Computational methods to study phenotype evolution and feature selection techniques for biological data under evolutionary constraints

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Christina Kratsch

geb. Tusche, aus Eisenach

Düsseldorf, Juli 2014

aus dem Institut für Informatik
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathemathisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent:	Prof. Dr. Alice C. McHardy
Korreferent:	Prof. Dr. Martin J. Lercher

Tag der mündlichen Prüfung:

Selbstständigkeitserklärung

Ich versichere an Eides Statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf erstellt worden ist. Arbeiten Dritter wurden entsprechend zitiert. Diese Dissertation wurde bisher in dieser oder ähnlicher Form noch bei keiner anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Düsseldorf, den ....

(Christina Kratsch)

Statement of authorship

I hereby certify that this dissertation is the result of my own work, and that no other person's work has been used without due acknowledgement. This thesis was created in accordance with the principles of good scientific practice of the Heinrich-Heine-University. It has not been submitted in the same or similar form to other institutions. I have not previously failed a doctoral examination procedure.

To Conny, Klaus, and Horst.

Summary

Viral pathogens like influenza A, or evolutionary diseases like cancer, are responsible for several millions of deaths every year. They pose an ongoing challenge to human health, because they continuously evolve and adapt to circumvent medical treatment and attacks from the human immune system. Computational genomics allows us to analyze the effects of genetic and phenotypic evolution, which can improve our understanding of such pathogens or medical conditions. The study of evolutionary dynamics may allow us to link genotypic to phenotypic evolution, to detect effects of positive selection or adaptation, and to reconstruct or predict changes in phenotypes.

This thesis presents three methods for the joint investigation of genotype and phenotype evolution. *AdaPatch* finds dense patches of residues under positive selection on the surface of a protein, and was used to select amino acid markers of influenza A haemagglutinin of subtypes H1 and H3. *AntiPatch* extends the approach and detects patches of high antigenic (i.e., phenotypic) impact for influenza A haemagglutinin of subtype H3. Both AdaPatch and AntiPatch provided insights into the genetic and antigenic evolution of influenza A, and we will discuss their relevance for vaccine design and disease surveillance in detail. *RidgeRace* reconstructs continuous ancestral character states along a phylogenetic tree with the help of ridge regression, and allows to infer phenotypic rates for single branches.

All three methods infer the evolutionary properties from the underlying data, and can easily be applied to other data sets. They were developed with a focus on influenza A evolution, and belong to a group of methods that take shared similarities into account. Such an approach is of particular importance when studying biomedical data sets of evolutionary closely related samples. We suggest that those methods may also be of help to study cancer data, and discuss example applications.

Zusammenfassung

Virale Krankheitserreger wie Influenza A oder evolutionäre Krankheiten wie Krebs verursachen jedes Jahr Millionen Todesfälle. Derartige Krankheiten stellen eine ganz besondere Herausforderung dar, da sie ständig evolvieren, um Therapieansätzen oder der menschlichen Immunabwehr auszuweichen. Bioinformatische Methoden erlauben es, derartige genotypische und phänotypische Evolution zu studieren, was unser Verständnis dieser Erreger oder Krankheiten deutlich verbessern kann. Detailierte Untersuchungen der evolutionären Dynamiken einer Krankheit machen es möglich, genotypische und phänotypische Entwicklungen zu vernetzen, die Auswirkungen von positiver Selektion zu verstehen, oder phänotypische Veränderungen zu rekonstruieren oder vorherzusagen.

Diese Doktorarbeit präsentiert drei Methoden, welche genotypische und phänotypische Evolution vereinen. *AdaPatch* identifiziert Bereiche unter besonders hoher positiver Selektion auf einem Protein. Die Methode wurde verwendet, um Aminosäuremarker auf der Oberfläche von Influenza A Hämagglutinin vom Subtyp H1 und H3 zu finden. *AntiPatch* erweitert diesen Ansatz und wurde genutzt, um Bereiche mit besonders hohem antigenischen (d.h. phänotypischen) Einfluss für Influenza A Hämagglutinin vom Subtyp H3 zu beschreiben. Sowohl AdaPatch als auch AntiPatch lassen Rückschlüsse auf die genetische und antigenische Evolution von Influenzaviren zu, deren Implikationen für die fortschreitende Beobachtung des Virus und die Impfstoffentwicklung wir detailiert diskutieren. *RidgeRace* ist eine Methode zur Rekonstruktion kontinuierlicher Werte von Vorfahren in einer Phylogenie mit Hilfe von Ridge-Regression, und erlaubt die Inferenz phänotypischer Raten auf einzelnen Baumästen.

Alle drei Methoden untersuchen die evolutionären Eigenschaften der zugrunde liegenden Daten und können leicht auf andere Anwendungsfälle angepasst werden. Sie wurden mit einem Fokus auf Influenzaviren entwickelt; wir zeigen jedoch, dass sie Teil einer Gruppe von Methoden sind, welche strukturelle Ähnlichkeiten zwischen Datenpunkten beachten und modellieren. Eine derartige Vorgehensweise kann insbesondere notwendig

sein, wenn biomedizinische Daten von evolutionär verwandten Datenpunkten zu untersuchen sind. Wir argumentieren, dass derartige Methoden deshalb auch für das Studium von Krebsdaten geeignet sein können, und diskutieren Beispiele dafür.

Acknowledgements

During the years dedicated to the creation of this thesis, I had the wonderful opportunity to meet many smart and motivated people. Some of them had a significant impact on this project, and I want to thank them for their help. I'd like to thank Alice McHardy for the guidance and motivation to start a thesis in a scientific field that seemed very exotic in the beginning, for the supervision and encouragement to struggle with the scientific challenges, and for the freedom to pursue my own interests and ideas. I thank Sebastian Konietzny for patiently explaining to me the very basics of biology and genomics (and the Bio-Plugin™, of course!), and my colleagues at the MPII Saarbrücken for their input and help. I want to thank Lars Steinbrück, for proof-reading this thesis, for the hours he spent to explain basic bioinformatics and phylogenetics to me, and for all the helpful discussions that improved this work. Also, his antigenic trees method had a considerable impact on this thesis. Finally, every person associated with the Algbio group helped me solving some smaller or bigger problems in some way or another, and I really appreciate the constructive atmosphere both in Saarbrücken and in Düsseldorf. I also like to thank my colleagues at the children's hospital in Düsseldorf, for allowing me to learn about the exciting medical and technical challenges in a new field.

I thank my husband for his interest in my work, his patience with my doubts and occasional frustration, and his ongoing encouragement to finish this work. And, at last, I'd like to thank my parents for always being entirely convinced that I can achieve anything I want in my life. Of course, any decent computer scientist knows that this is strictly impossible (at least from a rather theoretical point of view). Nevertheless, their encouragement has left an invaluable impact on my life.

Contents

I Introduction and Background **1**

1 Introduction **3**

2 Background on the study of evolution **9**

2.1 Natural selection 9

2.2 Phylogenetic techniques 12

2.3 Phylodynamic techniques 15

II Studying influenza evolution **17**

3 Introduction to influenza A viruses **19**

3.1 Epidemiology of influenza A viruses 19

3.2 Evolutionary dynamics of influenza A 21

3.3 Methods to study positive selection in influenza 24

3.4 Methods to study the antigenic evolution of influenza 27

4 AdaPatch for the detection of patches of sites under positive selection 31

4.1 Introduction 31

4.2 Haemagglutinin of seasonal influenza A, subtypes H1 and H3 32

4.3 HA and PB2 of the 2009 A/H1N1pdm influenza 37

4.4 Conclusions on the identified patches 38

4.5 Technical details of the analysis 39

5 AntiPatch for the detection of patches of sites of antigenic impact 45

5.1 Introduction 45

5.2 Haemagglutinin of seasonal influenza A, subtype H3 46

5.3 Conclusions on the identified patches 52

5.4 Technical details of the analysis 55

5.5 Comment on epistasis in influenza A haemagglutinin 59

III Studying phenotype evolution 63

6 Phenotypic evolution and ancestral character state reconstruction 65

6.1 Discrete characters 66

6.2 Continuous characters 67

7 Feature Selection in biomedical data under population constraints 71

7.1 Feature selection and bio-marker detection 72

7.2 Feature selection methods that include phylogenies 78

8 RidgeRace for ancestral character state reconstruction and inference of phenotypic rates 81

8.1 Introduction 81

8.2 Evaluation with simulated data 83

8.3 Exemplary application on a thaumarchaeota data set 85

8.4 Example application to cancer data 87

8.5 Technical details of the method 91

IV Synopsis and Outlook **97**

V Appendix **103**

A Projects involved in this thesis **105**

B AdaPatch: Supplement **109**

C AntiPatch: Supplement **111**

D RidgeRace: Supplement **117**

E Additional source code **121**

List of Figures **127**

List of Tables **129**

References **131**

Index **157**

Part I

Introduction and Background

CHAPTER 1

Introduction

Natural selection is the process that determines whether an allele or a biological trait in a population gradually becomes more or less frequent, based on its fitness advantage over competitors. It has shaped the face of our planet, and the theory of selection forms the basis of all modern understanding of biology. The introduction of the principle by Darwin, Lamarck, and others, and its refinement in later years, had a profound impact on our understanding of the origin and development of present-day species, on cell and molecular biology, and even on social and philosophical theories.

An understanding of the principles of natural selection is also of great importance for the study of diseases and pathogens affecting human health. This might be the case either because a disease or pathogen itself is able to change rapidly and thus evade the human immune system or attempts at therapy or prevention. Or it might be relevant because knowledge of the evolutionary history of the affected cell populations or the pathogens is required to unravel the genetic factors causing the disease. The

following three examples shall illustrate such cases and demonstrate the importance of such knowledge:

Influenza is a disease caused by a viral pathogen that circulates word-wide, causing up to five million cases of severe illness, and up to half a million deaths every year (WHO, 2009). Due to its great impact on global human health, it is one of the best-studied and monitored pathogens of our time. Of greatest interest are genomic changes in the viral genome associated with pathogenicity and transmissibility (McHardy and Adams, 2009). The virus mutates rapidly and thus continuously evades immunity of the human host population, which was acquired by previous infections or vaccination efforts. Understanding the complex evolutionary dynamics of the virus helps to identify these genetic changes relevant for the adaptation of the virus.

Cancer is a very diverse genomic disease that can affect any part of the body, and is responsible for more than seven million disease-related deaths every year (WHO, 2013a). It has been classified as an "evolutionary disease", since it is caused by initial genomic aberrations in single cells that result in uncontrolled proliferation, leading to the growth of tumors, and, potentially, metastases. During the course of the disease, additional genomic changes may be acquired by cells in the tumor, increasing its fitness (*i.e.*, ability for cell proliferation) to further adapt to the tissue environment or to evade therapy and attacks by the immune system. A *stratification* of cancer samples, *i.e.* an identification of functionally similar subtypes of a cancer, can help to develop and optimize individual treatments (TCGAN, 2008; Riester *et al.*, 2010). It is also of of essential importance to discriminate between functionally relevant *driver events* (mutations, copy number variations, genome rearrangements) that truly favor massive growth of a tumor, from functionally neutral *passenger events*, that are just distributed jointly with the drivers due to selective sweeps (Haber and Settleman, 2007; Illingworth and Mustonen, 2011).

Genome-wide association studies (GWAS) offer a third example (Hirschhorn and Daly, 2005; WTCCC, 2007; McCarthy *et al.*, 2008). These studies compare the presence or absence of single mutations in the human genome between a group of healthy and a group of affected individuals, and try to identify disease-related changes. An obvious

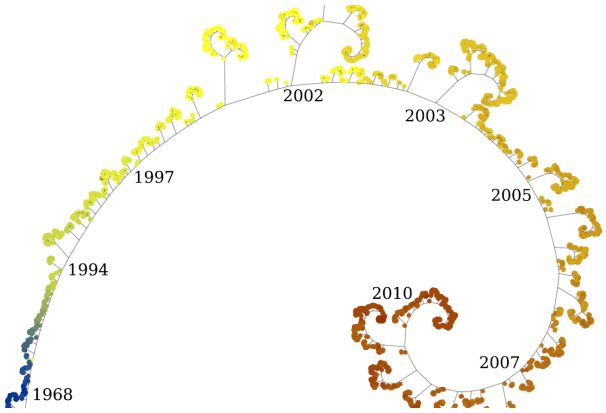


Figure 1.1: Example for a phylogenetic tree: Cytoscape plot (Shannon *et al.*, 2003) of all influenza A/H3N2 haemagglutinin sequences used in the AdaPatch article (Tusche *et al.*, 2012).

strategy might be to look for genetic aberrations present in the affected and missing in the healthy patients. However, such straightforward statistical tests, as they were applied in early GWAS experiments, can overlook the true disease-associated variants and might instead "learn" other discriminating features between target groups such as their geographical or ethnic origin. Shortly speaking, they fail to account for population stratification (see Balding (2006) for a tutorial on stratification and Price *et al.* (2010) for more recent developments).

All three examples share the necessity to consider the origin of the data as well as similarities between data samples that are caused by *shared inheritance*. Knowledge of the genetic (dis-)similarities between samples due to shared history, annotated with clinical or epidemiological information, can help to identify the geographical origin of a pathogen or genetic changes associated to a phenotype. Compared to the relatively slow process of human inheritance relevant for GWAS, the fingerprint of evolution is much more dominant in a typical sample of the rapidly evolving genomes of influenza viruses or cancer cells (see Nobusawa and Sato (2006) for details on influenza and Yates and Campbell (2012) for a discussion of human genome mutation rates elevated by cancer). Studying viral or cancer evolution therefore offers particularly promising examples to gain insights into the evolutionary history of genomic data.

This thesis presents methods to study viral and, in particular, human influenza virus evolution. It suggests ways to select phenotypically relevant genetic features and to

reconstruct ancestral phenotypes. *AdaPatch* (Tusche *et al.*, 2012) finds dense patches of residues under positive selection on the surface of a protein, and was applied to the influenza A protein haemagglutinin of subtypes H1 and H3. *AntiPatch* (Kratsch *et al.*, 2014, *in prep.*) extends the approach and detects patches of high antigenic impact for influenza A haemagglutinin of subtype H3. Both methods use a *phylogenetic tree* of the haemagglutinin sequences, *i.e.* a special tree-shaped clustering that describes a reconstruction of evolutionary relatedness for the strains in the data (see Figure 1.1 for an example). The identified markers are important to understand how influenza evades the human immune response and to predict the further development of the viral population, so that suitable vaccines can be produced in time.

In addition to AdaPatch and AntiPatch, we present a method to reconstruct *ancestral character states* of an arbitrary continuous phenotype for a collection of samples with shared inheritance. *RidgeRace* (Kratsch and McHardy, 2014, *in prep.*) extracts branch lengths from a phylogenetic tree and combines them with a ridge regression on the phenotypic leaf values to reconstruct ancestral phenotype values without requirement of an underlying model of phenotype evolution. It also provides estimates of the rate of phenotypic changes that can be used to judge the phenotypic impact of amino acid changes observed on the branches of the phylogeny.

All three methods include a reconstruction of the evolutionary history of the underlying data as guidance for improved feature selection or phenotype prediction. We suggest to consider such phylogeny-based methods from a broader perspective of statistical techniques that account for the non-independence of data. They have the potential to extract additional information, and to avoid statistical distortions. Finally, we argue that cancer as an evolutionary process (Merlo *et al.*, 2006) is a disease that offers many applications for such methods, and that future phylogeny-based studies of suitable cancer data might reveal important details on key players in the disease progression. We discuss the strengths and challenges of this approach with an application of RidgeRace to a data set of ovarian cancer samples, where we tried to detect somatic mutations associated with an increased patient survival time.

This thesis is organized into four parts: the first part comprises this introduction and a review of the principles of evolution and natural selection (Chapter 2). The second part is concerned with the evolutionary dynamics of influenza A viruses. It briefly reviews our current understanding of influenza epidemiology and evolution, and discusses methods to study the genetic and antigenic evolution of the virus (Chapter 3). It then presents AdaPatch (Chapter 4) and AntiPatch (Chapter 5). The third part is concerned with general methods to reconstruct and explain phenotype evolution. It briefly reviews methods for the reconstruction of discrete and continuous ancestral characters (Chapter 6), discusses problems with the selection of suitable phenotype-associated features (Chapter 7), and presents RidgeRace together with several applications (Chapter 8). The fourth part summarizes the contributions of this work and provides a short outlook.

CHAPTER 2

Background on the study of evolution

2.1 Natural selection

Evolution is "*the process by which different kinds of living organism are believed to have developed from earlier forms during the history of the earth*".[1] *Natural selection* is considered to be the basic driver of evolution, and is based on three simple observations: first, that there is variability in the survival and reproduction rates of (sub-)species, that, second, this variability, also described as "fitness", is dependent on the present environmental conditions, and that, third, part of this variability is heritable (Hurst, 2009). *Adaptation* refers to both the process and the result of change by which an organism or species improves its ability to live in its environment, improving fitness (Hurst, 2009; Dobzhansky, 1986). Although the principles of natural selection, first described by Darwin (1859), are well understood today, many questions remain in

[1] Oxford Dictionary, *http://oxforddictionaries.com/definition/english/evolution*, accessed 09/08/2013

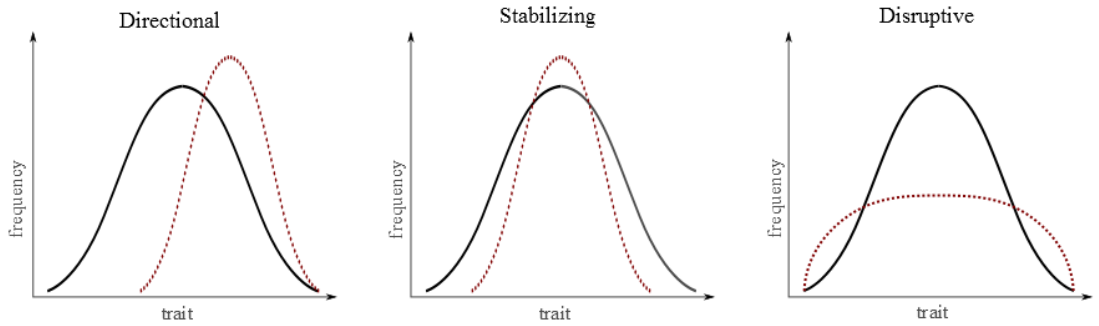


Figure 2.1: The three main modes of selection. The graphs show the relative frequency of a continuous trait before (solid black lines) and after the impact of selection (dotted red lines). Figure redrawn from Brodie *et al.* (1995), with permission from Elsevier.

open debate. Some are concerned with alleles that have only a small impact on fitness, selection acting on gene order or on silent mutations, the spread of deleterious variants, and the degree of influence of neutral genetic drift on fitness or a phenotype (Hurst, 2009). Barrick and Lenski (2013) review how evolutionary trait changes may occur either gradually or in distinct jumps, how mutation rates are influenced by genetic and environmental factors, and how population dynamics may influence the frequencies of beneficial alleles.

Several types of selection exist (Brodie *et al.*, 1995; Hurst, 2009):

Positive selection can also be termed "Darwinian selection", "selection for change", or, in the case of a quantitative phenotype, "directional selection". It refers to the increase in frequency of alleles in a population that confer fitness advantages to the individuals carrying them. Since a change in phenotype typically requires amino acid changes (instead of silent mutations), positive selection is thought to be acting on a genetic region when the ratio of non-synonymous to synonymous mutations in that region is larger than one (Yang and Bielawski, 2000), although smaller values can be sufficient when considering members of the same population (Kryazhimskiy and Plotkin, 2008).

Purifying selection is also known as negative or stabilizing selection. It refers to the process that maintains beneficial, well-adapted alleles in a population, re-

ducing variation. It is believed to be the most common mode of selection (see argumentation in Hurst, 2009, box 2).

Disruptive selection describes selection that favors or even increases diversity. It is very often equated with balancing selection, which maintains multiple phenotypes at the same frequency, "balancing" their relative ratio. An example is the case of a heterozygous locus in a haemoglobin gene associated with sickle cell anaemia, which also offers resistance to malaria (Allison, 1956). Other forms of disruptive selection can be the result of over-dominance, or of iterations of positive selection for new allelic variants in a host-pathogen-co-evolution (see Hurst (2009) for details).

Positive selection is the type most relevant for this thesis. As will be described in more detail in Chapter 3, influenza viruses co-evolve with their hosts. They are forced to change constantly, since protective antibodies of the host recognize and attack virions with surface structures similar to viruses from previous infections. Therefore, viral strains accumulating amino acid changes on their surface proteins (*antigens*) haemagglutinin (HA) and neuraminidase (NA) may have a selective advantage.

Figure 1.1 presents a Maximum Likelihood reconstruction (see Section 2.2) of the phylogeny of influenza A haemagglutinin sequences of subtype H3. It shows a "cactus-like" structure, which is typical for the continuous accumulation of changes in this gene: a surviving lineage of single dominant strains, the "trunk" of the tree, stretches from the earliest isolate towards present strains, with relatively short side branches of quickly extinct strains (Fitch *et al.*, 1997; Holmes *et al.*, 2005). Only strains on the trunk of the tree were able to acquire changes that rose to predominance in the viral population, replacing all other variants. The coloring of the nodes by year is a very simple application of phylodynamic methods (Section 2.3), and clearly shows that new dominant viral strain arise by accumulating only few amino acid changes that distinguish them from their predecessors (for a detailed discussion, see Nelson and Holmes, 2007).

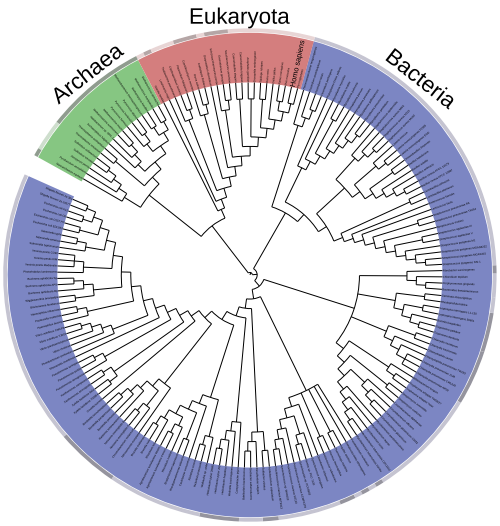


Figure 2.2: Example for a phylogenetic tree: a phylogenetic representation of the genetic relationships between the main sequenced phyla for all species on earth, created with the Tree Of Life Web project (Letunic and Bork, 2007, 2011).

2.2 Phylogenetic techniques for the visualization and reconstruction of evolution

The evolutionary relationships between species or members of a population are typically analyzed using phylogenetic inference (Felsenstein, 2004; Yang and Rannala, 2012). A *phylogeny* describes the inferred evolutionary history of a set of biological samples, with inner nodes representing speciation events or branching events within a population. Before the rise of genomic sequencing, phylogenies were created based on morphological, molecular or developmental traits (Nunn, 2011, chapter 2), while recent studies are typically based on suitable genetic data. Bininda-Emonds *et al.* (2007, 2008) have used phylogenetics to infer a "supertree" of mammals, and Figure 2.2 shows a phylogenetic representation of the genetic relationships between the main sequenced phyla on earth (Letunic and Bork, 2007, 2011). Note, however, that there is emerging evidence that this representation might be partially wrong (Williams *et al.*, 2013). Phylogenetic techniques are now an essential tool in biology and genomics, and used in a broad range of applications, Yang and Rannala (2012) offer a detailed list of possible examples.

The methodology used to infer phylogenies can broadly be categorized into four different groups:

Parsimony methods are character based, with characters typically representing nucleotides or amino acids. To score a given topology, maximum parsimony methods assign character states to its inner nodes in a way that minimizes the number of changes (discrete characters) or the difference (continuous characters) between neighboring (=related) nodes over all sites. The maximum parsimony tree is the topology that minimizes that tree score, and branch lengths are proportional to the reconstructed numbers of changes on them.

Distance matrix methods use matrices describing the pairwise distances between samples as input for a clustering. The most widely used method is the Neighbor Joining algorithm (Saitou and Nei, 1987). Distances between genomic sequences are usually computed by counting the pairwise differences (mutations, insertions, deletions). An appropriate model of evolution is then used to transform the observed differences into evolutionary distances.

Maximum Likelihood methods (ML methods) assign a likelihood to a given tree based on how well the topology and branch lengths fit the data under a predefined substitution model. For nucleotide sequences, these are continuous-time Markov models describing the single nucleotide frequencies and the substitution rates between nucleotide states. To find the maximum likelihood tree, one has to perform a heuristic search through the tree space and compute the likelihood for each candidate tree state consisting of the actual topology and the lengths of its branches. For a defined tree state, estimation is typically performed using Felsenstein's pruning algorithm (Felsenstein, 1981), and branch lengths can be further optimized with a Newton-Raphson approach (Felsenstein, 2004, chapter 16).

Bayesian methods extend the Maximum Likelihood approach. Whereas ML methods consider parameters like the substitution model or the tree topology as constant, Bayesian methods allow these parameters to follow a predefined prior distribution and estimate the posterior distribution of parameters and data, typically derived with Markov chain Monte Carlo algorithms.

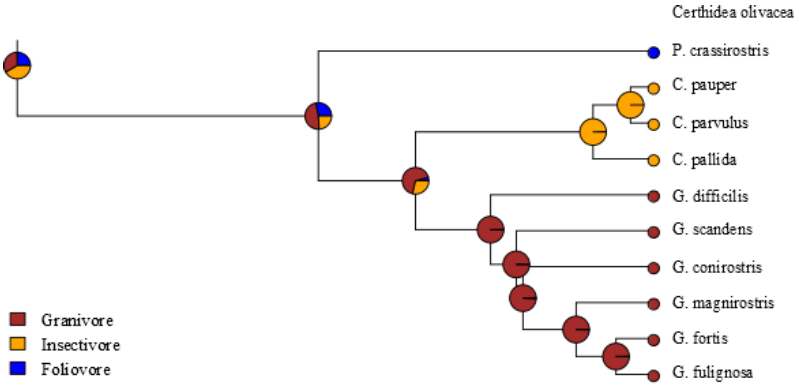


Figure 2.3: An exemplary phylogeny of 11 Galapagos finches, and a reconstruction of their diet. Example redrawn from Schluter *et al.* (1997), with a phylogeny downloaded from R-Sig-Phylo (2013). The authors inferred the tree with a distance matrix method, and we performed ancestral character state reconstruction using a maximum likelihood method with the ape function in R (Paradis *et al.*, 2004). The three finch groups *Geospiza*, *Camarhynchus-Platyspiza*, and *Certhidea* cluster separately, but the reconstruction of the diet of their ancestor is rather unstable (see discussion by Schluter *et al.*).

Phylogenetic techniques usually make assumptions on character evolution and compute at least indirect assignments of characters to ancestral nodes, which are required to estimate the quality of a topology. However, the reconstruction methods can also be used to derive interesting conclusions on the history of the data (see an example in Figure 2.3). Many techniques exist for discrete and continuous character reconstruction, we provide an overview in Chapter 6.

The four categories of phylogenetic techniques only broadly describe the multitude of available algorithms and software packages. The categories differ strongly in their complexity, interpretation, their need for model assumptions, and their ability to cope with practical problems, such as long branch attraction or missing data. A detailed discussion of these aspects and a comparison of the four categories can be found in Yang and Rannala (2012), and detailed explanations in Felsenstein (2004). In an older yet still widely cited review, Felsenstein (1996) discusses the adaptation of the first three categories to amino acid sequences.

2.3 Phylodynamic techniques that combine genetic and phenotype evolution

Phylodynamic techniques can be considered as an extension of phylogenetics, and were first discussed by Grenfell *et al.* (2004). With a focus on RNA viruses, the authors defined the purpose of phylodynamic methods "*to clarify how pathogen genetic variation, modulated by host immunity, transmission bottlenecks, and epidemic dynamics, determines the wide variety of pathogen phylogenies observed at scales that range from individual host to population*". Phylodynamic techniques combine the inference of phylogenies from genetic data with epidemiological annotations to gain insight into the dynamics of the underlying epidemiological process. A typical example is phylogeography, where sampling locations of samples are mapped and reconstructed on the tree to infer geographic migration patterns (Wallace *et al.*, 2007), or phylogenetic dating, where sampling times (or dating information available from fossil records) are mapped to the tree to date past events (Drummond *et al.*, 2006).

In general, phylodynamic techniques offer two advantages:

1. they naturally integrate genetic and epidemiological (or phenotypic) information, *e.g.* by representing genetic similarities in a phylogeny and mapping other data to the branches or inner nodes. Thus, they allow to draw conclusions on the connection between the different data types.

2. they take the statistical relatedness between samples due to shared inheritance into account. As will be discussed in Chapter 7, non-independentness of samples can be a source of both statistical bias or relevant information.

To track the antigenic evolution of influenza A viruses, Steinbrück and McHardy (2012) suggest *antigenic trees* to map antigenic distances to a phylogeny inferred on the involved strains (see also Section 3.4 and Chapter 5). The authors use such a mapping to compute antigenic weights for individual branches, and infer the antigenic weight of individual amino acid changes of influenza A haemagglutinin of subtype H3. In

the present work, we have extended this approach with our AntiPatch tool to further filter functionally neutral hitchhiker changes that occur on the same branches as those with truly high antigenic impact by clustering sites of high antigenic weight on the haemagglutinin protein structure. With the help of our RidgeRace tool, it is possible to extend the Antigenic Tree approach to arbitrary continuous (or ordered) phenotypes, and to reconstruct ancestral phenotypic values together with the phenotypic rate (or impact) for individual branches.

Part II

Studying influenza evolution

CHAPTER 3

Introduction to influenza A viruses

3.1 Epidemiology of influenza A viruses

Influenza is a viral disease of birds and mammals (Medina and García-Sastre, 2011), caused by a single-stranded, segmented, negative-sense RNA virus of the family *Orthomyxoviridae*. Three genera, influenza A, B, and C, exist, and all are present in the human host, circulating worldwide and in any age group (WHO, 2009). Seasonal influenza (types A and B) is responsible for the majority of infections, and poses a severe health risk for the human population, causing up to five million cases of severe illness, and up to half a million deaths every year (WHO, 2009). According to a study by Molinari *et al.* (2007), average annual influenza epidemics result in ca. 36,000 deaths and 200,000 hospitalizations in the United States alone, costing nearly 87 billion USD.

The influenza A genome consists of eight segments which encode up to fourteen proteins. Eleven of those (PB2, PB1, PA, HA, NP, NA, M1, M2, NS1, NS2, PB1-F2)

have been known for a long time (Webster *et al.*, 1992; Das *et al.*, 2010), whereas three additional proteins (PB1-N40, M42, PA-X) are the result of more recent research (Wise *et al.*, 2009, 2012; Jagger *et al.*, 2012).

The surface glycoproteins haemagglutinin (HA) and neuraminidase (NA) are responsible for the binding to the target cell and viral entry into that cell (HA), or the release of progeny from the infected cell (NA). Both are *antigens* recognized by the host immune system, and are thus the two proteins of the highest importance for influenza surveillance and vaccine design (Dormitzer *et al.*, 2011). Influenza A viruses are also classified based on their serotype, *i.e.* their combination of HA and NA proteins, for which 18 and 11 variants, respectively, exist in wild water birds or mammals (Webster *et al.*, 1992; Medina and García-Sastre, 2011). The most recent subtypes (H17, H18, N10, N11) have so far only been observed in bats (Tong *et al.*, 2012, 2013).

