.

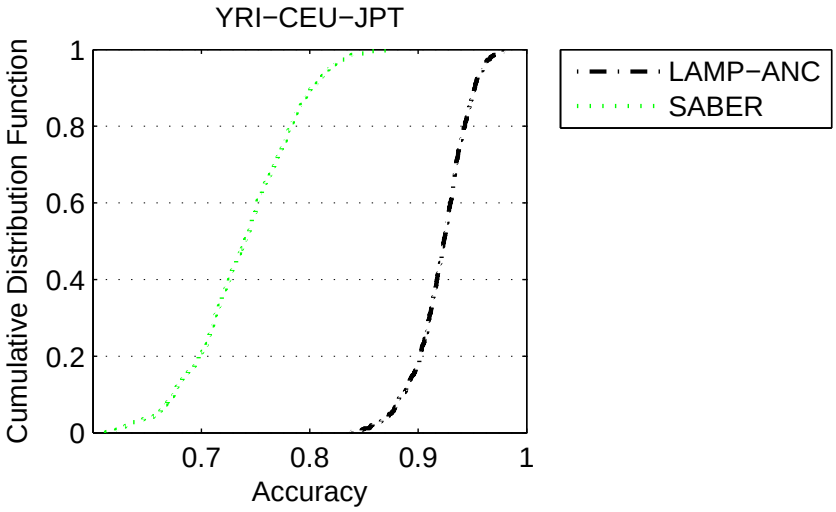

Figure 4.5: Cumulative Distribution Function (CDF) of the accuracy achieved per individual. The methods compared are `LAMP-ANC` and `SABER` for the YRI-CEU-JPT admixture. `LAMP` achieves an accuracy of at least 80% on all the individuals.

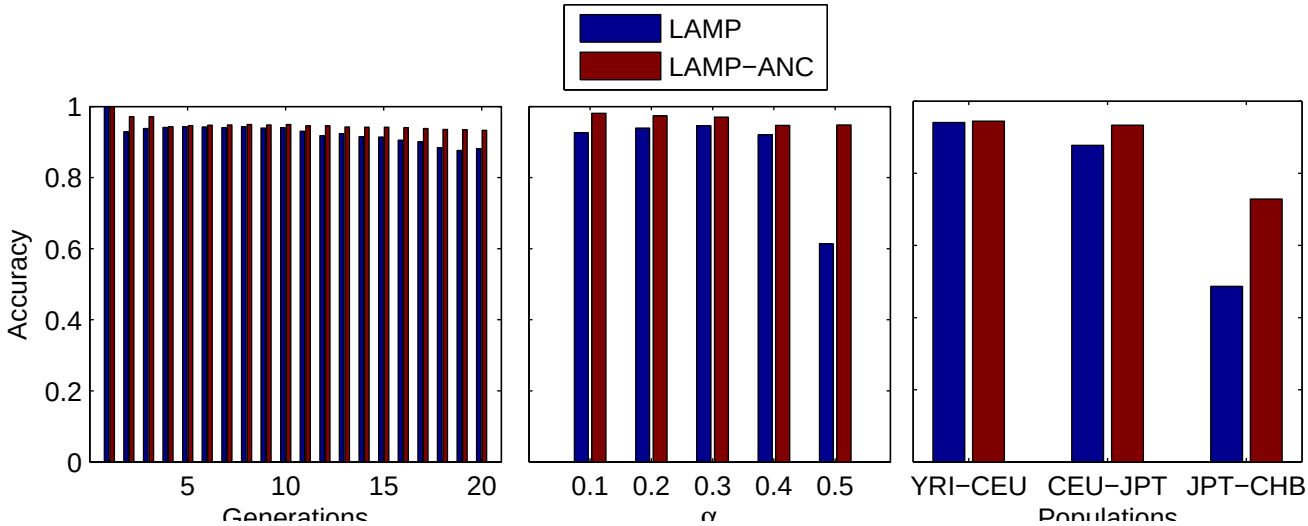

Figure 4.6: Accuracy of `LAMP` and `LAMP-ANC` with varying number of generations g, fraction of admixture α and populations. In each figure, inferring the ancestries becomes increasingly harder as we move from left to right. The difference in the accuracies between `LAMP` and `LAMP-ANC` shows the advantage conferred by a knowledge of the ancestral allele frequencies.

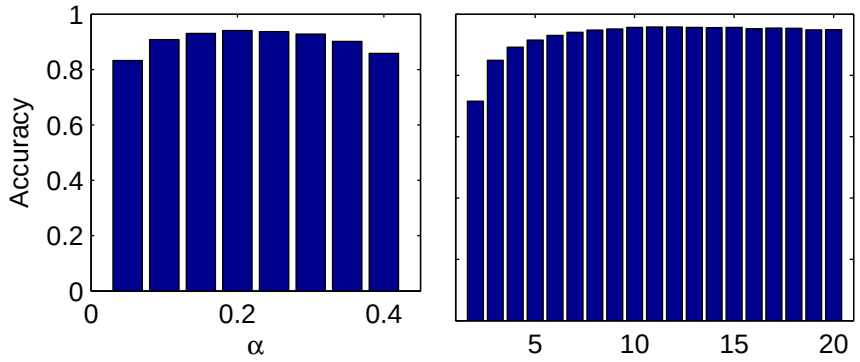

Figure 4.7: Robustness of LAMP estimates to uncertainty in the parameters - g, α. The accuracy of LAMP has been shown on the YRI-CEU dataset for different values of g and α with true values of $g = 7$ and $\alpha = 0.20$.

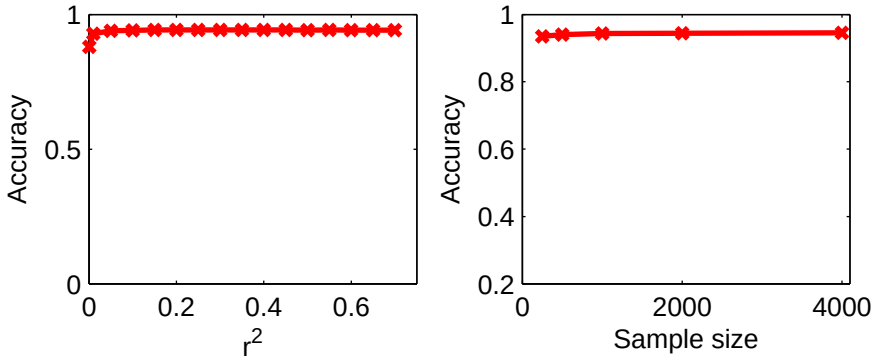

Figure 4.8: Robustness of LAMP estimates to the r^2 threshold used to discard SNPs and the sample size. The accuracy of LAMP has been shown on the YRI-CEU dataset for different values of g and α with true values of $g = 7$ and $\alpha = 0.20$.

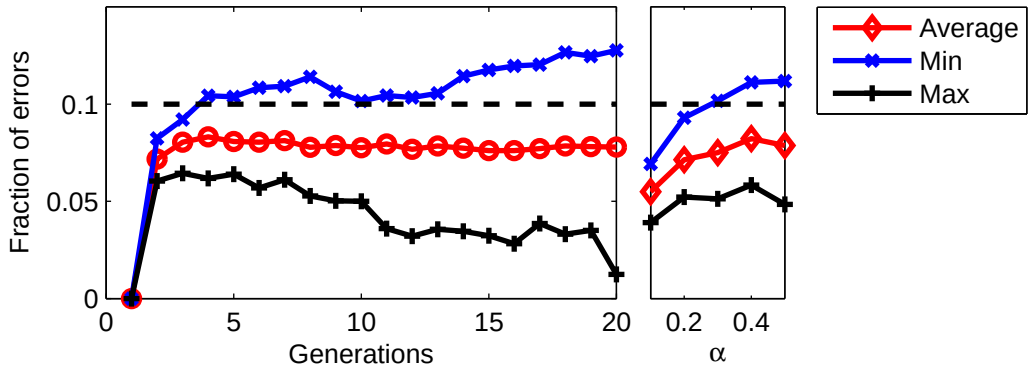

Figure 4.9: Empirical validation of the window length estimates. The window length is estimated in the Methods Section (Section 4.2.4). These estimates are based on a parameter ϵ, which represents the average desired fraction of errors incurred by the most accurate classification algorithm that can only return one of A_1A_1, A_1A_2, A_2A_2 for the entire window. The figure presents the actual average error rates for different values of g and α, run on the CEU-YRI dataset, with $\epsilon = 0.1$. Evidently, the actual average error rate falls within the desired error bound. The maximum and the minimum fraction of errors in a window are also shown.

Chapter 5

A probabilistic model for inferring local ancestry

5.1 Introduction

Given the genetic underpinnings of the ancestral origin problem it is natural to consider inference methods based on probabilistic models. Indeed, most previous work has made use of hidden Markov models (HMMs), where the states are the ancestral populations, the transitions roughly correspond to historical recombination events and the emission matrix models population-specific allele frequencies [Pritchard *et al.*, 2000; Falush *et al.*, 2003; Patterson *et al.*, 2004; Hoggart *et al.*, 2004]. Such Markovian models capture the linkage disequilibrum (LD) among alleles that arises due to admixture, but they fail to account for within-population linkage disequilibrium (the HMM assumes that alleles are independent once the ancestries are known). It is possible, however, to augment the HMM to include additional Markovian dependencies among the observed alleles to attempt to account for the latter form of LD; such a model has been referred to as a Markov Hidden Markov Model (MHMM) and has been implemented in the program SABER [Tang *et al.*, 2006].

In this chapter, we consider an augmented form of the HMM/MHMM framework for modeling admixture which includes explicit indicators for recombination events. Specifically, if a recombination event occurs between SNPs, then the ancestry of the SNPs are chosen independently; if recombination does not occur, then the ancestries are set equal. These explicit indicators serve several purposes. First, they make it possible to estimate the location of recombination events; the set of events is generally a strict superset of the set of change-of-ancestry events that are captured by the state sequence. The use of explicit indicators within an admixture model thus makes it possible to use admixture data to make inferences regarding historical recombinations and recombination rates. Second, recombination indicators can yield improvements in the estimates of haplotype frequencies. Note in particular that the MHMM used in SABER conditions on the ancestral state to decide

whether to use pairwise or singleton allele probabilities (if the state does not change, then the pairwise probabilities are used; otherwise singleton probabilities are used). However, haplotypes are broken up by recombination, not only by change of ancestry, and it would seem desirable to be able to condition on these more fine-grained events.

One of the goals of this chapter is thus to investigate the role of recombination indicators in HMM/MHMM models. Another goal is to consider more broadly whether the HMM/MHMM modeling and inference framework provides a practical computational solution to the problem of modeling of admixture and LD. In these models, inference of ancestry is tractable once its parameters are determined, but the need to estimate various hyperparameters is a challenging problem that has led researchers to Markov chain Monte Carlo (MCMC) sampling procedures. These procedures have desirable theoretical properties in the limit of large numbers of samples, but in practice they can be overly slow for realistic data sets.

To tackle the computational problem, [Sankararaman *et al.*, 2008] have recently presented a rather different, non-model-based approach to inferring locus-specific ancestries. This method (referred to as "LAMP") is based on running a window over the genome, computing the local ancestry of each individual within each window based on a local-likelihood model, and combining the results from the windows overlapping a given SNP using a majority vote. [Sankararaman *et al.*, 2008] have shown empirically that this approach provides estimates of ancestry that significantly improve on the HMM-based methods. This improvement may be due to the inadequacy of the Markovian assumptions, but it may also arise because the HMM models are being initialized randomly and the MCMC procedures are not mixing on a practical time scale.

To address this issue, note that practical applications of HMMs in other literatures, most notably the speech and signal processing literatures [Huang *et al.*, 2001], emphasize the critical need for effective initialization of parameter estimation procedures for HMMs. Practical inference for HMM-based admixture models may also require effective initialization. Accordingly, we investigate the possibility of using the solution from LAMP to initialize an HMM. Hill-climbing in likelihood from the LAMP solution may provide an effective way to retain the advantages of a model-based method while not sacrificing performance.

A final issue that we investigate concerns the modeling of background LD when the data are a dense set of SNPs. As alluded to earlier, the HMM does not attempt to model background LD. The MHMM models background LD via a simple first-order Markov chain that links neighboring alleles. To evaluate the adequacy of this model of background LD, we compare the MHMM to an alternative approach that prunes SNPs with a heuristic that discards highly-correlated SNPs and then uses these SNPs as input to an HMM.

Our experimental work focuses on the problem of inferring locus-specific ancestries in a population that is assumed to originate from two unknown ancestral populations [Sankararaman *et al.*, 2008; Falush *et al.*, 2003]. We also consider a less-studied scenario in which we assume that one of the ancestral populations is unknown, or its genotypes are not given, and

we wish to infer the allele frequencies in this population. This scenario may be appropriate in situations in which it is difficult to obtain external estimates of the allele frequencies of one of the ancestral populations. This is the case, for example, in many modern Caribbean populations (such as Puerto Ricans), where the native American population has vanished.

5.2 Methods

In this section, we describe the augmented HMM that serves as the basis of our experiments. We also describe an MHMM that incorporates a model of background LD along the lines of SABER [Tang *et al.*, 2006]. We then describe various forms of inference algorithms for these hidden Markov models, emphasizing the use of the EM procedure for parameter estimation.

5.2.1 Probabilistic Model

To simplify our presentation, let us assume that the number of populations that have been admixed is two (the notation is slightly more involved in the case of more than two populations but no new ideas are needed). Also, again for simplicity of presentation, we restrict our attention to haplotypes; genotypes can be handled in a straightforward manner as described in Appendix 5.4.1.

Let m denote the number of haplotypes, and let n denote the number of SNPs. Let X be the observed binary matrix of SNPs; i.e., $X_{i,j}$ is the jth SNP of the ith haplotype. Let $\boldsymbol{p}$ and $\boldsymbol{q}$ be the vectors of the allele frequencies in the ancestral populations. Hence, p_j is the probability to obtain '1' in the jth SNP in the first population and q_j is the corresponding probability in the second population. The matrix Z denotes the ancestry information of each haplotype at each SNP: $Z_{i,j} \in \{0, 1\}$ holds the ancestry of haplotype i at the jth SNP (0 if SNP j of haplotype i originated from the first population and 1 if it originated from the second). We use the matrix W to denote recombination events. Specifically, $W_{i,j}$ equals '1' if at least one recombination event occurred during the history of the admixture process in the ith haplotype in the interval between the $(j-1)$th SNP and the jth SNP, and '0' otherwise. The $(n-1)$-dimensional vector $\boldsymbol{\theta}$ denotes the probability of at least one such recombination event, with θ_j corresponding to the interval between the $(j-1)$th and the jth SNPs. The fraction of the first population in the ancestral population, which we call the *admixture fraction*, is denoted by α. Finally, g denotes the number of generations of the admixed process (in the sense that $\frac{1}{g-1}$ models the average length of ancestral chromosome blocks in the current admixed population).

Given the parameters g, α, $\boldsymbol{p}$, $\boldsymbol{q}$, and $\boldsymbol{\theta}$, we model a haplotype as being generated as follows: recombination points are generated on each chromosome based on a Poisson process whose rate parameter depends on g and the recombination rate r. This process corresponds to setting some of the W's to 1. Then the ancestries of the resulting chromosomal blocks

are determined independently for each block with α being the probability to choose the first ancestry. We assume that the mating is random across the populations. Given the ancestry at a particular position, an allele is generated using the corresponding ancestral allele frequency. We assume that the alleles are generated independently in a block.

We now describe the marginal and conditional distributions of the model. We assume a uniform prior over the interval $[0, 1]$ for each of the parameters α, $\boldsymbol{p}$, $\boldsymbol{q}$. The parameter g is assumed to be distributed uniformly over the interval $[g_{min}, g_{max}]$ for some $g_{max} > g_{min} > 1$. Given the ancestry and given the allele frequencies of a specific SNP j in haplotype i, $X_{i,j}$ is a Bernoulli random variable with distribution:

$$\Pr(X_{i,j} = 1 | Z_{i,j}, p_j, q_j) = \begin{cases} p_j & Z_{i,j} = 0 \\ q_j & Z_{i,j} = 1 \end{cases} . \tag{5.1}$$

The distribution of the ancestry of a specific SNP depends on the occurrence of a recombination event. On the occurrence of a recombination between SNPs j and $j-1$ of haplotype i, the ancestry $Z_{i,j}$ is chosen with probability α to be 0 and 1 otherwise. If there was no recombination, the ancestry stays the same as that at the previous SNP:

$$\Pr(Z_{i,j} | Z_{i,j-1}, W_{i,j}, \alpha) = \begin{cases} \delta(Z_{i,j}, Z_{i,j-1}) & W_{i,j} = 0 \\ (1-\alpha)^{Z_{i,j}} \alpha^{(1-Z_{i,j})} & W_{i,j} = 1 \end{cases} .$$

where $\delta(x, y) = 1, \text{iff } x = y$.

Since we assume that the recombination process is a Poisson process, the variables $W_{i,j}$ and $W_{i,k}$ are independent for $k \neq j$ and the probability for a specific location between SNPs $j-1$ and j to have at least one recombination depends solely on θ_j. For $j > 1$, we have $\Pr(W_{i,j} = 1 | \theta_j) = \theta_j$ and $\theta_j = 1 - e^{-(g-1) l_j r_j}$, where l_j is the distance between the $(j-1)$th SNP and the jth SNP and r_j is the recombination rate in that region. In our specific problem, θ_j is a deterministic function of g. In other situations, it may be more appropriate for g to parameterize a prior over θ_j.

Marginalizing over the recombination indicator $W_{i,j}$ we obtain the mixture distribution that is used as a transition matrix by programs such as STRUCTURE [Falush *et al.*, 2003] and SABER [Tang *et al.*, 2006].

5.2.2 Modelling Background LD

The HMM framework assumes that alleles are conditionally independent given the states and thus is not able to capture within-population LD. The MHMM model implemented in SABER [Tang *et al.*, 2006] attempts to capture such background LD by allowing additional dependencies directly between the observable $\boldsymbol{X}_i$ variables. The form of these dependencies differ depending on the ancestries $Z_{i,j-1}$ and $Z_{i,j}$. In particular, if these ancestries are the same, then a pairwise emission probability is used. If these ancestries are different, then

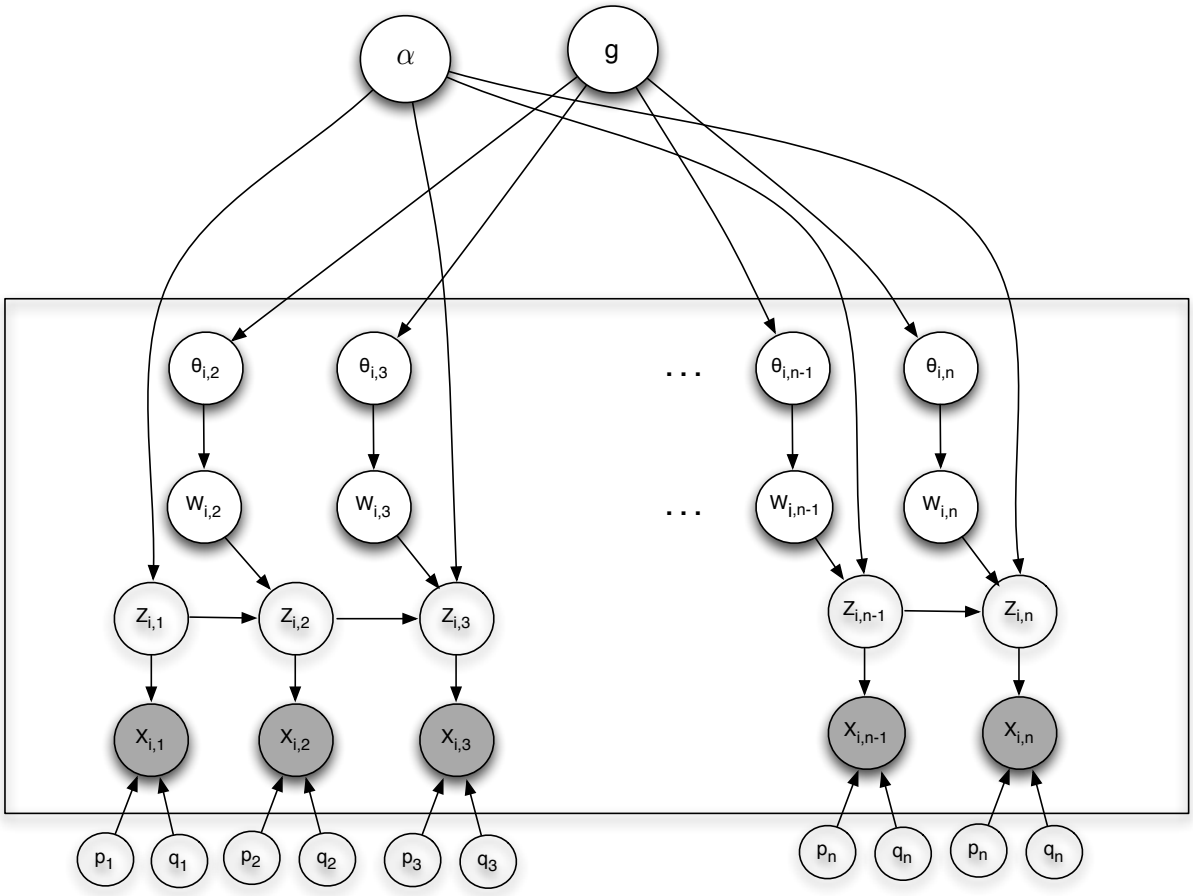

Figure 5.1: A graphical representation of the probabilistic model underlying SWITCH. The shaded circles correspond to observed random variables while the unshaded circles are unobserved random variables. The circles within the box correspond to a single individual; the circles outside the box are shared across the individuals. See Section 4.2.1 for details.

a singleton emission probability is used. SABER estimates the pairwise probabilities using ancestral haplotypes (which are assumed to be available).

Given that our model makes use of explicit recombination indicators $W_{i,j}$, we can condition on these variables instead of the ancestry variables $Z_{i,j}$. Formally, we define the following transition matrix for $j > 1$:

$$\begin{aligned} &\Pr(X_{i,j} = 1 | W_{i,j}, Z_{i,j}, X_{i,j-1}, p_j, q_j, p_{j-1,j}, q_{j-1,j}) \\ &= \begin{cases} Pr(X_{i,j} = 1 | Z_{i,j}, p_j, q_j), & \text{if} \quad W_{i,j} = 1 \\ \Pr(X_{i,j} = 1 | Z_{i,j}, X_{i,j-1}, p_{j-1,j}, q_{j-1,j}), & \text{otherwise} \end{cases} \end{aligned} \tag{5.2}$$

The transition matrix is defined so that if $W_{i,j} = 1$ (i.e., a recombination has occurred between SNPs $j-1$ and j), then the allele seen at position j is independent of the allele at position $j-1$. If $W_{i,j} = 0$, the SNPs at position $j-1$ and j belong to the same ancestral haplotype, and the emission probability of the allele at position j depends on the allele at $j-1$. Here $p_{j-1,j}$ and $q_{j-1,j}$ are the pairwise (conditional) SNP frequencies at positions $j-1$ and j in the haplotypes from the two respective populations.

Why do we condition on recombination events instead of ancestries (as in SABER)? Note that the conditioning in SABER ignores recombinations that do not change the ancestries. Such recombinations are expected to be common when the admixture fraction $\alpha << \frac{1}{2}$. In that case, assuming random mating, an expected fraction $\alpha^2 + (1-\alpha)^2$ of recombinations will not lead to a change in the ancestry. Ignoring such events can be problematic. Consider a scenario where the haplotype frequencies are estimated from an ancestral population. Assume that 00 and 11 are the only haplotypes present in this ancestral population. In the admixed population, a new haplotype, say 01, may arise due to a recombination event that is not accompanied by a change in the ancestry. By ignoring the recombination event and assuming that the two loci share a haplotype, the MHMM would assign a small probability (indeed, a zero probability in our example) to the new haplotype 01. On the other hand, in a model that conditions on the recombination indicators $W_{i,j}$, the new haplotype is assigned a frequency that is the product of the allele frequencies at the two loci.

5.2.3 Inference Problems

In this section, we focus on two inferential problems that can be framed within the HMM/MHMM formalism. In both problems, we seek the *maximum a posteriori* (MAP) estimates of a subset of the variables in the model and we find parameter estimates via the EM algorithm. For simplicity, we assume that the number of generations g is constant and known, and therefore $\boldsymbol{\theta}$ is known. This is often the case for admixed populations. The two problems that we consider are: (1) The admixture fraction is known, the allele frequencies are unknown, and the goal is to find the local ancestries for each SNP in each haplotype. The optimization problem is to find (W, Z) such that $\Pr(W, Z | X, \alpha, g)$ is maximized. We refer to this problem

as the *local ancestries inference problem*. (2) The allele frequencies are known for one of the ancestral populations, and the goal is to find the allele frequencies of the other as well as the admixture fraction. Here, the local ancestries are missing variables. The optimization problem is to find (q, α) such that $\Pr(\boldsymbol{q}, \alpha | X, \boldsymbol{p})$ is maximized. We refer to this problem as the *ancestral allele frequencies inference problem*.

5.2.3.1 Local Ancestries Problem

To compute the local ancestries, we would like to compute the MAP estimates of Z and W by solving the following optimization problem:

$$\arg\max_{Z,W} \log[\Pr(W, Z | X, \alpha, \boldsymbol{\theta})]. \tag{5.3}$$

In each iteration of EM, the updates to Z and W are computed by a Viterbi algorithm with the emission probabilities $\Pr(X_{i,j}|Z_{i,j}, p_j, q_j)$ replaced by an integral over p_j, q_j. The E-step involves computing the posterior probabilities of p_j, q_j; i.e., $\Pr(p_j, q_j | X_{\cdot,j}, Z^{(t)}_{\cdot,j})$. This can be done easily using Bayes' theorem. The M-step involves solving m separate optimization problems in $\boldsymbol{Z}_i, \boldsymbol{W}_i, i \in \{1, \ldots, m\}$ where $\boldsymbol{Z}_i$ denotes the vector of ancestries for the ith haplotype and $\boldsymbol{W}_i$ denotes the corresponding vector of recombination events:

$$\{\log[\Pr(Z_{i,1}|\alpha)] + I_{1,i}(Z_{i,1})\} + \sum_{j=2}^{n} \{I_{j,i}(Z_{i,j}) + f_{i,j-1,j}(Z_{i,j-1}, Z_{i,j}, W_{i,j})\} \tag{5.4}$$

where $f_{i,j-1,j}(Z_{i,j-1}, Z_{i,j}, W_{i,j}) \equiv \log[\Pr(Z_{i,j}|Z_{i,j-1}, W_{i,j}, \alpha)] + \log[\Pr(W_{i,j}|\theta_j)]$ corresponding to log transition probabilities and
$I_{j,i}(Z_{i,j}) \equiv \sum_{i=1}^{m} \sum_{j=1}^{n} \int \{\log[\Pr(X_{i,j}|Z_{i,j}, p_j, q_j)] \Pr(p_j, q_j | X_{\cdot,j}, Z^{(t)}_{\cdot,j}) dp_j dq_j\}$ are expectations of the log emission probabilities.

Generally, the values of $I_{j,i}(z)$ can be tabulated for each i, j, z by computing the integral over a grid on the $\{p_j, q_j\}$. For our setting, we have a uniform prior over p_j and q_j which permits the integral to be evaluated analytically as shown in Appendix 5.4.2. We can maximize (5.4) by dynamic programming. The values obtained for Z, W are then used to recompute the integrals $I_{j,i}(Z_{i,j})$ and the procedure is iterated.

5.2.3.2 Ancestral Allele Frequencies Problem

To compute the ancestral allele frequencies, we compute the MAP estimates of $\boldsymbol{q}$ and α:

$$\arg\max_{\boldsymbol{q},\alpha} \log \Pr(\boldsymbol{q}, \alpha | X, \boldsymbol{p}, \boldsymbol{\theta}) = \arg\max_{\boldsymbol{q},\alpha} \log \Pr(X | \boldsymbol{p}, \boldsymbol{q}, \alpha, \boldsymbol{\theta})$$

since we have a uniform prior on $\boldsymbol{q}$ and α. We assume g and $\boldsymbol{p}$ to be known. Let $\boldsymbol{q}^{(t)}$, $\alpha^{(t)}$ denote the current estimates of $\boldsymbol{q}, \alpha$. The EM algorithm produces new estimates $\boldsymbol{q}^{(t+1)}$, $\alpha^{(t+1)}$ that improve the objective function:

$$q_j^{(t+1)} = \frac{\sum_{i=1}^{m} X_{i,j} d_{i,j}(z)}{\sum_{i=1}^{m} d_{i,j}(1)}, \quad \alpha^{(t+1)} = \frac{\sum_{i=1}^{m} d_{i,1}(0) + \sum_{j=2}^{n} c_{i,j}(1,0)}{m + \sum_{i=1}^{m} \sum_{j=2}^{n} \sum_{z \in \{0,1\}} c_{i,j}(1,z)}$$

Here $c_{i,k}(w,z) \equiv \Pr(W_{i,k} = w, Z_{i,k} = z | X_i, \boldsymbol{q}^{(t)}, \alpha^{(t)}, \boldsymbol{p}, \boldsymbol{\theta})$ and $d_{i,j}(z) \equiv \Pr(Z_{i,j} = z | X_i, \boldsymbol{q}^{(t)}, \alpha^{(t)}, \boldsymbol{p}, \boldsymbol{\theta})$ are the posterior probabilities of (W, Z) and Z at the jth SNP of haplotype i respectively and are computed by an application of the forward-backward algorithm in the E-step.

These updates have an intuitive interpretation. At each position j, the new value of q_j is the fraction of SNPs that are 1 out of all SNPs belonging to the second population (weighted by their posterior probabilities). The update for α is the fraction of ancestries chosen from the first population whenever a new haplotype is chosen (weighted by their posterior probabilities).

5.3 Experiments

We have implemented the HMM and the EM algorithm that we have described in a program that we term "SWITCH." We have also implemented a program that we refer to as "SWITCH-MHMM" that is based on the MHMM. In this section, we describe experiments aimed at evaluating these procedures.

These experiments were run on datasets generated from HapMap data [http://www.hapmap.org]. We used SNPs found in the Affymetrix 500K GeneChip Assay® [http://www.affymetrix.com/products/arrays/specific/500k.affx] from chromosome 1 for each of the HapMap populations; i.e., Yorubans (YRI), Japanese (JPT), Han Chinese (CHB), and western Europeans (CEU). For a pair of populations, we simulated admixture by picking individuals from two ancestral populations in the ratio $\alpha : 1 - \alpha$. In each generation, individuals mate randomly and produce offspring. The rate of the recombination process is set to 10^{-8} per base pair per generation [Nachman and Crowell, 2000]. The mixing process is repeated for g generations. We generated datasets consisting of admixtures of YRI-CEU, CEU-JPT and JPT-CHB populations. We set g to 7 and α to 0.20 since these roughly correspond to the admixing process in African-American populations as estimated in [Patterson *et al.*, 2004; Falush *et al.*, 2003; Tian *et al.*, 2006]. For each of the problems, we use only genotype data. Since the HMM underlying SWITCH assumes that the SNPs are conditionally independent given the states, in the input to SWITCH we greedily remove SNPs that have a high correlation coefficient, $r^2 > 0.1$, with any other SNP. We refer to this usage of SWITCH as "uSWITCH." (When the entire set of SNPs is used, we refer to the usage simply as SWITCH). Ancestry estimates at the discarded SNPs were filled in from the

highly-correlated SNP that was retained.

The remainder of this section is organized as follows. In Section 5.3.1 we compare the performance of various methods on the local ancestries problem. The role of the inference algorithms and background LD models are discussed in Sections 5.3.2 and 5.3.3 respectively. The performance of methods on the problems of predicting recombination events and the ancestral allele frequencies problem are discussed in Sections 5.3.4 and 5.3.5 respectively.

5.3.1 Local Ancestries Problem

We first compare the estimates of the ancestries obtained from SWITCH to the estimates obtained from SABER and LAMP. In these experiments, the methods are given g and α. We consider two settings depending on whether the ancestral frequencies, $(\boldsymbol{p}, \boldsymbol{q})$, are available. Even when the frequencies of the ancestral populations are available, it is still advantageous to use the data to update the frequency estimates, which may have drifted from the ancestral frequencies.

When they are available, uSWITCH uses a maximum-likelihood classification based on these frequencies as initialization. We refer to this variation of uSWITCH as uSWITCH-ANC. SABER also requires the ancestral allele frequencies. The version of LAMP that uses ancestral frequencies is termed LAMP-ANC.

When the ancestral allele frequencies are not known, LAMP can still be used, as can uSWITCH. For the latter, we use the estimates of ancestries from LAMP to initialize the EM algorithm.

For each individual i and SNP j, each method finds an estimate $\hat{a}_{ij}^p \in \{0, 0.5, 1\}$ for the true ancestry a_{ij}^p. We measure the accuracy of a method as the fraction of triplets (i, j, p) for which $a_{ij}^p = \hat{a}_{ij}^p$. The first half of Table 5.1 compares the accuracies of SABER, LAMP-ANC and uSWITCH-ANC on 100 random datasets of YRI-CEU, CEU-JPT and JPT-CHB. uSWITCH-ANC improves significantly over LAMP-ANC and SABER on the YRI-CEU dataset. On the CEU-JPT, uSWITCH-ANC and LAMP-ANC have comparable performance, and both methods are more accurate than SABER. All methods perform poorly on the JPT-CHB dataset due to the closeness of the two populations. The second half of Table 5.1 compares the accuracies of uSWITCH and LAMP. On the YRI-CEU data, uSWITCH, with an accuracy of 96.0%, improves significantly over LAMP, which has an accuracy of 94.0% (Wilcoxon paired signed rank test p-value of 3.89×10^{-18}). Interestingly, uSWITCH improves significantly over LAMP-ANC even though the latter uses the ancestral allele frequencies. On the CEU-JPT and the JPT-CHB datasets, uSWITCH seems to have slightly higher accuracies than LAMP. We believe that using more informative priors on the variables $\boldsymbol{p}, \boldsymbol{q}$, should yield further improvements by improving the estimation of low-frequency alleles. These results indicate that the HMM is most useful when the mixing populations can be easily distinguished as is the case with the YRI-CEU admixture.