At present, most human infections are caused by endemic influenza viruses of subtypes H1N1 and H3N2. The swine-origin H1N1 influenza (*swine flu*, also called influenza A/H1N1pdm), is a recent replacement of the previously circulating H1N1 subtype in annual epidemics (Garten *et al.*, 2009). The transmission between humans occurs via droplet-infection through the air or direct contact (WHO, 2009). However, sporadic infections with subtypes H5N1, H7N9 or H9N2, as well as a new variant of H3N2 (labeled H3N2v, CDC, 2013b), have been observed in recent years (Belser *et al.*, 2009; CDC, 2013a; WHO, 2014), and were attributed to close contact with infected animals. Sporadic transmissions breaching the host-species-barrier may occur on animal farms or markets where humans live or work in close contact to potentially infected animals, in particular poultry and swine (Kuiken *et al.*, 2006). Within the human host population, influenza infection rates are particularly high during winter in the northern or southern hemisphere, and year-round, but modest in the tropics (Nelson and Holmes, 2007). Infections are usually mild, but certain groups (young children, elderly people, pregnant women, persons suffering from other medical conditions) are at risk for a more severe or even fatal course of disease.

3.2 Evolutionary dynamics of influenza A

Note: this section contains material from the introduction of Tusche et al. *(2012). Minor text modifications were made to improve readability and to fit the text into the structure of this work. See Chapter A for details on author contributions.*

Influenza A viruses are an example for co-evolution between a viral pathogen and its host. The viral surface proteins HA and NA are recognized by the antibodies of the (here: human) host immune system, leading to immunity of the infected individual against the the viral strain causing the infection, and against antigenically similar strains. Amino acid changes, particularly on the surface of the globular head region of NA and in the epitope and receptor binding regions of HA, increase the chances of a new virus to re-infect a previously immune person (Shih *et al.*, 2007; Cattoli *et al.*, 2011; Sandbulte *et al.*, 2011). The very high mutation rate (Drake, 1993; Nobusawa and Sato, 2006) and the fast air-borne spread via droplet infection allows influenza A to manifest altered strains quickly world-wide and thus to adapt to the newly acquired immunity in its host population (Pybus and Rambaut, 2009). The process of continuous accumulation of changes in the antigens is referred to as *antigenic drift*, and the co-evolutionary race between the acquired defense of the host immune system and the viral evasion by advantageous mutations is an example for the *Red Queen hypothesis* (Hurst, 2009). This analogy was introduced by van Valen, who described it as "*a race in which each species is competing in a zero-sum game against others; each game is a dynamic equilibrium between competing species where no species can ever win and new adversaries grinningly replace the losers*" (Clarke *et al.* 1994, citing van Valen 1973). Van Valen referred to an explanation of the red queen of hearts in Lewis Caroll's "Through the Looking-Glass": "*Now, here, you see, it takes all the running you can do, to keep in the same place*". The Red Queen dynamic leads to the aforementioned cactus-like phylogeny of influenza A/HA sequences (Figure 1.1). It shows a single trunk with the amino acid changes of the continuously adapting strains that became dominant, and displays relatively short side branches of strains of lower fitness (*i.e.*, lower antigenic divergence) (Fitch *et al.*, 1997; Holmes *et al.*, 2005).

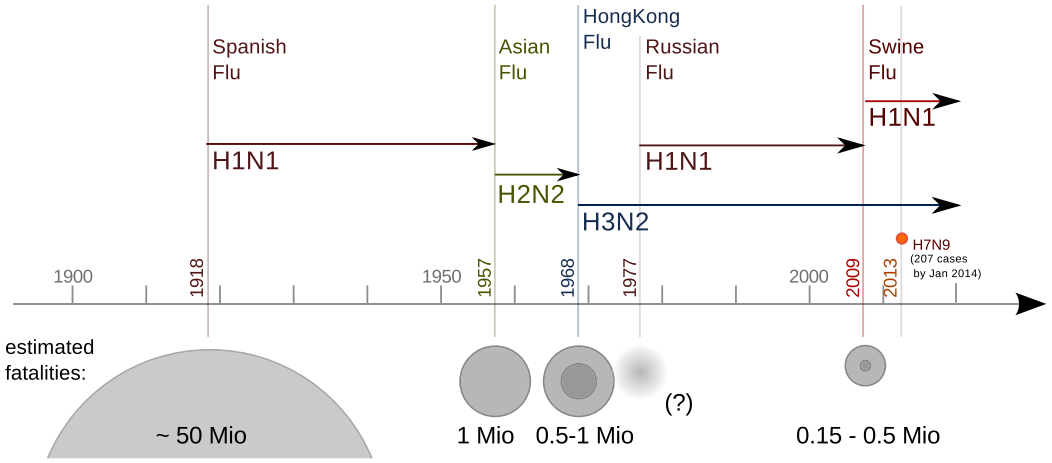


Figure 3.1: Influenza pandemics of the current and previous centuries. Figure adapted from McHardy (2013), fatality estimates according to Kilbourne (2006), Dawood *et al.* (2012), and WHO (2013b, 2014). No fatality estimate was available for the 1977 Russian Flu.

Configurational changes of multiple proteins of animal influenza A viruses are thought to be necessary to enable an efficient replication and transmission in human hosts (Kuiken *et al.*, 2006; Neumann and Kawaoka, 2006). A region of particular importance for this process is the receptor-binding site of the viral haemagglutinin. It enables attachment to different types of host-specific glycosydic bonds on surface epithelial cells in the host respiratory and gastrointestinal tracts (Glaser *et al.*, 2005; Neumann and Kawaoka, 2006). Furthermore, certain areas of the viral polymerase complex determine host range (Neumann and Kawaoka, 2006; Yamada *et al.*, 2010). Following the establishment of a virus within a novel host, additional adaptive changes are thought to optimize viral replication and rapid dispersal within the host population (Deem and Pan, 2009; Hensley *et al.*, 2009; Neumann *et al.*, 2009; Smith *et al.*, 2009).

Two additional mechanisms affect influenza adaptation: genome reassortment and receptor avidity. Since the viral genome is encoded on eight separate RNA segments, two or more viruses co-infecting the same cell can produce chimeric descendants with new combinations of their genome segments, possibly allowing the descendant to infect a different host.

A reassortment of viral genomes that leads to the generation of a new subtype may result in a dramatic change in antigenic potential if it leads to an alteration of the surface proteins (*antigenic shift*, Nelson and Holmes, 2007; Medina and García-Sastre, 2011). Since the host population is much less protected against the reassortant strain, antigenic shift can lead to large epidemic outbreaks or even pandemics. The 2009 swine-origin H1N1 influenza is a reassortant virus that stems from a swine co-infection with an Eurasian H1N1 swine virus (N1 and M segments) and a "classical" Eurasian/ North American swine virus (H1, PB2, PB1, PA, NP and NS segments). The latter virus was itself a triple reassortant of avian, human, and swine viruses (a detailed discussion on the origins of this subtype can be found in Smith *et al.*, 2009).

Similarly, the 1957 and 1968 pandemics have been the result of reassortment, leading to antigenic shifts and resulting in pandemics. The origin of the devastating Spanish Flu outbreak in 1918 that killed at least 40 million people is still unclear (Reid *et al.*, 2004). Medina *et al.* (2010) reported that the 2009 swine-origin H1N1 virus is antigenically similar to the 1918 virus, and that vaccination or previous exposure to the 2009 virus elicits cross-protective antibodies against the Spanish Flu. Figure 3.1 provides a schematic overview over all pandemics observed since the 1918 outbreak.

Besides reassortment, a second potential influence on the dynamics of human influenza evolution was recently described. Hensley *et al.* (2009) argued that, in an immune host, an increase of receptor binding strength offers a selective advantage for a virus, mainly achieved by increasing the number of positively charged amino acids on the viral surface close to the haemagglutinin receptor binding site. The authors suggest that *receptor avidity* might be the main phenotype under positive selection, and antigenic drift merely a side-effect. Yuan and Koelle (2013) used epidemiological models to predict and confirm parameters such as the number of mutations towards positively charged amino acids observed in specific host populations, but the primary phenotypic concepts driving influenza evolution are still under debate.

When a novel antigenic variant with higher fitness appears, it will spread very rapidly in the still weakly protected host population, replacing other variants and thus decreasing

the overall viral diversity. This process is referred to as *selective sweep*. The quick spread of a single strain will strongly increase the frequency of changes that distinguish the new strain from the previously predominant antigenic variant. But not all amino acid changes between antigenic clusters are necessarily relevant for adaptation. Only a few of the non-synonymous substitutions might be of actual phenotypic (=antigenic) relevance, whereas other behave functionally neutral. The latter are referred to as *hitchhikers*. The distinction between functionally relevant and hitchhiker changes is one of the ongoing challenges in influenza research, and a main focus of this work. We developed AdaPatch (Chapter 4) to identify regions under positive selection on the surface of influenza A haemagglutinin, and we extended the approach with AntiPatch (Chapter 5) to identify regions associated with a high antigenic impact. The following sections will review related methods and discuss the results inferred with these tools.

3.3 Methods to study positive selection in influenza

Note: this section contains material combined from the introductions of Tusche et al. *(2012) and Kratsch* et al. *(2014). Paragraphs were combined and minor text modifications were made to improve readability and to fit the text into the structure of this work. See Chapter A for details on author contributions.*

Human influenza A viruses continuously change antigenically. The associated genetic changes are mostly situated in the antibody-binding sites of the viral surface proteins HA and NA (Bush *et al.*, 1999; Smith *et al.*, 2004; Weinstock and Zuccotti, 2009; Steinbrück and McHardy, 2011), and allow a re-infection of previously infected or vaccinated individuals. A precise knowledge of the viral protein regions that are relevant for adaptation to a novel host or an increasingly immune population is therefore a crucial factor for the surveillance and prevention of seasonal and pandemic influenza A virus infections.

Multiple methods exist that search for functional regions of proteins, for example, on the basis of evolutionary conservation ratios (Pupko *et al.*, 2002; Glaser *et al.*, 2003; Nimrod *et al.*, 2005; Shazman *et al.*, 2007; Nimrod *et al.*, 2008; Ashkenazy *et al.*, 2010). However, regions under positive selection do not follow the assumption of

strong conservation made by these methods and can therefore not be detected this way. Other techniques predict the location of antibody-binding (epitope) sites based on structural and sequence information (Blythe and Flower, 2005; El-Manzalawy *et al.*, 2008; Rubinstein *et al.*, 2008, 2009; Lacerda *et al.*, 2010). However, besides epitope regions, receptor avidity changing sites or host-specificity determinants might play a similarly important role for the adaptive evolution of influenza A viruses (Hensley *et al.*, 2009).

Measures of positive selection can indicate sites that are relevant for the adaptation of influenza viruses to altering environmental conditions, which, for human influenza viruses, include an escape from immune recognition by antibodies generated in response to previous infections or vaccinations (Bush *et al.*, 1999; Medina and García-Sastre, 2011). Protein sites with a significantly increased ratio of non-synonymous to synonymous mutations on influenza A haemagglutinin (Fitch *et al.*, 1997; Bush *et al.*, 1999; Suzuki and Gojobori, 1999; Suzuki, 2004a) and regions under positive selection including T- and B-cell epitope sites (Suzuki, 2006) have been described. Sophisticated maximum likelihood approaches have been developed that estimate the degree of positive selection for individual sites (Nielsen and Yang, 1998; Yang, 2000; Kosakovsky Pond *et al.*, 2005), and extensions thereof allow the dN/dS ratio to vary along both sites and lineages (Kosakovsky Pond *et al.*, 2008; Yang and Nielsen, 2002; Nozawa *et al.*, 2009; Yang and dos Reis, 2011; Kosakovsky Pond *et al.*, 2011; Murrell *et al.*, 2012). A detailed evaluation of these methods was recently provided by Lu and Guindon (2013). Application of these techniques to influenza A haemagglutinin has determined overlapping sets of sites under positive selection (Yang, 2000; Kosakovsky Pond *et al.*, 2008; Murrell *et al.*, 2012; Nei, 2005). In addition to the determination of individual sites under selection, pairs of sites where changes might show epistatic interactions can be identified based on evolutionary distances in a phylogenetic tree (Kryazhimskiy *et al.*, 2011). Determination of sites under selection can also be improved by consideration of the protein structure, as the interaction of flexible macromolecules such as human antibodies and the viral surface proteins is likely to include multiple areas of both interacting proteins. Thus, sliding

sphere-shaped windows along the protein structure have been used to identify of sets of neighboring sites under selection (Suzuki, 2004b; Berglund *et al.*, 2005; Zhou *et al.*, 2008). Robinson *et al.* (2003) take the effects of solvent accessibility and pairwise interactions between amino acids into account using evolutionary models. In the AdaPatch method described in Chapter 4, we present an approach similar to the one of Robinson *et al.* (2003), but use a less complex evolutionary model and consider the spatial distribution of residues in a consecutive phase of our algorithm. Our approach does not require specification of a cluster radius, nor does it restrict the geometrical form of the inferred clusters.

In contrast to the discussed methods, we assume that not only mutations at individual sites, but also of multiple sites within a certain region of a protein can cause adaptive protein conformation changes. Shape and charge modifications within larger patches of residues on the protein surface are important for viral adaptation to structural changes in the interacting proteins of the host (see *e.g.*, Yamada *et al.* 2010). We therefore devised AdaPatch to detect dense patches of sites showing a large average positive selection, using dN/dS estimates of positive selection for individual sites, and information on the spatial distances between them. With this approach, we also included sites with a large, but not exceptionally large, dN/dS ratio. Such residues would be discarded by methods that identify top-ranking sites using a measure of selection. With our method, such residues were included if their spatial position supported the continuity of a patch. By searching for clusters of sites that are close to each other in the protein structure and consistently exhibit elevated dN/dS values, one might have greater statistical power to detect adaptive evolution in proteins compared to methods that test for elevated dN/dS ratios at individual sites.

As mentioned above, more advanced techniques can be used for estimating positive selection. We here rely on the dN/dS statistic to allow an easy understanding of the principles of our method, but the statistic can easily be exchanged with other measures. We evaluated our method by applying it to HA data for human influenza A viruses of the subtypes H3N2 and H1N1. These are particularly suited for evaluation, as large

numbers of sequences are available, and their interaction with the human host is very well studied. Additionally, we applied the method to HA and polymerase basic protein 2 (PB2) of swine-origin influenza virus (S-OIV) A/H1N1pdm, to study the more recent development of the virus.

3.4 Methods to study the antigenic evolution of influenza

Note: this section contains material from the introduction of Kratsch et al. *(2014). Minor text modifications were made to improve readability and to fit the text into the structure of this work. See Chapter A for details on author contributions.*

As discussed in Section 3.3, the surface proteins HA and NA are of particular importance for viral evolution and adaptation (McHardy and Adams, 2009). Amino acid changes on the surface of the globular head region of neuraminidase and in the epitope and receptor binding sites of haemagglutinin result in alterations of antigenicity, and reduced recognition by the host immune response (Shih *et al.*, 2007; Cattoli *et al.*, 2011; Sandbulte *et al.*, 2011). Furthermore, sites which directly affect receptor avidity have been described for influenza A/H1N1 viruses, which change as part of viral adaptation to hosts with different levels of immune protection (Hensley *et al.*, 2009). The rapid antigenic drift of influenza A viruses necessitates frequent updates of the seasonal influenza A virus vaccine (Dormitzer *et al.*, 2011) such that it includes antigenically similar strains to the predominant circulating antigenic variant (Russell *et al.*, 2008).

We have argued in Section 3.3 that regions under positive selection on the surface of HA (and, potentially, NA) are likely to be of importance for viral *antigenic evolution* and informative for surveillance of viral diversity and vaccine design. However, genetic changes do only approximately describe the immunological changes during influenza evolution. Despite the continuous accumulation of genetic changes, the antigenic evolution of influenza A has been shown to be punctuated (Smith *et al.*, 2004). Antigenic differences between viral strains can directly be assessed with haemagglutination inhibition (HI) assays, which determine the strength of interaction between a viral isolate and

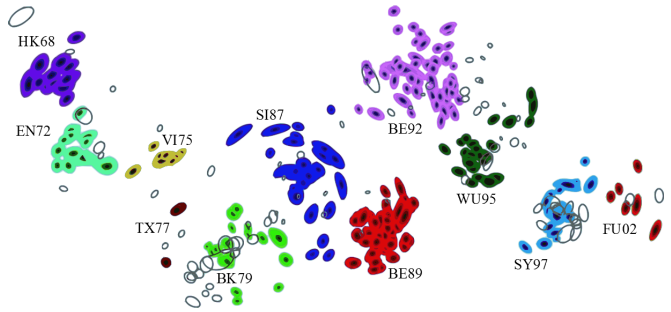


Figure 3.2: Antigenic cartography of influenza A (H3N2) from 1968 to 2003, showing the clustered nature of the antigenic evolution of influenza A. Figure adapted from Smith *et al.* (2004), with permission from the American Association for the Advancement of Science.

an antiserum elicited against another viral strain. The concentration-dependent inhibition of red blood cell agglutination by a viral isolate with an antiserum is determined in a series of dilution steps (Hirst, 1943), and used to define the *antigenic distance* between virus and serum. HI assays are routinely used within the global surveillance program of human influenza viruses of the World Health Organization, as the antigenic novelty of a given strain relative to past predominant strains is a relevant factor for future epidemic potential. Using antigenic distances, Smith *et al.* (2004) performed multidimensional scaling to visualize phenotypic similarities of viruses from 1968 to 2003 on a two-dimensional map. The result of this *antigenic cartography* showed clusters of antigenically similar strains for the same time period (Lapedes and Farber, 2001; Smith *et al.*, 2004, Figure 3.2).

The original definition of the epitopes recognized by antibodies (Wiley and Skehel, 1987) is broad and includes 131 out of 328 sites of the entire HA1 chain of haemagglutinin. It is likely that many of these epitope sites are of low relevance for viral immune evasion. Recent computational studies therefore aimed to identify key antigenicity-altering sites of the epitopes. Information gain (Huang *et al.*, 2009), multivariate linear models (Lee *et al.*, 2007) and similar scoring schemes (Liao *et al.*, 2008, 2013) have been used to estimate the association between amino acid changes and changes in antigenic type. Lees *et al.* (2011) used the genetic variability in "cells" on a three-dimensional grid on

the protein structure, with similar regression models being used to weigh substitutions in pre-selected clusters as predictors for antigenic distances. Sun *et al.* (2013) used ridge regression to infer antigenic weights for amino acid changes. Steinbrück and McHardy (2012) used HI assay data to infer antigenic weights for individual branches of a phylogenetic tree. From this *Antigenic Tree*, the antigenic weight of amino acid changes and the average impact of changes at individual protein sites can be determined. Koel *et al.* (2013) experimentally quantified the antigenic impact of changes at individual and pairs of sites involved in antigenic cluster transitions for HA in H3N2 viruses. They found that seven positions altered in past antigenic cluster transitions have had a significant antigenic impact. However, not all permutations of observed changes could be tested due to the required effort, and the authors note that changes at other sites may have collective effects on antibody binding, or they may be compensatory mutations that are necessary to retain function. Computational methods which link genetic to phenotypic information are not limited to exploring subsets of sites, but can also return predictions that might include *antigenic hitchhiker* changes. Hitchhikers are (near-)neutral changes introduced into a strain shortly before or after an antigenicity-altering change. As the strain then shows a significant change in antigenicity relative to other strains, the contributions of the individual amino acid changes to this antigenic difference cannot be distinguished from one another. Thus antigenic hitchhikers may falsely be determined as being relevant for the antigenic evolution of the virus. Similarly, epistatic effects may lead to a suppression or enhancement of the antigenic impact of individual changes and therefore to problems in determining the most relevant sites.

We extended the antigenic tree approach with AntiPatch (Chapter 5), a method to combine phylogenetic, antigenic and structural information for antigenic tree inference and the detection of patches of antigenicity-altering sites on the three-dimensional structure of haemagglutinin of human influenza A/H3N2 viruses. We identified six patches around the receptor binding site, which include residues of all five epitope regions and contain known relevant residues for the antigenic evolution of the virus.

CHAPTER 4

AdaPatch for the detection of patches of sites under positive selection

Note: this chapter contains material from several sections of Tusche et al. *(2012). Minor text modifications were made to improve readability and to fit the text into the structure of this work. See Chapter A for details on author contributions.*

4.1 Introduction

We have developed AdaPatch to identify patches of sites under selection on the surface of proteins, and used it for the analysis of haemagglutinin proteins of seasonal influenza A (H1 and H3), and for two proteins of the pandemic 2009 influenza A/H1N1pdm. AdaPatch consists of a graph-cut as optimization technique to find an optimal group of sites that show a strong signal of positive selection, and also a small spatial distance to other selected sites. Selected sites are then grouped into patches. Figure 4.1 provides a short schematic overview, and the following three sections will describe the results

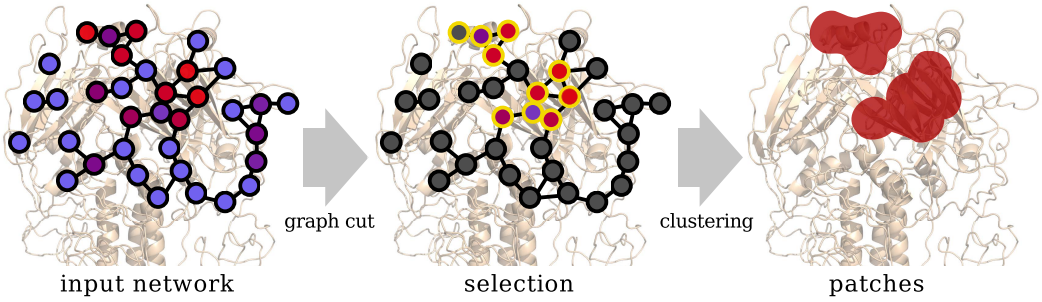


Figure 4.1: Schematic representation of AdaPatch clustering: in a first step, we find a measure of positive selection for all residues on the protein structure (indicated by circles, red and blue colors indicate high and low significance values), and determine their spatial distance. In a second step, relevant residues with high significance and small spatial distance are selected with the graph cut approach (yellow circles), and then grouped to patches in a final step (red areas).

we obtained with our analysis. Details of the method, together with all preprocessing steps, are described at the end of this chapter.

4.2 Haemagglutinin of seasonal influenza A, subtypes H1 and H3

We first applied AdaPatch to study the haemagglutinin proteins of subtypes H1 and H3. Our goal was to rediscover regions known to play an important role in the interaction of the virus with the host's immune system, since such regions are expected to comprise many important sites for adaptation. We therefore first considered known antigenic site regions on the HA of the human influenza A virus (Caton *et al.*, 1982; Wiley and Skehel, 1987) as our approximate reference for evaluation. We assigned each residue a p-value based on the derivation of its dN/dS ratio from the protein-wide average, and then considered its spatial position on the protein structure (see Section 4.5 for details). The clustering algorithm identified eight and nine dense patches of residues for subtypes H1 and H3, respectively (Figure 4.3, Table 4.2 and Table 4.3). They mostly consist of sites showing a substantial deviation from the expected value of the protein-wide dN/dS. In comparison, a site ranking based on $p-$value alone resulted only in a low sensitivity for

Setting	Recall (H1)	Precision (H1)	Recall (H3)	Precision (H3)
Graph Cut	0.53	0.49	0.25	0.94
PV 0.05	0.19	0.4	0.15	0.86
PV 0.1	0.19	0.4	0.17	0.81

Table 4.1: Precision and recall of different settings and approaches when put to the task of detecting influenza epitope sites.

discovering relevant sites, with only 6 of 32 (H1) or 19 of 131 (H3) known antigenic or receptor avidity-changing sites exhibiting a significant ($p < 0.05$) signal. To compare this approach with our method, we calculated the precision and recall for sites selected by setting a $p-$value ranking at θ=0.05 (PV 0.05) or θ=0.1 (PV 0.1) as a threshold, as well as calculating these characteristics for the sites in patches identified with our graph-cut approach. Our evaluation (Table 4.1) showed that including information on the spatial proximity of residues under selection and applying our clustering algorithm resulted in a significant improvement in recall (*i.e.*, a larger number of epitope sites being identified) while maintaining similar or better precision (meaning that a similar or lower number of non-epitope sites were inferred). In the light of a recently proposed hypothesis on the relevance of receptor avidity-changing sites (Hensley *et al.*, 2009), as opposed to the epitope sites of haemagglutinin in subtype H1 antigenic evolution, we also tested the value of these sites as a reference and compared these with the inferred patches of sites. The currently available data do not allow discrimination between these two hypotheses, as the reference sites of known receptor avidity-changing sites and antigenic sites overlap greatly (Figure 4.2). Still, residues 156 and 158, found to play the most significant role in receptor avidity, are included in the second patch identified for subtype H1.

Of the detected patches on the HA protein surface (Figure 4.3, Table 4.2 and Table 4.3), several include known epitope or receptor-avidity changing sites up to a fraction of 100%. The patches contain many sites that are relevant for antigenic evolution (Matrosovich *et al.*, 1997; Hay *et al.*, 2003; Lin *et al.*, 2004; Yamada *et al.*, 2010), including position 145, which has been shown experimentally to have a high

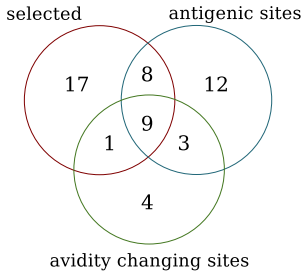


Figure 4.2: Overlap between selected epitope and avidity-changing sites. Venn diagram showing the overlap between subtype H1 residues in patches selected by the dN/dS graph-cut approach (red), the influenza A H1 epitope sites according to Caton *et al.* (1982) (blue), and avidity-changing sites according to Hensley *et al.* (2009) (green).

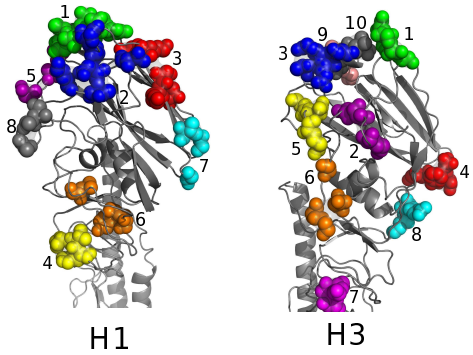


Figure 4.3: Patches under positive selection on seasonal HA of subtype H1 and H3 selected by the graph-cut algorithm. Patches are numbered according to Table 4.2 and Table 4.3.

antigenic impact (Smith *et al.*, 2004).

We also compared our results with similar techniques for predicting the properties of sites under positive selection or relevant for adaptive evolution. Our predictions match 7 of 13 sites inferred to be under positive selection by a maximum likelihood approach (Yang, 2000). However, 10 of these 13 sites are at least direct neighbors of those listed by our method, confirming its ability to find positively selected regions on the tertiary structure. Similar observations can be made for sites identified in Fitch *et al.* (1997), where five of six are matches or direct neighbors, and the sites discussed in Bush *et al.* (1999) (again, 10 of 13 matches).

Several related techniques combine biochemical and phylogenetic information to gain insights into the adaptive evolution of influenza A. It has recently been suggested that HA evolves by increasing the number of charged amino acids in regions recognized by the immune system, particularly in the dominant epitope (*i.e.*, the one with the highest proportion of amino acid mutations, see Pan *et al.* 2011). We therefore compared the number of charged and uncharged amino acids in the H1 and H3 consensus sequences for selected sites in the patches and sites lying outside the patches. Indeed, we found

Patch	Residues
1	187, 188, 189, 190, 192, 193, 196, 197, 198
2	129, 131, 132, 133, 156, 158
3	163, 165, 166, 244, 248
4	274, 275, 276
5	219, 225, 227
6	56, 81, 82
7	169, 173, 240
8	142, 144, 145

Table 4.2: Patches and residues selected for the influenza A haemagglutinin protein, subtype H1. Underlined numbers refer to known epitope sites according to Caton *et al.* (1982) and Supplementary Table B.1. All positions are given in H3 numbering (Aoyama *et al.*, 1991).

Patch	Residue
1	156, 157, 158, 159
2	188, 189, 192, 193
3	171, 172, 173, 174, 175
4	186, 220, 229
5	137, 140, 142, 144, 145
6	62, 91, 92, 94
7	53, 275, 276
8	196, 197, 198, 199
9	47, 48, 50

Table 4.3: Patches and residues selected for the influenza A haemagglutinin protein, subtype H3. Underlined numbers refer to known epitope sites (Wiley *et al.*, 1981; Wiley and Skehel, 1987; Suzuki, 2006), see Supplementary Table B.1. All positions are given in H3 numbering.

that the percentage of charged amino acids is much higher within patches (H1: 67%, H3: 67%) than outside patches (H1: 27%, H3: 28%). Finally, other authors suggest statistics based on rates of substitutions toward specific residues (Kosakovsky Pond *et al.*, 2008; Kryazhimskiy and Plotkin, 2008) or based on epistatic effects between pairs of sites (Kryazhimskiy *et al.*, 2011). The overlap between the predictions by both methods and ours is not large, possibly due to the different nature of the measured quantities and statistics, and because, as Kryazhimskiy *et al.* (2011) discusses, hitchhiking changes without selective impact might comprise a fraction of identified epistatic pairs, particularly among the trailing change of a pair. However, our simple criterion for positive selection can easily be exchanged for more advanced estimates for adaptive evolution, allowing a search for clusters of residues that show significantly elevated

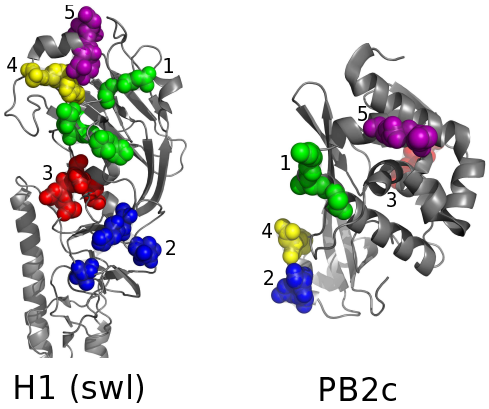


Figure 4.4: Patches on the 2009 swine-origin influenza A protein structures of HA and the C-terminal region of PB2, selected by the graph-cut algorithm. Patches are numbered according to Table 4.4 and Table 4.5.

Patch	Residue
1	135, 137, 140, 141, 142, 144, 145
2	53, 54, 56, 57, 276
3	63, 91, 92, 93, 94
4	186, 188, 189, 218
5	197, 198, 199, 200

Table 4.4: Patches and residues selected for the HA protein of the 2009 swine-origin influenza A/H1N1 virus

statistics of such properties.