Although the versions of uSWITCH have a factor of $5 - 10$ increase in running time

compared to LAMP, they still run under an hour on large datasets making them feasible for genome-scale problems.

5.3.2 Role of the Inference algorithm

To understand the impact of the inference algorithm and the initialization, we compared uSWITCH to STRUCTURE. While the model used in uSWITCH is the same as the model used in STRUCTURE when the recombination indicators W are integrated out, the inference algorithms differ. uSWITCH obtains the posterior mode of the ancestries Z using an EM algorithm with LAMP providing the initialization. STRUCTURE computes the posterior marginals of each $Z_{i,j}$ using an MCMC algorithm to integrate out the unknown parameters. To evaluate the output from STRUCTURE, we threshold the posterior mean to obtain the actual ancestry estimates; that is, position i, j is assigned $0, 1$ or 2 alleles from one of the populations depending on whether the posterior marginal $E(Z_{i,j}|X)$ lies in $[0, 0.5), [0.5, 1.5)$ or $(1.5, 2)$. We compared the ancestry estimates produced by the two methods on the YRI-CEU dataset. STRUCTURE was run for 10000 burn-in and 50000 MCMC iterations (see below for further discussion of this choice). The linkage model was used. STRUCTURE was run on non-overlapping sets of 4000 SNPs covering 36000 of the 38000 initial SNPs due to numerical instabilities when larger number of SNPs were used.

On the YRI-CEU dataset, uSWITCH achieved an accuracy of 97% while STRUCTURE achieved an accuracy of 84%. To isolate the reason for this difference, we evaluated MCMC algorithms which differ from STRUCTURE in varying degrees. First, we ran MCMC from a random starting point for 1000 iterations with 100 iterations of burn-in and used the posterior mean as the ancestry estimates. This yielded estimates with an accuracy of 91.13%. When the LAMP estimates were used as a starting point, the accuracy was 94.9%. This suggests that the chain has not mixed in our STRUCTURE runs. To test this suggestion formally, we simulated five such chains each from different random starting points. We then computed a multivariate potential scale reduction factor (PSRF) [Brooks and Gelman, 1998] for random sets of 100 p's and q's and found it to be consistently large (> 1.2). When the Markov chain is unable to converge quickly, the initialization influences the ancestry estimates. Given that the MCMC algorithms do not converge even after being run for several days (in particular, the STRUCTURE runs required a little less than three days while the other MCMC runs took about a day to run), good initialization becomes essential.

Two other differences between STRUCTURE and the MCMC algorithm that we implemented are that the latter discards correlated SNPs and fixes the hyperparameters. We modified the MCMC runs to retain the correlated SNPs and the accuracy falls to 74.9%. We conclude that the pruning of highly correlated SNPs can have a large impact on the accuracy of models that do not attempt to account for background LD. Another approach to this problem is to attempt to account for background LD via the MHMM approach; we discuss this approach in the following section.

5.3.3 Modelling background LD

As discussed earlier, we refer to our implementation of an MHMM model based on the recombination indicators $W_{i,j}$ as SWITCH-MHMM. We also implemented a version of the model based on the ancestries $Z_{i,j}$ instead of the recombination indicators. We refer to this model as "MHMM"; it is the same as the model underlying SABER. (Our implementation differs from SABER in the inference procedures that we used; in particular, the ancestry estimates were computed by a Viterbi algorithm.)

In the first scenario that we studied, both the MHMM and the SWITCH-MHMM were given the ancestral haplotypes. The ancestral haplotypes were used to estimate the pairwise SNP emission probabilities. The single SNP frequencies were estimated using LAMP-ANC. In this experiment, SWITCH-MHMM achieved an accuracy of 91.9%, while the MHMM yielded an accuracy of 88.9%. This demonstrates that improvements can be obtained by conditioning on recombination indicators instead of conditioning on ancestral states.

In a second scenario, the pairwise SNP emission probabilities were estimated directly from the admixed data. In this case, the accuracies of SWITCH-MHMM and MHMM were both 95.7%. It is interesting to note that these accuracies are higher than in the case that ancestral haplotypes were used to estimate parameters. This is presumably due to the fact that the estimates of haplotype frequencies are more accurate when estimated from the admixed population itself. Finally, we also measured the accuracy of ancestry estimates from SWITCH (i.e., when the entire set of SNPs was taken as input) and observed that the accuracy drops to 93.1%. This improvement in accuracies when background LD is taken into account has been observed before [Tang *et al.*, 2006]. However, the accuracy of uSWITCH is higher than SWITCH-HMM. Thus, the heuristic of removing highly correlated SNPs and then running SWITCH appears to be competitive, in practice, to the methods based on explicit (but simplified) models of background LD.

5.3.4 Predicting Recombinations

Another advantage of the use of the recombination indicators W is that they open the possibility of inference of historic recombinations created by the mixing process after the initial admixture event. While a change in the ancestry between two SNPs implies a recombination event, many recombination events do not result in a change in the ancestry. When α is small, this happens quite often. To study this issue, we measured the accuracy of uSWITCH in predicting such recombinations. If a predicted recombination falls within 5 Kbases of the SNPs flanking a true recombination, it is called a true positive. If multiple recombinations are predicted within this window, only one is counted as a true positive. False positives and false negatives are defined similarly. The *precision* and *recall* of the predictions are then computed as $Precision = \frac{TP}{TP+FP}$ and $Recall = \frac{TP}{TP+FN}$. We combine these numbers by taking a harmonic mean, reporting $F - score = \frac{2Precision \times Recall}{Precision+Recall}$.

As a baseline, we use a null model that predicts recombinations based on the exponentially-distributed lengths of the haplotypes. The total number of recombinations in the null model is set to the number of predicted recombinations and the precision and recall of the predictions are computed similarly.

On the YRI-CEU dataset, uSWITCH attains an $F - score$ of 70.8 while the null model attained an F-score of 52.8. uSWITCH was found to be consistently more accurate than the null model on the CEU-JPT and JPT-CHB datasets as well (data not shown).

We now consider models that attempt to account for background LD. For the MHMM model, since the model does not explicitly represent recombinations, the recombinations are inferred (naively) based on a change in the ancestry labels. The results are shown in Table 5.2. When we use the ancestral haplotypes to estimate parameters, the MHMM and SWITCH-MHMM achieve $F - scores$ of 35.0 and 41.5 respectively. Using the admixed data to estimate parameters, the two models achieve $F - scores$ of 78.0 and 79.3 respectively. We see that the explicit W variables allow more accurate prediction of recombinations in the admixed genomes. When we restrict attention to breakpoints (recombinations that change the ancestry), the difference between the models is diminished though the relative performance is the same.

As discussed in the previous section, SWITCH-MHMM (and the other models that incorporate background LD) has lower accuracy than uSWITCH which ignores background LD and uses a heuristic to prune correlated SNPs. However, SWITCH-MHMM predicts recombinations more accurately (while uSWITCH is more accurate in predicting breakpoints). This result suggests that models that incorporate background LD (albeit imperfectly) may be useful in inferring recombinations in admixed genomes.

5.3.5 Ancestral Allele Frequencies Problem

We now turn to the problem of inferring ancestral allele frequencies. To obtain a benchmark, we implemented a naive algorithm. The naive algorithm is given the true value of α (which is *not* available to the model). The idea behind the naive algorithm is as follows. For a position j with minor allele frequency f_j, and allele frequencies p_j and q_j in the two populations, if the number of individuals is large, f_j can be written as $f_j = (1-\alpha)p_j + \alpha q_j$. So we compute the allele frequency q_j at position j as $q_j = \max\left(\min(\frac{f_j-(1-\alpha)p_j}{\alpha}, 1), 0\right)$. We used two different estimates of α, yielding algorithms that we refer to as "Naive1" and "Naive2." Naive1 uses the value of $\alpha = 0.20$ which is the admixture fraction in the first generation of admixture. Naive2 uses an α measured from each dataset.

We calculated the L1 error (the sum of the absolute values of the errors) between the estimated $\hat{\boldsymbol{q}}$ and the true $\boldsymbol{q}$. The L1 error averaged over 100 datasets of YRI-CEU, CEU-JPT and JPT-CHB is shown in Table 5.3. We see that uSWITCH reduces the L1 error by about 30% in the YRI-CEU and the CEU-JPT datasets while there is no significant difference for the JPT-CHB dataset.

We also compared the ancestry estimates from uSWITCH with those from STRUCTURE on single instances of YRI-CEU, CEU-JPT and JPT-CHB datasets (the running time of STRUCTURE prohibited multiple runs). The L1 errors for uSWITCH are $7.1\%, 8.3\%$, and 12.7% on the respective datasets. STRUCTURE obtains errors of $25.8\%, 29.0\%$, and 25.2% respectively.

5.4 Discussion

Markovian models such as HMMs and MHMMs are a natural approach to admixture that aim to strike a balance between predictive performance and inferential complexity. We have explored several variations on the HMM/MHMM theme with the aim of identifying combinations of model specification, inference procedure and data preprocessing that are most effective in realizing this balance.

We have found that explicit indicators of recombination events can be useful. These indicators allow us to provide a more fine-grained version of the MHMM that allows new haplotypes to emerge when recombinations occur, and not only when ancestral state changes. We found that this approach yielded better estimates when haplotype emission probabilities are inferred from ancestral populations. Also, by making the recombination events explicit in our model, we are able to infer historic recombinations. While being interesting in and of themselves, these predictions may be helpful in allowing admixture data to be used in the inference of recombination hotspots.

HMM and MHMM models require the estimation of model hyperparameters. One approach to estimating these hyperparameters is to use MCMC algorithms, but these algorithms can be impractical on realistic datasets. We have shown that an EM-based approach starting with an accurate initialization (the non-model-based procedure LAMP) yielded high accuracy at reasonable cost. Indeed, this approach yielded the best results of any algorithm that we studied.

Our conclusions regarding background LD are mixed. If an MHMM model is to be used to attempt to capture background LD, then we recommend conditioning on explicit recombination indicators. On the other hand, we found that a heuristic approach, in which highly-correlated SNPs are discarded before running an HMM, yielded higher accuracy than the MHMM. One possible direction for future research is to consider richer MHMM models than the pairwise model considered here and in SABER.

Method	YRI-CEU	CEU-JPT	JPT-CHB
uSWITCH-ANC	97.6±0.3	94.5±0.8	66.4±2.7
LAMP-ANC	94.9±0.6	93.7±0.7	69.9±2.1
SABER	89.4±0.8	85.2±1.2	68.2±1.9
uSWITCH	96.0± 0.6	83.2±5.6	51.4±2.8
LAMP	94.0±0.8	82.9±5.5	50.6±2.5

Table 5.1: Accuracies of ancestry estimates averaged over 100 datasets. The methods are compared under two settings. When the ancestral allele frequencies are known, the methods compared are LAMP-ANC, uSWITCH-ANC, and SABER. When the ancestral allele frequencies are not known, the methods compared are uSWITCH and LAMP.

Model	Recombinations		Breakpoints	
	F-score	Precision/Recall	F-score	Precision/Recall
MHMM (anc)	35.0	21.5/95.1	12.2	6.5/99.9
SWITCH-MHMM(anc)	41.5	26.5/95.2	23.4	13.3/98.3
MHMM	78.0	87.0/70.0	49.5	33.5/94.8
SWITCH-MHMM	79.3	85.0/74.3	49.8	33.8/94.8
uSWITCH	74.5	88.7/64.2	53.7	33.8/92.5

Table 5.2: Accuracies of the different models on the prediction of recombinations and breakpoints. (anc) denotes the ancestral haplotypes were used to estimate parameters.

Appendix

5.4.1 Model for Genotype Data

It is straightforward to extend the model to handle genotype data. Since the SNPs are assumed to be independent, we can model the SNP at each position as a random variable that depends on the alleles in the corresponding haplotypes. We introduce random variables $Y_{i,j} \in \{0,1,2\}, i \in \{1,\ldots,\frac{m}{2}\}$ (assuming that m is even) representing the j- th SNP of the i-th genotype. The value of this SNP depends on the values of the j-th alleles in haplotypes $2i-1$ and $2i$:

$$\Pr(Y_{i,j}|X_{2i-1,j}, X_{2i,j}) = \delta(Y_{i,j} = X_{2i-1,j} + X_{2i,j}).$$

We now replace all X variables in previous equations with Y, and instead of Equation (5.1) we use $\Pr(Y_{i,j} = N|Z_{2i-1,j}, Z_{2i,j}, p_j, q_j)$, which can be calculated for each $N \in \{0,1,2\}$.

5.4.2 Analytical Computation of $I_{j,i}$

In this section, we show how the integrals $I_{j,i}(Z_{i,j})$ introduced in Section 5.2.3.1 can be analytically evaluated. Recall the definition of $I_{j,i}$:

$$I_{j,i}(Z_{i,j}) = \int \Big\{ \log[\Pr(X_{i,j}|Z_{i,j}, p_j, q_j)] \Pr(p_j, q_j|X_{\cdot,j}, Z^{(t)}_{\cdot,j}) dp_j dq_j \Big\}. \tag{5.5}$$

We define the following quantities:

$$\begin{aligned} \pi^{(t)}_{j,1} &= \sum_{i=1}^{m} X_{i,j} Z^{(t)}_{i,j} \qquad \pi^{(t)}_{j,0} = \sum_{i=1}^{m} (1 - X_{i,j}) Z^{(t)}_{i,j} \\ \xi^{(t)}_{j,1} &= \sum_{i=1}^{m} X_{i,j} (1 - Z^{(t)}_{i,j}) \qquad \xi^{(t)}_{j,0} = \sum_{i=1}^{m} (1 - X_{i,j})(1 - Z^{(t)}_{i,j}). \end{aligned} \tag{5.6}$$

The log likelihood in Equation (5.5) can be written as

$$\begin{aligned} &\Pr(X_{i,j}|Z_{i,j}, p_j, q_j) = (q_j^{X_{i,j}} (1 - q_j)^{1 - X_{i,j}})^{Z_{i,j}} \\ &\cdot (p_j^{X_{i,j}} (1 - p_j)^{1 - X_{i,j}})^{1 - Z_{i,j}}. \end{aligned}$$

Using the above expression, we can now write the posterior:

$$\begin{aligned} \Pr(p_j, q_j|X_{\cdot,j}, Z^{(t)}_{\cdot,j}) &\propto \Pr(X_{\cdot,j}|p_j, q_j, Z^{(t)}_{\cdot,j}) \Pr(p_j) \Pr(q_j) \\ &\propto \prod_{i=1}^{m} \Pr(X_{i,j}|p_j, q_j, Z^{(t)}_{i,j}) \Pr(p_j) \Pr(q_j) \\ &\propto p^{\pi^{(t)}_{j,1}} (1 - p)^{\pi^{(t)}_{j,0}} q^{\xi^{(t)}_{j,1}} (1 - q)^{\xi^{(t)}_{j,0}} \\ &= \frac{p^{\pi^{(t)}_{j,1}} (1 - p)^{\pi^{(t)}_{j,0}} q^{\xi^{(t)}_{j,1}} (1 - q)^{\xi^{(t)}_{j,0}}}{B(\pi^{(t)}_{j,1}, \pi^{(t)}_{j,0}) B(\xi^{(t)}_{j,1}, \xi^{(t)}_{j,0})}. \end{aligned}$$

Method	YRI-CEU	CEU-JPT	JPT-CHB
uSWITCH	7.7±0.5	8.5±0.6	11.7±1.3
Naive1	11.8±0.5	12.2±0.5	12.5±0.5
Naive2	11.8±1.2	12.3±1.2	12.6±1.2

Table 5.3: Average L1 error in the estimates of $\boldsymbol{q}$. The methods compared are uSWITCH (which estimates q and α jointly) and two naive algorithms that are given the true $\alpha = 0.20$ and α estimated from the data respectively.

Here $B(a, b)$ denotes the beta function $\int_0^1 x^a(1-x)^b dx$.

Substituting the above expression into Equation (5.5) we obtain:

$$\begin{aligned} I_{j,i}(Z_{i,j}) &= X_i Z_i J(\pi_{j,1}^{(t)}, \pi_{j,0}^{(t)}) \\ &+ (1 - X_i) Z_i J(\pi_{j,0}^{(t)}, \pi_{j,1}^{(t)}) \\ &+ X_i (1 - Z_i) J(\xi_{j,1}^{(t)}, \xi_{j,0}^{(t)}) \\ &+ (1 - X_i)(1 - Z_i) J(\xi_{j,0}^{(t)}, \xi_{j,1}^{(t)}), \end{aligned}$$

where $J(a, b) = \int_0^1 \log x x^a (1-x)^b dx$.