Additionally, we identified one patch in H1 without known epitope sites, but with similar evidence for positive selection as the other patches, which indicates its potential importance for antigenic evolution (Table 4.2 and Figure 4.3, patch 4). For both subtypes, one patch in HA overlaps with the receptor-binding site of the protein. This could be due to the overlap of the antigenic and receptor-binding regions. However, the receptor-binding site, particularly position 189, is also known to be relevant for adaptation to avian and human hosts (Matrosovich *et al.*, 1997; Sorrell *et al.*, 2009). Both the H1 and H3 of human influenza A viruses show evidence of selection acting upon the receptor-binding region when grown in eggs, due to the effects of egg adaptation (Robertson *et al.*, 1987; Gambaryan *et al.*, 1999). Therefore, part of the signal in the receptor-binding sites could also be due to the effects of egg cultivation.

Patch	Residue
1	586, 588, 590
2	714, 715
3	660, 661
4	709, 711
5	575, 578

Table 4.5: Patches and residues selected for the PB2 protein of the 2009 swine-origin influenza A/H1N1 virus.

4.3 HA and PB2 of the 2009 A/H1N1pdm influenza

As a second application, we analyzed data of 2009 S-OIV A/H1N1. The virus stems from a triple re-assortant swine virus, which recently acquired avian segments (Smith *et al.*, 2009). The molecular basis of its successful establishment in the human host is not fully understood yet. It has, in particular, been argued that lysine at position 627 of the PB2 protein of the viral polymerase complex, instead of the avian-like glutamic acid, is required for successful transmission and replication within mammals (Gabriel *et al.*, 2005). However, the 2009 H1N1 virus still has lysine at position 627 in PB2, which it has maintained since its descent from an originally avian lineage. A change at residue 591 has been proposed to compensate for the lack of lysine in 627, allowing its efficient replication in mammals (Yamada *et al.*, 2010). We searched for regions with evidence for positive selection and relevance for adaptation of PB2 since the introduction of the 2009 S-OIV into the human population, since the virus might have acquired changes in PB2 to further optimize replication and transmission in the novel host.

We identified five patches (Figure 4.4 and Table 4.5). The first is localized in a region around residue 591, which lends support to its relevance for mammalian and, in particular, human adaptation. To gain more insight, we allowed the method to also report patches containing only two residues. The resulting second patch surrounds residue 714, which is known to increase polymerase activity in mammals (Gabriel *et al.*, 2005).

We furthermore analyzed the genetic sequences and protein structure of the HA protein of 2009 S-OIV A/H1N1. We identified five patches of sites under positive selection. The first one (Figure 4.4 and Table 4.4) overlaps with the Ca_2 epitope site of

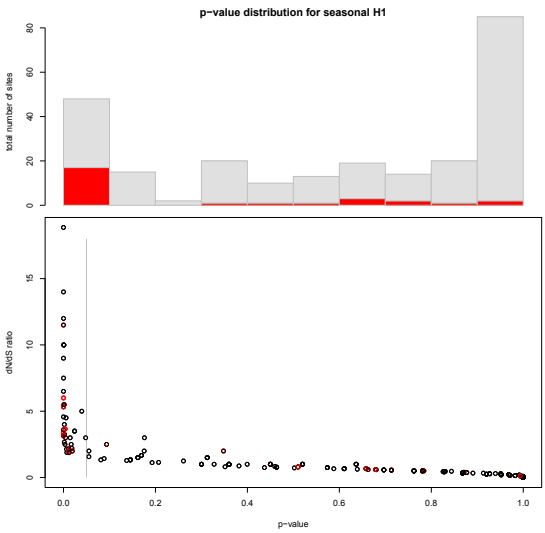


Figure 4.5: Epitope sites not under positive selection. The histogram displays the ratio of residues within the corresponding $p-$value intervals and demonstrates that many epitope sites feature insignificant $p-$values resulting from an average dN/dS ratio. Epitopic sites are marked in red. The lower plot shows the distribution of the $p-$values versus the dN/dS ratios for all residues of the H1 subtype.

seasonal H1 (Caton *et al.*, 1982). The remaining ones cluster densely at the head of the protein, indicating emerging areas of relevance for adaptation and antigenic evolution of the 2009 H1N1 virus.

4.4 Conclusions on the identified patches

Our analysis showed that AdaPatch increases the predictive accuracy relative to the commonly used approach of searching for individual sites with significantly deviating dN/dS statistics. This indicates that focusing on evolutionary change in larger regions, instead of individual sites, is helpful for revealing patches of residues that are important for adaptation, which together show a stronger signal of positive selection.

Still, the precision and recall values for detecting known epitope sites based on patches under positive selection are rather low overall, mostly at or below 50%, indicating that not all sites in the epitope regions are under positive selection and contributing to adaptation of the viral HA. Influenza A epitopes seem to be variable only in part (Figure 4.5) and probably change over time, thus diluting the overall signal of positive selection. Furthermore, receptor avidity-changing sites or host-specificity determinants may play a similarly important role in adaptive evolution, which lowers precision if

one considers only the epitope sites that are predicted to be evolving under positive selection.

We evaluated our method using influenza A viruses as they are very well studied and much is already known about the relevant sites for adaptive evolution. Still, our inferred patches might be more informative than individual sites for monitoring circulating viral strains for adaptive changes with relevance for transmission and spread in the human population. Our analyses of HA and PB2 identified many sites known to be relevant for antigenic drift or for the adaptation of influenza A to its host, improving its ability for infection, replication, and immune evasion. We therefore suggest analysis of the new patches identified in this study to determine the underlying causes of their consistent variability. We also suggest applying the method to other protein structures of rapidly evolving viruses with as yet unknown adaptive behavior in order to identify candidate regions that are important for virus-host interaction. Our software, AdaPatch, is available online[1], and can also be applied to analyze other viral proteins.

4.5 Technical details of the analysis

We implemented a graph-cut algorithm to cluster protein residues based on structural and evolutionary protein information. Our goal was to identify dense patches of spatially close residues on the protein surface that show significant signs of positive selection. Generally speaking, our algorithm includes residues in a patch if they show evidence for positive selection and are close to other patch residues. A patch is rated both by its average $p-$value and the density of sites under selection. Individual sites can compensate for a weaker signal of positive selection by being close to neighbors with a strong signal. Structural protein models were used to identify the spatial coordinates of individual residues. To measure positive selection for individual sites, ancestral character states were inferred from phylogenetic trees constructed from available genetic sequences for a particular protein. Subsequently, dN/dS statistics for each site were

[1] http://www.cs.uni-duesseldorf.de/AG/AlgBio

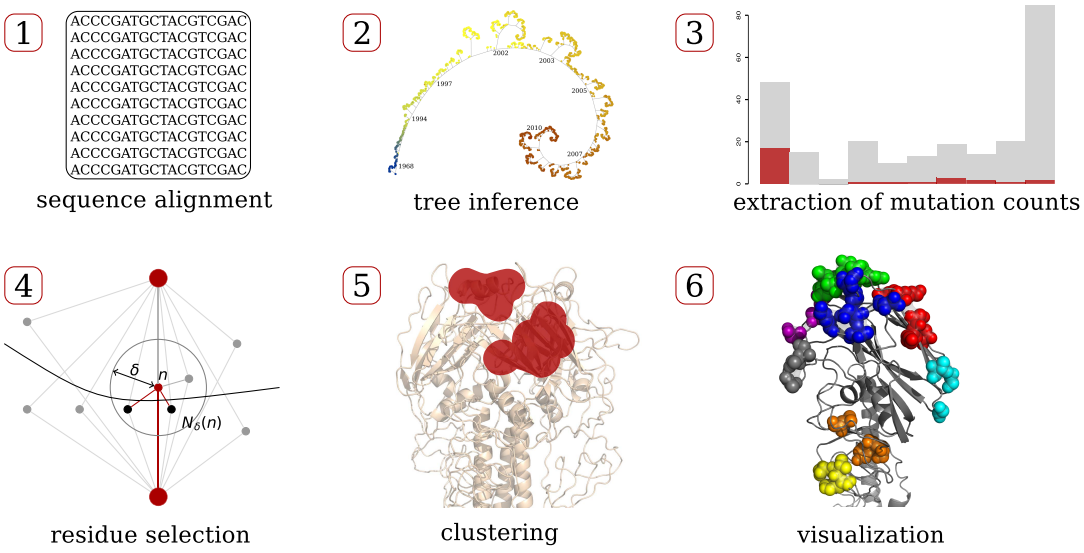


Figure 4.6: Work flow for predicting patches under positive selection.

Protein	S-OIV Sequence	PDB Code	Template Sequence	id
H1 (seas)	-	2wrgH,I	A/BrevigMission/1/1918	-
H3 (seas)	-	3hmgA,B	A/Aichi/2/1968	-
H1 (swl)	-	3al4A,B	A/California/04/2009	-
PB2cap	A/California/14/2009	2vqzA	A/Victoria/3/1975	94%
PB2c	A/California/14/2009	2vy6A	A/Victoria/3/1975	94%

Table 4.6: Sequence codes and PDB codes of selected templates.

calculated, according to the ratio of the number of synonymous and non-synonymous changes mapping to the tree edges (Bush *et al.*, 1999; Suzuki, 2006). After clustering, the identified patches were visualized on the protein structure. The complete process is shown in Figure 4.6.

Structural Models

HA structures of the human influenza A/H3N2 virus, the human influenza A/H1N1, and S-OIV A/H1N1 were downloaded from the RCSB Protein Data Bank (RCSB, 2013) (for identifier codes of structures, sequences, and templates, see Table 4.6). The analysis process was restricted to residues annotated in the PDB structure file and to sites found to be on the protein surface using the NetSurf software (Petersen *et al.*, 2009).

Structural models were generated for PB2 of the S-OIV isolate A/California/14/2009 (H1N1) based on the PB2 structures of PDB. To this end, the S-OIV PB2 sequence was compared with sequences of PB2 proteins with experimentally determined structure using Blast (Altschul *et al.*, 1990). For PB2, there was no single structural template that covered all protein domains. Therefore, two models were generated from two templates, one for the PB2cap and one for the PB2c domain. The highest sequence identity, the largest coverage of the S-OIV protein, and the quality according to resolution and free R-factor values were used as criteria to select the best matching structural templates for the PB2cap and PB2c domains. The S-OIV sequences were aligned to the templates with MODELLER (version 9v6) (Sali and Blundell, 1993). The alignments are expected to be reliable, given a sequence identity of 94% and a lack of insertions and deletions. Subsequently, the structural models were generated with MODELLER.

Sequence data, alignments, phylogenetic tree construction

Available HA sequences of the seasonal influenza A virus, subtypes H1N1 and H3N2, were downloaded from the GISAID EpiFlu database (GISAID, 2013). Only sequences longer than 1,500 bp were selected, resulting in 1,734 and 3,221 sequences for H1 and H3, respectively (Supplementary Table S2, Supplementary Material online[2]). Alignments of DNA and protein sequences were computed with MUSCLE (Edgar, 2004), and manually curated. Phylogenetic trees were inferred with PhyML v3.0 (Guindon and Gascuel, 2003) under the general time reversible (GTR) + I +Γ_4 model, with the frequency of each substitution type, the proportion of invariant sites (I), and the gamma distribution of among-site rate variation with four rate categories (Γ_4) estimated from the data. Subsequently, the tree topology and branch lengths of the maximum likelihood tree inferred with PhyML were optimized for 200,000 generations with Garli v0.96b8 (Zwickl 2006). Substitution events were inferred for the genome segment tree topologies from intermediates reconstructed with accelerated transformation (ACCTRAN; Felsenstein (2004)). The total number of substitutions occurring on all reconstructed internal

[2]http://mbe.oxfordjournals.org/content/suppl/2012/03/15/mss095.DC1/MSS095_TabS2.xls

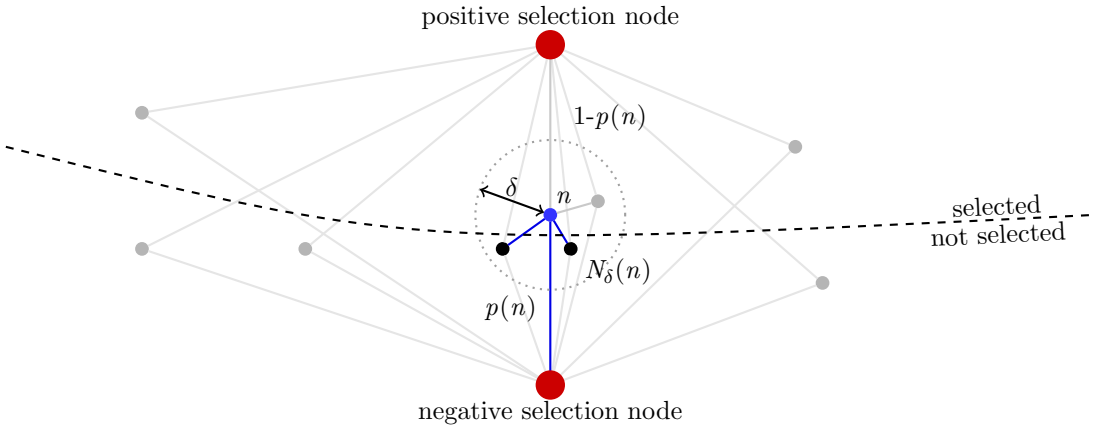


Figure 4.7: Schematic drawing of the graph-cut approach. The minimum cut minimizes the sum of weights of all edges cut by the line separating the positive and negative selection nodes. For a single node n, these are the lines shown in blue: the scaled distances to the non-selected neighbors in $N_{\delta(n)}$ and the connection to the other side (*i.e.*, the negative) selection node with the weight $p(n)$.

branches was then calculated for each site independently. These numbers were used to compute the dN/dS ratio for each codon site (Bush *et al.*, 1999; Suzuki, 2006). The ratios were transformed to $p-$values by a one-sided Fisher test for independence of the dN and dS values at an individual site and the mean values of the protein. $p-$values were corrected for the ranking comparison with the false discovery rate (Benjamini and Yekutieli, 2001) and used as a measure of selection for individual sites. Furthermore, 3,419 sequences of the PB2 protein and 7,373 sequences of the HA protein of the 2009 S-OIV A/H1N1 strains were downloaded from the GISAID EpiFlu database (also Supplementary Table S2). Phylogenetic trees were inferred using neighbor joining with PAUP (Swofford, 2003) under the GTR model. Sequence alignment and residue statistics were inferred as described above.

Structural clustering

Before clustering, all spatial coordinates were normalized to fit the protein structure into a hypercube of size 1. For clustering with a graph cut algorithm (Boykov *et al.*, 2002), we constructed a graph in which each node represents a residue in the protein. Edges

were added between all pairs of residues m and n for which the Euclidean distance $dist(m, n)$ was below a threshold δ and these edges were weighted according to their spatial distance (Figure 4.7). Weights were set to be in inverse exponential proportion to the Euclidean distance $dist(m, n)$, *i.e.*, the closer the residues were located relative to each other on the protein structure, the larger the weight of the corresponding edge. Therefore, nodes that are close to each other have a strong connection to each other. We then augmented the graph with two additional nodes, which we call the 'positive selection node' and the 'negative selection node', corresponding to 'source' and 'sink' nodes in a standard graph cut formulation. These two special nodes are connected to each residue node, with the weights equal to the p-value $p(n)$ of the residue n in the case of the negative selection node or $1 - p(n)$ in the case of the positive selection node. Thus, residues that have high dN/dS ratios (large $1 - p(n)$) have a strong connection with the positive selection node, while nodes with low dN/dS values (large $p(n)$) have a strong connection with the negative selection node. The two types of edges and edge weights were added to the graph to represent the spatial information for each residue (by adding distances to close neighbors) and the evolutionary evidence for selection (by encoding the p-value of the dN/dS ratios).

A *graph cut* will divide this graph in two parts, one containing the positive selection node, the other containing the negative one (Figure 4.7). A *minimum graph cut* is a graph cut that minimizes the sum E of the weights of the edges connecting these two parts:

$$E = \sum_{n \in Pos} p(n) + \alpha \sum_{n \in Neg} \overline{p}(n) + \beta \sum_{n \in Pos} \sum_{\substack{m \in Neg \\ m \in N_{\delta(n)}}} e^{-dist(m,n)},$$

where $\overline{p}(n) = 1 - p(n)$, Pos represents all nodes assigned to the positive selection half, Neg all nodes assigned to the negative one and $N_\delta(n)$ all neighbors of residue n within a distance less than δ. This means that the minimum cut will select residues to be in Pos if they show strong signs of positive selection (i.e., a low p-value) and if they separate well spatially from the residues in Neg.

The distance δ defines how many sites of a single residue are considered to be

neighbors. We set δ such that a residue has, on average, ten close neighbors. The factor β weighs this distance statistic. The smaller β, the more likely the method is to balance the residue evenly between the positive and the negative selection set halves according to the ratio 1:α (we set $\alpha = 1$). The larger β, the more expensive an even distribution becomes, and the more stringently the method searches for a small, exclusive set of residues that spatially separate well from the rest. Since the total distance statistic is dependent on the number of residues in the protein, β was set manually (details in Section B). However, we devised an improved and automated scheme to define β for the AntiPatch approach, described in Section 5.4. Finally, the selected residues were grouped into patches by merging all residues within a spatial distance d of each other into a set. The parameter d was set to represent the first quartile of all pairwise distances in the protein. We excluded outliers by filtering out all patches that contained two or less residues. Patches were identified for the H1 and H3 proteins of human influenza A viruses of the subtypes H1N1 and H3N2, respectively, and for the HA and PB2 proteins of the 2009 S-OIV virus of the subtype H1N1. Subsequently, we analyzed their enrichment with known epitope sites (Caton *et al.*, 1982; Wiley and Skehel, 1987) and receptor-avidity changing sites (Hensley *et al.*, 2009).

Evaluation and visualization

For evaluation, we calculated the precision (ratio of selected epitope sites to all selected residues) and recall (ratio of selected epitope sites to all epitope sites) of the inferred patches based on the epitope regions defined for subtypes H1 (Caton *et al.*, 1982) and H3 (Wiley *et al.*, 1981; Wiley and Skehel, 1987; Suzuki, 2006). See Supplementary Table B.1 for a list of epitope sites used as a reference for evaluation. The identified patches of all proteins were visualized with PyMOL v1.4 (Schrödinger, 2013).

CHAPTER 5

AntiPatch for the detection of patches of sites of antigenic impact

Note: this chapter contains material from several sections of Kratsch et al. *(2014). Minor text modifications were made to improve readability and to fit the text into the structure of this work. See Chapter A for details on author contributions.*

5.1 Introduction

AntiPatch is an extension of AdaPatch to include antigenic weights for the detection of patches of sites of antigenic impact. We combined a previous method for the inference of antigenic weights from HI titer matrices (*Antigenic Trees*, Steinbrück and McHardy 2012) with AdaPatch clustering as described in Chapter 4. Antigenicity-altering effects were inferred for individual HA residues from genetic sequences and HI data of haemagglutinin samples of seasonal H3N2, as described in Steinbrück and McHardy (2012). Subsequently, antigenicity-altering effects were mapped onto a three-dimensional protein structure.

Then, clusters of residues on the protein structure with large antigenic impact were identified with AdaPatch clustering, following the principle described in Figure 4.1 of the previous chapter. We compared the identified patches to protein sites of known relevance for altering antigenicity, sites involved in transitions between consecutive antigenic clusters, and other amino acid changes of known antigenic impact. The next two sections will describe the results of this study. The methodological details on data acquisition, preprocessing, and analysis are explained at the end of this chapter.

5.2 Haemagglutinin of seasonal influenza A, subtype H3

We used antigenic trees (Steinbrück and McHardy, 2012) to map the antigenic distances derived from HI titers between viral strains and reference antisera onto a phylogenetic tree inferred from the corresponding sequences of the HA1 subunit of HA for human influenza A/H3N2 viruses sampled between 1968 and 2003 (see Section 5.4). Unlike Steinbrück and McHardy (2012), only internal branches were considered, as we have previously observed a tendency towards systematic bias and noise, particularly for the antigenic weights of terminal branches, caused by single isolate variations. Subsequently, amino acid changes were reconstructed for the branches of the tree using ancestral character state reconstruction with PAML v4.5 (Yang, 2007). Antigenic weights for individual sites were inferred from these data as the mean of all antigenic weights of the branches with a mutation at this position. Sites with fewer than three branches contributing to their weight were penalized by reducing their weight (see Section 5.4).

Distribution of antigenic weights

We visualized the distribution of antigenic weights on the protein structure of HA of influenza subtype H3 (Figure 5.1a), and determined the relative solvent accessibility (RSA) for each residue. Sites in the epitopes had larger antigenic weights than sites outside these regions ($p = 5.5 \cdot 10^{-5}$, Kolmogorov-Smirnov test; H_1 was that epitope sites have larger antigenic weights). Still, 93 out of the 131 epitope sites had no antigenic

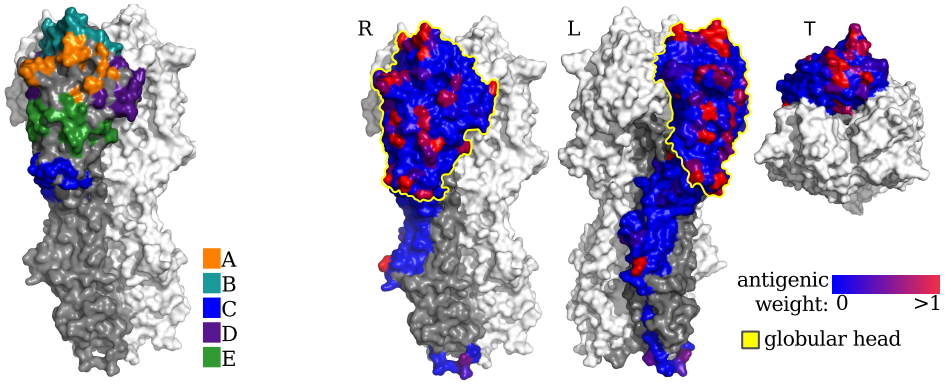


Figure 5.1: Antigenicity-altering effects of protein sites on the influenza A/H3N2 HA surface. **(a)** Location of the five epitope regions A-E, colored according to the legend. **(b)** Inferred antigenic weights for residues on the surface of the HA1 chain of HA protein for human influenza A/H3N2 viruses. The amount of the weight is shown with a bluered gradient for one subunit of the HA homotrimer in three different orientations. The globular head is indicated in yellow; HA2 is colored dark gray. It is evident that not all epitope sites have large antigenic weights and that sites with large antigenic weights are mostly found as spatial clusters within the epitope sites and not outside the epitopes. Sites with antigenic weights outside of the head are not supported by the known biology of the antibody-HA interactions and likely antigenic hitchhikers found in genotypephenotype inference.

impact assigned, and thus had no discernible relevance for immune evasion in the past. Within the epitope sites, a tendency for sites with weights to cluster seemed evident (Figure 5.1b) and many sites were found in the vicinity of the RBS in the globular head of the protein. There were also 23 buried sites with antigenic weights in the epitopes (RSA $< 20\%$). These sites might not be directly involved in antibody interactions, but could contribute to antigenic evolution by compensating for stability or fitness disadvantages caused by nearby antigenic changes. As expected, the sites outside the epitope regions were mostly assigned low antigenic weights: 167 out of 180 had no antigenic weight, including 49 of the 57 exposed sites. The 13 sites with antigenic weights assigned mostly lie outside the head region or within the stem, which has no known relevance for antigenic evolution. Thus these sites could be antigenic hitchhikers falsely identified as being relevant by genotype-phenotype inference, due to their co-occurrence

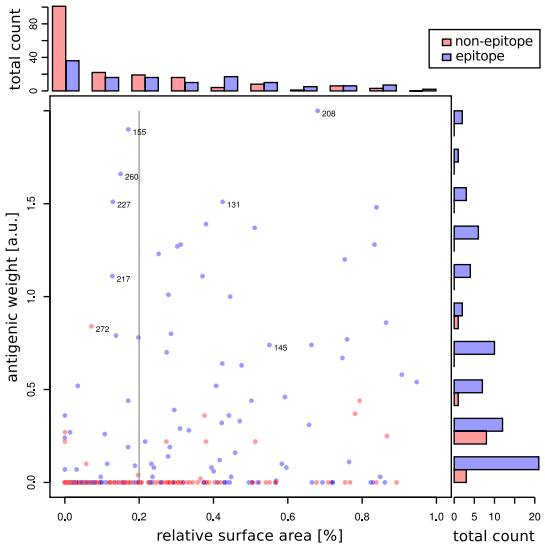


Figure 5.2: Distribution of antigenicity-altering effects versus the relative available surface area (RSA) for residues within the epitope regions (blue) and outside the epitope regions (red) on HA of human influenza A/H3N2 viruses. Residues with an RSA of more than 20% lie to the right of the dashed gray line. On the side of the plot are histograms showing the number of sites with particular antigenic weights or RSA values. The histogram for antigenic weights is reduced to the interval (0,2], as most sites have no antigenic weight.

Patch #	Sites	Epitope
1	131, 156, 158, 159, 189, 196, 155, 186, 217, 227	ABBBBB BBDD
2	137, 144, 145	AAA
3	50, 272, 276, 278	CC-C
4	62, 75	EE
5	207, 208	DD
6	174, 260	DE

Table 5.1: Identified patches and patch sites and their placement within the five epitope regions. Sites are enumerated according to the H3 numbering convention (Aoyama *et al.*, 1991).

with an antigenicity-altering change. Epitope sites had more surface exposure than sites outside the epitopes ($p = 1.9 \cdot 10^{-6}$, KS test; H_1: epitope values have a larger RSA value) and a smaller portion of buried sites (52 out of 131 epitope sites versus 123 out of 180 non-epitope sites; Figure 5.2). As there was no significant correlation between surface exposure and the antigenic weight ($r = 0.139$ with Pearsons correlation; $p = 0.215$) for sites with non-zero antigenic weights on HA1, we included both surface and buried residues in our subsequent analyses, as buried sites might be located close to key antigenic regions on the protein surface and changes here might indirectly affect antigenicity.

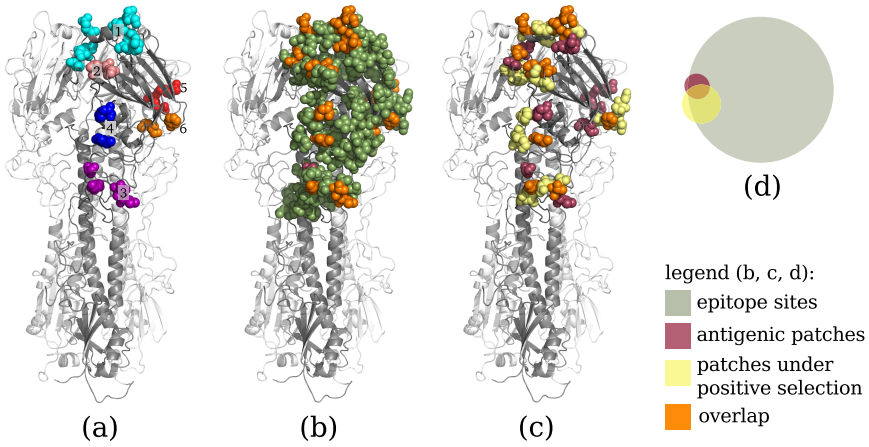


Figure 5.3: Location of the inferred antigenic patches on influenza A/H3N2 HA. **(a)** Antigenic patches on the HA homotrimer, numbered according to Table 5.1. **(b)** Location of epitope sites (dark green), antigenic patch sites (purple), and sites included in both sets (orange). Note that only residue 272 is not included in the overlap. **(c)** Comparison of antigenic patch sites (purple) to patches of sites under positive selection (Tusche *et al.*, 2012) (yellow) and to sites included in both sets (orange). **(d)** Venn diagram showing the overlap of antigenic patches, patches under selection, and epitope sites (see Supplementary Figure C.1 and Figure C.2 for details).

Inference of antigenic patches

We used AdaPatch clustering (Chapter 4, Tusche *et al.* 2012) to identify spatial clusters of sites with large antigenic weights on the three-dimensional structure of haemagglutinin. We thus identified six patches in the HA1 subunit of HA (Figure 5.3, Table 5.1). One large patch (patch 1) includes ten residues, and five patches have four or fewer residues each. With the exception of residue 272, all patch sites are part of the epitopes. Patch 1 is located on top of the protein head and includes residues 189, 196, and 227 of the RBS (Skehel, 2009). Patch 2 is located within epitope A and overlaps with the 130-loop of the RBS. It surrounds and includes residue 145, which has repeatedly been reported as being under positive selection (Bush *et al.*, 1999; Kosakovsky Pond *et al.*, 2008), and changes at this site have very large antigenic weight (Smith *et al.*, 2004; Lee *et al.*, 2007; Huang *et al.*, 2009; Liao *et al.*, 2008, 2013; Koel *et al.*, 2013). The other three patches are located within epitopes C, D, and E (Table 5.1).

Comparison to antigenic clusters

With the exception of positions 159, 186, and 227, all residues of patch 1 were altered in past transitions between consecutive antigenic clusters of influenza A/H3N2 during the studied time period (Smith *et al.*, 2004). Of the 23 sites included in the six antigenicity-altering patches, 18 are part of 45 sites that have changed in antigenic cluster transitions (Smith *et al.*, 2004) (Supplementary Figure C.1 and Figure C.2). Koel *et al.* (2013) used site-directed mutagenesis to confirm that in these 45 sites, positions 145, 155, 156, 158, 159, 189 and 193 had a strong antigenic impact in previous antigenic evolution. Position 145 is included in patch 2; all other sites, except for position 193, are part of patch 1, which indicates the importance of these particular two patches for antigenic evolution. Large antigenic weights were also inferred for the remaining five patch sites, namely residues 159, 186, 208, 227 and 272 (0.86, 0.79, 2.0, 1.5 and 0.8 antigenic units, respectively), indicating the relevance of these sites for antigenic evolution. In fact, residue 208 was assigned the largest antigenic weight of all sites (Figure 5.2). As the five sites were not part of the antigenic cluster transitions, these changes likely reflect antigenic variations between subsets of the strains which were never fixed in a new predominating antigenic variant. As evolution is a stochastic process, this does not preclude their relevance for antigenic evolution.

Patches in antigenic cluster transitions

To gain more detailed insight into the relevance of individual patches, we characterized their relevance in the ten antigenic cluster transitions between 1968 and 2002 (Smith *et al.*, 2004). Two (SI87-BE89, BE92-WU95) were accompanied by a single change at position 145 of patch 2; four had several changes in patch 1 (TX77-BA79, BA79-SI87), or in patch 1 and additional patches (WU95-SY97, SY97-FU02). The remaining four showed changes in patches 1, 2, and other patches (Supplementary Table C.1). Thus, changes in either patch 1 or 2 seem to consistently accompany antigenic cluster transitions, while changes in the other patches occur more sporadically.