Notice that in our setting a and b are non-negative integers. So we can compute $J(a, b)$ by performing a Binomial expansion on $(1-x)^b$ and integrating each term:

$$\begin{aligned} J(a,b) &= \int_0^1 \log x x^a \left\{ \sum_{r=0}^{b} \binom{b}{r} (-1)^r x^r \right\} dx \\ &= \sum_{r=0}^{b} \binom{b}{r} (-1)^r \int_0^1 dx \log x x^{a+r} \\ &= \sum_{r=0}^{b} \binom{b}{r} (-1)^{r+1} \frac{1}{(a+r+1)^2}. \end{aligned}$$

Chapter 6

Genomic privacy

6.1 Introduction

One of the major challenges in genome-wide association studies is that of achieving desired levels of statistical power for detecting weak associations while maintaining control on false positive rates. Power can be enhanced by combining data across studies in meta-analysis or replication studies. Such methods require data to flow freely in the scientific community, however, and this raises privacy concerns. Until recently, many studies have pooled individuals together, making the allele frequencies of each SNP in the pool publicly available. It has been implicitly assumed that releasing such summary data provides a secure way to share a study's results without compromising privacy. It therefore came as a major surprise when Homer et al. [Homer *et al.*, 2008] recently showed that high-density SNP arrays can be used to accurately identify the presence of individual genotypes in a mixture of DNA even when their DNA is present in small concentrations. Although aimed at applications in forensics, their findings raised the possibility that the presence of individual genotypes can be inferred from summary data, and this has led to the removal of formerly publicly available summary data from previous studies as a conservative means of protecting the privacy of human subjects [Gilbert, 2008].

However, for many applications (see, e.g., [Barrett *et al.*, 2008; Zeggini *et al.*, 2008; Cooper *et al.*, 2008]), it is sufficient to have access to the summary data for only a subset of the SNPs (*exposed* SNPs), and it thus seems desirable to investigate whether some appropriately defined level of privacy can still be maintained if the number of exposed SNPs is sufficiently small. Establishing guarantees of this kind requires understanding how this number varies as a function of factors such as the allele frequencies of the SNPs, the number of individuals in the pool, and, of particular importance, the method used to detect the individual in the pool. Indeed, an analysis of this kind was pursued by [Homer *et al.*, 2008], who proposed a particular detection method and estimated the statistical power of detecting an individual in a sample of exposed SNPs using that method. But while an analysis of any

specific method provides an estimate of power, it does not rule out the possibility that some other method yields a larger power, and is, thus, unable to provide any guarantee that the power of detection is below some acceptable level. What is needed is an upper bound on the power achievable *by any method*.

In this chapter, we address this issue, providing guidelines as to which set of SNPs can be safely exposed for a given pool size, maximal allowable power β and false positive level α. Our approach is based on casting the problem as a statistical hypothesis testing problem for which the *likelihood ratio test (LR-test)* attains the maximal power achievable ([Lehmann, 2005]). This provides a guarantee that it is safe to expose a set of SNPs for which the LR-test does not achieve sufficient power. Moreover, our empirical results show that the LR-test is more powerful than the method suggested by [Homer *et al.*, 2008], especially when α is small. Finally, our theoretical and empirical results lead to a conclusion that is qualitatively different than that of [Homer *et al.*, 2008] in that we find that the power achieved by considering whole-genome datasets is in fact limited.

6.2 Methodology

6.2.1 Model assumptions

In association studies, individuals in the pool are assumed to be chosen randomly from a pure population. For a pool of n individuals we expose m SNPs, for which the allele frequencies in the population and the pool are $p_1, \ldots, p_m$, and $\hat{p}_1, \ldots, \hat{p}_m$, respectively. In the models we consider, we assume that the SNPs are *independent*. This is motivated by the fact that, in practice, the SNPs that we choose to expose can be selected sufficiently far apart on the chromosome that they can be considered independent; moreover, this assumption makes our theoretical analysis tractable. We also assume that the SNP-allele frequencies are bounded away from zero and one; i.e., there exists $a > 0$ such that $a \leq p_j \leq 1 - a, j \in \{1, \ldots, m\}$. This is a natural assumption because it is usually the case that only those SNPs whose minor alleles are sufficiently well represented in the population are considered in association studies; moreover, the exposed SNPs can be explicitly selected to have a prespecified minimal minor allele frequency.

6.2.2 Hypotheses

To construct a likelihood ratio test, we must first specify the models corresponding to the null and the alternative hypotheses respectively. Since the SNPs are assumed independent, we describe the model for a single SNP.

Null hypothesis: We assume that the pool is constituted of n individuals drawn independently from a reference population, in which the SNP-allele frequency is p (two alleles are drawn independently for each individual). We assume that the pool frequency for that SNP

is obtained by averaging the binary values of the alleles of all individuals, so that $2n\hat{p}$ is a binomial random variable, $\text{Bin}(2n, p)$. The two alleles of the individual of interest, i.e., the individual whose genotype is being tested for presence in the pool, are drawn independently from a Bernoulli variable with parameter p, since, under the null, that individual is drawn independently of the pool from the same reference population.

Alternative hypothesis: We assume that the pool is constituted of the individual of interest whose alleles are drawn from a Bernoulli variable with parameter p, which is merged with a pool of $n-1$ individuals obtained as under the null. Thus $\hat{p}$ is the average of $2n-2$ alleles of the $n-1$ individuals in the pool and the two alleles of the individual of interest. For moderately large n the model can be approximated by a simpler model which consists in sampling a pool of size n, computing the allele frequency in the pool, and drawing the two alleles of the individual of interest as Bernoulli with parameter $\hat{p}$.

6.2.3 The LR-test

For an individual with genotype $(x_1, \ldots, x_m) \in \{0, 1, 2\}^m$, the LR-test is based on the log likelihood ratio statistic:

$$\bar{L} = \sum_{j=1}^{m} \sum_{k=0}^{2} \mathbf{1}_{\{x_j = k\}} \log \frac{\hat{\pi}_j^k}{\pi_j^k}, \tag{6.1}$$

where $\mathbf{1}_{\{x_j=k\}}$ is 1 if $x_j = k$ and 0 otherwise, and π_j^k and $\hat{\pi}_j^k$ are the genotype frequencies in the population and in the pool, derived from p_j and $\hat{p}_j$, respectively, under an assumption of Hardy-Weinberg equilibrium.

We note that the LR-test is an abstract test that cannot be constructed exactly in practice since it requires knowledge of the population allele frequencies p_j. In practice, these frequencies can only be estimated from an independent reference dataset drawn from the same population. We therefore differentiate the *exact LR-test* from the *approximate LR-test*, in which an estimate of the allele frequencies is substituted for p_j. An important property of the exact test is that the Neyman-Pearson lemma [Lehmann, 2005] guarantees that the power of any test, whether based on known population frequencies or not, cannot be better than that of the exact LR-test. By analytically characterizing the power of the exact LR-test, for large pools and common SNPs, we can bound the power β of any test as a function of m, n, and α.

6.2.4 Detection in a single pool vs. discrimination between pools

Our experiments demonstrate that there is a discrepancy between the power achieved by the LR-test and the power achieved by the test described in [Homer *et al.*, 2008]. This stems from the fact that the two experiments were based on different hypotheses. In [Homer *et al.*, 2008], the pool was assumed to be generated by random sampling of n individuals from the

distribution defined by $p_1, \ldots, p_m$. The alternative hypothesis in our study and in [Homer *et al.*, 2008] is that the pool contains the tested individual. Where the two studies differ is in the definition of the null hypothesis: in our case, under the null hypothesis the tested individual is randomly picked from the general population, while in [Homer *et al.*, 2008], the individual is assumed to be randomly sampled from the finite reference dataset. This seemingly subtle difference between the two null hypotheses leads to quite different results because the finite reference dataset is small (currently < 5000).

Consider, for example, an extreme case in which the population consists of 10 individuals of which 5 are in the pool and the rest in the reference dataset; it is easy to detect any particular individual in this case. Detection is harder when the population consists of 1 million individuals of which 5 are in the pool. It becomes even harder if, out of these 1 million individuals, only a random set of 5 are available in the reference dataset, which corresponds to the situation occurring in practice.

In practice, we show that the power attained using the null hypothesis of [Homer *et al.*, 2008] more than doubles at a false positive rate of 10^{-6}

Figures 6.4 and 6.5 compare the two settings. When all the independent common SNPs in the WTCCC are used, for a false positive level of 10^{-6} the approximate LR test attains a power of 1 and 0.88 under the alternate setting of [Homer *et al.*, 2008] and the original setting respectively. The ROC curves can be extrapolated to lower false positive levels using Equation 6.4 corrected for a finite reference dataset. We then see that for a false positive level of 10^{-6}, the power of the approximate LR test is 0.96 and 0.31 under the two settings. Figure 6.5 shows the same trend when only 10000 independent common SNPs are used— for discrimination between two pools, the power more than doubles at a false positive level of 10^{-6}. We have also theoretically analyzed the power in this alternate setting (see Supplementary Note accompanying [Sankararaman *et al.*, 2009]) and we can show that when the size of the reference dataset is the same as the size of the pool, the number of SNPs needed drops by a factor of four in this setting.

6.2.5 Summary of the Analysis.

For clarity, we will give an overview of the analysis for a haploid individual; the case of genotypes is slightly more technical: we refer the interested reader to the supplementary note accompanying [Sankararaman *et al.*, 2009]. For haploids, the LR-test is

$$\bar{L} = \sum_{j=1}^{m} \left[x_j \log \frac{\hat{p}_j}{p_j} + (1 - x_j) \log \frac{1 - \hat{p}_j}{1 - p_j} \right]. \tag{6.2}$$

The Neyman-Pearson lemma guarantees that no test can have larger power than the likelihood ratio test. Thus, characterizing the power of the LR test, as a function of the pool size n, the number of independent SNPs m and a tolerable false positive rate α, determines

the largest power β achievable for any test given (m, n, α); conversely, it also determines the maximal value m so that no (α, β)-test can be obtained for a pool of size n.

The exact LR-test cannot be constructed in practice since it requires knowledge of the true SNP-allele frequencies. In practice, the test performed will be the approximate LR-test, where the allele frequencies in the population are estimated from a reference dataset. Nonetheless, we analyze the exact LR-test because it provides an upper bound on the power of any test, whether it uses the true frequencies or not.

If n is larger than 100 and the minor allele frequency is greater than 0.05, the exact LR statistic can be shown to be very well approximated under both hypotheses by the simpler statistic

$$\sum_{j=1}^{m} \left[\frac{1}{\sqrt{n}} \frac{x_j - p_j}{\sqrt{p_j(1-p_j)}} Z_j - \frac{1}{2n} \frac{(x_j - p_j)^2}{p_j(1-p_j)} Z_j^2 \right], \tag{6.3}$$

where Z_j are standard Gaussian variables. The statistic in Equation 6.3 can be analyzed easily: provided n is moderately large, each term in the sum has mean $\mu_0 = -\frac{1}{2n}$ under the null and $\mu_1 = +\frac{1}{2n}$ under the alternative and variance $\sigma_0^2 = \sigma_1^2 = \frac{1}{n}$ in both cases. But, for m moderately large and MAF not too small (MAF > 0.05), the distribution of the exact LR statistic is itself approximately Gaussian and, for a Gaussian test, the relationship between sample size m, power β, and false positive rate α is $m\mu_0 + z_\alpha \sigma_0 \sqrt{m} = m\mu_1 - z_{1-\beta}\sigma_1\sqrt{m}$. In our case this yields the fundamental relation given in Equation 6.4. Note that this result is independent of the allele frequencies provided MAF > 0.05.

As a consequence, for pools of size greater than 100, if $m \leq (z_\alpha + z_{1-\beta})^2 n$, any test of level α is guaranteed to have power no larger than β. For small pools, μ_0, μ_1, σ_0^2 and σ_1^2 can be computed algorithmically and the power can still be computed exactly, even though no simple analytical expression is available (the power would depend in this case on all SNP frequencies).

A virtue of having reduced the analysis of the LR-test to the analysis of Equation (6.3) is that it can be used to obtain insight into the behavior of the LR-test under various interesting alternative scenarios (More details on these scenarios can be found in the Supplementary Note accompanying [Sankararaman *et al.*, 2009]):

- In general, the frequencies p_j need be estimated. The power drops due to the estimation procedure and we can characterize the drop in power for the approximate LR-test. In particular, if the reference dataset used to estimate p_j has the same size as the pool, the number of SNPs needed to reach the same power is doubled.

- The LR-test for the detection of an individual in a pool is very similar to the test for discrimination between two pools, though the test for discrimination between pools has higher power. We analyze this case with the same tools and show that if both pools have the same size, the necessary number of SNPs is halved. Combined with the

drop in power due to estimating p_j mentioned above, detection of an individual in a pool needs four times as many SNPs as discriminating between two pools.

- Genotyping errors only decrease the power of the optimal test. This is intuitively clear since genotyping errors would be expected to make it harder to match an individual's genotype to that in the pool. We can show analytically that this is true under a very broad assumption that the errors are generated by the same mechanism under the null and the alternative hypothesis.

- Detecting a relative of the individual of interest in the pool is also done at the expense of a drop in power. We show analytically that for a given power, detecting a sibling instead of the individual of interest requires four times as many SNPs.

6.3 Experiments

6.3.1 Experimental setup

We compared the power of the approximate LR-test and the power of the statistic used in [Homer *et al.*, 2008] empirically. We also compared the empirical results to the theoretical prediction, using a variant of Equation (6.4) with a correction for the finite size of the reference dataset (see the supplementary information accompanying [Sankararaman *et al.*, 2009] for the derivation of the correction). The only change arising from the correction is that the factor $\frac{1}{n}$ in Equation (6.4) is replaced by $\frac{1}{n}(1 - \frac{n}{\tilde{n}})$ where $\tilde{n}$ is the sum of the number of individuals in the pool and the reference dataset. We use this corrected version of Equation (6.4) whenever we compare the approximate LR to our theoretical calculations.

To evaluate the power of the approximate LR-test empirically, we created pools containing $n = 1000$ genotypes. Allele frequencies were computed for the pool. Individuals not part of the pool were used as a reference dataset to estimate the population allele frequencies. Under the null hypothesis (where the individual is not present in the pool), we pick an individual from the pool, remove the contribution of the individual to the allele frequencies for the pool and then compute the statistic for this individual; note that under the null, the individual is neither present in the pool nor in the sample of individuals outside the pool. Under the alternative hypothesis, we simply pick an individual from the pool and compute the statistic for this individual.

6.3.2 Experiments on simulated data

The ROC curves display the power of the approximate LR test, the power of the statistic used in [Homer *et al.*, 2008] and the theoretical power for the approximate LR-test. We first computed the ROC curves on simulated data; we simulated independent SNPs for values of

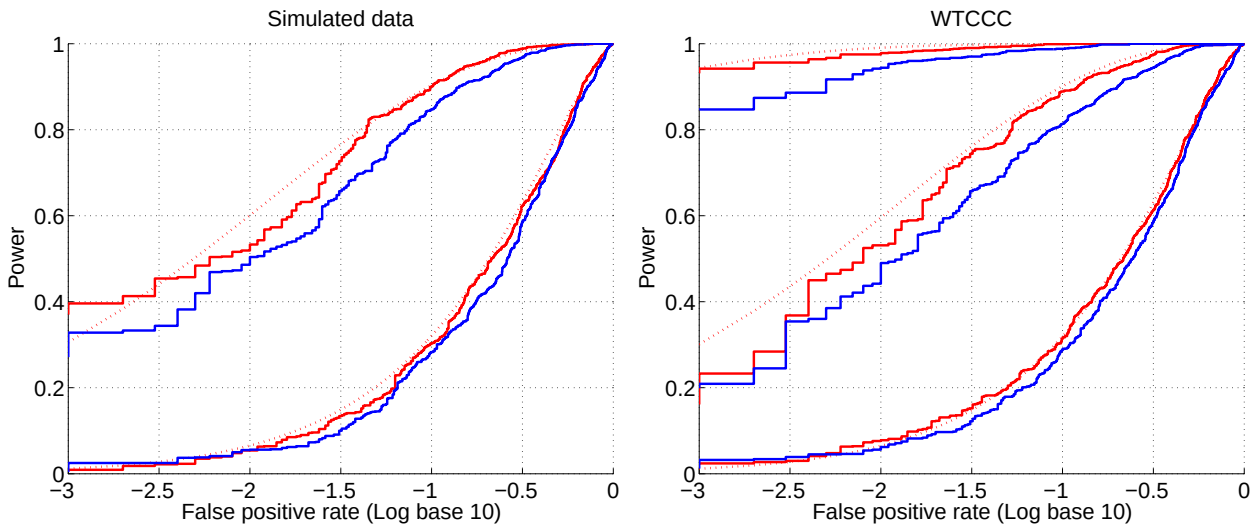

Figure 6.1: ROC curves comparing the LR-test with a plug-in allele frequency estimate, its theoretical power (denoted "LR theory"), computed using a modified version of Equation (6.4) corrected to account for the use of the plug-in estimate, and the statistic proposed by Homer et al. on a pool of size $n = 1000$. (Left) ROC curves for simulated data with $m = 1000, 10000$ exposed SNPs. (Right) ROC curves on the WTCCC data with $m = 1000, 10000$, and 33138 SNPs (the total set of independent SNPs). The LR-test performs significantly better than the test of Homer et al. Nonetheless, the power stays below 0.95 for a false positive level of 10^{-3} even when all the independent SNPs are used. Note the close agreement between the empirical and the theoretical results.