We then studied the genetic evolution of influenza A haemagglutinin of subtype H3

after 2002. A phylogeny created from haemagglutinin sequences sampled since 1968 shows the typical 'cactus-like' structure with a single main trunk ranging from the 1968 root of the tree to 2014 viral sequences; a single larger branch holds parts of strains sampled between 2011 and 2013 (see Supplementary Figure S2). Of the 34 sites that feature amino acid changes on the trunk of the phylogeny or on the 2011 branch, 13 appear in patches. Of the 13 sites, 10 appear in patch 1 or 2. With exception of eight branches, of which four consist only of mutually inverse mutations (e.g. D53N, N53D in 2009), all branches on the trunk that show amino acid changes also contain patch sites. These observations support the validity of our patch sites also for viral strains after 2002, i.e. strains that are not covered by the HI titer data used for patch inference.

Comparison to patches under selection

A comparison of sites included in antigenic patches to those in patches of sites under positive selection (Tusche *et al.*, 2012) showed that only 13 of the 23 sites in antigenic patches were also included in patches of 35 sites under positive selection (Figure 5.3, Supplementary Figure C.1 and Figure C.2). Therefore, a substantial fraction of sites under selection and antigenic sites are distinct from one another. Though there was a significant overlap of the two sets of sites within the head region ($hypergeom(N = 230, K = 35, n = 23, k = 13)$; H_0: sites are sampled from the same process; $p = 9.9{\cdot}10^{-7}$), the antigenic weights in antigenic patches were significantly different from those in the patches under positive selection (two-sided KolmogorovSmirnov-test; H_0: both samples were derived from the same distribution; $p = 0.0025$), indicating that alterations of antigenicity do not explain all the sites in patches under positive selection. Possibly, sites that do not alter antigenicity but directly affect binding avidity to sialic acid residues on the host cell surfaces are under selection, which has recently been described as affecting influenza A/H1N1 evolution in mouse models (Hensley *et al.*, 2009). In fact, six sites (192, 193, 197, 198, 199, and 229) that are found only in patches under positive selection but not in the antigenic patches are located in the RBS.

5.3 Conclusions on the identified patches

Knowledge of sites that are implicated in the antibody binding of influenza A/H3N2 viruses is of substantial importance for viral surveillance and vaccine design (Gershoni *et al.*, 2007; Lees *et al.*, 2011). Here, we combined genetic, evolutionary, phenotypic, and structural information to detect regions of large antigenic impact on the protein structure of HA of human influenza A/H3N2 viruses. Antigenic weights for individual sites were inferred by mapping antigenic distances derived from HI titers onto a phylogeny inferred for the HA1 subunit of HA of the respective viral isolates. A subsequent clustering of sites based on antigenic weight and spatial proximity revealed six areas on the viral HA which have been the most relevant for changing the antigenicity over a period of 35 years of viral evolution. Antigenic tree inference is able to assign antigenic weights to individual sites, but cannot distinguish between sites with true antigenic impact and hitchhiker sites that feature changes on the same branches of the phylogeny. By removing all sites with antigenic weights located far apart from spatial clusters of antigenic sites on HA, we eliminated changes that likely have no real antigenic effects (such as changes in the stem region of HA), and were potentially identified due to their coevolution with antigenicity-altering changes in the antigenic tree inference.

The six identified patches are located in the protein head region, mostly close to the RBS and within the known epitope regions (Figure 5.3). They include many previously described antigenicity-altering sites and sites altered in antigenic cluster transitions of human influenza A/H3N2 viruses. The ten antigenic cluster transitions of the studied time period all included amino acid changes in either the first or second antigenic patch, or both. This is a sparser hypothesis for the requirements of an antigenic cluster transition than previous ones, which stated that changes in epitopes A and B are required for an antigenic cluster transition (Huang and Yang, 2011; Wilson and Cox, 1990).

Computational methods that select sites based on their correlation with antigenic distances or location on the protein structure have been suggested before (Lee *et al.*,

2007; Liao *et al.*, 2008; Huang *et al.*, 2009; Lees *et al.*, 2011), though none has attempted yet to reconstruct antigenic areas on the protein structure as a replacement for the broadly defined epitope regions by joint consideration of spatial, antigenic, evolutionary, and genetic information. Lees *et al.* (2011) mapped sphere-shaped clusters of amino acid substitutions between predominant strains onto a grid of the HA protein structure to predict antigenic distances, which indicated a large number (76) of potentially relevant sites. Here, we clustered all HA residues jointly, based on their antigenic weights and location on the protein structure without restricting our attention to a particular set of substitutions or a specific shape of cluster. Sun *et al.* (2013) used ridge regression to infer antigenic weights based on genetic and antigenic profiles, identifying 39 antigenicity-associated sites on the protein surface, but did not consider their spatial distances. Of the total 23 AntiPatch sites reported here, 14 were found in their study. Of the other nine residues, five were involved in antigenic cluster transitions and were assigned above-average antigenic weights (an average of 1.2 a.u. for the nine sites; the average of all sites in HA1 is 0.13 a.u.). Huang *et al.* (2009) used a decision tree to detect "antigenic critical positions" that were relevant for classifying antigenic clusters, without consideration of the protein structure. They describe 11 sites, which include positions 137, 145, 156, 158, and 189 in patches 1 and 2, and positions 62, 260, and 278 in patches 3, 4, and 6. Residues 155 and 156 have been confirmed as being responsible for the antigenic cluster transition to A/Fujian-like viruses in 2003 (Jin *et al.*, 2005). In the underlying study on antigenic tree inference (Steinbrück and McHardy, 2012), a strict criterion was used to define relevance and seven sites with antigenic weights that were larger than one antigenic unit were described (positions 112, 137, 144, 155, 156, 189, and 208). Here, we used an updated method for antigenic weight inference, which excluded terminal branches (see Section 5.4), and identified several more sites with an antigenic weight above or close to one antigenic unit. Six of the seven sites we described before are located in the patches; five of them in patch 1 or 2, which again stresses the importance of these regions for antigenic evolution. We consider that our approach here allows improved insight into the antigenically relevant features of the viral HA, as it

also reveals sites with lesser antigenic weights as being relevant, if these are additionally supported by spatial clustering, which is in agreement with the underlying model of molecular interactions between the viral surface proteins and host antibodies.

We observed a strikingly small overlap between sites in the antigenic patches (13 of 23 patch sites) with patches in the sites under selection that we identified before (Chapter 4, Tusche *et al.* 2012). In comparison to other studies, four out of the 18 sites found by Bush *et al.* (1999) are included in the antigenic patches (residues 145, 156, 158, 186); patch sites 137, 155, 196, and 276 were also reported to be under positive selection elsewhere (Fitch *et al.*, 1997; Kosakovsky Pond *et al.*, 2008) (Supplementary Figure C.1 and Figure C.2). Earlier, Smith *et al.* (2004) described that sites involved in antigenic cluster transitions and sites under selection (Bush *et al.*, 1999) seem distinct; however, the authors compared sites under positive selection from a different period to the one studied based on antigenic measurements. Here, we directly compared antigenic patches with patches under selection identified for the same period of time and find that this is still the case. The lack of overlap is also apparent when comparing other studies reporting sites under positive selection (Bush *et al.*, 1999; Kosakovsky Pond *et al.*, 2008; Murrell *et al.*, 2012) with studies reporting sites of antigenic impact (Koel *et al.*, 2013). These findings raise the question on the type of selective advantage provided by sites that are under positive selection but do not have observed antigenic impact. Possibly, such sites directly affect receptor avidity, instead of altering antigenicity, or else that binding to negatively charged cell surface structures plays a role. More refined models of the factors shaping influenza virus evolution need to be considered. Determining sites or patches associated with other phenotypes, such as the host receptors binding avidity, protein stability or binding to the negatively-charged phospholipids on the cell surface, could provide further insight into the different mechanisms shaping the evolution of human influenza A/H3N2 viruses.

Antibodies interact with influenza haemagglutinin in three ways: they disrupt viral attachement to sialic acids on the host cell surface, they prevent release of viral offspring, and they block viral fusion with the host cell (DiLillo *et al.*, 2014; Laursen

and Wilson, 2013). Only the first interaction, which is mediated via the HA1 domain of HA, is associated with a haemagglutination effect and can be detected with an HI assay. We therefore restricted our study to HA1. HI titers are commonly used to estimate antigenic characteristics of circulating viral strains within the global surveillance network of the World Health Organization (Russell *et al.*, 2008). However, the titers may be imprecise or show variable results (Steinbrück and McHardy, 2012; WHO, 2011), and measurements might be influenced by the effects of egg adaptations (Lin *et al.*, 2012) or neuraminidase activity (Sandbulte *et al.*, 2011). To reduce potential effects of egg adaptation on the measurements, we excluded terminal branches from our antigenic tree analysis and thus considered only amino acid changes supported by two or more viral strains. For the identification of further antigenically relevant regions in the influenza A neuraminidase protein, it will be straightforward to include measurements characterizing NA alterations and information on the structure of NA. Our method could easily be adapted to similar phenotypic measures, such as data from a recently described neutralization assay (Terletskaia-Ladwig *et al.*, 2013), if a suitable distance matrix comparing the viral strains and a structural model were available.

5.4 Technical details of the analysis

Our method comprised three steps : First, antigenic weights were inferred for individual HA residues from HA sequences and HI data using antigenic trees, with the data and methodology adapted from an earlier article (Steinbrück and McHardy, 2012). Subsequently, antigenic weights were mapped onto the three-dimensional protein structure of the influenza subtype H3. Next, clusters of residues on the protein structure with large antigenic impact were identified with a graph-cut clustering approach as described before (Chapter 4, Tusche *et al.* 2012). We compared the identified patches to protein sites of known relevance for altering antigenicity, sites involved in antigenic cluster transitions, and other amino acid changes of known antigenic impact. In the following sections, we describe each step in more detail.

Spatial coordinates and surface accessibility

The protein structure model (PDB identifier H3MG) of HA of human influenza A/H3N2 viruses was obtained from the RCSB database (RCSB, 2013). The coordinates of the C_α atoms were used to represent the spatial coordinates of the corresponding amino acid residues. To classify residues as exposed or buried, the relative solvent accessibility (RSA) was computed by estimating the accessible surface area with CCP4 (Lee and Richards, 1971; Winn *et al.*, 2011) and normalization with the respective maximum surface area (Chothia, 1976). Residues with an RSA of 20% or more were defined as exposed, following Chen and Zhou (Chen and Zhou, 2005). To determine the influence of the protein structure on our results, we repeated our complete analysis with an influenza structure based on a more recent viral strain (PDB identifier 2YP7), a structure of an HA trimer in connection with a neutralizing antibody (1QFU), and a structure predicted from the consensus of all 258 sequences used in our antigenic tree inference (prediction was performed with the Phyre 2 webserver (Kelley and Sternberg, 2009)). The identified patches were identical for the different structures. We also found that the root mean square deviation between these structures and the 3HMG model is very small: 0.43 Å for 2YP7, 0.714 Å for 1QFU, and 0.6 Å for the consensus structure, as determined on the Ca, N, and O atoms of the protein head with the 'super' command in Pymol v1.4 (Schrödinger, 2013), without additional refinement cycles. For the graph-cut clustering, residue coordinates were normalized so that the largest dimension of the protein was of length one, to ensure the normalized variance of input variables in the optimization.

Inference of antigenic weights

Antigenic weights for each residue were inferred following a method similar to that of a previous study (Steinbrück and McHardy, 2012). In short, HI assay data of Smith *et al.* (Smith *et al.*, 2004) were used and the associated HA1 sequences were downloaded from the GISAID EpiFlu database (GISAID, 2013). The collection comprised 258 seasonal human influenza A (H3N2) virus isolates from 1968 to 2003. The sequence data were used to compute a phylogenetic tree with PhyML (Guindon and Gascuel, 2003)

and Garli (Zwickl, 2006) under the GTR+I+Γ_4 model inferred with Modeltest (Posada, 2008). To root the tree, a related avian sequence, A/duck/33/1980, was used and subsequently removed from the study. Ancestral sequences were reconstructed for all internal nodes of the tree using PAML v4.5 under the $JTT + \Gamma_4 + F$ model (Jones *et al.*, 1992; Yang, 2007) inferred with ProtTest (Darriba *et al.*, 2011). Amino acid changes between parent and child node sequences were then mapped to the branches of the tree. Assay data were normalized as described by Smith et al.(Smith *et al.*, 2004). The resulting normalized antigenic distances between pairs of influenza antigens and antisera were used to infer antigenic weights for the branches of the phylogeny with non-negative least squares optimization (Cavalli-Sforza and Edwards, 1967).

To account for the asymmetric nature of HI assay data, we inferred antigenic 'up' and 'down' weights for branches on the path between antigen-antisera pairs in the tree, in accordance with the directionality of the measurements, from an antigen to the root of the tree (up) or from the root to an antiserum (down) (see also Figure 2 from Steinbrück and McHardy (2012)). To assess the antigenic weight of a particular amino acid position, we used the average of all available antigenic 'up' or 'down' weights of the branches with an amino acid change at this position. In contrast to an earlier study (Steinbrück and McHardy, 2012), only internal branches were considered, as we observed a tendency towards systematic bias and highly variable antigenic weights being assigned to the terminal branches, caused by single isolate variations. To avoid assigning large weights to positions based on little data and to penalize the estimated weight for a lack of data, we divided the weight of each branch with a reconstructed change for this position by the total number of amino acid changes on the branch. If three or fewer branches contributed antigenic weights to an amino acid position, we considered the estimate for this position to be less reliable.

Further details both on the data and antigenic tree and a list of GISAID identifiers for all sequences can be found in (Steinbrück and McHardy, 2012). In our previous study, we decided to use maximum likelihood for reconstructing ancestral sequences. We found that ancestral sequence reconstructions of maximum likelihood and parsimony

methods are very similar, and that maximum likelihood reconstruction provides an intermediate between accelerated and delayed transition in case of ties with maximum parsimony. Also, a recent study reported that the use of Bayesian methods to integrate phylogenetic uncertainty into ancestral sequence reconstruction is unnecessary, and that both maximum likelihood and Bayesian methods perform very similarly (Hanson-Smith *et al.*, 2010).

Clustering and visualization

We used the antigenic weights of individual residues and the spatial coordinates from the protein structure model as input for a graph-cut based clustering to infer dense patches of residues with a large antigenic impact. As described previously (Chapter 4, Tusche *et al.* 2012), this divides the set of all analyzed residues into a relevant (Pos) and an irrelevant (Neg) part. The graph-cut approach was applied to a graph where each node n represents a protein site. Two additional nodes P and N represent the Pos and the Neg sets. Each residue n is connected with an edge to both P and N, and to all its neighboring residues on the protein structure within a distance of δ. Edges to P and N are weighted with $a(n)$ and $\overline{a}(n)$, respectively, and edges between residues m and n are weighted with the proximity $exp(-dist(m,n)$, where $dist(m,n)$ measures the Euclidean distance. For a residue n, we set $a(n)$ to be equal to the antigenic weight of that residue, and $\overline{a}(n) = \hat{a} - a(n)$, with $\hat{a}$ being the largest antigenic weight in the data. The measure $exp(-dist(m,n)$ is large if residues m and n are close to each other. The graph-cut divides the set of all residues by placing the nodes into the positive class that (1) have a large antigenic weight $a(n)$, (2) are close to other nodes in the positive class, and (3) are far away from nodes in the negative class. This is achieved by searching for a set of edges with minimal costs which, when removed, cut all paths between P and N:

$$\left(\sum_{n \in Pos} \sum_{\substack{m \in Neg \\ m \in D_{\delta(n)}}} e^{-dist(m,n)} \right) + \beta \left(\sum_{n \in Pos} \overline{a}(n) + \sum_{n \in Neg} a(n) \right),$$

where $D_\delta(n)$ corresponds to the neighboring residues within a distance of δ to n, and β determines the size of the positive and negative classes. Based on the value of β, the result gradually changes between two extreme (and undesirable) cases: It either assigns all sites to one of the two classes (β=0, *Neg* in our implementation) or distributes sites between *Pos* and *Neg* (large β), such that $\sum_{n \in Pos} \overline{a}(n) + \sum_{n \in Neg} a(n)$ is minimal. The latter case ignores spatial distances and we refer to it as saturation. To determine the optimal value for β, we performed a parameter search. Since the value of β required for saturation is dependent on the size of the protein, the total number of residues and their antigenic weights, we iteratively increased β by one and defined β^{sat} as the first β for which we did not observe updates in assignments for 1,000 iterations (our implementation returns a warning and an empty result when β reaches 5,000). We then set the final value for $\beta = 0.5\beta^{sat}$, i.e. we set β to the value returning a compromise between the two extreme cases as a result. We set δ for $D_\delta(n)$ to 22 Å, which is half the distance of the largest epitope in the HA of subtype H3 (epitope D has a maximum diameter of $\sim$ 44 Å). Subsequently, the selected residues were grouped into patches if their distance was less than 0.5δ. Any remaining single residues were discarded. The resulting patches were visualized on the original protein structure with PyMOL v1.4 (Schrödinger, 2013).

5.5 Comment on epistasis in influenza A haemagglutinin

Kryazhimskiy *et al.* note that, although epistatic pairs in HA are likely to be within epitope sites, their spatial distance on the structure is not smaller than that between randomly selected sites. We observed that epistatic sites were indeed closer to each other than expected, which supports the use of spatial clustering in AdaPatch and AntiPatch. For this thesis, we considered pairs of non-synonymous mutations that occurred within a branch length distance of less than or equal to 0.1 on a phylogeny of influenza A haemagglutinin sequences of subtype H1,[1] measured their Euclidean

[1]The H1 phylogeny is identical to the one used by Tusche *et al.* (2012).

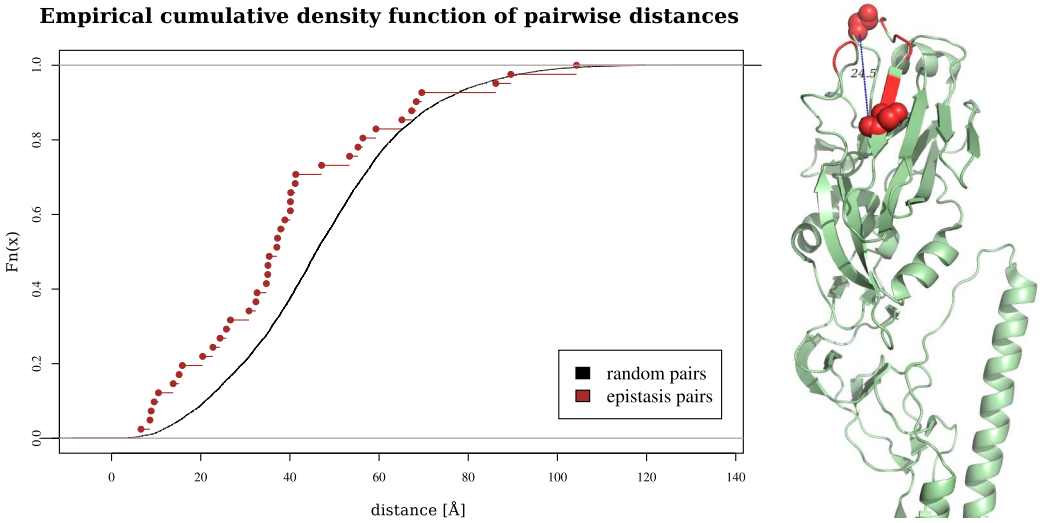


Figure 5.4: **(a)** Empirical cumulative distribution of pairwise distances, measured in Ångstrøm, for pairs of sites on the protein structure of influenza A haemagglutinin of subtype H1. Shown in black are sites that were randomly selected on the structure, shown in red are sites considered to be under epistasis following a simple branch length criterion. **(b)** The largest distance between residues of the epitope site Sa of seasonal H1 is 24.5 Å, measured between the Cα atoms of residues 158 and 167.

distance on the three-dimensional protein structure (PDB identifier 1RUZ, RCSB 2013), and compared the inferred distances to those of 1000 randomly drawn pairs on the structure. Figure 5.4a shows that epistasis pairs tend to be significantly closer to each other than random pairs (Kolmogorov-Smirnov-Test, $p = 9,3 \cdot 10^{-4}$). In addition, the median distance for epistatic pairs (21.40 Å, median random pairs: 30.25 Å) is well within a typical expansion size for epitope sites (Figure 5.4b shows epitope site "Sa" as an example, plotted on a single haemagglutinin homomer using the 1RUZ structure with Pymol (Schrödinger, 2013)).

We believe that the approach of Kryazhimskiy *et al.* (2011) might be biased, as the authors consider only consecutive branches to find pairs of epistatic sites, defining a "leading" and a "trailing" site. In our analysis, we also included sites that (a) appear together on the same branch, or that (b) appear within a molecular distance of 0.1 on the tree, regardless of the number of branching events between them. We believe this

approach to be more precise, since this number of branching events is mainly governed by the sample size, and should have no influence on the determination of co-evolution between sites.

We have developed AdaPatch and AntiPatch, two methods that search for dense clusters of residues on the major surface protein HA that are relevant for adaptation and co-evolution with the host, as well as immune evasion of human influenza A viruses. The main motivation for both projects was to develop a method that provides higher accuracy in selecting highly relevant sites by including additional structural information. A second aim was to create a method that is able to handle the bias caused by potential interactions between residues. We believe that the coarser scale of *regions* instead of single residues might be less prone to epistasis effects. Both methods take the *spatial* distance between residues into account to estimate their potential "relatedness", and the dense patches they find indeed comprise residues from epitope sites to a very high degree. Future extensions of the method could easily include other measures like the epistasis statistic of Kryazhimskiy *et al.* (2011) to search for clusters of strongly co-evolving sites under positive selection (AdaPatch) or of antigenic impact (AntiPatch).

Part III

Studying phenotype evolution

CHAPTER 6

Phenotypic evolution and ancestral character state reconstruction

As described in Section 2.2, shared ancestry between species or population samples can be expressed with the help of phylogenetic trees. The genetic sequences or phenotypes of such samples can be represented as arrays of discrete or continuous characters. Models of genetic or phenotype evolution are necessary for many applications, *e.g.*

- to estimate the goodness-of-fit of a phylogeny to the data,
- to reconstruct ancestral character states,
- to provide input and reconstructed data for phylodynamic methods (Section 2.3), or
- as model assumptions for *comparative methods* that analyze correlations between two phenotypes while taking phylogenetic relationships into account.

Models are typically motivated either from a statistical perspective, or based on the maximum parsimony principle. This chapter briefly summarizes the main methods to model phenotype evolution for the purpose of *ancestral character state reconstruction* (ACR). It serves as an introduction for the next two chapters, which will discuss feature selection techniques using phylogenies, and introduce RidgeRace, a method for continuous ACR and feature selection developed by the author.

6.1 Discrete characters

The reconstruction of discrete ancestral characters is of high relevance for phylogenetic methods, especially since genetic sequences can be considered as vectors of discrete characters. At least an implicit sequence assignment to inner nodes of a phylogeny is required to measure the fit of the phylogeny to the data. Ancestral states are also considered for studies of adaptation and evolution. Such studies ask when, in which lineages, or how often a discrete trait was gained or lost (Coddington, 1988; Baum and Larson, 1991; Brooks and McLennan, 1991). In statistical models describing DNA sequence evolution along the branches of a tree, each character position is typically assumed to have an equal chance of changing per unit of time, and a Markov model is used to describe the chances of transitions from one nucleotide state to another. Different parameterizations of the distributions take different properties or assumptions into account, such as different base probabilities or a difference between transitions and transversions (see Felsenstein, 2004, chapters 11 and 13). More complex models of genetic sequence evolution allow the rates of change do differ between individual sites of the sequence, individual branches of the tree, or both (Anisimova, 2012). The statistical model can then be used to estimate the likelihood of a concrete sequence assignment to inner nodes for given leaf sequences (Felsenstein, 2004; Yang, 2007). The same principles can be applied to model the evolution of amino acid sequences, or for binary or arbitrary discrete characters (Schluter *et al.*, 1997; Pagel, 1999b). Maximum parsimony methods for tree inference minimize the number of changes on the branches of the tree, and thus

require ancestral sequences to infer those. The two main algorithms for this purpose are the Fitch and the Sankoff algorithm (Sankoff, 1975; Fitch, 1971; Felsenstein, 2004). In their original form, both algorithms are primarily concerned with the estimation of a tree score, *i.e.* an estimate of the smallest number of changes required by that topology, but can easily be extended to report the actual sequences for inner nodes (Felsenstein, 2004). The reconstruction techniques were created mainly for genomic sequences, but can be applied to all characters that follow the Wagner parsimony assumption, *i.e.* characters that allow arbitrary changes between states in all directions. Other forms of parsimony that allow only directed changes or that require an order on character states are discussed by Felsenstein (2004)(chapter 7), and implemented in McClade (Maddison and Maddison, 2000) and Mesquite (Maddison and Maddison, 2011).

The choice of one method over the other depends on a wide range of factors, discussed by Schluter *et al.* (1997) and in Nunn (2011) (chapter 3). Both maximum likelihood and maximum parsimony approaches however suffer from the same weakness: they take a phylogeny as input and ignore the uncertainty underlying the topology and branch lengths. Bayesian methods either find an approximate solution to this problem by applying single-tree-methods to each tree of a Bayesian posterior sample (Ekman *et al.*, 2008), or integrate tree and ancestor inference directly (Huelsenbeck and Bollback, 2001; Ronquist, 2004; Pagel and Meade, 2006). However, there is also evidence that Bayesian methods do not necessarily improve reconstruction accuracy (Hanson-Smith *et al.*, 2010).

6.2 Continuous characters

For continuous characters, *Brownian Motion* (BM) is the default model to describe evolution over time (Felsenstein, 1985). The main assumption of the model is that the phenotypic character change dX is directly proportional to the molecular distance

(branch length) dB covered in time t:[1]

$$dX(t) = \sigma\, dB(t). \tag{6.1}$$

Character evolution along a phylogeny is modeled as a random walk starting at the root of the tree at time zero with the mean value $X(0)$. Every time the path down to a leaf bifurcates, two dependent processes are created that share the path from the root to their most recent common ancestor. According to the BM model, character values for inner or leaf nodes are drawn from a univariate normal distribution $\mathcal{N}(X_p, \sigma B(t))$, where X_p denotes the character value of the respective father node and $B(t)$ the branch length covered between father and son. Several extensions of the BM model exist that model trends, directional selection, adaptive radiation or other influences (Nunn, 2011, chapter 5).

Methods for the reconstruction of continuous ancestral characters are very often connected to, or even a side product of *comparative methods* (Harvey and Pagel, 1991). Comparative studies measure morphological or physiological traits, behavioral or metabolic properties, or environmental conditions, and compare them across a range of species to search for signals of adaptation (see Nunn (2011) for a wide range of example applications). Felsenstein's seminal paper on comparative methods (Felsenstein, 1985) noted that straightforward statistical approaches such as linear regression can fail to estimate the correlation between two traits correctly, since they do not consider the statistical dependence of samples that share a common biological ancestor (see Section 7.1). After introducing Brownian Motion, he suggested the use of *phylogenetic independent contrasts* (PICs) to estimate correlations between traits without any bias. PICs are differences between brother nodes in the phylogeny, scaled with the branch lengths between them. A basic bottom-up averaging algorithm to reconstruct internal node values as side product of the PICs algorithm led to the development of a maximum

[1] For the remainder of the document, we will denote real-valued random variables with upper-case letters, *e.g.* X. Vectors and matrices will be indicated by bold typesetting, *e.g.* $\mathbf{v}$ for a vector, and upper-case $\mathbf{M}$ for a matrix.

likelihood method for character reconstruction, and the mathematically equivalent maximum parsimony reconstruction (Schluter *et al.*, 1997). A powerful extension of the PICs framework, *Generalized Least Squares* regression (GLSR), was introduced by Grafen (1989). Years later, Martins and Hansen (1997) suggested it as a way for reconstruction that allows a multitude of extensions of the Brownian Motion process (see *e.g.* Hansen, 1997; Blomberg *et al.*, 2003; Butler and King, 2004; Freckleton and Harvey, 2006). A detailed review of the topic was provided by Felsenstein (2004)(chapter 24), the GLSR framework and its applications are discussed by Martins and Hansen (1997); Cunningham *et al.* (1998) offers a worked GLSR example, and Section 7.2 of this thesis deals with some mathematical details.

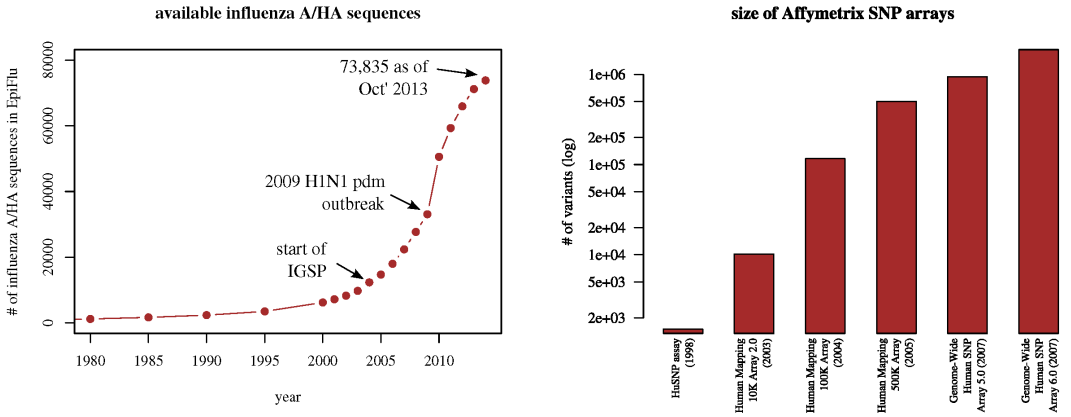
Despite their ongoing development, methods for continuous *ancestral character state reconstruction* have received a lot of criticism. Almost two decades ago, Schluter *et al.* (1997) provided an algorithm to compute confidence intervals on Maximum Likelihood reconstructions under a Brownian motion model, and showed that in some cases the resulting uncertainty might be so large as to render the reconstructions useless. Poor fidelity for reconstructed values has also been found in some studies comparing estimates with known fossil data (Finarelli and Flynn, 2006). However, ACR lead to very interesting results (Chang *et al.*, 2002; Lemey *et al.*, 2009; Nunn, 2011), and algorithms mentioned above are implemented in widely-used software packages, for example ape (Paradis *et al.*, 2004), geiger (Harmon *et al.*, 2008), phytools (Revell, 2012), Mesquite (Maddison and Maddison, 2011), BayesTrait/Continuous (Pagel, 1999a), PAUP (Swofford, 2003), and contml (Felsenstein, 1993). Even given the possibly large error rates, ancestral character state reconstructions makes sense in the context of additional evidence or as a side product of ecological studies that estimate phenotypic rates or compare different models of phenotype evolution.

CHAPTER 7

Feature Selection in biomedical data under population constraints

As indicated by the examples made in the introduction to this thesis, genomic applications are diverse and numerous. The increasing interest in genomic data and the availability of modern sequencing techniques (Figure 7.1) led to a drastic increase in the number of both available genomic samples and potential features. The term "feature" is used here broadly for whatever measurement serves as input to a method. Features can be discrete like the nucleotide or amino acid character states, or the absence or presence of a gene or trait, or they could be continuous, originating *e.g.* from continuous measurements of an environmental factor or trait, or from the expression of a gene.