$n = 1000$ and $m = 1000, 10000$. The allele frequencies were picked independently from a Beta distribution fitted to allele frequencies in the range $[0.05, 0.95]$ found in the HapMap CEU population. The reference dataset consisted of 2000 individuals drawn from the same allele frequency distribution. The results (Figure 6.1) show the close agreement between the theoretical and empirical curves for the LR-test. Further, the LR-test is consistently more powerful than the statistic proposed by [Homer *et al.*, 2008], particularly at low false positive levels. To test the statistical significance of this difference, we performed a Wilcoxon signed rank test on the AUC (area under the ROC curve) of 100 bootstrap replicates. We found that the AUC of the LR-test was significantly greater than the statistic in [Homer *et al.*, 2008] for $m = 1000$ and $m = 10000$ (in both cases, the p-value was 3.9×10^{-18}). Importantly, for a group of size 1000, the power is low even with 10000 independent SNPs—less than 0.50 at a 10^{-3} false positive level.

6.3.3 Experiments on the WTCCC data

We constructed pools of size 1000 from the WTCCC control dataset consisting of 2937 individuals. There were 3004 individuals from the 58C and the UKBS control groups. We retained 2937 individuals after removing individuals with more than 3% missing data, related individuals and individuals with non-European ancestry.

We set $a = 0.05$ and retained only the set of independent SNPs (we used a p-value on r^2 of 10^{-5}). This gave us a set of 33,138 autosomal SNPs from the original set of 462,386 SNPs. It is striking to see that the power stays below 0.95 for a false positive level of 10^{-3} even when all the independent SNPs are used (Figure 6.1). Computing the power at lower p-values using Equation (6.4), we see that the power to detect an individual at a false positive level of 10^{-6} is only about 0.47 even when all the independent SNPs are used. The latter contradicts the results of [Homer *et al.*, 2008], who find that the power to detect individuals is high even with a false positive level of 10^{-6}. This discrepancy can be attributed to different formulations of the hypothesis testing problem. In particular, [Homer *et al.*, 2008] test whether an individual is present in the pool or alternatively in the reference dataset, while we test whether an individual is present in the pool or alternatively in the larger underlying population. Although the null hypothesis of [Homer *et al.*, 2008] may be of interest in forensics applications, we argue that our formulation is more relevant to the discussion of privacy issues.

If the entire set of 358,053 SNPs with MAF above 0.05 is used (this set includes the set of 33,138 independent SNPs), Figure 6.2 shows a small reduction in the power when the same approximate LR test—which assumes independence—is used. However, when the SNPs are no longer independent, there is a potential risk that linkage disequilibrium could be exploited to design a more powerful test.

6.3.4 Genotyping errors

Thus far we have assumed that there are no genotyping errors for either the individual or the pool. In practice, genotyping errors occur in 0.1%-1% of the SNPs so that even when the tested individual is actually present in the pool, the tested genotype might differ from the genotype present in the pool. Intuitively, genotyping errors should reduce the power of the best detection method available, since noise is introduced, and this can be proved theoretically (see the supplementary note accompanying [Sankararaman *et al.*, 2009]). Empirically, when we randomly add genotyping errors to the set of 33,138 SNPs, we observe that the power decreases with the rate of genotyping errors (see Figure 6.3).

6.3.5 Detecting relatives in a pool

The LR-test can be extended to test for the presence of a specific relative of the tested individual. This scenario is similar to identifying a genotype in the presence of errors. The modified test is parameterized by γ, the probability that the relative and the tested individual share an allele. Thus, $\gamma = 1$ reduces to the case where the test detects the individual (or an identical twin), $\gamma = \frac{1}{2}$ denotes a test that detects siblings or parents, and so on. We used the independent SNPs (33,138) in the WTCCC dataset and evaluated the power to detect individuals with different degrees of relatedness to the tested individual ($\gamma = 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}$). We observe a sharp decrease in power when we go from $\gamma = 1$ to $\gamma = \frac{1}{2}$ even though all the SNPs were used (Figure 6.6). At a false positive rate of 10^{-3}, the power decreases from around 0.95 for $\gamma = 1$ to 0.22 for $\gamma = 0.5$ to 0.03 for $\gamma = 0.25$.

6.3.6 Transferrability across populations

Our analysis provides a population-independent bound on power, i.e., the power computed from Equation (6.4) does not depend on the allele frequencies and hence, should be the same across different populations. In a further experiment, we evaluated this aspect of our analysis by repeating our experiments on the YRI population from the HapMap. Since the number of YRI individuals in the HapMap is relatively small, we simulated a dataset of 3000 individuals by sampling from the YRI allele frequencies at independent SNPs with MAF > 0.05. We computed power for a pool of size 1000 individuals for $m = 1000, 10000$ and 33,138 SNPs (the number obtained from the WTCCC data). The results shown in Figure 6.7 confirm our analysis. A caveat is that the number of independent SNPs with small MAF may differ across the populations. This would affect the total number of SNPs that can potentially be exposed for a given population.

6.4 Discussion

We have analytically characterized the power of the LR-test when m common SNPs in linkage equilibrium are exposed in a pool of n individuals. In this case, the relation between m,n, α and β can be described as

$$z_\alpha + z_{1-\beta} \approx \sqrt{\frac{m}{n}}, \tag{6.4}$$

where z_x is the $100(1-x)$-percentile of the normal distribution. (so that the probability for a normally distributed variable to have a value larger than z_x is exactly x)

Equation (6.4) is valid for large pools ($n > 100$) and for common SNPs (minor allele frequency > 0.05). It provides an upper bound on the number of SNPs that can be safely exposed for a particular choice of false positive rate and power. Note that Equation (6.4)

implies that m, the allowed number of exposed SNPs, is linear in n for fixed α and β, and importantly, that the power of the test does not depend on the allele frequencies $p_1, \ldots, p_m$, as long as the minor allele frequencies (MAFs) are sufficiently large.

The conditions necessary for our analysis to hold suggest the following prefiltering protocol to obtain a set of SNPs that can potentially be exposed: remove all SNPs with MAF ≤ 0.05 and retain a subset of SNPs in linkage equilibrium. We then use the LR-test to determine the set of exposed SNPs.

In using this bound, several issues should be kept in mind. First, our analysis assumes that the exposed SNPs are in linkage equilibrium. When the exposed SNPs are in linkage disequilibrium, the power of the LR-test is reduced (see Figure 6.2); nonetheless, under these circumstances, there is a potential risk that one could leverage the linkage disequilibrium in order to get better power from a different test. We thus recommend that dependent SNPs not be exposed until this issue can be studied rigorously. Second, Equation (6.4) is based on the assumptions of common SNPs and large pools (MAF > 0.05 and $n > 100$). The presence of rare SNPs may improve the power of the LR-test or other tests, and thus jeopardize privacy. We have studied the effect of pool size empirically using both simulated data and real summary data (see Methods), and found that Equation (6.4) is accurate for $n > 100$ for a Caucasian population. However, unless it is clear that the assumptions of common SNPs and large pools are met, we would recommend that Equation (6.4) be used as a rough guide and that final decisions regarding the set of exposed SNPs should be based on an empirical computation of the power of the LR-test.

To this end, we have implemented a tool, *SecureGenome*, that takes as input a genotype dataset (including the individuals' genotypes) together with a ranking of the SNPs, greedily removes SNPs that are in linkage disequilibrium, and determines the number of highly ranked SNPs that can be safely exposed. The program outputs this value along with the power of the LR-test evaluated both empirically and theoretically. This tool can serve as a practical guide to allow researchers to develop a consensus that takes into account both privacy and the need to leverage data collected throughout the community.

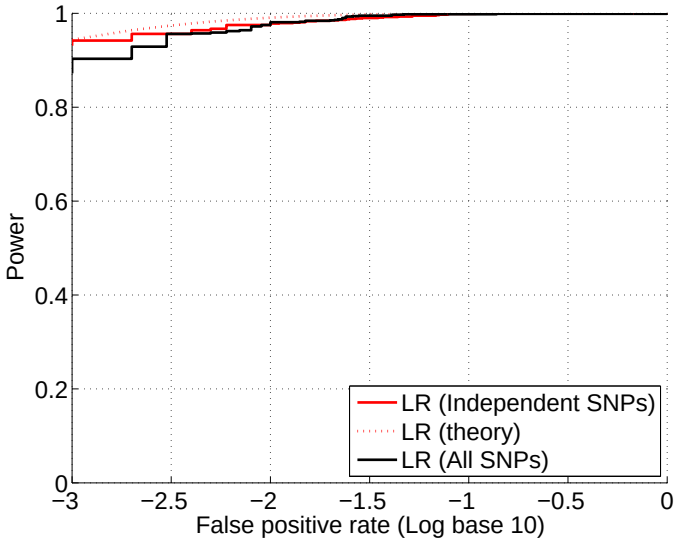

Figure 6.2: **ROC curve of the approximate LR test, constructed under a model of independent SNPs, applied on all 358,053 dependent SNPs from the WTCCC data**. The power of the test decreases slightly when SNPs in linkage disequilibrium are included. We compare the power of the approximate LR test applied on the 358,053 SNPs from the WTCCC data (the set of all SNPs published in the WTCCC study with MAF > 0.05) to the test applied on the set of 33138 independent, common SNPs.

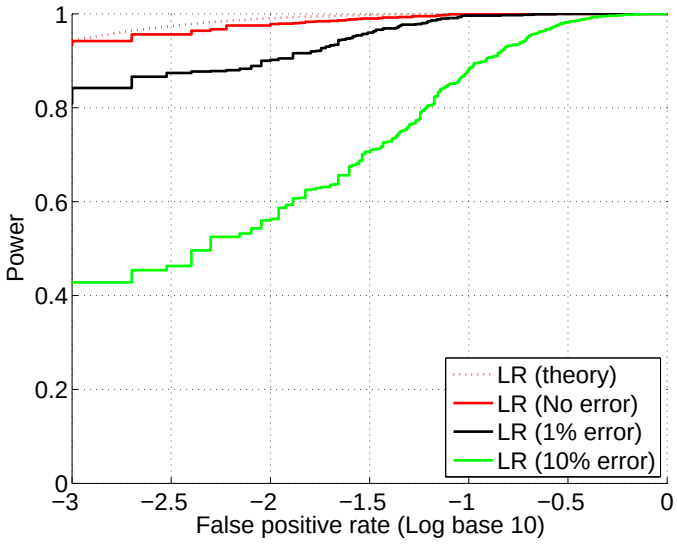

Figure 6.3: **ROC curves for the approximate LR-test under different genotyping error rates.** As the genotyping error rate increases from 1% to 10%, the power of the LR-test (constructed with the assumption of no genotyping errors) decreases significantly. We used the set of all 33138 independent, common SNPs in the WTCCC dataset.

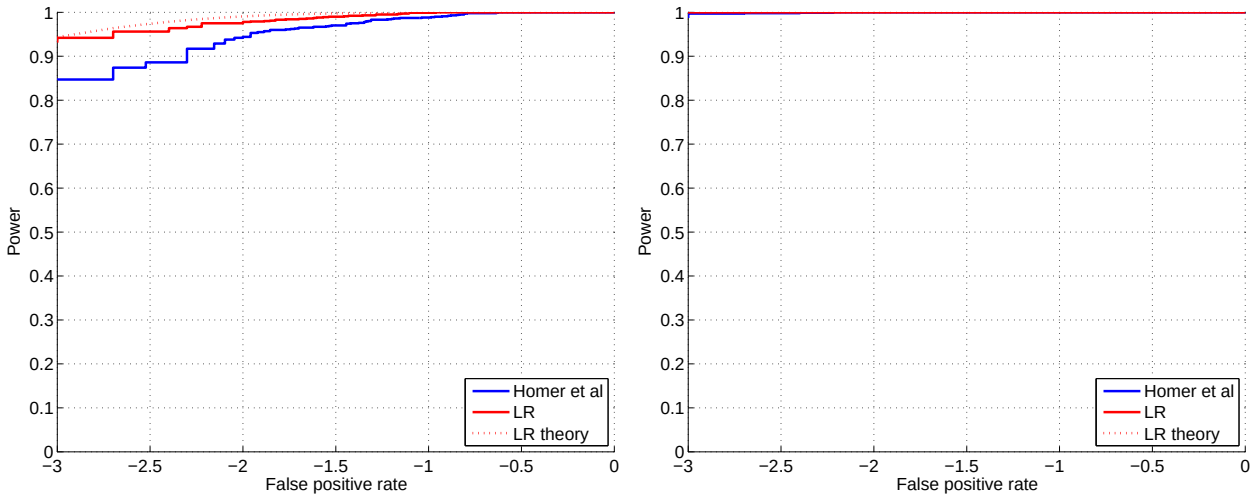

Figure 6.4: **ROC curves comparing the power attained by the approximate LR-test and by the statistic used in Homer et al. [Homer *et al.*, 2008] when applied to all 33138 independent, common SNPs in the WTCCC data, under two different settings**. (Left) In the setting studied in this paper, individuals in the pool and the finite reference dataset, and the individual of interest are all sampled independently from the same distribution under the null hypothesis. Under the alternative hypothesis, the tested individual is randomly sampled from the pool. (Right) The setting considered in [Homer *et al.*, 2008] has the identical alternative hypothesis whereas, under the null hypothesis, the individual is randomly sampled from the finite reference dataset. Both the LR statistic and the statistic proposed in [Homer *et al.*, 2008] are markedly more powerful in the second setting. When extrapolated to a false positive level of 10^{-6}, the power (both theoretical and empirical) is less than 0.5 (0.30 for the approximate LR-test and 0.47 for the LR theory) in the first setting. In contrast, at the same false positive level, the approximate LR test and 0.99 for the LR theory attain a power of 0.95 and 0.99 respectively, in the second setting.

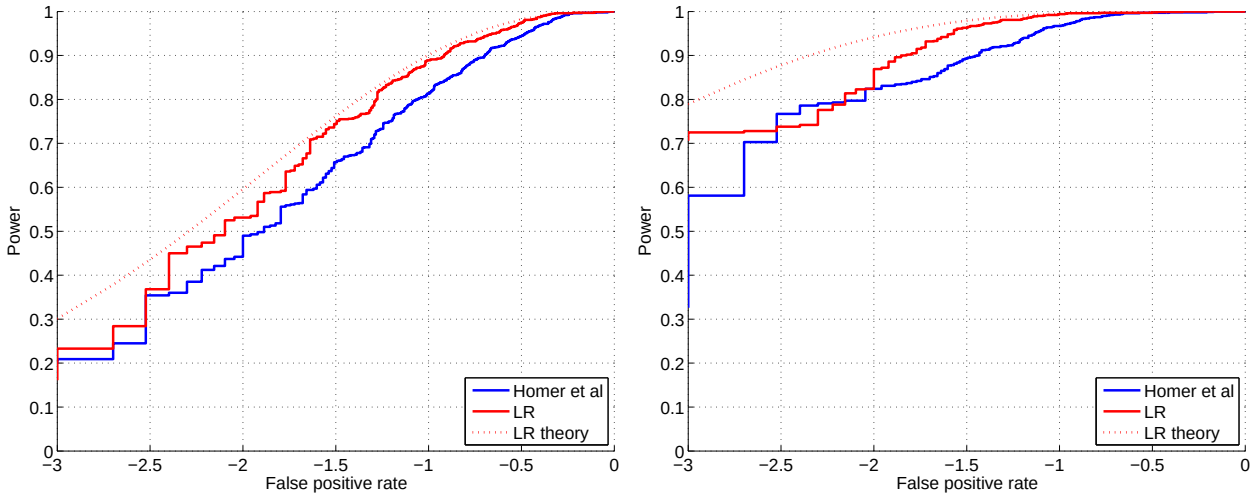

Figure 6.5: **ROC curves comparing the power attained by the approximate LR-test and by the statistic used in Homer et al [Homer *et al.*, 2008] when applied to a subset of 10000 independent, common SNPs from the WTCCC data, under two different settings** (see the caption of Figure 6.4 for descriptions of the settings). For a false positive level of 10^{-3}, the power of either test is at least doubled in the second setting.