Due to this increase both in sample size and in number of features, the selection of relevant features is a now very typical task of genomic studies. Obviously, the concrete meaning of "relevance" is problem-specific, but typically refers to the predictive value



(a) **(b)**

Figure 7.1: Exemplary statistics showing increasing numbers of available feature and sample counts over the last decades: **(a)** number of influenza A haemagglutinin (HA) sequences of arbitrary hosts isolated before January 1st of a given year, approximated by number of entries available in the EpiFlu database (GISAID, 2013). IGSP refers to the Influenza Genome Sequencing Project (Ghedin *et al.*, 2005). **(b)** Number of single nucleotide variants on the human genome observable with previous and state-of-the-art Affymetrix platforms (Affymetrix, 2013) (note the logarithmic y-scale).

of the feature for a certain biological function or trait, such as the existence or severity of a disease, the pathogenicity of a viral strain, or its antigenic behavior. While the first part of this thesis presented an introduction to natural selection and discussed biological implications of our AdaPatch and AntiPatch studies, this chapter will note some methodological details of feature selection techniques like Ada- and AntiPatch. Our goal is to point out some technical difficulties that arise in the study of biomedical data, which often comprise evolutionary related samples, and to very briefly summarize methods that deal with such difficulties.

7.1 Feature selection and bio-marker detection

Basic machine learning concepts

Two very central problems in supervised machine learning are classification and regression, *i.e.* the prediction of a discrete or continuous outcome y from a collection of

features $x_i, i = 1, \ldots, P$, with the help of a predictive function or algorithm:

$$y = f(x_1, x_2, \ldots, x_P).$$

The features x_i could refer to the ith character of the sequence of a sample, or may be determined from a collection of measurements. As a recent example, Aguas and Ferguson (2013) use random forests for feature selection to identify amino acid variants associated to the host species of RNA viruses like influenza A and SARS. For influenza, they successfully recover several sites known to be relevant for host adaptation in the HA and PB2 proteins. Many other examples are discussed by Saeys *et al.* (2007).

In the classical machine learning setting, *feature selection* techniques aim to reduce the potentially vast set of input features to those features of highest relevance. According to Guyon (2003) in their seminal paper on feature selection, the process has three main purposes: (1) to improve the performance of the predictor by avoiding noise and overfitting, (2) to provide faster and more cost-effective predictors, and (3) to understand and interpret the process that generated the data, and to visualize it better. The authors classify available techniques roughly as filter approaches and subset selection approaches, the latter including wrappers and embedded selection methods. Filter methods only consider single features and find some measure of similarity to the target variable (*e.g.* correlation, mutual information, accuracy of a single-variable predictor), but fail to account for covariance between features. Multivariate schemes like "maximum relevance minimum redundancy" (Peng *et al.*, 2005) or correlation based feature selection (Hall, 1998) try to overcome this problem.

Wrappers for feature selection refer to brute-force approaches that more or less enumerate all possible subsets of the available feature collection and determine which subset achieves the highest prediction quality on a training data set using a heuristic search framework. They are highly flexible in the choice of predictor function or algorithm, and naturally consider covariance between features, but they are typically computationally expensive and might overlook redundant but relevant features. Schemes

like genetic algorithms (Li *et al.*, 2001), recursive feature elimination (Guyon *et al.*, 2002), sequential forward selection/ sequential backward elimination or combinations thereof (Kohavi and John, 1997) provide simple heuristics to guide the search and to reduce the search space of all subsets of available features.

Finally, embedded approaches refer to classifiers or regression methods that explicitly include a feature selection step in their algorithm. The most prominent examples are decision trees (Hastie *et al.*, 2009, chapter 10), weight vectors of (linear) support vector machines or similar methods (Guyon *et al.*, 2002), or linear least squares regression with regularization terms, in particular using an L2 norm or an L1 norm (Hastie *et al.*, 2009, chapter 3). The latter has inspired various methods that allow to encode special structures within the learning problem, *e.g.* to select for sparse groups of features (Friedman *et al.*, 2010), hierarchical feature groups (Zhao *et al.*, 2006), multiple connected output variables (*i.e.*, multi-task learning, Obozinski *et al.*, 2009), and tree-structured output groups (Kim and Xing, 2010).

Some challenges of feature selection in biomedical data

Very often in biological or medical studies, some form of feature selection or at least simple feature ranking is a crucial part of the presented result (a detailed review of typical examples and practical solutions was provided by Saeys *et al.*). The most obvious problems in a data-driven analysis are usually quantitative: too many features, and/ or too few or too many samples. In general, extremely high dimensionality requires high computational costs, and too many irrelevant features may introduce noise that disturbs the inference. Such problems can be addressed by preprocessing with simple and efficient filters (*e.g.* Ferreira and Figueiredo, 2012) or by predictors that are sparse by definition, such as support vector machines or support vector regression (Hastie *et al.*, 2009, chapter 12). However, the literature discussing problems of high dimensionality in supervised machine learning is extensive, and a detailed review is beyond the scope of this thesis. This section is concerned with statistical properties that are typical for collections of *evolutionary related samples* like influenza A sequence data and similar

examples. Two assumptions are usually made in supervised learning: independence of features, and independence of samples.

First violated assumption: independence of features

As mentioned in Section 6.1, a very typical assumption of phylogenetic reconstruction and tree inference techniques is the independence of characters in a sequence. However, this is very likely not true, and redundancy in features might be even crucial in consecutive studies selecting single loci of functional relevance, *e.g.* amino acid positions on an influenza A surface protein associated with host adaptation. Guyon (2003) shows that highly correlated or redundant variables might still be complimentary for classification, and even that variables that are completely useless (or, presumably, functionally irrelevant) alone might be important when combined with others.
Correlation between features is ubiquitous in biological settings, and has long been referred to as *epistasis*. In his review on the subject, Phillips (2008) notes that the term has several meanings, referring, among others, to

- molecular interactions between proteins, and the genetic phenomena associated with disruptions of these interactions, Phillips refers to this type as "functional epistasis", or simply "protein-protein interactions".
- the blocking of the phenotypic impact of one allele by another. This is the traditional usage of the term as defined by Bateson (1909), and is named "compositional epistasis" by Phillips.
- the statistical deviation from the expected additive phenotypic result after a combination of alleles at two loci, referred to as "statistical epistasis".

Epistasis is a very commonplace phenomenon, and also highly relevant for influenza evolution. In a recent study, Tria *et al.* (2013) use an epidemiological model to show that epistatic effects can explain certain phylogenetic properties of HA or even the composition of antigenic clusters. Kryazhimskiy *et al.* (2011) provide a detailed review on the subject, and discuss several examples for epistasis in influenza.

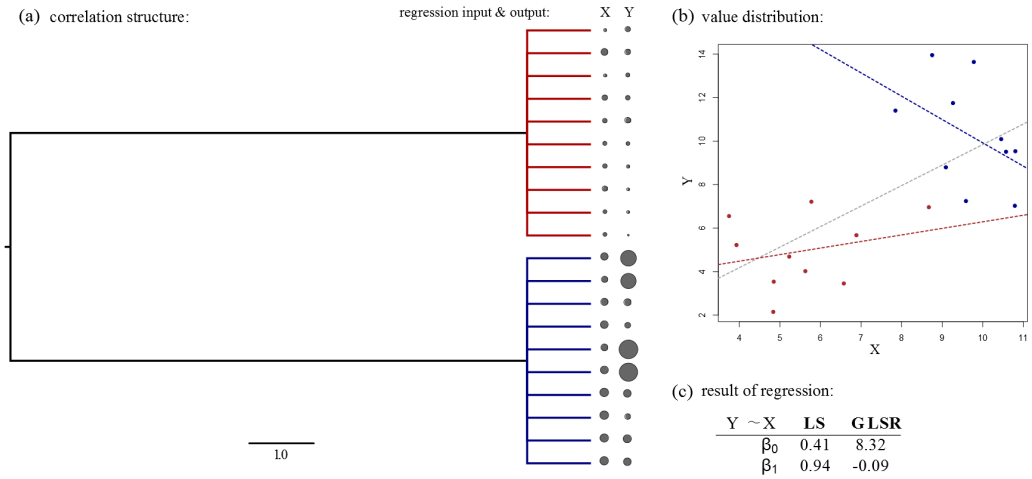
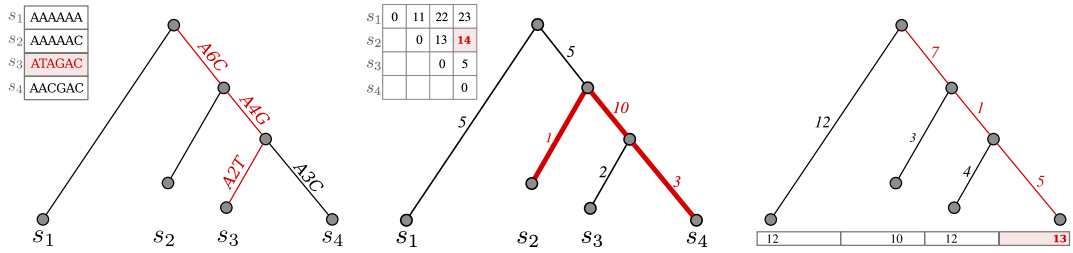


Figure 7.2: Example for phylogenetic correlation, in the sense of the example given by Felsenstein (1985). **(a)** Two traits x and y are measured and compared for two groups of hypothetical species (red and blue), far apart from each other, but with high intra-group relatedness. **(b)** When compared ignoring phylogenetic structure, x and y seem highly correlated, although they are in fact the result of independent draws from Normal distributions. **(c)** Linear regression wrongly infers high correlation ($\beta_1 = 0.94$) between x and y, while GLSR takes the long distance between the two clusters into account and correctly infers almost no dependence ($\beta_1 = -0.09$).

Second violated assumption: independence of samples

The second assumption regarding the independence of samples is violated in data sets representing closely related species or individual members of a population. Since related samples share a long way of their evolutionary (and, therefore, phenotypic) development, they are more similar than expected under the assumption of statistical independence. This might affect studies of both discrete or continuous characters. An example for the discrete case might be the study of changes in allele frequencies over time when considering alignments of (highly related) influenza A sequences (see the "frequency diagrams" in Shih *et al.*, 2007). Since sequence sample counts are biased, *e.g.* due to different laboratories or countries contributing more samples than others, and not only due to some strains being more fit than others, counting alleles based on the number of sequences without considering their similarity can produce wrong



(a) sequence mapping **(b)** distance mapping **(c)** phenotype mapping

Figure 7.3: Examples for tree mapping procedures, with example branch mappings highlighted in red: **(a)** nucleotide sequences (discrete characters) are associated with leaf nodes, and differences (*i.e.*, mutations) between samples are mapped to the branches of the tree using ancestral character state reconstruction (Chapter 6); **(b)** continuous characters represented in a phenotypic distance matrix, and **(c)** a vector of phenotypic measurements are mapped to the branches.[1]

estimates of allele frequencies. As a continuous case example, standard linear least squares regression assumes all samples to be independent and, typically, distributed with the same variance σ^2:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta + \epsilon \\ \mathbf{y} &\sim \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbf{I}) \end{aligned}$$

for some weight vector β, random noise ϵ, and $\mathbf{I}$ being the identity matrix (Kuan, 2013, chapter 3). As described in the previous section, $\mathbf{y}$ could be a vector of phenotypic measures, and the weights in β could be used as indicators for the relevance of features (columns) in the matrix $\mathbf{X}$, *e.g.* genes or sequence positions. However, if the assumption of independent and identically distributed samples is violated, the least squares estimate of β might be highly biased (Hastie *et al.*, 2009, chapter 3). A simple example has been given by Felsenstein (1985) in the context of independent contrasts, Figure 7.2 provides a modified explanation based on this example.

As a general strategy to address dependence between samples due to shared inheritance (and assuming the underlying phylogeny is known sufficiently well), it is advisable to use methods that find a way to take the phylogeny into account, typically by "mapping the differences" between samples to the branches of the tree. Figure 7.3 and Section 7.2 will explain such methods in more detail.

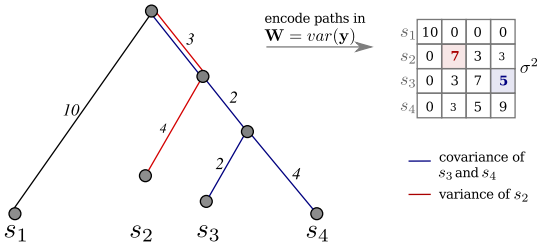


Figure 7.4: Exemplary usage of a covariance matrix for generalized least squares regression to include phylogenetic structure. Under a Brownian motion model, (co-) variance is proportional to branch length, so the total sums of path lengths correspond to the entries of $\mathbf{W}$.

7.2 Feature selection methods that include phylogenies

As described in the previous Section 7.1, it is important to account for sampling bias and phylogenetic relatedness when studying variation in genetic data. We already briefly suggested the "phylogenetic method" as a solution: we use phylogenetic trees and map differences between samples to the branches of the tree. We suggest to generally favor this strategy over simple alignments and the assumption of independent samples.

For discrete characters (in the form of genetic sequences), Steinbrück and McHardy (2011) suggest a suitable example to estimate unbiased allele frequencies from a phylogeny of influenza A viruses. By reporting alleles that show drastic increases in frequency over time, the authors provide a feature selection method that reports amino acid changes of relevance for viral adaptation. Similarly, other methods use the phylogeny in maximum likelihood models to estimate ratios of non-synonymous to synonymous mutations (dN/dS ratios) of individual sites, or estimate the ratio using ancestral character state reconstruction (*e.g.* Yang 2000; Kosakovsky Pond *et al.* 2005, 2008, but also AdaPatch). They then use the ratio to identify sites under positive selection.

For continuous characters, *Generalized Least Squares* regression (GLSR) allows to solve linear regression by including a matrix $\mathbf{W} = var(\mathbf{y})$ into the solution to the least

[1]Note that, for the sake of simplicity, Figures (b) and (c) represent error-free mappings, while real settings allow only approximate correspondence between input data and sums of branch weights.

squares optimization problem:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta \\ \hat{\beta}_{GLSR} &= \left(\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}^{-1}\mathbf{y}. \end{aligned} \tag{7.1}$$

Pagel (1997) suggested to encode the phylogeny using a matrix $\mathbf{W}$. The covariance $\mathbf{W}_{i,j}$ between two samples i and j is proportional to the branch length of the path they share in a phylogeny, and the variance $\mathbf{W}_{i,i}$ of a sample i is its distance from the root (see Figure 7.4 for an example). In a similar fashion, it is possible to reconstruct ancestral character states using GLSR. Martins and Hansen (1997) suggest to infer all ancestral states $\hat{\mathbf{A}}$ as weighted average of leaf contributions, with weights according to the covariance, or branch length, between ancestral node and the respective leaf:

$$\begin{aligned} \hat{\mathbf{A}} &= \mathbf{W}\mathbf{y} + \epsilon \\ \mathbf{W} &= cov[\mathbf{A}, \mathbf{y}]var[\mathbf{y}]^{-1}. \end{aligned} \tag{7.2}$$

As in the original regression setting, this provides a flexible formulation of the actual evolutionary process, and typical models of continuous character evolution (Hansen, 1997; Blomberg *et al.*, 2003; Butler and King, 2004; Freckleton and Harvey, 2006) can easily be plugged in. Thus, the GLSR principle allows to create regression models predicting a continuous outcome, to use its weight vector β for feature selection, and to reconstruct ancestral characters.

A second approach based on an algorithm to infer branch lengths on a given tree topology (Cavalli-Sforza and Edwards, 1967) was adapted by Steinbrück and McHardy (2012) to determine amino acid sites or changes of high antigenic impact. This approach was applied in one study of this thesis (Kratsch *et al.*, 2014, *in prep.*) and served as a motivation for a second study on a related method (Kratsch and McHardy, 2014, *in prep.*). To create *antigenic trees*, Steinbrück and McHardy (2012) use an antigenic distance matrix determined with the help of haemaglutination inhibition assays (Hirst, 1943) that measures the ability of an antiserum to inhibit red blood cell agglutination induced by a viral antigen. They map the distance matrix to the branches of a tree to

minimize the difference between the given distances $\mathbf{D}$ and the estimated distances $\hat{\mathbf{E}}$ between two samples i and j, while $\hat{\mathbf{E}}$ is a sum over branch weights v_k:

$$\begin{aligned} \hat{\mathbf{E}} &= \arg\min_{\mathbf{E}} \sum_{i=1}^{n} \sum_{j:j\neq i} (D_{i,j} - E_{i,j})^2 \\ E_{i,j} &= \sum_{k} x_{i,j,k} v_k. \end{aligned}$$

Here, $x_{i,j,k}$ is an indicator variable that equals one if branch k is on the path between leaves i and j in the phylogenetic tree, and zero otherwise. This results in a search for optimal branch weights $v_k \geq 0$ (Figure 7.3b shows an example mapping). The weights inferred for a single branch can then be assigned to amino acid changes reconstructed on the phylogeny to judge their impact. If a set of changes coincides with a high weight (*e.g.* the change *A4G* in Figure 7.3a lies on a branch with weight *10* in Figure 7.3b), this indicates that one or more of these changes or the genomic positions featuring them have a high phenotypic relevance. In this sense, the principle underlying antigenic trees allows a feature selection approach that takes phylogenetic structure into account. However, the antigenic tree approach has several disadvantages: first, it ignores the fact that long branches should generally account for higher antigenic distances than shorter weights - antigenic trees ignore branch lengths completely. Second, the technique requires a distance matrix between samples, while many biomedical samples provide only a phenotype vector (one measurement per sample). Third, the least squares optimization (Equation 7.2) tends to assign the majority of weights to the extant branches, since this is the easiest way to optimize the reconstruction error. In the next chapter, we present RidgeRace, which is primarily a method designed to reconstruct ancestral character states, but which can also be used as a feature selection method similar to antigenic trees, overcoming these three limitations.

CHAPTER 8

RidgeRace for ancestral character state reconstruction and inference of phenotypic rates

Note: this chapter contains material from several sections of Kratsch and McHardy (2014). Minor text modifications were made to improve readability and to fit the text into the structure of this work. See Chapter A for details on author contributions.

8.1 Introduction

Many biological studies investigate the ancestral states of one or several discrete and continuous characters on a phylogenetic tree (Chapter 6 reviews current methods). Typical examples are the absence, presence or state of genes or traits, environmental preferences of different species, measures of morphology or physiology, or of behavioral or metabolic properties (a comprehensive collection of examples can be found in Nunn,

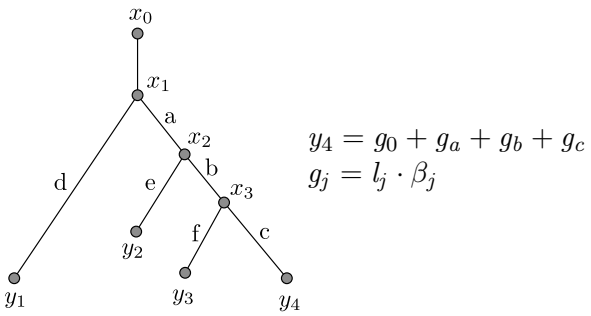


Figure 8.1: Model of phenotype evolution on a phylogenetic tree. The observed continuous character values at nodes y_i are the result of a sum of contributions on ancestral branches. A virtual branch "above" the root node x_1 is contributing the global phylogenetic mean, *i.e.* the ancestral state of x_1.

2011). Comparative methods aim to determine alleles at different loci correlating with each other or with a trait, and thus often require the reconstruction of ancestral values (Elliot, 2013). Such reconstructions are also of interest when fossil records cannot be retrieved, or when the phenotype of interest cannot be determined from the fossil tissue, *e.g.* when studying environmental conditions for a particular species.

RidgeRace (Ridge Regression for Ancestral Character Estimation) is a method inspired by the least-squares optimization technique of Cavalli-Sforza and Edwards (1967). RidgeRace does not assume fixed evolutionary rates at specific lineages predefined by the user, or a particular model of rate change over time. It considers continuous phenotypic measurements as continuous characters defined at the terminal nodes of a phylogeny. The measurements are treated as sample observations y_i which are the result of a linear regression on the branch lengths l_j of the phylogeny (Figure 8.1, see Section 8.5 for details). While the original BM model allows only a global constant rate σ, we allow phenotypic rates β_j to vary at every branch j. Using ridge regression, *i.e.* least squares regression with L_2 norm regularization, we estimate branch-wise rates and ancestral characters simultaneously by inferring a regression model that best describes the phenotypes observed at the terminal nodes.

In a simulation study, we evaluate variations of Brownian Motion on randomly created trees and show that our method performs equally well or better than established

implementations of two state-of-the-art reconstruction algorithms (Section 8.2). The branch weights β_j can be interpreted as phenotypic rates and provide insight into particularly interesting areas of the phylogeny (see Section 8.3). RidgeRace can be used for reconstructions of ancestral character states of continuous characters when no definite assumptions can be made about the type of evolutionary process, or when the assumption of a model for phenotypic evolution is not appropriate at all. The latter might for example be the case in studies that rely only on a hierarchical clustering of samples instead of phylogenies. RidgeRace can also be used as comparative method to judge the phenotypic impact of *e.g.* genetic changes or other types of events associated with branches of the phylogeny. When discrete data is provided in addition to the continuous phenotype and the underlying phylogeny, our implementation reconstructs genetic changes to the inner branches of the tree, and identifies those changes that occur on branches with a particularly high phenotypic change. We will demonstrate this with an example application for a cancer subtype stratification in Section 8.4. The last section of this chapter explains the technical details of RidgeRace and the preprocessing steps for the analyzed data.

8.2 Evaluation with simulated data

To evaluate the suitability of our method for ancestral character state reconstruction, we randomly created phylogenetic trees of increasing number of leaves $N \in \{30, 100, 200, \ldots, 500\}$. We evaluated two different settings, which we named the simple Brownian motion setting (BMS) and the extended setting using multiple regimes (ESMR). BMS refers to a standard simulation of Brownian Motion beginning at the root, creating ancestral values for each inner node. ESMR is an extension of BMS that randomly divides the tree into κ regimes of Brownian Motion with different variation parameters. Three different methods were compared for evaluation. We used the maximum likelihood method ("REML") (Felsenstein, 1985) and the generalized least squares method (Martins and Hansen, 1997) provided by the function `ape::ace` as

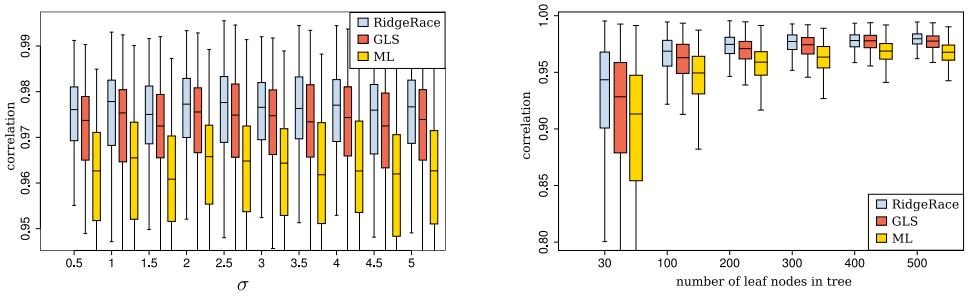
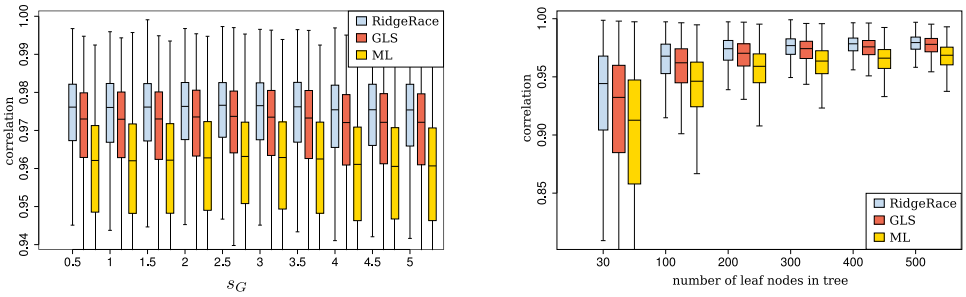


Figure 8.2: Pearson's correlation between inferred ancestral characters and true simulated values, when using maximum likelihood reconstruction (yellow), GLS (red), and RidgeRace (blue). The plot shows **(a)** the dependence of performance on the standard deviation σ of the BM process or **(b)** when increasing the number of leaf nodes in the tree.

well as our ridge regression method. For each tree and each character assignment, we provided the tree structure and the leaf node assignment to the reconstruction method, which created a prediction for the assignment of inner nodes. To estimate the correctness of a reconstruction method, we computed Pearsons correlation between the predicted values and the true simulated values at those inner nodes. Our evaluation showed that RidgeRace performs similarly or up to 3 percent points better than other state-of-the-art techniques. For the BMS evaluation, Figure 8.2 shows that all three methods are able to reconstruct ancestral states very well, achieving correlation values between 85 and 98 percent, even for very small trees or very large variation values. However, RidgeRace showed consistently better correlation values than the two reference methods. Performance was independent of the variation parameter for all methods (Figure 8.2a), but did increase with the size of the tree (Figure 8.2b). A similar observation was made for ESMR with variable rates. Similarly to the simple BMS, performance was independent of the size of range from which standard deviations for the tree regimes were drawn, and again the correlation between predicted and true values increased with the number of nodes in the tree (Figure 8.3). RidgeRace achieved correlation values consistently higher than the two other methods in all settings. When increasing the number of regimes, performance dropped slightly for all three methods, with the slope



(a) (b)

Figure 8.3: Pearson's correlation between inferred ancestral characters and true simulated values, with colors analogous to Figure 8.2. The plot shows (a) the dependence of performance on increasing the interval $\mathcal{U}(0, s_G)$ from which the rates of the BM processes of each of the κ single regimes are drawn. Figure (b) shows performance when increasing the number of leaf nodes in the tree.

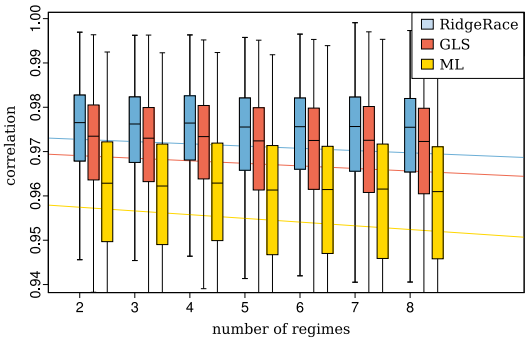


Figure 8.4: Pearson's correlation between inferred ancestral characters and true simulated values, with colors analogous to Figure 8.2. The plot shows the dependence of performance on the increasing number of regimes in the tree. Straight lines indicate a linear fit between the two variables.

of the linear fit being almost zero (see Figure 8.4, but still smallest (= slowest) for RidgeRace (RR: $-5.06 \cdot 10^{-4}$, GLS: $-5.78 \cdot 10^{-4}$, ML: $-8.38 \cdot 10^{-4}$, estimated using the R-funcion `stats::lm` (R Core Team, 2012)).

8.3 Exemplary application on a thaumarchaeota data set

One advantage of RidgeRace over other methods is that it allows the phenotypic rate β_i to change on every branch of the tree. Typically, the phenotypic rate is assumed to be constant. In more complex models, phenotypic rates may vary, but have to be

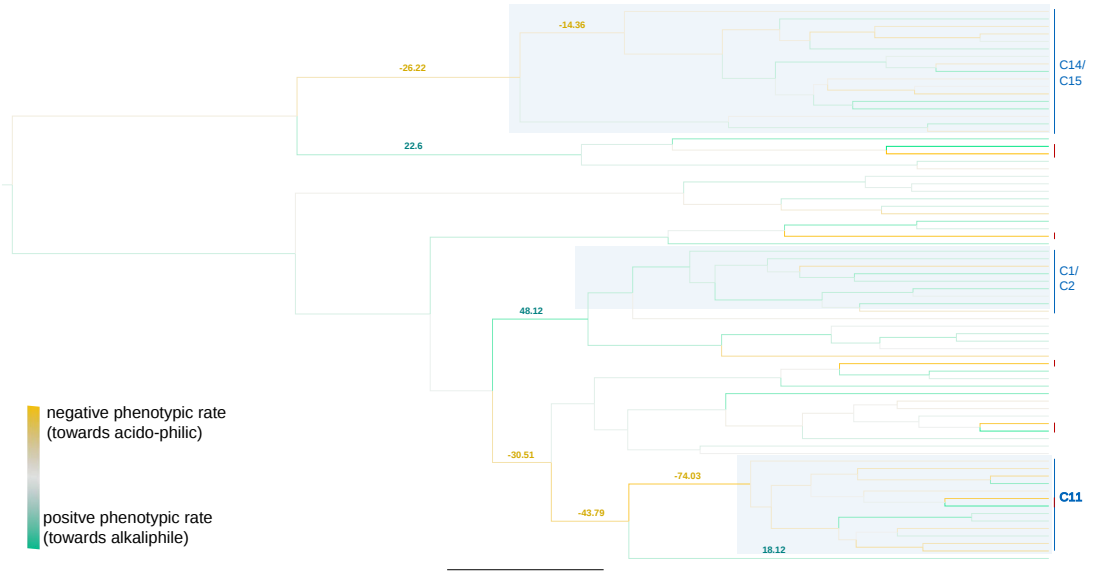


Figure 8.5: Visualization of inferred phenotypic rates (parameter β) for a RidgeRace reconstruction of thaumarchaeota *amoA* sequences (Gubry-Rangin and Hai, 2011). Strong orange or green colors indicate high positive or negative rates. Rates are particularly high directly after speciation to one of the phenotypically most specialized clusters (indicated by blue bars on the right side of the figure, absolute rates indicated directly on branches), or at certain leaf notes with strongly diverging phenotype (indicated by red bars).

known in advance to reconstruct ancestral character states[1] (Harmon *et al.* (2010) offers a review of such models). Other techniques test for changes in the phenotypic rate in predefined regimes of the tree (McPeek, 1995; O'Meara *et al.*, 2006; Revell, 2008), but do not reconstruct ancestral character states beside the root state, and again require specific a priori assumptions on the locations of the regimes.

In a recent study on the ecology and evolutionary history of terrestrial thaumarchaeota (Gubry-Rangin *et al.*, 2014, *in prep.*), we have analyzed the influence of the pH value of soil as an environmental factor that controls the adaptation of a specific lineage of archaea capable of ammonia oxidation (see Gubry-Rangin and Hai (2011) and Nicol *et al.* (2008) for details on the ecology and terrestrial distribution of thaumarchaeota).

[1] In fact, the GLSR approach is flexible enough to re-define the correlation structure provided by the phylogenetic tree in a completely arbitrary manner, allowing the inclusion of all kinds of branch transformations. However, it still requires the user to know the degree and position of the deviations from the basic model.