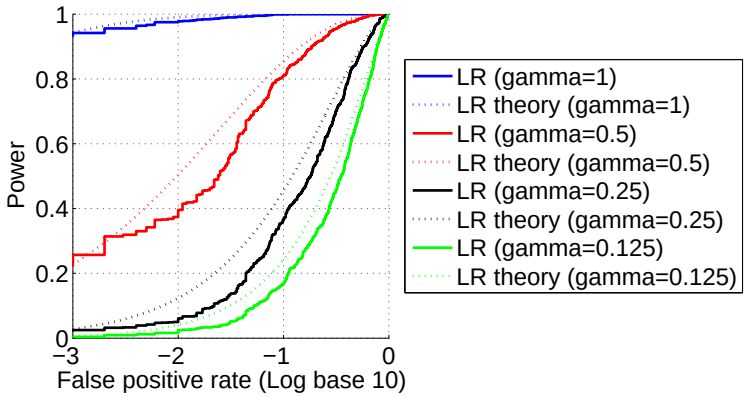

Figure 6.6: **ROC curve of the approximate LR-test on the task of detecting relatives.** The power to detect relatives is considerably smaller relative to the power to detect the tested individual: it decreases significantly even when testing first-order relationships ($\gamma = \frac{1}{2}$ for siblings and parents). We used the set of all 33138 independent, common SNPs from the WTCCC data.

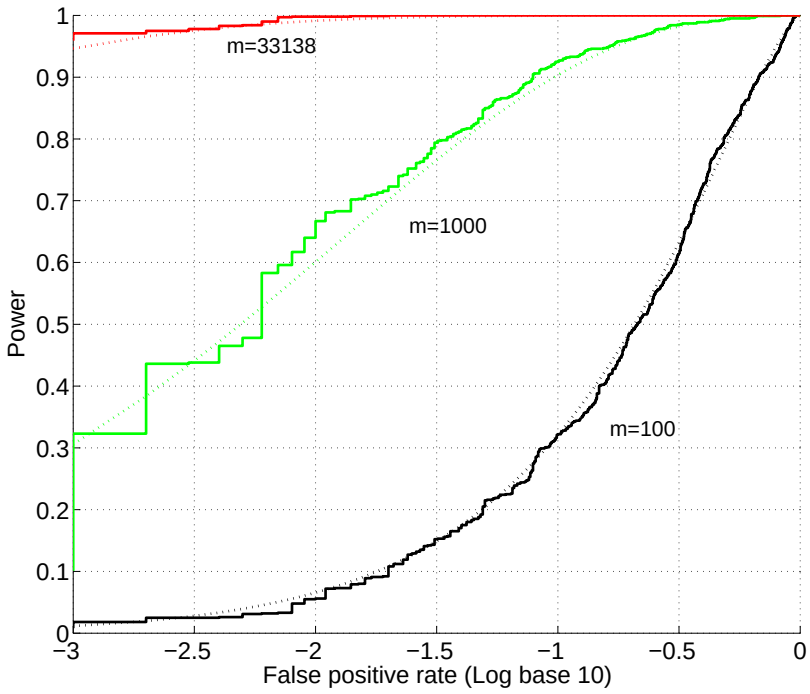

Figure 6.7: The power attained by the LR-test is not population-specific. The power of the LR-test, with a plug-in allele frequency estimate, computed for $m = 1000, 10000$ and 33138 on the HapMap YRI dataset, closely matches its theoretical power, computed using a modified version of Equation (6.4) corrected to account for the use of the plug-in estimate. Note that the theoretical power does not depend on the specific allele frequencies of a population.

Chapter 7

Conclusions

7.1 Contributions of this thesis

In this thesis, we have focused on three important statistical problems that arise in the study of human genetic variation.

- At the molecular level, we have focused on the problem of predicting functional residues in proteins. In chapter 2, we have described INTREPID, a phylogenomic method, that infers functionally important residues based on the sequence information alone.

  The primary innovation in INTREPID is its use of the phylogeny of the family to infer the evolutionary pressures on positions within different subgroups. INTREPID infers functionally important positions through a traversal of the phylogeny from the root to the target protein located at a leaf; at each point on this path and for each position in the multiple sequence alignment, INTREPID computes a positional conservation score based on Jensen-Shannon (J-S) divergence between the distribution of amino acids at that position and a background distribution. Positional scores are adjusted to take into consideration the scores of other positions within the same subtree; thus positional scores for a subtree containing highly similar sequences will be small, even though individual positions may be highly conserved. By contrast, a position that is highly conserved within a subtree that is otherwise highly variable will have a high JS divergence. Each position is then assigned the maximal JS score achieved over all nodes on the path. Positions that are conserved across the entire family achieve their maximum score at the root, whereas other positions will achieve their maximum at some distance from the root. Since even catalytic residues are not always perfectly conserved across a family (if, for instance, sequences with divergent functions are included in the analysis, or due to alignment errors), this tree traversal enables INTREPID to exploit the information in highly divergent datasets. On the task of catalytic residue prediction, INTREPID was found to be more accurate than other conservation-based functional

residue prediction methods such as ConSurf and Evolutionary Trace as well as scoring functions that do not use a phylogenetic tree. Further, the sensitivity of INTREPID was found to increase as the alignments became more divergent.

In chapter 3, we have addressed the problem of functional residue prediction using evolutionary and structural information. We described a statistical method, Discern, that brings together three important ideas. First, DISCERN uses an evolutionary modeling approach (specifically, the INTREPID phylogenomic method) to infer the degree to which residues are under selective pressure. Second, we incorporate information from the structural neighborhood of a residue including features (such as sequence conservation, charge, solvent accessibility, etc.) computed for structurally proximal residues. Third, and critically, we use statistical sparsification methods (specifically, L_1 regularization) to cope with the fact that our statistical model is based on a large number of redundant, noisy features. Without such regularization, we find that our method overfits—in particular the inclusion of information from structural neighbors leads to a decrease in accuracy. With regularization, we obtain a significant increase in accuracy. Regularization allows us to find a signal within the large set of candidate features that can be used to describe the structural and evolutionary neighborhood of an amino acid. On a homology-reduced subset of manually curated enzymes from the Catalytic Site Atlas, DISCERN attains improvements in recall of 12-20% over reported results on the task of catalytic residue prediction.

- At the population level, we have focused on the problem of inferring locus-specific ancestries in admixed populations. In chapter 5, we have described a Bayesian hidden Markov model (HMM) that describes the admixture process. Inference in this model is intractable because the model parameters are unknown. Existing approaches for inference use Markov Chain Monte Carlo (MCMC) algorithms which are computationally expensive and do not converge on a time scale of several days. Instead, we have proposed directly maximizing the likelihood using an iterative Expectation-Maximization (EM) algorithm.

  The EM algorithm requires an initial solution. Random initializations result in low accuracies because of the high dimensionality of the solution space. We propose a fast and accurate initialization procedure, LAMP, that exploits the structure of the admixed genome (Chapter 4). LAMP divides the genome into overlapping windows. The windows are chosen to be long enough to be informative about the ancestry but short enough so that few windows have a breakpoint. Classification within a window is relatively easy and classification across overlapping windows are combined by a majority vote. This gives us an initial estimate of the ancestries which are then improved by the EM algorithm.

  Results on simulated admixtures show that our implementation of this model, SWITCH,

can accurately estimate ancestries even when the ancestral genotypes are not available. Further, SWITCH can run efficiently on genome-scale datasets e.g., locus-specific ancestries of 500 genotype from human chromosome 1 can be inferred in under 30 min. Further, the model incorporates other parameters of biological interest and these can be inferred as well. As an example, we demonstrate that the model can be used to infer the allele frequencies of ancestral populations and these estimates can be used to reconstruct extinct populations such as the native American populations that were ancestral to present-day Latinos.

- As association studies become a widely adopted tool, it is increasingly important for the generated data to be shared across studies. While such sharing can improve the power to detect associations, there is a danger that the privacy of study participants might be compromised.

  In Chapter 6, we have addressed the issue of genomic privacy by providing guidelines as to which set of SNPs can be safely exposed for a given pool size, maximal allowable power β and false positive level α. Our approach is based on casting the problem as a statistical hypothesis testing problem for which the *likelihood ratio test* attains the maximal power achievable ([Lehmann, 2005]). This provides a guarantee that it is safe to expose a set of SNPs for which the LR-test does not achieve sufficient power. Our theoretical and empirical results lead to a conclusion that is qualitatively different than the result obtained in [Homer *et al.*, 2008] in that we find that the power achieved by considering whole-genome datasets is in fact limited.

7.2 Future Directions

In this section, we outline directions for future research.

7.2.1 Functional residue prediction

- We have focused on catalytic residue prediction because of the availability of a large dataset of enzymes annotated with experimentally verified catalytic residues. However, the methodology underlying Discern can be applied to predicting other classes of functional residues such as ligand-binding residues, allosteric residues, specificity-determinants and residues in interaction interfaces. Allosteric residues are a specially interesting and difficult class for our methodology because of their weaker conservation signals and the lack of large-scale annotations.

  Similarly, specificity-determining residues are difficult to detect because they show conservation within subtrees of the family tree but vary across these subtrees. Methods for identifying these residues have relied primarily on these conservation patterns.

However, these residues are, often, close to the active site on the 3D structure. Thus, we can imagine a two step procedure in which we predict the catalytic residues and then use proximity to the catalytic residues as a feature to predict the specificity-determinants. The success of this approach will depend on a large-scale dataset of annotated specificity-determinants. A good predictor for specificity-determinants can, in turn, provide features for catalytic residue prediction.

- A test of the usefulness of computational functional residue prediction methods would be their ability to make novel predictions which can then be verified experimentally. It would be important to understand if the mutation of predicted functional residues results in alleles with reduced functional abilities. A first step in this direction involves evaluating predicted functional residues in human proteins against resources that document the functional impact of mutations in these proteins, such as OMIM [http://www.ncbi.nlm.nih.gov/omim].

  There are a number of families of enzymes that are believed to be catalytically inactive, e.g., some kinase families lack key catalytic residues and are designated as pseudokinases. Functional residue prediction methods can provide clues about the presence of novel catalytic residues in these enzymes and could provide insights into alternate mechanisms of catalysis.

- Functional residue prediction methods can also be used to improve protein function prediction. Such a method would be particularly useful in cases where homologues with known function have low sequence similarity to the protein of interest. In such cases, it is difficult to confidently assert that the two distantly related proteins have similar function. On the other hand, conservation of key functional residues across the two proteins increases the likelihood that they perform the same function. Such a prediction method would assign a score to the alignment of the two proteins. The score could be computed as a weighted sum of the log likelihood of an ancestral residue mutating to the aligned pair of residues in the two sequences. The weight of a position would be determined by the probability that this position is functionally important. Such a method has been proposed [George *et al.*, 2005] using catalytic residue annotations from the Catalytic Site Atlas (CSA). However, the use of CSA limits the applicability of their method.

  An interesting test case for these enhanced function prediction methods is the set of proteins whose 3D structures are being solved as part of structural genomics initiatives [Chandonia and Brenner, 2006]. As of December 2009, these initiatives have solved the structure of more than 250 domains of unknown function; of these only 41% can be annotated based on literature and structural homology leaving more than half the domains with no reliable functional annotation [http://kb.psi-structuralgenomics.org/update/2009/12/full/fa_psisgkb.2009.54.html].

7.2.2 Population structure and association studies

- While we can now efficiently and accurately infer locus-specific ancestries in admixed populations, our studies have been restricted to recent admixtures (admixtures that occurred over the last 500 years). Ancestry inference in ancient admixtures is a harder problem because historic recombination events have broken down the stretches of the chromosome that share ancestry. Such admixtures require models that can faithfully capture the patterns of variation across short stretches of the genome as well as efficient inference algorithms in these models.

- Current methods for association mapping in admixed populations use the signal from the ancestries (admixture mapping) or the signal from the alleles (association mapping) in isolation. Tests that combine both signals can improve the statistical power to detect weak associations. The ancestries are informative of associations across large regions while the alleles are informative across smaller regions – a consequence of the fact that recombinations within each of the ancestral populations have broken down correlations over a longer period of time than the recombinations during admixture. Thus, we can use the ancestry signal to rapidly identify a large region and then use the allelic signal to narrow down the association within these regions. While it is conceivable that many variants of these hybrid tests would be sensitive to even modest associations, combining these signals optimally is an open question.

- The model underlying SWITCH can be used to infer the allele frequencies of the ancestral populations. This problem is of great interest as it can be used to reconstruct the allelic spectrum of currently extinct populations, such as the Taino who are ancestors of modern-day Puerto Ricans.

7.2.3 Genomic Privacy

- There are a number of questions that need to be investigated in the context of genomic privacy. Our current analysis holds for common, independent SNPs. What privacy guarantees can be given when rare or dependent SNPs need to be exposed? Privacy mechanisms developed in the databases community, like the notion of differential privacy [Dwork, 2006], require sufficiently perturbing the summary data to prevent detection. Notions such as differential privacy provide strong guarantees but render the summary data useless as they require a large amount of noise to be added to each SNP (Each SNP needs to be perturbed with noise drawn from a distribution with standard deviation $O(m)$ where m is the total number of exposed SNPs). A reason for the large amount of noise is that differential privacy provides a distribution-free privacy guarantee. One approach around this problem is to formulate a notion of privacy that

is dependent on the allele frequency distribution. Such an approach could resolve the limitations associated with differential privacy.

Analyzing a utility-constrained notion of privacy where SNPs are perturbed while maximizing the detection power presents yet another interesting direction. Analyses such as these can relax the current requirement of common, independent SNPs thereby allowing rare and dependent SNPs to be exposed. A further important direction is to understand how to share data in the case where multiple phenotypes are present for each genotype. Treating each phenotype as an independent study risks exposing summary statistics that could then be combined to increase the power of detection.

- Privacy risks are not restricted to data from association studies alone. Heterogeneous, often publicly-available, databases present newer risks. For instance, geneological databases can be combined with genotype datasets to identify the genotypes [Gitschier, 2009]. Social networks can often be used to identify these family relationships when geneological databases are not available. Similarly, hospital health records can potentially be linked to genotype data enabling an attacker to obtain the identity associated with a genotype.

  Thus, there are two major challenges facing the scientific community: i) understanding the privacy risks involved in sharing genomic data and devising algorithms that would enable secure sharing, and ii) developing an infrastructure that would give researchers easy access to this data. This, in turn, requires integrating the different data sources and establishing an efficient mechanism to query these sources. Organizations such as the Wellcome Trust and companies such as 23andMe provide access to genotype and phenotype data, hospital databases store longitudinal phenotype data from patients, while independent researchers and drug companies aim to combine these data to identify novel associations.

  Further, the nature of genetic data is such that the privacy of an individual is closely tied to the privacy of his/her parents, siblings and other relatives. As the number of sources of data and the number of individuals in these databases grow, we may have to reconsider the very notion of genomic privacy.

Bibliography

[Aloy *et al.*, 2001] P. Aloy, E. Querol, F. X. Aviles, and M. J. Sternberg. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.*, 311(2):395–408, August 2001.

[Altschul *et al.*, 1997] SF Altschul, TL Madden, AA Schaffer, J Zhang, Z Zhang, W Miller, and DJ Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17):3389–3402, 1997.

[Apweiler *et al.*, 2004] R. Apweiler, A. Bairoch, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, and L.S. Yeh. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, 32:D115–119, Jan 2004.

[Bagley and Altman, 1995] S. C. Bagley and R. B. Altman. Characterizing the microenvironment surrounding protein sites. *Protein Sci.*, 4(4):622–635, 1995.

[Baker and Sali, 2001] D. Baker and A. Sali. Protein structure prediction and structural genomics. *Science*, 294(5540):93–96, October 2001.

[Barrett *et al.*, 2008] J. C. Barrett, S. Hansoul, D. L. Nicolae, J. H. Cho, R. H. Duerr, J. D. Rioux, S. R. Brant, M. S. Silverberg, K. D. Taylor, M. M. Barmada, A. Bitton, T. Dassopoulos, L. W. Datta, T. Green, A. M. Griffiths, E. O. Kistner, M. T. Murtha, M. D. Regueiro, J. I. Rotter, L. P. Schumm, A. H. Steinhart, S. R. Targan, R. J. Xavier, C. Libioulle, C. Sandor, M. Lathrop, J. Belaiche, O. Dewit, I. Gut, S. Heath, D. Laukens, M. Mni, P. Rutgeerts, A. Van Gossum, D. Zelenika, D. Franchimont, J. P. Hugot, M. de Vos, S. Vermeire, E. Louis, L. R. Cardon, C. A. Anderson, H. Drummond, E. Nimmo, T. Ahmad, N. J. Prescott, C. M. Onnie, S. A. Fisher, J. Marchini, J. Ghori, S. Bumpstead, R. Gwilliam, M. Tremelling, P. Deloukas, J. Mansfield, D. Jewell, J. Satsangi, C. G. Mathew, M. Parkes, M. Georges, and M. J. Daly. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.*, 40:955–962, Aug 2008.