An additional RidgeRace analysis of the pH preferences of thaumarchaeota samples using a phylogeny inferred on *amoA* gene sequences (Figure 8.5) reconstructed the pH value of the root of the tree, *i.e.* the common ancestor of all thaumarchaeota, to 6.18, a value very similar to the reconstruction of 6.3 under a Brownian Motion model performed by the authors using the `ape` package in R (R Core Team, 2012; Paradis *et al.*, 2004)[2]. It also revealed that the phenotype (pH preference) has often evolved quicker on ancestral than on more recent branches of the tree, and in particular on branches directly after the separation of certain highly specialized pH clusters, such as the three main abundant clusters of terrestrial thaumarchaeota (marked by blue bars in Figure 8.5). This might indicate a particularly high speed of adaptation. RidgeRace assigned similarly high rates to a few samples with a pH preference that strongly deviated from the mean of their clade (marked by red bars). This can be considered an artifact of the method, but it could also be used as an indicator of interesting phenotypic outliers.

8.4 Example application to cancer data

Cancer as an evolutionary disease

According to the World Health Organization, cancer is a leading cause of disease-related deaths worldwide, responsible for 7.6 million deaths in the year 2008 (WHO, 2013a). The term "cancer" describes a variety of different diseases that may affect any part of the human body, with lung, stomach, liver, colon, and breast cancer being responsible for most of the cancer-associated deaths. The causes for cancer are not entirely described yet, but certain behavioral factors strongly contribute to the disease risk: about 30% of cancer deaths are related to high body mass index, low fruit and vegetable intake, lack of physical activity, and tobacco or alcohol consumption (WHO, 2013a).

Cancer is initiated by transformations in single cells that are caused by external factors such as physical, chemical, or biological carcinogens, or by a deficiency of cellular repair mechanisms (*e.g.* due to high age). The disease is the result of a complex interplay

[2]We performed Maximum Likelihood ratio tests to confirm that BM is indeed the most suitable model and that no signal of evolutionary trend is present in the data.

of genetic preconditions, external influences and interaction with the immune system. The main genetic factors (*hallmarks*) underlying the disease are summarized in two seminal papers by Hanahan and Weinberg (2000, 2011), which can be considered two of the most influential publications in the field (but see also the critique of Lazebnik 2010). At the time of creation of this thesis, the earlier paper has been cited more than 17,000 times[3]. For a wide variety of cancer types, recent studies identified genes that are significantly associated with cancer risk, onset, and progression (TCGAN, 2008, 2012a,b, 2013; Kandoth *et al.*, 2013; Röhr *et al.*, 2013).

Tumors are a heterogeneous population of cells that are the result of a shared process of evolution. In 1967, Nowell presented the hypothesis of cancer as an evolutionary disease, and discussed the interplay between cancer therapy and the evolution of tumor cell subpopulations. By now, a large number of observations have provided evidence for this theory. Several studies confirmed that the degree of genetic diversity in a tumor cell population is a good predictor for malignancy (Maley *et al.*, 2006). In 2006, Merlo *et al.* suggested to include the genetic instability preceding this diversity as an additional hallmark of cancer, and the concept was included as an "enabling characteristic" in the 2011 article by Hanahan and Weinberg. A multitude of known cancer diseases exists with a highly varying pathology and a similarly varying heterogeneity of involved pathways. However, there might also be a strong difference in tumor histologies between patients suffering from the same subtype (Yates and Campbell, 2012). Even within a single tumor tissue, sub-tissues show differing copy number profiles (Podlaha *et al.*, 2012).

RidgeRace for integrating cancer study data

To demonstrate a possible application of RidgeRace integrating phenotypic and genotypic data, we studied an ovarian cancer data set, created by the TCGA research network, and recently analyzed with Network Based Stratification (Hofree *et al.*, 2013). Hofree *et al.* argue that somatic mutations are likely to contain the causal drivers of tumor

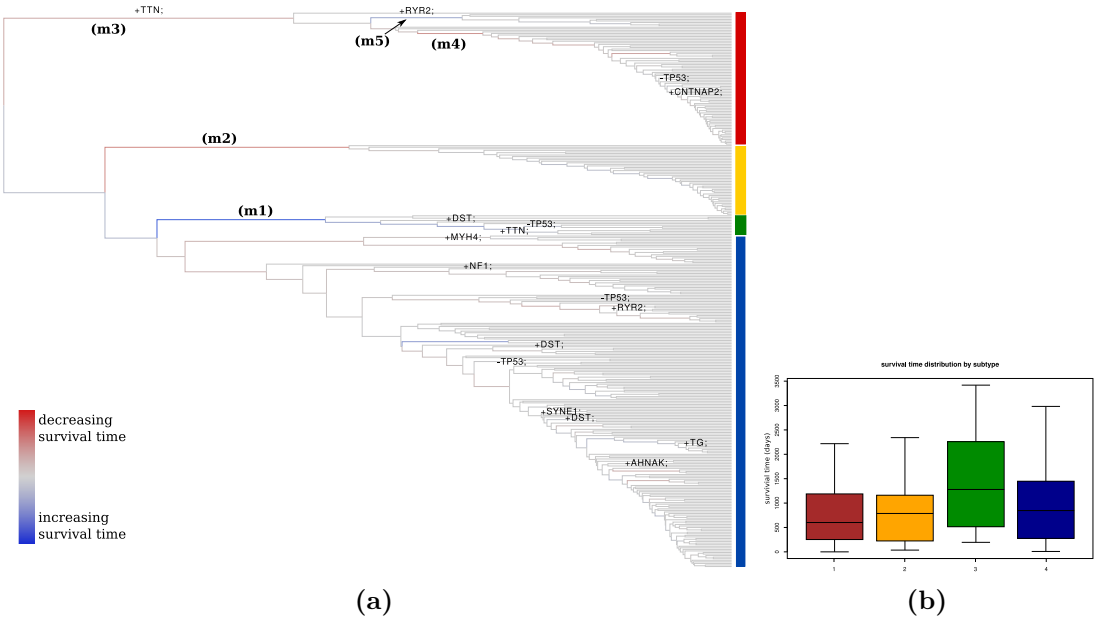
[3]according to Google Scholar, accessed 09/09/2013.

progression, and that this type of data provides a promising source of information to identify clinically relevant sub-clusters. These sub-clusters are identified with methods finding groups of samples with significant changes in their allele frequency profile, a process described as *stratification*. Hofree *et al.* note that tumors are very heterogeneous, and genetic profiles are sparse and vary strongly between patients, making clustering and stratification a challenging task. Network based stratification is a new clustering method that smooths genetic profiles with the help of gene interaction networks, and the authors show that it produces clinically meaningful clusterings. We use a data set and the software provided by the authors (NBS, version 0.2, available from the authors website[4]) to reconstruct a hierarchical clustering on somatic mutation data of ovarian cancer samples, creating a tree structure (Figure 8.6a).

Although it was not possible to determine if our inferred clustering was completely identical to the one of Hofree *et al.*, we similarly found that patients assigned to the smallest of the four subtypes showed an increased survival time (Figure 8.6b, green cluster). A RidgeRace analysis of patient survival time as a phenotype consistently showed a strong positive increase in rate at the branch leading to that cluster (Figure 8.6a, marker *m1*). Similarly, RidgeRace inferred a decrease in survival time for the branch leading to the yellow cluster (branch *m2*). Branch *m3* was associated with a rather small decrease in survival time, because the red cluster splits in distinct two subtypes with a successive second increase (branch *m5*) or a decrease (branch *m4*) in survival time, with branch *m4* leading to the majority of the red cluster, which had the lowest survival time of all four clusters.

As suggested above, the RidgeRace reconstruction can be combined with the reconstruction of discrete genetic events. We mapped the binary data encoding the absence or presence of non-synonymous mutations in a selection of genes to the tree (see Section 8.5 for technical details). The mapping confirmed the diverse nature of the somatic mutations. Only *P53* was found to be altered in almost all patients, and was reconstructed to have mutated at the root of the tree. Beside *P53*, only *TTN* was

[4]http://chianti.ucsd.edu/ mhofree/wordpress/?page_id=26



(a) **(b)**

Figure 8.6: RidgeRace application to a clustering on somatic mutations inferred for a ovarian cancer data set. Colors on the side of the tree indicate subtypes inferred with Network Based Stratification (Hofree *et al.*, 2013). Branches are colored according to the phenotypic rate parameter β, thickness of branches is proportional to the number of nodes below them. Branches leading to leaf nodes were colored grey for improved visibility. Labels *m1* to *m5* indicate branches with strong changes in patient survival time. Changes in the absence or presence of mutations in selected genes are indicated on all branches with 4 or more children.

reconstructed to appear on a higher level node, it is "gained" (mutated) at branch *m3*, and present in 83 of 85 patients of the red cluster. *RYR2* is gained on branch *m5* and present in 9 out of 85 patients of the red cluster. Besides these changes, no change appears on a branch higher than five levels below the root.

It is obvious that, although RidgeRace was able to reconstruct the main clusters of the phenotype distribution, no significant association between genetic aberrations and change in survival rate was found for this data. However, this study provides many insights for future improvements:

- The survival rate as a phenotype might be a very biased measurement, since it is based on the time of diagnosis and the (potential) death of the patient. A very late onset of therapy or simply a survival of the patient might influence this

measurement.

- The hierarchical clustering itself may be a source of bias: although Hofree *et al.* 2013 convincingly argue that their method produces clinically meaningful subclusters, the technique is based on a large number of parameters, among them the final number of main clusters, and the underlying gene interaction network (we used the parameters inferred by the authors). Hofree *et al.* suggest that future improvements of NBS may consider other alterations than non-synonymous mutations, or consider the length of genes.

- The subtypes identified by NBS may indeed represent the correct *genetic* stratification of the patients, nevertheless, their survival time may be dependent on many other factors, *e.g.* patient age and type of received therapy. Since RidgeRace essentially performs a regression on the patient data, such information can easily be included as additional covariates (features), controlling for the influence of such factors and providing insight into their relevance relative to the genetic factors.

8.5 Technical details of the method

RidgeRace weight inference

RidgeRace is primarily intended as a method to estimate ancestral character states on a phylogenetic tree. As in the original BM model, we consider the leaf values to be the result of a weighted sum of intermediate contributions g_i created along the tree, beginning at the root (see Figure 8.1). The contributions represent the gain or loss in character value on each branch of the tree, so that, for example, the character value of sample y_4 can be described as

$$y_4 = g_0 + g_a + g_b + g_c,$$

where a, b, c represent the branches in the tree, and g_0 holds a bias term representing the original contribution of the root node. The contribution g_j of a single branch j can

be seen in analogy to the formulation of BM: the gain or loss in phenotype is dependent on the length l_j of branch j and of the speed β_j of the process, in analogy to the variance term σ^2 in the BM model:

$$g_j = l_j \cdot \beta_j.$$

One can then write the solution for the vector of leaf phenotypes $\mathbf{y}$ in matrix form:

$$\hat{\mathbf{y}} = \mathbf{L}\beta, \tag{8.1}$$

where

$$\mathbf{L}_{i,j} = \begin{cases} l_j & \text{if branch } j \text{ is on the way from the root to sample } i \\ 1 & \text{if } j = 0 \\ 0 & \text{else} \end{cases}$$

and β is a vector of length equal to the number of branches in the phylogeny, including a single virtual branch above the root to account for its original contribution g_0. Note that this scheme allows an easy inclusion of measurements at inner nodes, *e.g.* from fossil records. It is also suited to account for multiple measurements at single nodes simply by adding additional rows to $\mathbf{y}$ and $\mathbf{L}$.

Ridge regression is a simple extension of ordinary least squares regression. As ordinary least squares, ridge regression also aims to minimize the squared error term, but adds a quadratic regularization penalty on large values of the weight vector β. A tuning paramenter λ controls the relative impact of both terms. The regularization does not only help to reduce the variance of the model, it also acts as an integrated paramter selection method for overparameterized models (see Gareth *et al.* (2013) for details). We use Ridge regression to estimate a vector $\hat{\beta}$ that explains the known observations $\mathbf{y}$ best:

$$\hat{\beta} = \arg\min_{\beta} \sum_i (y_i - (\mathbf{L}\beta)_i)^2 + \lambda \sum_j \beta_j^2, \tag{8.2}$$

and the text book solution (see, *e.g.* Hastie *et al.*, 2009) to this optimization problem

is

$$\hat{\beta} = \left(\mathbf{L}^T\mathbf{L} + \lambda\mathbf{I}\right)^{-1}\mathbf{L}^T\mathbf{y}. \tag{8.3}$$

Equation 8.2 shows how the optimization tries to balance the leaf reconstruction error versus a regularization term. This second term forces the correlated β_j and therefore the gains g_j to be distributed evenly across the tree, towards a globally constant rate (see Hastie *et al.* (2009) for a discussion on the influence of regularization terms). Without this term, a trivial but undesirable solution to the optimization would set the gain at each terminal branch equal to the according terminal node value, leaving all other gains empty and making ancestral reconstruction impossible.

For a given estimate of $\hat{\beta}$, the vector $\hat{\mathbf{a}}$ containing the phenotypic reconstruction of all inner nodes can then be computed analogous to Equation 8.1:

$$\hat{\mathbf{a}} = \mathbf{L}'\hat{\beta}, \tag{8.4}$$

where

$$\mathbf{L}'_{i,j} = \begin{cases} l_j & \text{if branch } j \text{ is on the way from the root to ancestor } i \\ 1 & \text{if } j = 0 \\ 0 & \text{else} \end{cases}$$

Note that this formulation is very similar to the generalized least squares method proposed by Martins and Hansen (1997). They similarly suggest to infer ancestral character states as weighted average of leaf contributions, with weights according to the covariance between an ancestor and a leaf (see equ. (10) in Martins and Hansen (1997), and Cunningham *et al.* (1998) for a worked example):

$$\hat{\mathbf{a}} = \mathbf{W}\mathbf{y} + \epsilon \tag{8.5}$$

$$\mathbf{W} = cov[\mathbf{a}, \mathbf{y}]var[\mathbf{y}]^{-1}, \tag{8.6}$$

where the covariance between an inner node a and a leaf node y is defined as $\sigma^2 t(a, y)$, with $t(a, y)$ being the distance between the root of the tree and the most recent common

ancestor of a and y (see Figure 7.4 for an example). RidgeRace differs in the sense that it allows to estimate a weight β_j for every branch instead of assuming a constant rate σ^2, or, more general, predefined covariances between nodes. Extensions of the simple GLS approach under the Brownian motion model use more complex matrices $\mathbf{W}$. However, the design of $\mathbf{W}$ has to be defined in advance based on specific model assumptions, whereas RidgeRace offers to estimate rates independently.

Estimation of regularization weight

The regularization weight parameter λ in Equation 8.2 balances the impact of accuracy at the leaves versus the complexity of the model and variance of β. To find the optimal value of λ, we performed a leave-one-out iteration over all leaves of the tree. To estimate the goodness-of-fit of a particular λ_0, we iteratively removed a single leaf x from the tree, estimated β on the remaining tree, and used the rate of the branch leading to the father of x as approximation for the branch rate of the missing node. The leave-one-out error for x is defined as the difference between the inferred phenotypic value for x and the actual value according to the input data. The leave-one-out error for a particular λ_0 is the sum over all leave-one-out errors for all leaves. Iterating $\lambda_0 \in \{10^{-5}, 10^{-4}, \ldots, 10^{+2}\}$, we selected the final λ to be the one that minimizes the leave-one-out error.

Simulation study

We created random trees with an increasing number N of leaves using the function `rtree` in the R-package textttape (R Core Team, 2012; Paradis *et al.*, 2004). For a first evaluation (BMS), we simulated Brownian motion with variation σ^2 along the branches of the tree, resulting in a character assignment for every inner or leaf node. The parameter σ^2 was iteratively increased in each round of simulation. In the extended setting (ESMR), we simulated changing rates of evolution in the tree by dividing the tree into κ different regimes, and in every regime r_i, a new rate σ_i^2 was drawn at random from a global uniform distribution in the interval $\mathcal{U}(0, s_G)$. To make the process increasingly more variable and difficult, the size of that interval s_G was iteratively

increased with every simulation, thus allowing larger σ_i to be drawn. The simulation of Brownian motion in each regime was performed using the respective σ_i^2, resulting in a character assignment for all nodes. This process was repeated several times and for different parameters σ^2, s_G, κ and N. See Chapter D and supplementary Figure D.1 and Figure D.2 for details on the simulation algorithms and parameter settings. The random tree and the simulated values obtained at the leaf nodes were provided as input to RidgeRace, and an implementation of the ML and GLSR algorithms in the `ape` function for ancestral character state estimation (`ace`). Obtained reconstructed values were mapped back to the inner nodes of the tree and compared with the simulated ones using Pearson's correlation coefficient (leaf nodes were excluded).

Analysis of thaumarchaeota data

The pH preferences of 425 thaumarchaeota samples of 16 subtypes and a phylogeny based on their *amoA* genes was obtained from a collaboration project (Gubry-Rangin *et al.*, 2014, *in prep.*). PH preferences were mapped to the leaves of the phylogeny and ancestral values were reconstructed with RidgeRace. Inferred phenotypic rates β_j were visualized using FigTree (Rambaut, 2013). In the collaboration project, we used the given phylogeny for maximum likelihood ratio tests to infer the most suitable model of continuous phenotype evolution. Among others, we compared Brownian Motion to models containing Ornstein-Uhlenbeck processes with different numbers of regimes, an early burst, and a trend model (Harmon *et al.*, 2008). The ML ratio tests inferred BM to be the best fitting model.

Preprocessing of cancer data

A binary matrix describing the absence or presence of non-synonymous mutations in 9850 genes for 325 patients was taken from the supplementary data provided by Hofree *et al.* (2013). As indicated by the authors in their article and supplementary material, we used their software with 4 clusters and the HM network, and default parameters, creating 1000 bootstrap samples. We then inferred a hierarchical clustering (average linkage) on

	100	200	300	400	500
RidgeRace	4+0	29+1	135+1	1074+7	3372+24
GLS	1	2	3	7	10
REML	2	6	15	38	123

Table 8.1: Comparison of average running times in seconds for RidgeRace and the `APE` implementations of GLS and REML, shown for trees of different size, ranging from 100 to 500 leaf nodes. RidgeRace running time is provided as the running time required for full λ inference plus time required for ACR.

the bootstrap similarity matrix with NBS methods and used the inferred topology as input tree for RidgeRace. Information on patients survival rate was downloaded from the TCGA database

(TCGAN, 2011). Phenotypic rates were inferred with RidgeRace as described above. The binary genetic profile of each patient was mapped to the leaf nodes and reconstructed to inner nodes with the Sankoff algorithm implemented in RidgeRace, using a simple 0/1 cost matrix and the ACCTRAN principle in case of ambiguities. "Mutations", *i.e.* changes in genetic profile, were then reconstructed on the branches of the tree. Finally, the tree was visualized using FigTree (Rambaut, 2013).

System requirements

RidgeRace requires only minimal system ressources (<100 MB RAM). The C++ implementation relies on the `boost ublas` library (BOOST, 2014) to solve the ridge optimization (Equation 8.3). The running time of a full RidgeRace inference is larger than the time required by comparable methods (Table 8.1, measured using an Intel Xeon X5660 with 2.8 GHz), but still within the range of a few minutes. The large majority of the running time for RidgeRace is consumed by estimating the λ parameter, performing a leave-one-out iteration over all leaf nodes of the tree and testing $\lambda \in \{10^{-6}, 10^{-4}, \ldots, 10^{+2}\}$. Decreasing the evaluation range for λ or performing the leave-one-out iteration only on a subset of the leaf nodes can considerably decrease running time for larger trees.

Part IV

Synopsis and Outlook

Synopsis and outlook

The main objective of this thesis was to study the evolution of influenza A viruses and to detect regions and markers on the influenza haemagglutinin protein that are associated with immune evasion. To this end, two methods were contributed that identify regions under selection or of high antigenic impact. The main observations made by the two studies are the following:

- Both types of methods find functionally relevant regions that form distinct clusters on the globular head region of the protein (see Figure 5.3).
- Both AdaPatch and other methods (Bush *et al.*, 1999; Suzuki, 2006) observe that sites under positive selection and epitope sites do not overlap very well, many epitope sites are conserved, and many non-epitope sites feature high dN/dS ratios. Although AdaPatch performs better at identifying epitope sites than a simple dN/dS-based method, we concluded that positive selection is no optimal indicator for sites of antigenic relevance (compare Supplementary Figure C.1).
- To overcome this problem, AntiPatch directly identified patches of antigenic

impact. However, we again observed only a limited overlap with epitope sites. We concluded that the epitope site definition (of influenza haemagglutinin of subtype H3) may be very broad, and that AntiPatch provides a more specific selection of regions of antigenic impact. In particular, our results are in agreement with other authors who note that only very few sites are of high impact for the antigenic evolution (Smith *et al.*, 2004; Koel *et al.*, 2013). However, in contrast to these methods, AntiPatch offers the advantage to consider (practically) all sites of the protein, not only a predefined selection.

- A comparison between the patches selected by both methods indicated that the overlap between sites under positive selection and antigenicity-altering sites is surprisingly small. We concluded that this could be an indicator showing that besides directional selection for change in antigenicity, other mechanisms shape influenza evolution.

In addition, this thesis contributed RidgeRace, a method to reconstruct continuous ancestral character states without requiring explicit assumptions on the underlying model of evolution. The phenotypic rates inferred by the method can be used to judge the phenotypic impact of events reconstructed on the underlying phylogeny. We provided two examples to demonstrate an application of the method.

From a biologists perspective, AdaPatch and AntiPatch are two methods to study the genetic and antigenic evolution of influenza A with a focus on the haemagglutinin protein, which is the protein most relevant for viral immune evasion and, thus, for vaccine design. Our observations reveal potential problems in the definition of epitope sites and gaps in our current understanding of the reasons for positive selection in influenza A haemagglutinin. A more precise specification of sites relevant for human host adaptation may contribute to current process of vaccine development. To predict the genetic and antigenic viral evolution for the biannual vaccine update, a susceptible-infected-recovered (SIR) model has recently gained wide attention (Luksza and Lässig, 2014). The authors consider amino acid changes within epitope regions as candidates to

decrease cross-immunity, and thus strictly advantageous; changes outside epitope regions are assumed to reduce protein stability and are considered deleterious. Particularly the results of our AntiPatch study indicate that such models can be improved further.

The graph cut clustering principle can easily be applied to other proteins or measurements. In particular, it could be promising to compare the known haemagglutinin patches with patches associated to receptor avidity (Hensley *et al.*, 2009) or similar traits to clearly identify substructures on the protein and their respective function(s). There is also evidence that the influence of the influenza A neuraminidase protein has been underrated so far (Sandbulte *et al.*, 2011), and neuraminidase evolution, as well as neuraminidase inhibition assays (WHO, 2012), could reveal interesting aspects. An analysis of other influenza proteins, in particular those discovered only recently (Wise *et al.*, 2009, 2012; Jagger *et al.*, 2012) might provide further insights into the ongoing adaptation of the virus to improve replication within the human body and release of viral offspring. As a more general extension, the graph cut clustering principle can be extended even further to combine any measure of relatedness with any measure of phenotypic relevance. In future extensions and applications in other contexts, relatedness could be described in terms of epistasis (Kryazhimskiy *et al.*, 2011), in terms of some (*e.g.* expression-) profile similarity, or in terms of functional similarity (see, for example, the use of the STRING database in the network clustering of Hofree *et al.*, 2013).

For general sequence data of related individuals or species, RidgeRace allows a way to reconstruct continuous ancestral character states that is competitive to state-of-the-art methods, and provides easy to interpret estimates of phenotypic rate along the branches of the phylogeny that can be correlated with other observations. It can be applied in any context where a tree-like topology and annotation for the leaves is available, and we suggested two potential applications. We discussed literature arguing that cancer is a process of evolution, which is, in some properties, comparable to influenza A. We argued that the use of methods aware of the hidden similarities can be of advantage, and suggested to apply RidgeRace to cancer data to identify genetic aberrations like SNPs or CNVs associated with the pathogenicity of the disease. Our first attempt

at an application to a clustering of (evolutionary unrelated) cancer data samples (see Section 8.4, but also Klesper (2013) for a related idea) revealed many difficulties, mostly due to the extremely sparse nature of the somatic mutation data, and due to potentially high levels of noise in the phenotype measurements, and we discussed a number of potential improvements. Nevertheless, we are convinced that RidgeRace can be a helpful tool in a wide variety of biological questions.

From a computer scientists perspective, we contributed methods that consider the shared inheritance of biological samples. As a second, minor objective, this thesis aimed to explain the statistical problems introduced by this very typical property of biomedical data, and to show how some existing methods and the ones introduced in the thesis account for these problems. We argued that, irrespective of the actual type of data, it can be a good general approach to map the inferred differences between two samples to a phylogeny (or clustering topology), instead of comparing them directly: AdaPatch maps mutations to the phylogeny to estimate dN/dS ratios, AntiPatch maps continuous distance matrices (using a method by Steinbrück and McHardy, 2012), and RidgeRace infers differences implicitly when only single phenotype measurements are provided.

Part V

Appendix

APPENDIX A

Projects involved in this thesis

Tusche C, Steinbrück L, and McHardy AC (2012). Detecting Patches of Protein Sites of Influenza A Viruses under Positive Selection. *Molecular Biology and Evolution*, **29**(8):2063–2071 .

We use a graph cut approach to find dense patches of sites under positive selection on the protein structure of several influenza A proteins.

Discussed in Chapter 4.

75% Adapted & implemented the method (graph cut clustering), performed research (with LS & AM), analyzed results (with AM), wrote manuscript (with AM).

Kratsch C, Klesper L, Steinbrück L, and McHardy AC (2014). Determination of antigenicity-altering patches of sites on the hemagglutinin of human influenza A/H3N2 viruses. pages 1–29 .

Project under Review at *Plos Computational Biology*. We extended the AdaPatch approach to find patches of sites of antigenic relevance.

Discussed in Chapter 5.

50% Implemented the algorithm (initial development by CK, extension by LK, refinement by CK), performed research (with AM), analyzed & interpreted results (with AM), performed additional experiments, wrote manuscript (with AM).

Kratsch C and McHardy AC (2014). RidgeRace: Ridge regression for continuous ancestral character estimation on phylogenetic trees. *Bioinformatics*, **30**(17):i527–i533 .

Presented at ECCB 2014. We describe a method to perform ancestral character state reconstruction for continuous characters without requirement of any model specifications.

Discussed in Chapter 8.

80% Implemented the method, performed research (with AM), prepared data & analyzed example results, wrote manuscript (with AM).

Gubry-Rangin C, Macqueen D, Kratsch C, McHardy AC, and Prosser J (2014). Evolutionary history of terrestrial thaumarchaea. in preparation .

Minor project, in preparation. We provide a review of the evolutionary development and ecology of thaumarchaeota.

Extension of the article is discussed as example in Section 8.3.

15% Helped with tree creation, performed model selection and ancestral reconstruction, documented method in manuscript (main manuscript by CGR).

Röhr C, Kerick M, Fischer A, *et al.* (2013). High-Throughput miRNA and mRNA Sequencing of Paired Colorectal Normal, Tumor and Metastasis Tissues and Bioinformatic Modeling of miRNA-1 Therapeutic Applications. *PloS One*, **8**(7):e67461 .

Minor project. We performed profiling of miRNA and mRNA expression in matching normal, tumor and metastasis tissues of eight colorectoral cancer patients and identified miRNAs significantly associated with the disease.

Mentioned as example in Section 8.4.

$< 10\%$ Helped with sequencing analysis, performed statistical tests for miRNA expression.

APPENDIX B

AdaPatch: Supplement

Supplemental Text:

Additional information on selection of parameters

As described in Chapter 4, the minimum graph cut is a graph cut that minimizes the sum E of the weights of the edges connecting the positive and the negative selection halves:

$$E = \sum_{n \in Pos} p(n) + \alpha \sum_{n \in Neg} \overline{p}(n) + \beta \sum_{n \in Pos} \sum_{\substack{m \in Neg \\ m \in N_{\delta(n)}}} e^{-dist(m,n)},$$

The factor $\beta \in [0, 1]$ weighs the influence of the distance statistic. For $\beta = 0$ (and very small β), the method assigns sites with small p-value to the positive selection halve, and sites with high p-value to the negative selection halve, so that the two resulting p-value sums show a ratio of approximately 1:α (we set $\alpha = 1$). The method classifies based on p-value alone and ignores the distance of neighboring sites.

For large β, the influence of the distance of residues becomes more and more important, and the method searches for a small, dense set of residues that spatially separate well from the rest. For very large β, assigning residues to different sides becomes too expensive, so that all residues are assigned to one (here by default: the negative) selection side. Since the absolute values of the p-value and the distance statistics are dependent on the number of residues in the protein, β has to be set manually. We searched for suitable values by setting $\beta = 1/x, x \in \mathbb{N}$. As an example, for the seasonal influenza H1 data, we observed that values of $\beta > 1/30$ led to empty selections, whereas smaller values led to the final selection of all reported 35 residues. Decreasing β below 1/30 added more and more residues to the positive selection side, up to 90 residues for $\beta < 1/150$, but those additions were always removed by the final clustering step that grouped patches and removed those with two or less residues.

Supplemental Table

Hemagglutinin antigenic sites used for evaluation

H1	Sa	128, 129, 158, 160, 162, 163, 165-167
	Sb	156, 159, 192, 193, 196, 198
	Ca1	169, 173, 182, 207, 240, 273
	Ca2	140, 143, 145, 224, 225
	Cb	78, 79, 81-83, 122
H3	A	122, 124, 126, 130-133, 135, 137, 138, 140, 142-146, 150, 152, 168
	B	128, 129, 155-160, 163-165, 186-190, 192-194, 196-198
	C	44-48, 50, 51, 53, 54, 273, 275, 276, 278-280, 294, 297, 299, 300, 304, 305, 307-312
	D	96, 102, 103, 117, 121, 167, 170-177, 179, 182, 201, 203, 207-209, 212-219, 226-230, 238, 240, 242, 244, 246-248
	E	57, 59, 62, 63, 67, 75, 78, 80-83, 86-88, 91, 92, 94, 109, 260-262, 265

Table B.1: Antigenic sites for influenza A haemagglutinin, of subtypes H1 and H3 (Caton *et al.*, 1982; Wiley *et al.*, 1981)

APPENDIX C

AntiPatch: Supplement

Supplementary Figures

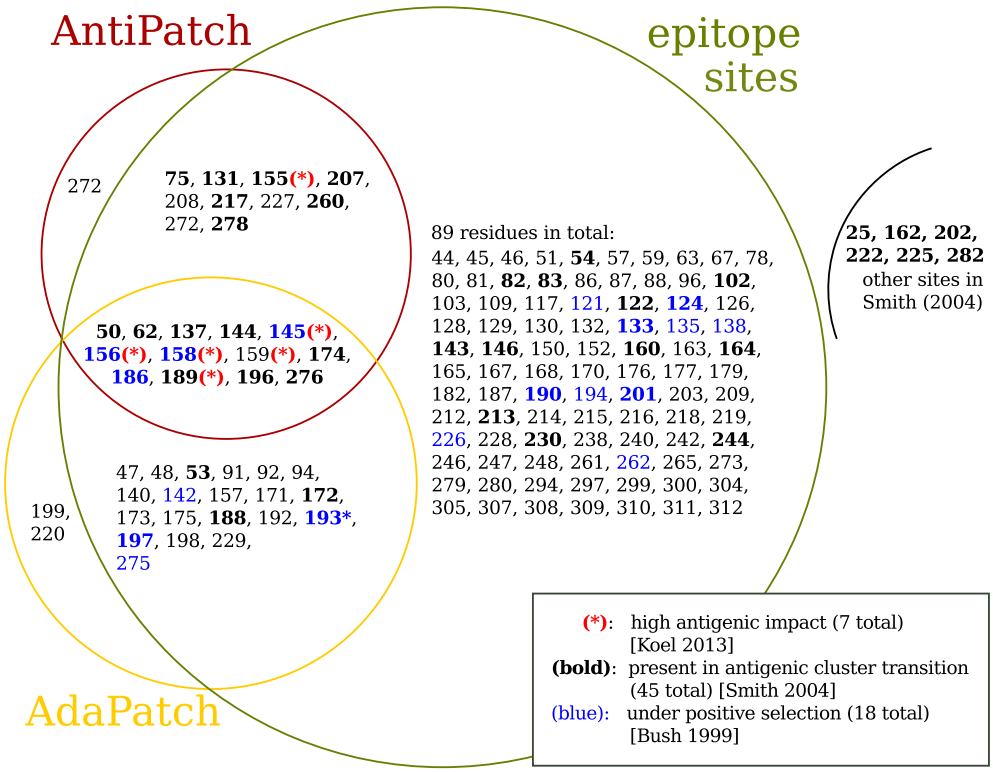


Figure C.1: Venn diagrams in analogy to Figure 5.3D. Starred sites are of large antigenic impact according to Koel *et al.* (2013); bold sites are present in antigenic cluster transitions according to Smith *et al.* (2004); blue sites are under positive selection according to Bush *et al.* (1999). All numbers are according to H3 numbering (Aoyama *et al.*, 1991)

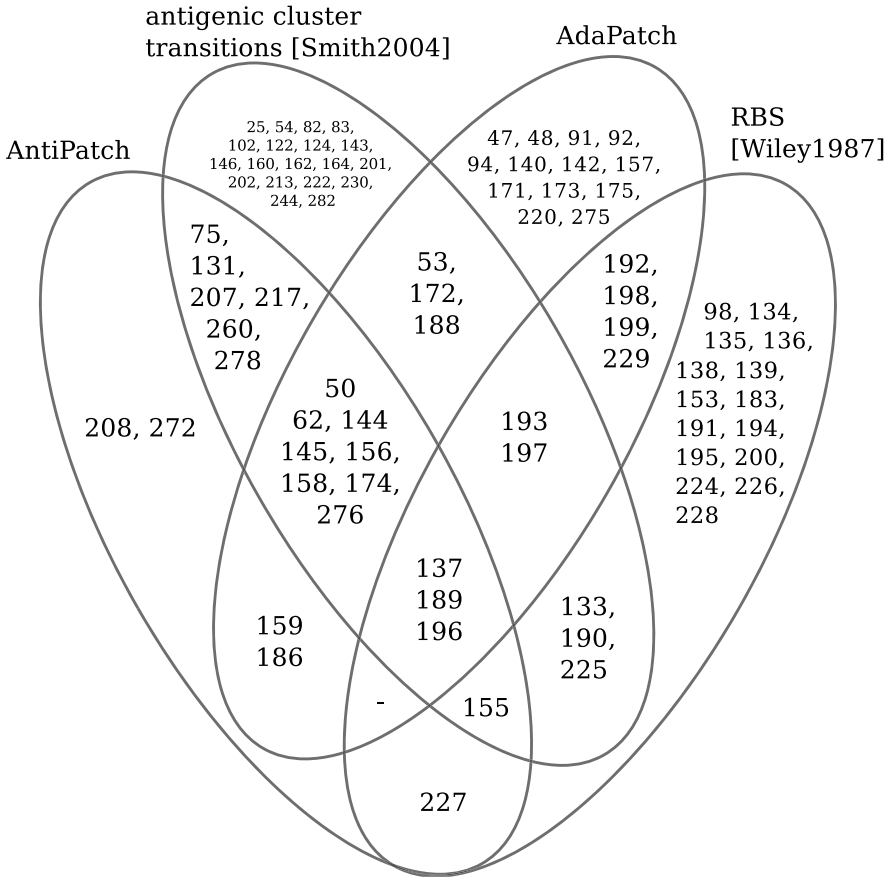


Figure C.2: Venn diagram showing the overlap of different kinds of annotation of influenza A/haemagglutinin sites, comparing sites from antigenic patches to patches of sites under selection, sites involved in antigenic cluster transitions (Steinbrück and McHardy, 2012) and sites on the RBS. All numbers are according to H3 numbering (Aoyama *et al.*, 1991).

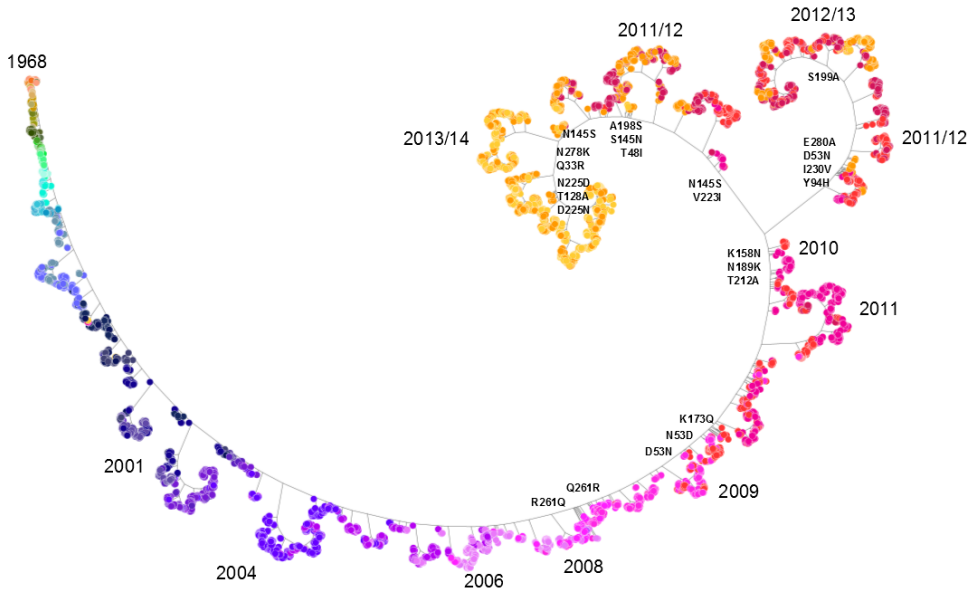


Figure C.3: Phylogeny showing the genetic evolution of influenza A haemagglutinin sequences of subtype H3 since 1968. A single trunk line (and a major branch around 2011) shows the evolution of haemagglutinin since 1968. Amino acid changes occuring on the trunk and the major branch line are indicated for the subtree holding strains that occured after 2002, and changes on patch sites are colored in green. Note that branches are scaled for visualization purposes only and do not represent molecular distances.

Supplementary Tables

cluster transition	sites involved	patches involved
HK68-EN72	122,144,155,188,207	1,2,5
EN72-VI75	137,145,164,189,193, 53,278,174,102,213,217,230	1,2,3,6
VI75-TX77	137,158,164,193,50, 53,174,201, 213, 230, 82, 260	1,2,3,6
TX77-BA79	133,143,146,156,160, 197,53,54,172, 217, 244,162,82	1
BA79-SI87	124,155,189	1
SI87-BE89	145	2
BE89-BE92	133,145,156,190,262	1,2
BE92-WU95	145	2
WU95-SY97	62,156,158,196,276	1,3,4
SY97-FU02	131,155,156,50,75, 83,25,202,222,225	1,3,4

Table C.1: Changes between consensus sequences representing consecutive predominating and antigenic variants of human influenza A7H3N2 viruses according to (Smith *et al.*, 2004), and patches involved according to Table 5.1.

APPENDIX D

RidgeRace: Supplement

Parameter settings

Globally constant rate

To estimate the performance of RidgeRace, we iteratively created random trees of size $N = \{30, 100, 200, \ldots, 500\}$ leaf nodes for $L = 10$ times. The ancestral grand mean at the root was set to $\mu_r = 0$. For each tree, we simulated Brownian motion along the tree and repeated each simulation step $M = 10$ times for the same tree. Every simulation was governed by a variance σ^2, ranging from $\sigma_2^2 = 0.5$ to $\sigma_e^2 = 5.0$ and incremented by steps of $\sigma_c^2 = 0.5$. After each round of BM simulation, ancestral values were defined for all internal nodes as well as all leaf nodes. We provided the leaf values as input to the inference algorithms for ML, GLS and RidgeRace, and compared their inferred output to the true simulated values hidden from them. RidgeRace requires a parameter λ, which we globally defined to be $\lambda = 10^{-5}$.

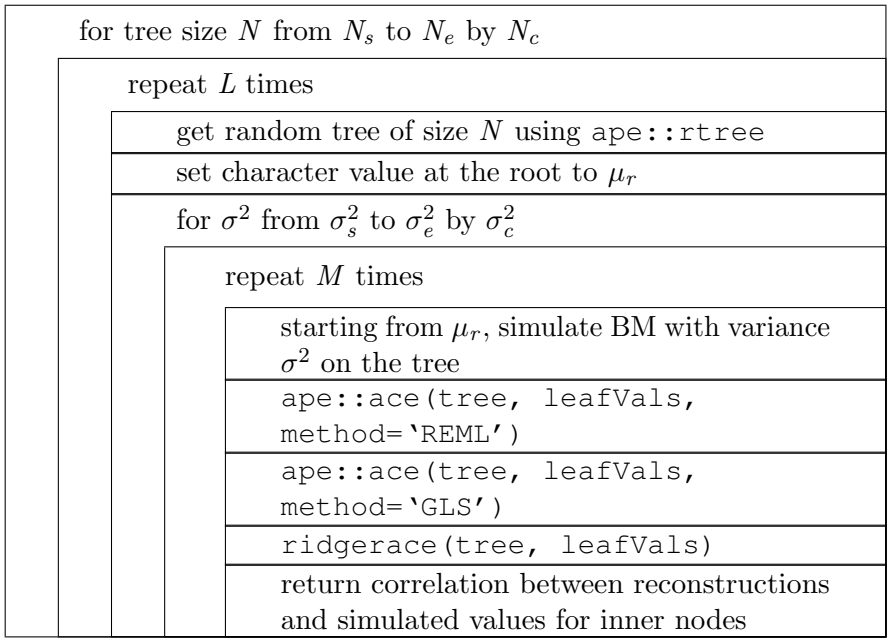


Figure D.1: Block diagram visualizing the iterative structure of the simulation and evaluation algorithm for the simple constant rate case.

Variable rate

The extended simulation and evaluation algorithm is very similar to the simple case of a globally constant rate, and the respective parameter settings are identical. However, the obsolete parameters σ_s^2, σ_e^2, and σ_c^2 are replaced by $s_G^{min} = 1.5$, $s_G^{max} = 5.0$ and $s_G^c = 0.5$, describing the maximum variance in the regimes from which the rate σ_i^2 is drawn for each regime i. In each simulation step, the tree was randomly divided into κ regimes, up to a maximum of $K = 8$, and for each regime a BM process was simulated, again repeating the simulation $M = 10$ times.

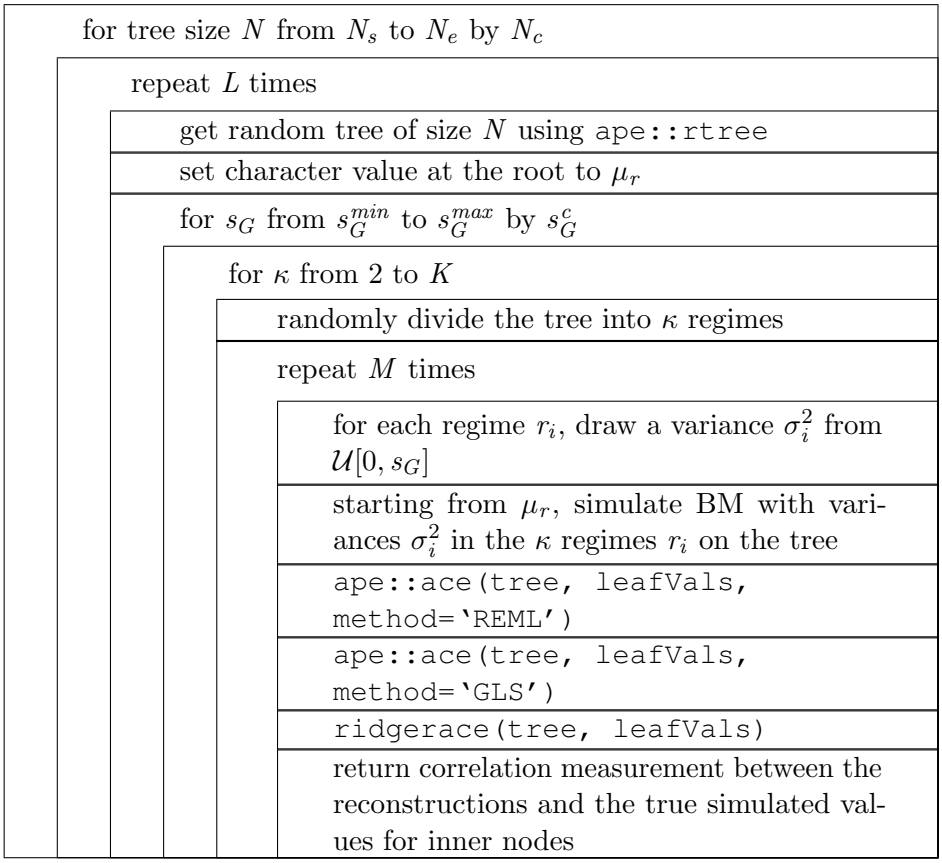


Figure D.2: Block diagram visualizing the iterative structure of the simulation and evaluation algorithm for the extended case with evolutionary rates randomly drawn for each regime from $\mathcal{U}[0, s_G]$.

APPENDIX E

Additional source code

This chapter contains instructions how to recreate some of the results not discussed in the three articles associated with the thesis. Software for AdaPatch, AntiPatch and RidgeRace can be found in the supplement of the respective papers.

ACR on diet of Galapagos finks

This R code can be used to create a basic version of the tree shown in Figure 2.3.

Listing E.1: R code to reconstruct ancestral diet of finches

```
# download tree from
# http://www.r-phylo.org/wiki/HowTo/DataTreeManipulation
geotree <- read.nexus("Geospiza.nex")
g <- "Granivore"
i <- "Insectivore"
f <- "Foliovore"

diet <- c(g,g,g,g,g,g,i,i,i,f,i)
names <- c("fuliginosa", "fortis", "magnirostris",
  "difficilis", "conirostris", "scandens",
  "parvulus", "pauper", "pallida",
  "Platyspiza", "olivacea")
drop <- ! geotree$tip.label %in% names
tree <- drop.tip(geotree, geotree$tip.label[drop])

rec <- ace(x = diet, phy = tree, type = "discrete")

plot(tree, label.offset=0.1)
cols <- c("blue", "brown", "orange")
nodelabels(pie=rec$lik.anc, cex=0.8, piecol=cols)

legend( x="bottomleft", fill=cols, legend=c(f, g, i))
```

Epistasis and spatial distance

This Java code can be used to extract pairs of sites from a phylogeny that appear within a branch length of 0.1 or less, as discussed in Section 7.1 and shown in Figure 5.4a. A tree can only be practically traversed in a recursive fashion. However, the "path" of length 0.1 can stretch over several branches/ nodes/ function calls. It is therefore necessary to keep track of all mutations seen within a distance of less than 0.1 units. The algorithm assumes mutations to occur in the middle of a branch, and follows five steps:

1. Copy the list of all epistasis interactions found so far.
2. Assume there are N mutations occurring on the current branch between the current node N and its father. Add all $N \times (N-1)$ pairwise interactions as epistasis events with distance zero.
3. For a mutation m on the current branch and a mutation X in `ancestralMuts`, check if distance of X plus 1/2 of current branch length is smaller than 0.1. If so, add to epistasis interactions.
4. To all old mutations in `ancestralMuts`, add a distance of current branch length.
5. Add all current mutations with distance 1/2 branchLength to the list.

The final result is stored in `ancestralMuts` when calling `extractEpistasisInfo` for the root of the tree. This code builds on an implementation of the *PhyloTreeTools* library by Lars Steinbrück, which defines the Node class and takes care of user input, internal sequence mapping and inference of mutation. The method was applied to the H3N2 tree inferred for the AdaPatch study.

Listing E.2: Java Code to extract epistatic pairs

```
import java.util.ArrayList;

public class EpistasisInfoReader {

  Node root;
  double thresh = 0.1f;

  // focus on the basic length of influenza sequences
  int SEQUENCE_MAX = 550;
  int SEQUENCE_MIN = 20;

  /**
  OTHER METHODS (initialization, printing, ...)
  */

  /**
   * Update the given EpistasisPairArray based on the mutations found on the
   * current node and the mutations found so far.
   * @param ancestralMuts    all mutations that occurred on upwards branches
   *        (= towards the root)
   * @param e   all valid epistasis interactions observed so far
   * @param n   the current node/branch
   */
  //EpistasisPair[][] extractEpistasisInfo(Node n, ArrayList<Mutation> ancestralMuts) throws Exception {
  void extractEpistasisInfo(Node n, ArrayList<Mutation> ancestralMuts) throws Exception {


    // step 1
    // the list of mutations on this branch:
    ArrayList<Mutation> allMutPos = new ArrayList<Mutation>();

    // step 2a: find out all positions mutating on this branch
    char curChar, fatChar;
    // but do this only if it's not the root
    if (n.parent != null) {
      for (int i = SEQUENCE_MIN; i < n.aaSequence.length(); i++) {

        if (i>=SEQUENCE_MAX ) break;

        curChar = n.aaSequence.charAt(i);
        fatChar = n.parent.aaSequence.charAt(i);

        if (curChar == '-' || fatChar == '-') {
          continue;
        }

        // if mutation
        if (curChar != fatChar) {
          Mutation m = new Mutation();
          m.dist = 0;
          m.pos = (i+1);
          m.from = fatChar;
          m.to = curChar;
          allMutPos.add(m);
```

```
    }
  }
}

System.err.println("branchLength: " + n.distance + "\tnrOfMuts: " + allMutPos.size());

// add all pairwise interactions
// between mutations on the current branch
for (int i = 0; i < allMutPos.size(); i++) {
  for (int j = 0; j < allMutPos.size(); j++) {
    if (i == j) {
      System.err.println("mutation: " + allMutPos.get(i).pos);
    }

    Mutation mut1 = allMutPos.get(i);
    Mutation mut2 = allMutPos.get(j);

    if (mut1.pos == mut2.pos) continue;

    EpistasisPair ep = new EpistasisPair(mut1.pos,mut2.pos);

    ep.distance.add(new Double(0.0));
    ep.x = mut1.pos;
    ep.y = mut2.pos;

    ep.fromCharX.add(new Character(mut1.from));
    ep.toCharX.add(new Character(mut1.to));
    ep.fromCharY.add(new Character(mut2.from));
    ep.toCharY.add(new Character(mut2.to));

    ep.write();
  }
}

// step 3 and 4
ArrayList<Mutation> updatedAncestralMuts = new ArrayList<Mutation>();
Mutation ancMut;

// run through all ancestral mutations that are still relevant at
// this point of the tree
for (int i = 0; i < ancestralMuts.size(); i++) {
  ancMut = ancestralMuts.get(i);

  // if this is a relevant ancestral mutation, it is a candidate
  // for step 3, and maybe for step 4
  double newDist = ancMut.dist + (n.distance / 2);
  if (newDist < this.thresh) { // still close enough?

    // step 3:
    // then add epistasis events between m and all current
    // mutations to res;
    for (int j = 0; j < allMutPos.size(); j++) {
      Mutation curMut = allMutPos.get(j);

      if (ancMut.pos == curMut.pos) continue;
```

```
            EpistasisPair ep = new EpistasisPair(ancMut.pos,curMut.pos);
            ep.distance.add(new Double(newDist));
            ep.x = ancMut.pos;
            ep.y = curMut.pos;
            ep.fromCharX.add(new Character(ancMut.from));
            ep.toCharX.add(new Character(ancMut.to));
            ep.fromCharY.add(new Character(curMut.from));
            ep.toCharY.add(new Character(curMut.to));

            ep.write();
          }

          // step 4: increase distance for ancestral muts
          // maybe the sum will be > thresh already, but we'll check
          // that in the child node
          ancMut.dist += n.distance;

          // if the distance to ancMut is still small enough, it might
          // still be relevant for another epistasis interaction
          // further down the tree
          if (ancMut.dist < this.thresh) {
            updatedAncestralMuts.add(ancMut);
          }
        }
      }

      // step 5
      for (int j = 0; j < allMutPos.size(); j++) {
        Mutation curMut = allMutPos.get(j);
        curMut.dist = n.distance / 2;
        updatedAncestralMuts.add(curMut);
      }

      // recursively call for children
      for (int i = 0; i < n.numChilds; i++) {
        extractEpistasisInfo(n.childArray[i], updatedAncestralMuts);

      }
    }
  }
```

List of Figures

1.1 Example for a phylogenetic tree: influenza A 5

2.1 The three main modes of selection 10
2.2 Example for a phylogenetic tree: tree of life 12
2.3 Exemplary ancestral character state reconstruction for 11 finches 14

3.1 Influenza pandemics 22
3.2 Antigenic cartography of influenza A (H3N2) from 1968 to 2003 28

4.1 AdaPatch: general workflow 32
4.2 AdaPatch: overlap between selected epitope and avidity-changing sites . 34
4.3 AdaPatch: patches under positive selection on seasonal HA 34
4.4 AdaPatch: patches on HA and PB2 (2009 H1N1pdm) 36
4.5 AdaPatch: epitope sites not under positive selection 38
4.6 AdaPatch: work flow for predicting patches under positive selection . . 40
4.7 AdaPatch: schematic drawing of the graph-cut approach 42

5.1 AntiPatch: antigenic weights on HA 47
5.2 AntiPatch: distribution of antigenic weights and RSA on HA 48
5.3 AntiPatch: position of antigenic patches on influenza H3 49
5.4 AntiPatch: influenza HA - distances of pairs in epistasis 60

7.1 Increasing feature and sample counts 72
7.2 Example for phylogenetic correlation 76
7.3 Examples for tree mapping procedures 77
7.4 Exemplary usage of GLSR 78

8.1 RidgeRace: model of phenotype evolution on a phylogenetic tree 82
8.2 RidgeRace: increasing σ or number of leaves, constant tree-wide σ . . . 84
8.3 RidgeRace: increasing s_G or number of leaves, variable tree-wide σ . . . 85
8.4 RidgeRace: increasing number of regimes, variable tree-wide σ 85
8.5 RidgeRace: phenotypic rates for thaumarchaeota *amoA* 86
8.6 RidgeRace: application to ovarian cancer 90

C.1 AntiPatch: Venn diagram showing annotation overlaps (epitopes) 112
C.2 AntiPatch: Venn diagram showing annotation overlaps (transitions) . . 113
C.3 AntiPatch: genetic evolution of H3 since 1968 114

D.1 RidgeRace: simple setting algorithm 118
D.2 RidgeRace: extended setting algorithm 119

List of Tables

4.1 AdaPatch: precision and recall 33
4.2 AdaPatch: patches and residues selected for seasonal H1 35
4.3 AdaPatch: patches and residues selected for seasonal H3 35
4.4 AdaPatch: patches and residues selected for H1 (2009 H1N1pdm) 36
4.5 AdaPatch: patches and residues selected for PB2 (2009 H1N1pdm) . . . 37
4.6 AdaPatch: sequence codes and PDB codes of selected templates 40

5.1 AntiPatch: identified patches 48

8.1 RidgeRace: comparison of average running times 96

B.1 AdaPatch: influenza A antigenic sites 110

C.1 AntiPatch: patches in antigenic cluster transitions 115

References

Affymetrix (2013). Product information.
URL http://www.affymetrix.com/estore/

Aguas R and Ferguson NM (2013). Feature selection methods for identifying genetic determinants of host species in RNA viruses. *PloS Computational Biology*, **9**(10):e1003254.

Allison AC (1956). The sickle-cell and haemoglobin C genes in some African populations. *American Journal of Human Genetics*, **21**(1):67–89.

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ (1990). Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3):403–410.

Anisimova M (2012). Parametric models of codon evolution. In Cannarozzi GM and Schneider A, eds., *Codon Evolution - Mechanisms and Models*, chapter 2, pages 12–33. Oxford University Press, New York.

Aoyama T, Nobusawa E, and Kato H (1991). Comparison of Complete Amino Acid

Sequences among 13 Serotypes of Hemagglutinins and Receptor-Binding Properties of Influenza A Viruses Indirect immunofluorescence. *Mutagenesis*, **485**:475–485.

Ashkenazy H, Erez E, Martz E, Pupko T, and Ben-Tal N (2010). ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acid Research*, **38**(3):1–5.

Balding DJ (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, **7**(10):781–791.

Barrick JE and Lenski RE (2013). Genome dynamics during experimental evolution. *Nature Reviews Genetics*, **14**(12):827–839.

Bateson W (1909). *Mendel's principles of heredity.* Cambridge Univ. Press, Cambridge.

Baum DA and Larson A (1991). Adaptation Reviewed: A Phylogenetic Methodology for Studying Character Macroevolution. *Systematic Biology*, **40**(1):1–18.

Belser JA, Bridges CB, Katz JM, and Tumpey TM (2009). Past, present, and possible future human infection with influenza virus A subtype H7. *Emerging Infectious Diseases*, **15**(6):859–865.

Benjamini Y and Yekutieli D (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, **29**(4):1165–1188.

Berglund AC, Wallner B, Elofsson A, and Liberles DA (2005). Tertiary windowing to detect positive diversifying selection. *Journal of Molecular Evolution*, **60**(4):499–504.

Bininda-Emonds ORP, Cardillo M, Jones KE, *et al.* (2007). The delayed rise of present-day mammals. *Nature*, **446**(7135):507–12.

Bininda-Emonds ORP, Cardillo M, Jones KE, *et al.* (2008). The delayed rise of present-day mammal. Erratum. *Nature*, **456**(274).

Blomberg SP, Garland T, and Ives AR (2003). Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. *Evolution*, **57**(4):717–745.

Blythe MJ and Flower DR (2005). Benchmarking B-cell epitope prediction: underperformance of existing methods. *Protein Science*, **14**(1):246–248.

BOOST (2014). ublas library, http://www.boost.org.

Boykov Y, Veksler O, and Zabih R (2002). Fast approximate energy minimization via graph cuts. *Pattern Analysis and Machine Learning*, **23**(11):1222–1239.

Brodie ED, Moore AJ, and Janzen FJ (1995). Visualizing and quantifying natural selection. *Trends in Ecology & Evolution*, **10**(8):313–8.

Brooks DR and McLennan DA (1991). *Phylogeny, Ecology, and Behavior: A Research Program in Comparative Biology*.

Bush RM, Fitch WM, Bender CA, and Cox NJ (1999). Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Molecular Biology and Evolution*, **16**(11):1457–1465.

Butler MA and King AA (2004). Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. *American Naturalist*, **164**(6):683–695.

Caton AJ, Brownlee GG, Yewdell JW, and Gerhard W (1982). The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). *Cell*, **31**(2 Pt 1):417–427.

Cattoli G, Milani A, Temperton N, *et al.* (2011). Antigenic drift in H5N1 avian influenza virus in poultry is driven by mutations in major antigenic sites of the hemagglutinin molecule analogous to those for human influenza virus. *Journal of Virology*, **85**(17):8718–8724.

Cavalli-Sforza LL and Edwards AW (1967). Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics*, **19**(3 Pt 1):233–57.

CDC (2013a). CDC information on Avian Influenza A (H7N9) Virus, accessed 8/20/13. URL `http://www.cdc.gov/flu/avianflu/h7n9-virus.htm`

CDC (2013b). Fact Sheet: Protect Yourself Against H3N2v, accessed 8/20/13. URL http://www.flu.gov/about_the_flu/h3n2v/

Chang BSW, Jönsson K, Kazmi MA, Donoghue MJ, and Sakmar TP (2002). Recreating a functional ancestral archosaur visual pigment. *Molecular Biology and Evolution*, **19**(9):1483–9.

Chen H and Zhou HX (2005). Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. *Nucleic Acid Research*, **33**(10):3193–9.

Chothia C (1976). The nature of the accessible and buried surfaces in proteins. *Journal of Molecular Biology*, **105**(1):1–12.

Clarke DK, Duarte EA, Elena SF, *et al.* (1994). The red queen reigns in the kingdom of RNA viruses. *PNAS*, **91**(11):4821–4.

Coddington JA (1988). Cladistic tests of adaptational hypotheses. *Cladistics*, **4**(1):3–22.

Cunningham CW, Omland KE, and Oakley TH (1998). Reconstructing ancestral character states: a critical reappraisal. *Trends in Ecology & Evolution*, **13**(9):361–366.

Darriba D, Taboada GL, Doallo R, and Posada D (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics (Oxford, England)*, **27**(8):1164–5. URL http://www.ncbi.nlm.nih.gov/pubmed/21335321

Darwin C (1859). *On the Origin of Species by Means of Natural Selection*. Murray, London.

Das K, Aramini JM, Ma LC, Krug RM, and Arnold E (2010). Structures of influenza A proteins and insights into antiviral drug targets. *Nature Structural & Molecular Biology*, **17**(5):530–538.

Dawood FS, Iuliano AD, Reed C, *et al.* (2012). Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. *The Lancet infectious diseases*, **12**(9):687–95.

Deem MW and Pan K (2009). The epitope regions of H1-subtype influenza A, with application to vaccine efficacy. *Protein Engineering*, **22**(9):543–546.

DiLillo DJ, Tan GS, Palese P, and Ravetch JV (2014). Broadly neutralizing hemagglutinin stalk-specific antibodies require FcγR interactions for protection against influenza virus in vivo. *Nature Medicine*, **20**(2):143–151.
URL `http://www.jimmunol.org/content/192/1_Supplement/140.12.short`

Dobzhansky T (1986). On Some Fundamental Concepts of Darwinian Biology. In *Evolutionary Biology*, pages 1–34. Springer US.

Dormitzer PR, Galli G, Castellino F, *et al.* (2011). Influenza vaccine immunology. *Immunological Reviews*, **239**(1):167–177.

Drake JW (1993). Rates of Spontaneous Mutation Among RNA Viruses. *PNAS*, **90**(9):4171–4175.

Drummond AJ, Ho SYW, Phillips MJ, and Rambaut A (2006). Relaxed phylogenetics and dating with confidence. *PloS Biology*, **4**(5):e88.

Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acid Research*, **32**(5):1792–1797.

Ekman S, Andersen HL, and Wedin M (2008). The limitations of ancestral state reconstruction and the evolution of the ascus in the Lecanorales (lichenized Ascomycota). *Systematic Biology*, **57**(1):141–56.

El-Manzalawy Y, Dobbs D, and Honavar V (2008). Predicting linear B-cell epitopes using string kernels. *Journal of Molecular Recognition*, **21**(4):243–255.