[Bartlett *et al.*, 2002] G. J. Bartlett, C. T. Porter, N. Borkakoti, and J. M. Thornton. Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.*, 324(1):105–121, November 2002.

[Bate and Warwicker, 2004] P. Bate and J. Warwicker. Enzyme/non-enzyme discrimination and prediction of enzyme active site location using charge-based methods. *J. Mol. Biol.*, 340:263–276, Jul 2004.

[Besag, 1986] Julian Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(3):259–302, 1986.

[Bonnen *et al.*, 2006] P. E. Bonnen, I. Pe'er, R. M. Plenge, J. Salit, J. K. Lowe, M. H. Shapero, R. P. Lifton, J. L. Breslow, M. J. Daly, D. E. Reich, et al. Evaluating potential for whole-genome studies in Kosrae, an isolated population in Micronesia. *Nat. Genet.*, 38:214–217, 2006.

[Boykov and Kolmogorov, 2004] Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(9):1124–1137, 2004.

[Boykov *et al.*, 2001] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:2001, 2001.

[Brenner *et al.*, 2000] S.E. Brenner, P. Koehl, and M. Levitt. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.*, 28:254–256, Jan 2000.

[Brooks and Gelman, 1998] Stephen P. Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, 1998.

[Campbell *et al.*, 2005] Catarina D. Campbell, Elizabeth L. Ogburn, Kathryn L. Lunetta, Helen N. Lyon, Matthew L. Freedman, Leif C. Groop, David Altshuler, Kristin G. Ardlie, and Joel N. Hirschhorn. Demonstrating stratification in a European American population. *Nat. Genet.*, 37:868–872, 2005.

[Capra and Singh, 2007] John A. Capra and Mona Singh. Predicting functionally important residues from sequence conservation. *Bioinformatics*, 23(15):1875–1882, 2007.

[Capra and Singh, 2008] John A. Capra and Mona Singh. Characterization and prediction of residues determining protein functional specificity. *Bioinformatics*, 24(13):1473–1480, 2008.

[Casari *et al.*, 1995] G. Casari, C. Sander, and A. Valencia. A method to predict functional residues in proteins. *Nat. Struct. Biol.*, 2(2):171–178, February 1995.

[Chakraborty and Weiss, 1988] R. Chakraborty and K. M. Weiss. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc. Natl. Acad. Sci. U.S.A.*, 85:9119–9123, Dec 1988.

[Chandonia and Brenner, 2006] J. M. Chandonia and S. E. Brenner. The impact of structural genomics: expectations and outcomes. *Science*, 311:347–351, Jan 2006.

[Chandonia *et al.*, 2004] J. M. Chandonia, G. Hon, N. S. Walker, L. Lo Conte, P. Koehl, M. Levitt, and S. E. Brenner. The ASTRAL Compendium in 2004. *Nucleic Acids Res.*, 32(Database issue), January 2004.

[Chernoff, 1952] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.*, 23:493–507, 1952.

[Cho *et al.*, 2006] Hoon Cho, Lingyu Huang, Adel Hamza, Daquan Gao, Chang-Guo Zhan, and Hsin-Hsiung Tai. Role of glutamine 148 of human 15-hydroxyprostaglandin dehydrogenase in catalytic oxidation of prostaglandin E2. *Bioorganic and medicinal chemistry*, 14(19):6486–6491, 2006.

[Church *et al.*, 2009] George Church, Catherine Heeney, Naomi Hawkins, Jantina de Vries, Paula Boddington, Jane Kaye, Martin Bobrow, Bruce Weir, and P3G Consortium. Public access to genome-wide data: Five views on balancing research with privacy and protection. *PLoS Genet*, 5(10):e1000665, 10 2009.

[Clayton *et al.*, 2005] D. G. Clayton, N. M. Walker, D. J. Smyth, R. Pask, J. D. Cooper, L. M. Maier, L. J. Smink, A. C. Lam, N. R. Ovington, H. E. Stevens, et al. Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat. Genet.*, 37:1243–1246, 2005.

[Collins *et al.*, 1997] Francis S. Collins, Mark S. Guyer, and Aravinda Chakravarti. Variations on a Theme: Cataloging Human DNA Sequence Variation. *Science*, 278(5343):1580–1581, 1997.

[Collins-Schramm *et al.*, 2003] H. E. Collins-Schramm, B. Chima, D. J. Operario, L. A. Criswell, and M. F. Seldin. Markers informative for ancestry demonstrate consistent megabase-length linkage disequilibrium in the african american population. *Hum Genet*, 113(3):211–219, August 2003.

[Cooper *et al.*, 2008] J. D. Cooper, D. J. Smyth, A. M. Smiles, V. Plagnol, N. M. Walker, J. E. Allen, K. Downes, J. C. Barrett, B. C. Healy, J. C. Mychaleckyj, J. H. Warram, and

J. A. Todd. Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nat. Genet.*, 40:1399–1401, Dec 2008.

[Davis and Goadrich, 2006] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. ICML '06: Proceedings of the 23rd International Conference on Machine Learning, pages 233–240, New York, 2006. ACM.

[Del Sol Mesa *et al.*, 2003] A. Del Sol Mesa, Florencio Pazos, and Alfonso Valencia. Automatic Methods for Predicting Functionally Important Residues. *J. Mol. Biol.*, 326(4):1289–1302, February 2003.

[Devlin and Roeder, 1999] B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, December 1999.

[Donald and Shakhnovich, 2005] J. E. Donald and E. I. Shakhnovich. Determining functional specificity from protein sequences. *Bioinformatics*, 21(11):2629–2635, June 2005.

[Durbin *et al.*, 1998] R. Durbin, S. Eddy, A. Krogh, and G. Mitchsion. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[Dwork, 2006] Cynthia Dwork. Differential privacy. In *in ICALP*, pages 1–12. Springer, 2006.

[Edgar, 2004] R. C. Edgar. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5(1), August 2004.

[Eisenberg *et al.*, 1982] D. Eisenberg, R. M. Weiss, T. C. Terwilliger, and W. Wilcox. Hydrophobic moments and protein structure. *Faraday Symp. Chem. Soc.*, 17:109–120, 1982.

[Elcock, 2001] A. H. Elcock. Prediction of functionally important residues based solely on the computed energetics of protein structure. *J. Mol. Biol.*, 312(4):885–896, September 2001.

[Falush *et al.*, 2003] D. Falush, M. Stephens, and J. K. Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587, August 2003.

[Felsenstein, 1993] J. Felsenstein. PHYLIP (Phylogeny Inference Package) version 3.5c. *Distributed by the author. Department of Genetics, University of Washington, Seattle.*, 1993.

[Fetrow and Skolnick, 1998] J.S. Fetrow and J. Skolnick. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J. Mol. Biol.*, 281:949–968, Sep 1998.

[Frayling, 2007] T. M. Frayling. Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nat. Rev. Genet.*, 8:657–662, Sep 2007.

[Freedman *et al.*, 2004] M. L. Freedman, D. Reich, K. L. Penney, G. J. McDonald, A. A. Mignault, N. Patterson, S. B. Gabriel, E. J. Topol, J. W. Smoller, C. N. Pato, et al. Assessing the impact of population stratification on genetic association studies. *Nat. Genet.*, 36:388–393, 2004.

[Freedman *et al.*, 2006] M. L. Freedman, C. A. Haiman, N. Patterson, G. J. McDonald, A. Tandon, A. Waliszewska, K. Penney, R. G. Steen, K. Ardlie, E. M. John, I. Oakley-Girvan, A. S. Whittemore, K. A. Cooney, S. A. Ingles, D. Altshuler, B. E. Henderson, and D. Reich. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc. Natl. Acad. Sci. U.S.A.*, 103:14068–14073, Sep 2006.

[George *et al.*, 2005] Richard A. George, Ruth V. Spriggs, Gail J. Bartlett, Alex Gutteridge, Malcolm W. MacArthur, Craig T. Porter, Bissan Al-Lazikani, Janet M. Thornton, and Mark B. Swindells. Effective function annotation through catalytic residue conservation. *Proc. Nat. Acad. Sci. USA*, 102(35):12299–12304, 2005.

[Gilbert, 2008] Natasha Gilbert. Researchers criticize genetic data restrictions. *Nature*, 2008.

[Gitschier, 2009] J. Gitschier. Inferential genotyping of Y chromosomes in Latter-Day Saints founders and comparison to Utah samples in the HapMap project. *Am. J. Hum. Genet.*, 84:251–258, Feb 2009.

[Glaser *et al.*, 2006] F. Glaser, R. J. Morris, R. J. Najmanovich, R. A. Laskowski, and J. M. Thornton. A method for localizing ligand binding pockets in protein structures. *Proteins*, 62(2):479–488, February 2006.

[Goldstein, 2009] David B. Goldstein. Common Genetic Variation and Human Traits. *N Engl J Med*, 360(17):1696–1698, 2009.

[Greenshtein and Ritov, 2004] E. Greenshtein and Y. Ritov. Persistence in high-dimensional predictor selection and the virtue of overparametrization. *Bernoulli*, 10:971–988, 2004.

[Gutteridge *et al.*, 2003] A. Gutteridge, G. J. Bartlett, and J. M. Thornton. Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J. Mol. Biol.*, 330(4):719–734, July 2003.

[Haldane, 1919] J. B. S. Haldane. The combination of linkage values and the calculation of distance between the loci of linked factors. *J. Genet.*, 8:299–309, 1919.

[Hanis *et al.*, 1986] C. L. Hanis, R. Chakraborty, R. E. Ferrell, and W. J. Schull. Individual admixture estimates: disease associations and individual risk of diabetes and gallbladder disease among mexican-americans in starr county, texas. *Am J Phys Anthropol*, 70(4):433–441, August 1986.

[Hannenhalli and Russell, 2000] S.S. Hannenhalli and R.B. Russell. Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.*, 303:61–76, Oct 2000.

[Hastie *et al.*, 2001] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2001.

[Helgason *et al.*, 2005] Agnar Helgason, Bryndís Yngvadóttir, Birgir Hrafnkelsson, Jeffrey Gulcher, and Kári Stefánsson. An Icelandic example of the impact of population structure on association studies. *Nat. Genet.*, 37:90–95, 2005.

[Henikoff and Henikoff, 1992] S Henikoff and J G Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Nat. Acad. Sci. USA*, 89(22):10915–10919, 1992.

[Hennig *et al.*, 1998] M. Hennig, A. D'Arcy, I.C. Hampele, M.G. Page, C. Oefner, and G.E. Dale. Crystal structure and reaction mechanism of 7,8-dihydroneopterin aldolase from Staphylococcus aureus. *Nat. Struct. Biol.*, 5:357–362, May 1998.

[Hirschhorn and Daly, 2005] J. N. Hirschhorn and M. J. Daly. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.*, 6:95–108, 2005.

[Hirschhorn, 2009] Joel N. Hirschhorn. Genomewide Association Studies – Illuminating Biologic Pathways. *N Engl J Med*, 360(17):1699–1701, 2009.

[Hoggart *et al.*, 2004] C. J. Hoggart, M. D. Shriver, R. A. Kittles, D. G. Clayton, and P. M. McKeigue. Design and analysis of admixture mapping studies. *Am J Hum Genet*, 74(5):965–978, May 2004.

[Hoggart *et al.*, 2008] Clive J. Hoggart, John C. Whittaker, Maria De Iorio, and David J. Balding. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet*, 4(7):e1000130, Jul 2008.

[Homer *et al.*, 2008] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genet*, 4(8):e1000167, 08 2008.

[Hosmer and Lemeshow, 2000] David W. Hosmer and Stanley Lemeshow. *Applied Logistic Regression*. John Wiley, New York, September 2000.

[Huang and Schroeder, 2006] B. Huang and M. Schroeder. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.*, 6:19, 2006.

[Huang *et al.*, 2001] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2001. Foreword By-Reddy, Raj.

[Hubbard and Thornton, 1993] S.J. Hubbard and J.M. Thornton. A computer algorithm to calculate surface accessibility. Department of Biochemistry and Molecular Biology, University College, London, 1993.

[Innis *et al.*, 2004] C.Axel Innis, A.Prem Anand, and R. Sowdhamini. Prediction of functional sites in proteins using conserved functional group analysis. *J. Mol. Biol.*, 337(4):1053 – 1068, 2004.

[Johnson *et al.*, 2007] Margaret A. Johnson, Maurice W. Southworth, Torsten Herrmann, Lear Brace, Francine B. Perler, and Kurt Wuthrich. NMR structure of a KlbA intein precursor from Methanococcus jannaschii. *Protein Sci.*, 16(7):1316–1328, 2007.

[Kabsch and Sander, 1983] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, Dec 1983.

[Kalinina *et al.*, 2004] Olga V. Kalinina, Andrey A. Mironov, Mikhail S. Gelfand, and Aleksandra B. Rakhmaninova. Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein Sci.*, 13(2):443–456, 2004.

[Ko *et al.*, 2005] J. Ko, L. F. Murga, Y. Wei, and M. J. Ondrechen. Prediction of active sites for protein structures from computed chemical properties. *Bioinformatics*, 21 Suppl 1, June 2005.

[Koh *et al.*, 2007] Kwangmoo Koh, Seung-Jean Kim, and Stephen Boyd. An interior-point method for large-scale L1-regularized logistic regression. *J. Mach. Learn. Res.*, 8:1519–1555, 2007.

[Kolmogorov and Zabih, 2002] Vladimir Kolmogorov and Ramin Zabih. What energy functions can be minimized via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:65–81, 2002.

[Kraft and Hunter, 2009] Peter Kraft and David J. Hunter. Genetic Risk Prediction – Are We There Yet? *N Engl J Med*, 360(17):1701–1703, 2009.

[Krishnamurthy *et al.*, 2007] Nandini Krishnamurthy, Duncan Brown, and Kimmen Sjolander. Flowerpower: clustering proteins into domain architecture classes for phylogenomic inference of protein function. *BMC Evolutionary Biology*, 7(Suppl 1):S12, 2007.

[Lafferty *et al.*, 2001] John Lafferty, Andrew McCallum, and Fernando Pereira. Probabilistic models for segmenting and labeling sequence data. Proc. 18th International Conf. on Machine Learning, pages 282–289, San Francisco, CA, 2001. Morgan Kaufmann.

[Landau *et al.*, 2005] M. Landau, I. Mayrose, Y. Rosenberg, F. Glaser, E. Martz, T. Pupko, and N. Ben-Tal. ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.*, 33(Web Server issue), July 2005.

[Lander and Schork, 1994a] E. S. Lander and N. J. Schork. Genetic dissection of complex traits. *Science*, 265:2037–2048, 1994.

[Lander and Schork, 1994b] ES Lander and NJ Schork. Genetic dissection of complex traits. *Science*, 265(5181):2037–2048, 1994.

[Lander, 1996] Eric S. Lander. The New Genomics: Global Views of Biology. *Science*, 274(5287):536–539, 1996.

[Landgraf *et al.*, 2001] R. Landgraf, I. Xenarios, and D. Eisenberg. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.*, 307(5):1487–1502, April 2001.

[Laurie and Jackson, 2005] Alasdair T. Laurie and Richard M. Jackson. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics*, 21(9):1908–1916, May 2005.

[Lawrence *et al.*, 2005] M.C. Lawrence, P. Iliades, R.T. Fernley, J. Berglez, P.A. Pilling, and I.G. Macreadie. The three-dimensional structure of the bifunctional 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase/dihydropteroate synthase of Saccharomyces cerevisiae. *J. Mol. Biol.*, 348:655–670, May 2005.

[Lehmann, 2005] E. L. Lehmann. *Testing Statistical Hypotheses*. Springer Texts in Statistics, New York, NY, 2005.

[Levy *et al.*, 2007] Samuel Levy, Granger Sutton, Pauline C Ng, Lars Feuk, Aaron L Halpern, Brian P Walenz, Nelson Axelrod, Jiaqi Huang, Ewen F Kirkness, Gennady Denisov, Yuan Lin, Jeffrey R MacDonald, Andy Wing Chun Pang, Mary Shago, Timothy B Stockwell, Alexia Tsiamouri, Vineet Bafna, Vikas Bansal, Saul A Kravitz, Dana A

Busam, Karen Y Beeson, Tina C McIntosh, Karin A Remington, Josep F Abril, John Gill, Jon Borman, Yu-Hui Rogers, Marvin E Frazier, Stephen W Scherer, Robert L Strausberg, and J. Craig Venter. The diploid genome sequence of an individual human. *PLoS Biol*, 5(10):e254, 09 2007.

[Li and Stephens, 2003] Na Li and Matthew Stephens. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics*, 165(4):2213–2233, 2003.

[Lichtarge *et al.*, 1996] O. Lichtarge, H. R. Bourne, and F. E. Cohen. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, 257(2):342–358, March 1996.

[Lin and Wong, 1990] J. Lin and S. K. M. Wong. A new directed divergence measure and its characterization. *Int. J. Gen. Syst.*, 17(1):73–81, 1990.