Elliot M (2013). Identical inferences about correlated evolution arise from ancestral state reconstruction and independent contrasts. *arXiv*, pages 1–31.

Felsenstein J (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, **17**(6):368–76.

Felsenstein J (1985). Phylogenies and the comparative method. *American Naturalist*.

Felsenstein J (1993). PHYLIP (Phylogeny Inference Package).

Felsenstein J (1996). Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods in Enzymology*, **266**:418–427.

Felsenstein J (2004). *Inferring Phylogenies*. Sinauer Associates, Inc, Sunderland, Massachusetts.

Ferreira AJ and Figueiredo MA (2012). Efficient feature selection filters for high-dimensional data. *Pattern Recognition Letters*, **33**(13):1794–1804.

Finarelli JA and Flynn JJ (2006). Ancestral state reconstruction of body size in the Caniformia (Carnivora, Mammalia): the effects of incorporating data from the fossil record. *Systematic Biology*, **55**(2):301–13.

Fitch WM (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology*, **20**(4):406–416.

Fitch WM, Bush RM, Bender CA, and Cox NJ (1997). Long term trends in the evolution of H3 HA1 human influenza type A. *PNAS*, **94**(15):7712–7718.

Freckleton RP and Harvey PH (2006). Detecting non-Brownian trait evolution in adaptive radiations. *PloS Biology*, **4**(11):e373.

Friedman J, Hastie T, and Tibshirani R (2010). A note on the group lasso and a sparse group lasso. *arXiv*, pages 1–8.

Gabriel G, Dauber B, Wolff T, *et al.* (2005). The viral polymerase mediates adaptation of an avian influenza virus to a mammalian host. *PNAS*, **102**(51):18590.

Gambaryan AS, Robertson JS, and Matrosovich MN (1999). Effects of egg-adaptation on the receptor-binding properties of human influenza A and B viruses. *Virology*, **258**(2):232–239.

Gareth J, Witten D, Hastie T, and Tibshirani R (2013). *An introduction to statistical learning*.
URL `http://link.springer.com/content/pdf/10.1007/978-1-4614-7138-7.pdf`

Garten RJ, Davis CT, Russell CA, *et al.* (2009). Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. *Science*, **325**(5937):197–201.

Gershoni JM, Roitburd-Berman A, Siman-Tov DD, Tarnovitski Freund N, and Weiss Y (2007). Epitope mapping: the first step in developing epitope-based vaccines. *BioDrugs*, **21**(3):145–156.

Ghedin E, Sengamalay NA, Shumway M, *et al.* (2005). Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature*, **437**(7062):1162–6.

GISAID (2013). EpiFlu database.
URL `http://platform.gisaid.org`

Glaser F, Pupko T, Paz I, *et al.* (2003). ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**(1):163–164.

Glaser L, Stevens J, Zamarin D, *et al.* (2005). A single amino acid substitution in 1918 influenza virus hemagglutinin changes receptor binding specificity. *Virology*, **79**(17):11533–11536.

Grafen A (1989). The phylogenetic regression. *Philosophical Transactions of the Royal Society of London B (Biological Sciences)*, **326**(1233):1–39.

Grenfell BT, Pybus OG, Gog JR, *et al.* (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, **303**(5656):327–332.

Gubry-Rangin C and Hai B (2011). Niche specialization of terrestrial archaeal ammonia oxidizers. *PNAS*, **108**(52):21206–21211.

Gubry-Rangin C, Macqueen D, Kratsch C, McHardy AC, and Prosser J (2014). Evolutionary history of terrestrial thaumarchaea. in preparation.

Guindon S and Gascuel O (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, **52**(5):696–704.

Guyon I (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, **3**:1157–1182.

Guyon I, Weston J, Barnhill S, and Vapnik V (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, **46**(1-3):389–422.

Haber DA and Settleman J (2007). Drivers and passengers. *Nature*, **446**(March):145–146.

Hall MA (1998). *Correlation-based Feature Selection for Machine Learning*. Ph.D. thesis, The University of Waikato.

Hanahan D and Weinberg RA (2000). The Hallmarks of Cancer. *Cell*, **100**:57–70.

Hanahan D and Weinberg RA (2011). Hallmarks of Cancer : The Next Generation. *Cell*, **144**(5):646–674.

Hansen TF (1997). Stabilizing selection and the comparative analysis of adaptation. *Evolution*, **51**(5):1341–1351.

Hanson-Smith V, Kolaczkowski B, and Thornton JW (2010). Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Molecular Biology and Evolution*, **27**(9):1988–99.

Harmon LJ, Losos JB, Jonathan Davies T, *et al.* (2010). Early bursts of body size and shape evolution are rare in comparative data. *Evolution*, **64**(8):2385–96.

Harmon LJ, Weir JT, Brock CD, Glor RE, and Challenger W (2008). GEIGER: investigating evolutionary radiations. *Bioinformatics*, **24**(1):129–131.

Harvey PH and Pagel MD (1991). *The Comparative Method in Evolutionary Biology*. Oxford University Press.

Hastie T, Tibshirani R, and Friedman J (2009). Elements of Statistical Learning. page 763.

Hay AJ, Lin YP, Gregory V, and Bennet M (2003). WHO Collaborating Centre for Reference and Research on Influenza, Annual Report. Technical report, National Institute for Medical Research, London.

Hensley SE, Das SR, Bailey AL, *et al.* (2009). Hemagglutinin receptor binding avidity drives influenza A virus antigenic drift. *Science*, **326**(5953):734–736.

Hirschhorn JN and Daly MJ (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, **6**(2):95–108.

Hirst GK (1943). Studies of the antigenic differences among strains of Influenza A by means of red cell agglutination. *Journal of Experimental Medicine*, **78**(5):407–423.

Hofree M, Shen JP, Carter H, Gross A, and Ideker T (2013). Network-based stratification of tumor mutations. *Nature Methods*, (September).

Holmes EC, Ghedin E, Miller N, *et al.* (2005). Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. *PloS Biology*, **3**(9):e300.

Huang JW, King CC, and Yang JM (2009). Co-evolution positions and rules for antigenic variants of human influenza A/H3N2 viruses. *BMC Bioinformatics*, **10**:1–10.

Huang JW and Yang JM (2011). Changed epitopes drive the antigenic drift for influenza A (H3N2) viruses. *BMC Bioinformatics*, **12 Suppl 1**:S31.

Huelsenbeck JP and Bollback JP (2001). Empirical and hierarchical Bayesian estimation of ancestral states. *Systematic Biology*, **50**(3):351–66.

Hurst LD (2009). Fundamental concepts in genetics: genetics and the understanding of selection. *Nature Reviews Genetics*, **10**(2):83–93.

Illingworth CJR and Mustonen V (2011). Distinguishing driver and passenger mutations in an evolutionary history categorized by interference. *Genetics*, **189**(3):989–1000.

Jagger BW, Wise HM, Kash JC, *et al.* (2012). An overlapping protein-coding region in influenza A virus segment 3 modulates the host response. *Science*, **337**(6091):199–204.

Jin H, Zhou H, Liu H, *et al.* (2005). Two residues in the hemagglutinin of A/Fujian/411/02-like influenza viruses are responsible for antigenic drift from A/-Panama/2007/99. *Virology*, **336**(1):113–119.

Jones DT, Taylor WR, and Thornton JM (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences*, **8**(3):275–282.

Kandoth C, McLellan MD, Vandin F, *et al.* (2013). Mutational landscape and significance across 12 major cancer types. *Nature*, **502**(7471):333–9.

Kelley LA and Sternberg MJE (2009). Protein structure prediction on the Web: a case study using the Phyre server. *Nature protocols*, **4**(3):363–71.

Kilbourne ED (2006). Influenza pandemics of the 20th century. *Emerging Infectious Diseases*, **12**(1):9–14.

Kim S and Xing E (2010). Tree-guided group lasso for multi-task regression with structured sparsity. *Proceedings of the 27th International Conference on Machine Learning*.

Klesper L (2013). *A method to detect somatic mutations associated with differential expression in glioblastoma multiform cancer*. Master thesis, Heinrich Heine University Düsseldorf.

Koel BF, Burke DF, Bestebroer TM, *et al.* (2013). Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science*, **342**(6161):976–9.

Kohavi R and John GH (1997). Wrappers for feature subset selection. *Artificial Intelligence*, **97**(1):273–324.

Kosakovsky Pond SL, Frost SDW, and Muse SV (2005). HyPhy: hypothesis testing using phylogenies. *Statistical Methods in Molecular Evolution*, **21**(5):676–679.

Kosakovsky Pond SL, Murrell B, Fourment M, *et al.* (2011). A random effects branch-site model for detecting episodic diversifying selection. *Molecular Biology and Evolution*, **28**(11):3033–43.

Kosakovsky Pond SL, Poon AFY, Leigh Brown AJ, and Frost SDW (2008). A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza A virus. *Molecular Biology and Evolution*, **25**(9):1809–1824.

Kratsch C, Klesper L, Steinbrück L, and McHardy AC (2014). Determination of antigenicity-altering patches of sites on the hemagglutinin of human influenza A/H3N2 viruses. pages 1–29.

Kratsch C and McHardy AC (2014). RidgeRace: Ridge regression for continuous ancestral character estimation on phylogenetic trees. *Bioinformatics*, **30**(17):i527–i533.

Kryazhimskiy S, Dushoff J, Bazykin GA, and Plotkin JB (2011). Prevalence of epistasis in the evolution of influenza A surface proteins. *PloS Genetics*, **7**(2):e1001301.

Kryazhimskiy S and Plotkin JB (2008). The population genetics of dN/dS. *PloS Genetics*, **4**(12):e1000304.

Kuan CM (2013). Introduction to Econometric Theory, Lecture Notes and Slides (Fall 2013).
URL `homepage.ntu.edu.tw/~ckuan`

Kuiken T, Holmes EC, McCauley J, *et al.* (2006). Host species barriers to influenza virus infections. *Science*, **312**(5772):394–397.

Lacerda M, Scheffler K, and Seoighe C (2010). Epitope Discovery with Phylogenetic Hidden Markov Models. *Molecular Biology and Evolution*, **27**(5):1212–1220.

Lapedes AS and Farber R (2001). The geometry of shape space: application to influenza. *Journal of Theoretical Biology*, **212**(1):57–69.

Laursen NS and Wilson IA (2013). Broadly neutralizing antibodies against influenza viruses. *Antiviral Research*, **98**(3):476–83.
URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3987986&tool=pmcentrez&rendertype=abstract`

Lazebnik Y (2010). What are the hallmarks of cancer? *Nature Reviews Cancer*, **10**(4):232–233.

Lee MS, Chen MC, Liao YC, and Hsiung CA (2007). Identifying potential immunodominant positions and predicting antigenic variants of influenza A/H3N2 viruses. *Vaccine*, **25**(48):8133–8139.

Lee B and Richards FM (1971). The interpretation of protein structures: estimation of static accessibility. *Journal of Molecular Biology*, **55**(3):379–400.

Lees WD, Moss DS, and Shepherd AJ (2011). Analysis of antigenically important residues in human influenza A virus in terms of B-cell epitopes. *Journal of Virology*, **85**(17):8548–8555.

Lemey P, Rambaut A, Drummond AJ, and Suchard MA (2009). Bayesian phylogeography finds its roots. *PloS Computational Biology*, **5**(9):e1000520.

Letunic I and Bork P (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, **23**(1):127–8.

Letunic I and Bork P (2011). Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acid Research*, **39**(Web Server issue):W475–8.

Li L, Weinberg CR, Darden TA, and Pedersen LG (2001). Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, **17**(12):1131–42.

Liao YC, Lee MS, Ko CY, and Hsiung CA (2008). Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus. *Bioinformatics*, **24**(4):505–512.

Liao YC, Lin HH, and Lin CH (2013). Monitoring the antigenic evolution of human influenza A viruses to understand how and when viruses escape from existing immunity. *BMC Research Notes*, **6**(1):227.

Lin YP, Gregory V, Bennett M, and Hay A (2004). Recent changes among human influenza viruses. *Virus research*, **103**(1-2):47–52.

Lin YP, Xiong X, Wharton Sa, *et al.* (2012). Evolution of the receptor binding properties of the influenza A(H3N2) hemagglutinin. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(52):21474–9.
URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3535595&tool=pmcentrez&rendertype=abstract`

Lu A and Guindon S (2013). Performance of Standard and Stochastic Branch-Site Models for Detecting Positive Selection among Coding Sequences. *Molecular Biology and Evolution*.

Luksza M and Lässig M (2014). A predictive fitness model for influenza. *Nature*.

Maddison DR and Maddison WP (2000). *MacClade 4*. Sinauer Associates, Incorporated.

Maddison WP and Maddison DR (2011). Mesquite: a modular system for evolutionary analysis.
URL `http://mesquiteproject.org`

Maley CC, Galipeau PC, Finley JC, *et al.* (2006). Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nature genetics*, **38**(4):468–73.

Martins EP and Hansen TF (1997). Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *American Naturalist*, **149**(4):646–667.

Matrosovich MN, Gambaryan AS, Teneberg S, *et al.* (1997). Avian influenza A viruses differ from human viruses by recognition of sialyloligosaccharides and gangliosides and by a higher conservation of the HA receptor-binding site. *Virology*, **233**(1):224–234.

McCarthy MI, Abecasis G, Cardon LR, *et al.* (2008). Genome wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Genetics*, **9**:356–369.

McHardy AC (2013). Lecture on "Phylogenies and Viral Evolution" (Algorithmen für phylogenetische Rekonstruktionen und deren Anwendungen in der Virusforschung).

McHardy AC and Adams B (2009). The role of genomics in tracking the evolution of influenza A virus. *PloS Pathogens*, **5**(10):e1000566.

McPeek MA (1995). Testing hypotheses about evolutionary change on single branches of a phylogeny using evolutionary contrasts. *American Naturalist*.

Medina RA and García-Sastre A (2011). Influenza A viruses: new research developments. *Nature Reviews Microbiology*, **9**(8):590–603.

Medina RA, Manicassamy B, Stertz S, *et al.* (2010). Pandemic 2009 H1N1 vaccine protects against 1918 Spanish influenza virus. *Nature communications*, **1**:28.

Merlo LMF, Pepper JW, Reid BJ, and Maley CC (2006). Cancer as an evolutionary and ecological process. *Nature Reviews Cancer*, **6**(12):924–935.

Molinari NA, Ortega-Sanchez IR, Messonnier ML, *et al.* (2007). The annual impact of seasonal influenza in the US: measuring disease burden and costs. *Vaccine*, **25**(27):5086–5096.

Murrell B, Wertheim JO, Moola S, *et al.* (2012). Detecting individual sites subject to episodic diversifying selection. *PLoS Genetics*, **8**(7):e1002764.

Nei M (2005). Selectionism and neutralism in molecular evolution. *Molecular Biology and Evolution*, **22**(12):2318–2342.

Nelson MI and Holmes EC (2007). The evolution of epidemic influenza. *Nature Reviews Genetics*, **8**(3):196–205.

Neumann G and Kawaoka Y (2006). Host range restriction and pathogenicity in the context of influenza pandemic. *Emerging Infectious Diseases*, **12**(6):881–886.

Neumann G, Noda T, and Kawaoka Y (2009). Emergence and pandemic potential of swine-origin H1N1 influenza virus. *Nature*, **459**(7249):931–939.

Nicol GW, Leininger S, Schleper C, and Prosser JI (2008). The influence of soil pH on the diversity, abundance and transcriptional activity of ammonia oxidizing archaea and bacteria. *Environmental Microbiology*, **10**(11):2966–78.

Nielsen R and Yang Z (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, **148**(3):929–936.

Nimrod G, Glaser F, Steinberg D, Ben-Tal N, and Pupko T (2005). In silico identification of functional regions in proteins. *Bioinformatics*, **21 Suppl 1**:i328–37.

Nimrod G, Schushan M, Steinberg DM, and Ben-Tal N (2008). Detection of functionally important regions in "hypothetical proteins" of known structure. *Structure (London, England : 1993)*, **16**(12):1755–63.

Nobusawa E and Sato K (2006). Comparison of the mutation rates of human influenza A and B viruses. *Journal of virology*, **80**(7).

Nowell PC (1976). The clonal evolution of tumor cell populations. *Science*, **194**(4260):23–8.

Nozawa M, Suzuki Y, and Nei M (2009). Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *PNAS*, **106**(16):2–7.

Nunn CL (2011). *The comparative approach in evolutionary anthropology and biology*. The University of Chicago Press, Chicago and London.

Obozinski G, Taskar B, and Jordan MI (2009). Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, **20**(2):231–252.

O'Meara BC, Ané C, Sanderson MJ, and Wainwright PC (2006). Testing for different rates of continuous trait evolution using likelihood. *Evolution*, **60**(5):922–33.

Pagel MD (1997). Inferring evolutionary processes from phylogenies. *Zoologica Scripta*, **26**(4):331–348.

Pagel MD (1999a). Inferring the historical patterns of biological evolution. *Nature*, **401**(6756):877–884.

Pagel MD (1999b). The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Systematic Biology*, **48**(3):612–622.

Pagel MD and Meade A (2006). Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *American Naturalist*, **167**(6):808–25.

Pan K, Long J, Sun H, Tobin GJ, and Nara PL (2011). Selective Pressure to Increase Charge in Immunodominant Epitopes of the H3 Hemagglutinin Influenza Protein. *Journal of Molecular Evolution*, **72**(1):90–103.

Paradis E, Claude J, and Strimmer K (2004). APE: analyses of phylogenetics and evolution in R language, v. 3.0-8. *Bioinformatics*, **20**:289–290.

Peng H, Long F, and Ding C (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE transactions on pattern analysis and machine intelligence*, **27**(8):1226–38.

Petersen B, Petersen TN, Andersen P, Nielsen M, and Lundegaard C (2009). A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Structural Biology*, **9**:51.

Phillips PC (2008). Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, **9**(11):855–867.

Podlaha O, Riester M, De S, and Michor F (2012). Evolution of the cancer genome. *Trends in Genetics*, **28**(4):155–63.

Posada D (2008). jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution*, **25**(7):1253–6.

Price AL, Zaitlen NA, Reich D, and Patterson N (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, **11**(7):459–463.

Pupko T, Bell RE, Mayrose I, Glaser F, and Ben-tal N (2002). 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:S71—-7.

Pybus OG and Rambaut A (2009). Evolutionary analysis of the dynamics of viral infectious disease. *Nature Reviews Genetics*, **10**(8):540–550.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

R-Sig-Phylo (2013). Wikipage.
URL `http://www.r-phylo.org/wiki/HowTo/DataTreeManipulation`

Rambaut A (2013). FigTree v1.4.
URL http://tree.bio.ed.ac.uk/software/figtree/

RCSB (2013). Protein Data Bank.
URL http://www.rcsb.org/

Reid AH, Taubenberger JK, and Fanning TG (2004). Evidence of an absence: the genetic origins of the 1918 pandemic influenza virus. *Nature Reviews Microbiology*, **2**(11):909–14.

Revell LJ (2008). On the analysis of evolutionary change along single branches in a phylogeny. *American Naturalist*, **172**(1):140–7.

Revell LJ (2012). phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, **3**:217–223.

Riester M, Stephan-Otto Attolini C, Downey RJ, Singer S, and Michor F (2010). A differentiation-based phylogeny of cancer subtypes. *PloS Computational Biology*, **6**(5):e1000777.

Robertson JS, Bootman JS, Newman R, *et al.* (1987). Structural changes in the haemagglutinin which accompany egg adaptation of an influenza A(H1N1) virus. *Virology*, **160**(1):31–37.

Robinson DM, Jones DT, Kishino H, Goldman N, and Thorne JL (2003). Protein evolution with dependence among codons due to tertiary structure. *Molecular Biology and Evolution*, **20**(10):1692–1704.

Röhr C, Kerick M, Fischer A, *et al.* (2013). High-Throughput miRNA and mRNA Sequencing of Paired Colorectal Normal, Tumor and Metastasis Tissues and Bioinformatic Modeling of miRNA-1 Therapeutic Applications. *PloS One*, **8**(7):e67461.

Ronquist F (2004). Bayesian inference of character evolution. *Trends in Ecology & Evolution*, **19**(9):475–81.

Rubinstein ND, Mayrose I, Halperin D, *et al.* (2008). Computational characterization of B-cell epitopes. *Molecular Immunology*, **45**(12):3477–3489.

Rubinstein ND, Mayrose I, and Pupko T (2009). A machine-learning approach for predicting B-cell epitopes. *Molecular Immunology*, **46**(5):840–847.

Russell CA, Jones TC, Barr I, *et al.* (2008). Influenza vaccine strain selection and recent studies on the global migration of seasonal influenza viruses. *Vaccine*, **26**:D31—-D34.

Saeys Y, Inza In, and Larrañaga P (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**(19):2507–17.

Saitou N and Nei M (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**(4):406–25.

Sali A and Blundell TL (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, **234**(3):779–815.

Sandbulte MR, Westgeest KB, Gao J, *et al.* (2011). Discordant antigenic drift of neuraminidase and hemagglutinin in H1N1 and H3N2 influenza viruses. *PNAS*, **108**(51):20748–20753.

Sankoff D (1975). Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, **28**(1):35–42.

Schluter D, Price T, Mooers AO, and Ludwig D (1997). Likelihood of ancestral states in adaptive radiation. *Evolution*, **51**(6):1699–1711.

Schrödinger (2013). The PyMOL Molecular Graphics System, Version 1.4, LLC.

Shannon P, Markiel A, Ozier O, *et al.* (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, **13**(11):2498–504.

Shazman S, Celniker G, Haber O, Glaser F, and Mandel-Gutfreund Y (2007). Patch Finder Plus (PFplus): a web server for extracting and displaying positive electrostatic patches on protein surfaces. *Nucleic Acid Research*, **35**(Web Server issue):W526–30.

Shih AC, Hsiao TC, Ho MS, and Li WH (2007). Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. *PNAS*, **104**(15):6283–6288.

Skehel J (2009). An overview of influenza haemagglutinin and neuraminidase. *Biologicals*, **37**(3):177–178.

Smith DJ, Lapedes AS, de Jong JC, *et al.* (2004). Mapping the antigenic and genetic evolution of influenza virus. *Science*, **305**(5682):371–376.

Smith GJD, Vijaykrishna D, Bahl J, *et al.* (2009). Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature*, **459**(7250):1122–1125.

Sorrell EM, Wan H, Araya Y, Song H, and Perez DR (2009). Minimal molecular constraints for respiratory droplet transmission of an avian-human H9N2 influenza A virus. *PNAS*, **106**(18):7565–7570.

Steinbrück L and McHardy AC (2011). Allele dynamics plots for the study of evolutionary dynamics in viral populations. *Nucleic Acid Research*, **39**(1):e4.

Steinbrück L and McHardy AC (2012). Inference of genotype-phenotype relationships in the antigenic evolution of human influenza A (H3N2) viruses. *PloS Computational Biology*, **8**(4):e1002492.

Sun H, Yang J, Zhang T, Long L, and Jia K (2013). Using Sequence Data To Infer the Antigenicity of Influenza Virus. *mBio*, **4**(4):e00230–13.

Suzuki Y (2004a). New methods for detecting positive selection at single amino acid sites. *Journal of Molecular Evolution*, **59**(1):11–19.

Suzuki Y (2004b). Three-dimensional window analysis for detecting positive selection at structural regions of proteins. *Molecular Biology and Evolution*, **21**(12):2352–2359.

Suzuki Y (2006). Natural selection on the influenza virus genome. *Molecular Biology and Evolution*, **23**(10):1902–1911.

Suzuki Y and Gojobori T (1999). A method for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution*, **16**(10):1315–1328.

Swofford DL (2003). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.

TCGAN (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**(7216):1061–8.

TCGAN (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**(7353):609–15.

TCGAN (2012a). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**(7407):330–7.

TCGAN (2012b). Comprehensive molecular portraits of human breast tumours. *Nature*, **490**(7418):61–70.

TCGAN (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, **499**(7456):43–9.

Terletskaia-Ladwig E, Meier S, and Enders M (2013). Improved high-throughput virus neutralisation assay for antibody estimation against pandemic and seasonal influenza strains from 2009 to 2011. *Journal of Virological Methods*, **189**(2):341–347.

Tong S, Li Y, Rivailler P, *et al.* (2012). A distinct lineage of influenza A virus from bats. *PNAS*, **109**(11):4269–74.

Tong S, Zhu X, Li Y, *et al.* (2013). New world bats harbor diverse influenza A viruses. *PloS Pathogens*, **9**(10):e1003657.

Tria F, Pompei S, and Loreto V (2013). Dynamically correlated mutations drive human Influenza A evolution. *Scientific Reports*, **3**:2705.

Tusche C, Steinbrück L, and McHardy AC (2012). Detecting Patches of Protein Sites of Influenza A Viruses under Positive Selection. *Molecular Biology and Evolution*, **29**(8):2063–2071.

van Valen L (1973). A new evolutionary law. *Evolutionary Theory*, (1):1–30.

Wallace RG, Hodac H, Lathrop RH, and Fitch WM (2007). A statistical phylogeography of influenza A H5N1. *PNAS*, **104**(11):4473–4478.

Webster RG, Bean WJ, Gorman OT, Chambers TM, and Kawaoka Y (1992). Evolution and ecology of influenza A viruses. *Microbiological Reviews*, **56**(1):152–179.

Weinstock DM and Zuccotti G (2009). The evolution of influenza resistance and treatment. *Journal of the American Medical Association*, **301**(10):1066–1069.

WHO (2009). Influenza Fact Sheet No 211, accessed 09/03/2013.
URL `http://www.who.int/mediacentre/factsheets/fs211/en/index.html`

WHO (2011). Manual for the laboratory diagnosis and virological surveillance of influenza. *Geneva: WHO press.*, page 151 p.

WHO (2012). Laboratory methodologies for testing the antiviral susceptibility of influenza viruses: Neuraminidase inhibitor (NAI).
URL `http://www.who.int/influenza/gisrs_laboratory/antiviral_susceptibility/nai_phenotyping/en/index.html`

WHO (2013a). Cancer Fact Sheet No 297.
URL `http://www.who.int/mediacentre/factsheets/fs297/en/`

WHO (2013b). History of influenza pandemics.
URL `http://www.euro.who.int/en/health-topics/communicable-diseases/influenza/pandemic-preparedness/about-pandemic-influenza/history-of-influenza-pandemics`

WHO (2014). WHO Risk Assessment of human infection with avian influenza A H7N9 virus.
URL http://www.who.int/entity/influenza/human_animal_interface/RiskAssessment_H7N9_21Jan14.pdf

Wiley DC and Skehel JJ (1987). The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. *Annual Review of Biochemistry*, **56**:365–394.

Wiley DC, Wilson IA, and Skehel JJ (1981). Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. *Nature*, **289**(5796):373.

Williams TA, Foster PG, Cox CJ, and Embley TM (2013). An archaeal origin of eukaryotes supports only two primary domains of life. *Nature*, **504**(7479):231–236.

Wilson IA and Cox NJ (1990). Structural basis of immune recognition of influenza virus hemagglutinin. *Annual Review of Immunology*, **8**:737–771.

Winn MD, Ballard CC, Cowtan KD, *et al.* (2011). Overview of the CCP4 suite and current developments. *Acta Crystallographica Section D: Biological Crystallography*, **67**(Pt 4):235–242.

Wise HM, Foeglein A, Sun J, *et al.* (2009). A complicated message: Identification of a novel PB1-related protein translated from influenza A virus segment 2 mRNA. *Journal of Virology*, **83**(16):8021–8031.

Wise HM, Hutchinson EC, Jagger BW, *et al.* (2012). Identification of a novel splice variant form of the influenza A virus M2 ion channel with an antigenically distinct ectodomain. *PloS Pathogens*, **8**(11):e1002998.

WTCCC (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**(7145):661–678.

Yamada S, Hatta M, Staker BL, *et al.* (2010). Biological and Structural Characterization of a Host-Adapting Amino Acid in Influenza Virus. *PloS Pathogens*, **6**(8):e1001034.

Yang Z (2000). Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *Journal of Molecular Evolution*, **51**(5):423–432.

Yang Z (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**(8):1586–1591.

Yang Z and Bielawski J (2000). Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, **15**(12):496–503.

Yang Z and Nielsen R (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution*, **19**(6):908–17.

Yang Z and Rannala B (2012). Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*, **13**(5):303–14.

Yang Z and dos Reis M (2011). Statistical properties of the branch-site test of positive selection. *Molecular Biology and Evolution*, **28**(3):1217–28.

Yates LR and Campbell PJ (2012). Evolution of the cancer genome. *Nature Reviews Genetics*, **13**(11):795–806.

Yuan HY and Koelle K (2013). The evolutionary dynamics of receptor binding avidity in influenza A: a mathematical model for a new antigenic drift hypothesis. *Philos Trans R Soc Lond B Biol Sci*, **368**(1614):20120204.

Zhao P, Rocha G, and Yu B (2006). Grouped and hierarchical model selection through composite absolute penalties. Technical report, Department of Statistics University of California, Berkeley.

Zhou T, Enyeart PJ, and Wilke CO (2008). Detecting clusters of mutations. *PloS One*, **3**(11):e3765.

Zwickl D (2006). *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion.* Ph.D. thesis, The University of Texas at Austin.

Index

dN/dS, 10, 26
influenza
    A/H1N1pdm, 20

AdaPatch, vii, viii, 6, 24, 31, 45, 99
adaptation, 9, 27
ancestral character states, 6
    reconstruction, 66, 69, 77, 81, 83
antigenic, 5
    cartography, 28
    distance, 28, 46
    drift, 21, 27, 39
    evolution, 7, 15, 27, 33, 100
    hitchhiker, 24, 29, 47
    shift, 23
Antigenic Tree, 15, 29, 45, 79
antigens, 11, 20, 21, 24
AntiPatch, vii, viii, 6, 24, 29

Brownian Motion, 67

cancer, 4, 87, 101
comparative methods, 65, 68
continuous characters, 67

discrete characters, 66
driver events, 4, 88

epistasis, 59, 75
evolution, 9
    continuous, 67
evolutionary related samples, 74

feature selection, 73
    embedded, 73, 74
    filter approaches, 73
    subset selection approaches, 73
    wrappers, 73
fitness, 4, 100

generalized least squares, 69, 78
Genome-wide association studies, 4
graph cut, 43, 101

hallmarks, 88

independence of features, 75
independence of samples, 76
influenza, 4, 19, 31, 46
    A/H1N1pdm, 20, 37

metastasis, 4

natural selection, 3, 9
    disruptive, 11
    positive, 10, 21
    purifying, 10, 24

passenger events, 4
phylodynamic techniques, 11, 15
phylogenetic independent contrasts, 68
phylogenetic inference, 12
    Bayesian, 13
    distance matrices, 13
    maximum likelihood, 11, 13
    parsimony methods, 13
phylogeny, 6, 12
proliferation, 4

reassortment, 23
receptor avidity, 23, 25, 27
Red Queen hypothesis, 21
RidgeRace, vii, viii, 6, 82

selective sweep, 24
shared inheritance, 5, 6, 15, 77, 102
stratification, 4, 5, 89

tumor, 4, 88