[Lohmueller *et al.*, 2003] Kirk E. Lohmueller, Celeste L. Pearce, Malcolm Pike, Eric S. Lander, and Joel N. Hirschhorn. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.*, 33:177–182, 2003.

[Mao *et al.*, 2007] Xianyun Mao, Abigail W. Bigham, Rui Mei, Gerardo Gutierrez, Ken M. Weiss, Tom D. Brutsaert, Fabiola Leon-Velarde, Lorna G. Moore, Enrique Vargas, Paul M. McKeigue, et al. A genome-wide admixture mapping panel for hispanic/latino populations. *Am J Hum Genet*, 80(6), 2007.

[Marchini *et al.*, 2004] Jonathan Marchini, Lon R. Cardon, Michael S. Phillips, and Peter Donnelly. The effects of human population structure on large genetic association studies. *Nat Genet*, 36(5):512–517, May 2004.

[Mayrose *et al.*, 2004] I. Mayrose, D. Graur, N. Ben-Tal, and T. Pupko. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.*, 21(9):1781–1791, September 2004.

[McKeigue, 2005] Paul M. McKeigue. Prospects for admixture mapping of complex traits. *The American Journal of Human Genetics*, 76(1):1 – 7, 2005.

[Meister, 1974] A Meister. *The Enzymes*, volume 10. Academic Press, New York, 3rd edition, 1974.

[Mihalek *et al.*, 2004] I. Mihalek, I. Res, and O. Lichtarge. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol.*, 336(5):1265–1282, March 2004.

[Mirny and Gelfand, 2002] Leonid A Mirny and Mikhail S Gelfand. Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J. Mol. Biol.*, 321(1):7–20, 2002.

[Mooney *et al.*, 2005] Sean D. Mooney, Mike Hsin-Ping Liang, Rob DeConde, and Russ B. Altman. Structural characterization of proteins using residue environments. *Proteins: Struct., Funct., Bioinf.*, 61(4):741–747, 2005.

[Murzin *et al.*, 1995] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247(4):536–540, April 1995.

[Myers *et al.*, 2005] S. Myers, L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–324, October 2005.

[Nachman and Crowell, 2000] Michael W. Nachman and Susan L. Crowell. Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1):297–304, September 2000.

[Nakatsu *et al.*, 1998] T. Nakatsu, H. Kato, and J. Oda. Crystal structure of asparagine synthetase reveals a close evolutionary relationship to class II aminoacyl-tRNA synthetase. *Nat. Struct. Biol.*, 5:15–19, Jan 1998.

[Nimrod *et al.*, 2005] G. Nimrod, F. Glaser, D. Steinberg, N. Ben-Tal, and T. Pupko. In silico identification of functional regions in proteins. *Bioinformatics*, 21 Suppl 1, June 2005.

[Ondrechen *et al.*, 2001] Mary J. Ondrechen, James G. Clifton, and Dagmar Ringe. THEMATICS: A simple computational predictor of enzyme function from structure. *Proc. Nat. Acad. Sci. USA*, 98(22):12473–12478, October 2001.

[Ota *et al.*, 2003] M. Ota, K. Kinoshita, and K. Nishikawa. Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. *J. Mol. Biol.*, 327:1053–1064, Apr 2003.

[Panchenko *et al.*, 2004] A. R. Panchenko, F. Kondrashov, and S. Bryant. Prediction of functional sites by analysis of sequence and structure conservation. *Protein Sci.*, 13(4):884–892, April 2004.

[Parra *et al.*, 1998] E. J. Parra, A. Marcini, J. Akey, J. Martinson, M. A. Batzer, R. Cooper, T. Forrester, D. B. Allison, R. Deka, R. E. Ferrell, et al. Estimating african american admixture proportions by use of population-specific alleles. *Am J Hum Genet*, 63(6):1839–1851, December 1998.

[Patterson *et al.*, 2004] N. Patterson, N. Hattangadi, B. Lane, K. E. Lohmueller, D. A. Hafler, J. R. Oksenberg, S. L. Hauser, M. W. Smith, S. J. O'Brien, D. Altshuler, M. J. Daly, et al. Methods for high-density admixture mapping of disease genes. *Am J Hum Genet*, 74(5):979–1000, May 2004.

[Pazos and Sternberg, 2004] F. Pazos and M. J. Sternberg. Automated prediction of protein function and detection of functional sites from structure. *Proc. Nat. Acad. Sci. USA*, 101:14754–14759, Oct 2004.

[Pei *et al.*, 2006] Jimin Pei, Wei Cai, Lisa N. Kinch, and Nick V. Grishin. Prediction of functional specificity determinants from protein sequences using log-likelihood ratios. *Bioinformatics*, 22(2):164–171, 2006.

[Peters *et al.*, 1996] K. P. Peters, J. Fauck, and C. Frommel. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.*, 256(1):201–213, February 1996.

[Petrova and Wu, 2006] Natalia Petrova and Cathy Wu. Prediction of catalytic residues using support vector machine with selected protein sequence and structural properties. *BMC Bioinformatics*, 7(1):312, 2006.

[Pirovano *et al.*, 2006] Walter Pirovano, K. Anton Feenstra, and Jaap Heringa. Sequence comparison by sequence harmony identifies subtype-specific functional sites. *Nucl. Acids Res.*, 34(22):6540–6548, 2006.

[Porter *et al.*, 2004] C. T. Porter, G. J. Bartlett, and J. M. Thornton. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, 32(Database issue), January 2004.

[Price *et al.*, 2006] Alkes L. Price, Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, July 2006.

[Price *et al.*, 2007] Alkes L. Price, Nick Patterson, Fuli Yu, David R. Cox, Alicja Waliszewska, Gavin J. McDonald, Arti Tandon, Christine Schirmer, Julie Neubauer, Gabriel Bedoya, et al. A genomewide admixture map for latino populations. *Am J Hum Genet*, 80(6), 2007.

[Pritchard and Rosenberg, 1999] Jonathan K. Pritchard and Noah A. Rosenberg. Use of unlinked genetic markers to detect population stratification in association studies. *The American Journal of Human Genetics*, 65(1):220 – 228, 1999.

[Pritchard *et al.*, 2000] Jonathan K. Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, June 2000.

[Pupko *et al.*, 2002] T. Pupko, R. E. Bell, I. Mayrose, F. Glaser, and N. Ben-Tal. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, 18 Suppl 1, 2002.

[Rabiner, 1989] Lawrence Rabiner. A tutorial on hmm and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[Reich *et al.*, 2005a] D. Reich, N. Patterson, P. L. De Jager, G. J. McDonald, A. Waliszewska, A. Tandon, R. R. Lincoln, C. DeLoa, S. A. Fruhan, P. Cabre, O. Bera, G. Semana, M. A. Kelly, D. A. Francis, K. Ardlie, O. Khan, B. A. Cree, S. L. Hauser, J. R. Oksenberg, and D. A. Hafler. A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. *Nat. Genet.*, 37:1113–1118, Oct 2005.

[Reich *et al.*, 2005b] David Reich, Nick Patterson, Philip L. De Jager, Gavin J. Mcdonald, Alicja Waliszewska, Arti Tandon, Robin R. Lincoln, Cari Deloa, Scott A. Fruhan, Philippe Cabre, et al. A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. *Nature Genetics*, 37(10):1113–1118, September 2005.

[Risch and Merikangas, 1996] Neil Risch and Kathleen Merikangas. The Future of Genetic Studies of Complex Human Diseases. *Science*, 273(5281):1516–1517, 1996.

[Sander and Schneider, 1991] C. Sander and R. Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9:56–68, 1991.

[Sankararaman and Sjölander, 2008] Sriram Sankararaman and Kimmen Sjölander. INTREPID–INformation-theoretic TREe traversal for Protein functional site IDentification. *Bioinformatics*, 24(21):2445–2452, 2008.

[Sankararaman *et al.*, 2008] Sriram Sankararaman, Srinath Sridhar, Gad Kimmel, and Eran Haplerin. Estimating local ancestry in admixed populations. *American Journal of Human Genetics*, 8(2):290–303, 2008.

[Sankararaman *et al.*, 2009] Sriram Sankararaman, Guillaume Obozinski, Michael I Jordan, and Eran Halperin. Genomic privacy and limits of individual detection in a pool. *Nature Genetics*, 41:965 – 967, 2009.

[Segal *et al.*, 2003] M.R. Segal, K.D. Dahlquist, and B.R. Conklin. Regression approaches for microarray data analysis. *J. Comput. Biol.*, 10:961–980, 2003.

[Setakis *et al.*, 2006] Efrosini Setakis, Heide Stirnadel, and David J. Balding. Logistic regression protects against population structure in genetic association studies. *Genome Res.*, 16(2):290–296, February 2006.

[Shenkin *et al.*, 1991] P.S. Shenkin, B. Erman, and L.D. Mastrandrea. Information-theoretical entropy as a measure of sequence variability. *Proteins*, 11:297–313, 1991.

[Shevade and Keerthi, 2003] S.K. Shevade and S.S. Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19:2246–2253, Nov 2003.

[Shriver *et al.*, 1997] M. D. Shriver, M. W. Smith, L. Jin, A. Marcini, J. M. Akey, R. Deka, and R. E. Ferrell. Ethnic-affiliation estimation by use of population-specific dna markers. *Am J Hum Genet*, 60(4):957–964, April 1997.

[Smith *et al.*, 2004] M. W. Smith, N. Patterson, J. A. Lautenberger, A. L. Truelove, G. J. McDonald, A. Waliszewska, B. D. Kessing, M. J. Malasky, C. Scafe, E. Le, et al. A high-density admixture map for disease gene discovery in african americans. *Am J Hum Genet*, 74(5):1001–1013, May 2004.

[Songyang *et al.*, 1993] Z. Songyang, S.E. Shoelson, M. Chaudhuri, G. Gish, T. Pawson, W.G. Haser, F. King, T. Roberts, S. Ratnofsky, and R.J. Lechleider. SH2 domains recognize specific phosphopeptide sequences. *Cell*, 72:767–778, Mar 1993.

[Southworth *et al.*, 2000] M.W. Southworth, J. Benner, and F.B. Perler. An alternative protein splicing mechanism for inteins lacking an N-terminal nucleophile. *EMBO J.*, 19:5019–5026, Sep 2000.

[Sridhar *et al.*, 2007] Srinath Sridhar, Satish Rao, and Eran Halperin. An efficient and accurate graph-based method to detect population substructure. *Proceedings of Research in Computational Molecular Biology (RECOMB)*, 2007.

[Tang *et al.*, 2006] H. Tang, M. Coram, P. Wang, X. Zhu, and N. Risch. Reconstructing genetic ancestry blocks in admixed individuals. *Am J Hum Genet*, 79(1):1–12, July 2006.

[Tian *et al.*, 2006] C. Tian, D. A. Hinds, R. Shigeta, R. Kittles, D. G. Ballinger, and M. F. Seldin. A genomewide single-nucleotide-polymorphism panel with high ancestry information for african american admixture mapping. *Am J Hum Genet*, 79(4):640–649, October 2006.

[Tian *et al.*, 2007] Chao Tian, David A. Hinds, Russell Shigeta, Sharon G. Adler, Annette Lee, Madeleine V. Pahl, Gabriel Silva, John W. Belmont, Robert L. Hanson, William C. Knowler, et al. A genome-wide snp panel for mexican american admixture mapping. *Am J Hum Genet*, 80(6), 2007.

[Tibshirani, 1996] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. B Stat. Meth.*, 58:267–288, 1996.

[Todd *et al.*, 2001] Annabel E. Todd, Christine A. Orengo, and Janet M. Thornton. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.*, 307(4):1113–1143, April 2001.

[Todd *et al.*, 2002] A.E. Todd, C.A. Orengo, and J.M. Thornton. Plasticity of enzyme active sites. *Trends Biochem. Sci.*, 27:419–426, Aug 2002.

[Tong *et al.*, 2008] Wenxu Tong, Ronald J. Williams, Ying Wei, Leonel F. Murga, Jaeju Ko, and Mary Jo Ondrechen. Enhanced performance in prediction of protein active sites with THEMATICS and support vector machines. *Protein Sci.*, 17(2):333–341, 2008.

[van de Geer, 2008] Sara A van de Geer. High-dimensional generalized linear models and the lasso. *Ann. Stat.*, 36(2):614–645, 2008.

[Waksman *et al.*, 1993] Gabriel Waksman, Steven E. Shoelson, Nalin Pant, David Cowburn, and John Kuriyan. Binding of a high affinity phosphotyrosyl peptide to the Src SH2 domain: Crystal structures of the complexed and peptide-free forms. *Cell*, 72(5):779–790, March 1993.

[Wang and Samudrala, 2006] Kai Wang and Ram Samudrala. Incorporating background frequency improves entropy-based residue conservation measures. *BMC Bioinformatics*, 7(1):385, 2006.

[Wang *et al.*, 1998] David G. Wang, Jian-Bing Fan, Chia-Jen Siao, Anthony Berno, Peter Young, Ron Sapolsky, Ghassan Ghandour, Nancy Perkins, Ellen Winchester, Jessica Spencer, Leonid Kruglyak, Lincoln Stein, Linda Hsie, Thodoros Topaloglou, Earl Hubbell, Elizabeth Robinson, Michael Mittmann, Macdonald S. Morris, Naiping Shen, Dan Kilburn, John Rioux, Chad Nusbaum, Steve Rozen, Thomas J. Hudson, Robert Lipshutz, Mark Chee, and Eric S. Lander. Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. *Science*, 280(5366):1077–1082, 1998.

[Wang *et al.*, 2006] Yi Wang, Yue Li, and Honggao Yan. Mechanism of dihydroneopterin aldolase: functional roles of the conserved active site glutamate and lysine residues. *Biochemistry*, 45(51):15232–15239, 2006.

[Youn *et al.*, 2007] Eunseog Youn, Brandon Peters, Predrag Radivojac, and Sean D. Mooney. Evaluation of features for catalytic residue prediction in novel folds. *Protein Sci.*, 16(2):216–226, 2007.

[Zeggini *et al.*, 2008] E. Zeggini, L. J. Scott, R. Saxena, B. F. Voight, J. L. Marchini, T. Hu, P. I. de Bakker, G. R. Abecasis, P. Almgren, G. Andersen, K. Ardlie, K. B. Bostrm, R. N. Bergman, L. L. Bonnycastle, K. Borch-Johnsen, N. P. Burtt, H. Chen, P. S. Chines, M. J. Daly, P. Deodhar, C. J. Ding, A. S. Doney, W. L. Duren, K. S. Elliott, M. R. Erdos, T. M. Frayling, R. M. Freathy, L. Gianniny, H. Grallert, N. Grarup, C. J. Groves, C. Guiducci, T. Hansen, C. Herder, G. A. Hitman, T. E. Hughes, B. Isomaa, A. U. Jackson, T. Jrgensen, A. Kong, K. Kubalanza, F. G. Kuruvilla, J. Kuusisto, C. Langenberg, H. Lango, T. Lauritzen, Y. Li, C. M. Lindgren, V. Lyssenko, A. F. Marvelle, C. Meisinger, K. Midthjell, K. L. Mohlke, M. A. Morken, A. D. Morris, N. Narisu, P. Nilsson, K. R. Owen, C. N. Palmer, F. Payne, J. R. Perry, E. Pettersen, C. Platou, I. Prokopenko, L. Qi, L. Qin, N. W. Rayner, M. Rees, J. J. Roix, A. Sandbaek, B. Shields, M. Sjgren, V. Steinthorsdottir, H. M. Stringham, A. J. Swift, G. Thorleifsson, U. Thorsteinsdottir, N. J. Timpson, T. Tuomi, J. Tuomilehto, M. Walker, R. M. Watanabe, M. N. Weedon, C. J. Willer, T. Illig, K. Hveem, F. B. Hu, M. Laakso, K. Stefansson, O. Pedersen, N. J. Wareham, I. Barroso, A. T. Hattersley, F. S. Collins, L. Groop, M. I. McCarthy, M. Boehnke, and D. Altshuler. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.*, 40:638–645, May 2008.

[Zhao and Yu, 2006] Peng Zhao and Bin Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.

[Zhu *et al.*, 2005a] X. Zhu, A. Luke, R. S. Cooper, T. Quertermous, C. Hanis, T. Mosley, C. C. Gu, H. Tang, D. C. Rao, N. Risch, et al. Admixture mapping for hypertension loci with genome-scan markers. *Nat Genet*, 37(2):177–181, February 2005.

[Zhu *et al.*, 2005b] X. Zhu, A. Luke, R. S. Cooper, T. Quertermous, C. Hanis, T. Mosley, C. C. Gu, H. Tang, D. C. Rao, N. Risch, and A. Weder. Admixture mapping for hypertension loci with genome-scan markers. *Nat. Genet.*, 37:177–181, Feb 2005.

[Ziv and Burchard, 2003] E. Ziv and E. G. Burchard. Human population structure and genetic association studies. *Pharmacogenomics*, 4(4):431–441, July 2003.