Doctoral Thesis

Understanding the spread and adaptation of infectious diseases using genomic sequencing data

Author(s):
du Plessis, Louis

Publication Date:
2016

Permanent Link:

https://doi.org/10.3929/ethz-a-010866338 →

DISS. ETH NO. 23660

UNDERSTANDING THE SPREAD AND ADAPTATION OF INFECTIOUS DISEASES USING GENOMIC SEQUENCING DATA

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

LOUIS DU PLESSIS

M.Sc. ETH Zurich, Zurich, Switzerland

born on 21.01.1985

citizen of Murten, FR and South Africa

accepted on the recommendation of

Prof. Dr. Tanja Stadler, (examiner)
Prof. Dr. Sebastian Bonhoeffer, (co-examiner)
Prof. Dr. Tracy Heath, (co-examiner)

2016

SUMMARY

This thesis is concerned with the question of how genome sequencing data may be used to shed light on infectious disease dynamics. I approach this question from several different angles, incorporating concepts from evolutionary biology, epidemiology, phylogenetics and phylodynamics. The first chapter gives a broad overview of the topic and introduces the most important concepts. The rest of the thesis is divided into two parts, respectively investigating the spread and adaptation of infectious diseases.

The first part (Chapters 2–5) deals with inferring the dynamics of epidemic spread from pathogen genetic data collected during an outbreak. Provided that a sufficient amount of genetic variation accumulates in the pathogen population over the course of an outbreak, this signature can be extracted with phylodynamic inference methods. I focus on models using a type of stochastic branching process called a birth-death process to model the spread of an epidemic. Throughout the course of this first part I refer to the 2013–2016 West African Ebola virus (EBOV) epidemic. Chapter 2 gives a brief review of available phylodynamic models and tools and how they may be used to gain insights into epidemic spread.

In Chapter 3, sequencing data, collected during the first month of the EBOV epidemic in Sierra Leone, is used to infer key epidemiological parameters. This chapter was published in October 2014, while the epidemic was still spreading. Although the short sampling period made it impossible to detect temporal changes due to the impact of control measures, it is shown that a number of different phylodynamic models return similar results. Moreover, the results are consistent with estimates from surveillance data, providing an independent verification of those results and showing that phylodynamic inferences are robust and reliable.

Chapter 4 builds on the results of Chapter 3 to present a general framework for real-time phylodynamics with birth-death models. To illustrate this framework, I perform a phylodynamic analysis on EBOV sequencing data from Guinea, Liberia and Sierra Leone, sampled over the course of almost a full year and covering the most intense phase of the epidemic. I show that a relatively simple model is powerful enough to robustly infer temporal changes in epidemiological parameters under a variety of different prior assumptions. Importantly, these inferences are also robust to the period over which data are sampled, illustrating the possibility of using this framework to give reliable real-time feedback during an outbreak.

In order to obtain reliable estimates, it is crucially important that the underlying model used in the birth-death process is a realistic approximation of the disease dynamics. The last chapter in this part, Chapter 5, investigates the likelihood-surface of a more realistic model that explicitly accounts for saturation effects due to the declining number of susceptible individuals. Using a number of simulations, I show that it is possible to infer epidemiological parameters under this model, and that these inferences are robust to mild sampling violations. However, in cases where stochasticity plays a large role, parameter correlations reduce the specificity of estimates, necessitating strong prior assumptions.

The second part of this thesis (Chapters 6 and 7) is concerned with the trajectories followed by an adapting population under different evolutionary pressures and environments, for instance a pathogen population under different drug exposures. The fitness landscape is a mapping from genotypes to fitness values that determines the course of adaptation, as

selection drives populations toward fitness peaks. As selection acts, it also leaves a signature in the genome, which may be extracted from the genome to determine the type of selection to which loci are exposed.

The high-dimensionality of genotype space makes it impossible to exhaustively sample all variants of biologically meaningful sequences. Chapter 6 investigates the ability of regression models to reconstruct a fitness landscape from a sparse sample of pathogen genotypes for which fitness measurements are available. Since we lack a complete real fitness landscape, I simulate quasi-empirical RNA fitness landscapes of comparable complexity. I show that while it is generally impossible to accurately reconstruct the complete fitness landscape, it is possible to achieve a remarkably good description of the local landscape, provided that an appropriate sampling scheme is used.

Although I mainly concentrate on virus populations, the concepts used in this thesis can be readily extended to species evolution and are equally applicable to host genomes as they evolve over many generations to escape from pathogen pressures. Chapter 7 investigates signatures of selection in the immune systems of 5 bee species, ranging across a social gradient, from highly eusocial honeybees through primitively eusocial bumblebees to the solitary leaf-cutting bee. The results show differential selective pressures between clades, which may be indicative of divergent pressures exerted by pathogens across social contexts.

Finally, in Chapter 8, I discuss the results obtained in this thesis within the context of recent developments in the field and suggest directions for future research.

ZUSAMMENFASSUNG

Diese Dissertation befasst sich mit der Frage, wie Genomsequenzdaten genutzt werden können, um Aufschluss über die Dynamik von Infektionskrankheiten zu erhalten. Ich untersuche diese Frage aus verschiedenen Blickwinkeln und integriere dabei Konzepte aus der Evolutionsbiologie, der Epidemiologie, der Phylogenetik und der Phylodynamik. Das erste Kapitel gibt einen Gesamtüberblick über das Thema und stellt die wichtigsten Konzepte vor. Der Rest der Dissertation besteht aus zwei Teilen, von denen der erste die Ausbreitung und der zweite die Anpassung von Infektionskrankheiten betrachtet

Der erste Teil (Kapital 2–5) befasst sich damit, wie anhand von Genomsequenzen des Krankheitserregers Rückschlüsse auf die Dynamik einer Epidemie gezogen werden können. Vorausgesetzt dass sich im Verlauf eines Ausbruchs genügend genetische Vielfalt in der Erregerpopulation akkumuliert hat, kann das im Genom entstandene Muster mit phylodynamischen Inferenzmethoden extrahiert werden. Ich konzentriere mich dabei auf Modelle, die für die Beschreibung der Ausbreitung einer Epidemie einen Typ von stochastischem Verzweigungsprozess, den sogenannten Geburts- und Todesprozess, verwenden. In diesem ersten Teil der Dissertation nehme ich durchgängig auf die Ebola-Epidemie (EBOV-Epidemie) Bezug, die von 2013 bis 2016 in West Afrika stattgefunden hat. Kapitel 2 gibt einen kurzen Überblick über die zur Verfügung stehenden Modelle und Methoden der Phylodynamik und wie diese verwendet werden können, um Erkenntnisse über die Ausbreitung einer Epidemie zu gewinnen.

In Kapitel 3 werden Genomsequenzdaten, die im ersten Monat der EBOV-Epidemie in Sierra Leone gesammelt wurden, verwendet, um wichtige Kennzahlen der Epidemie zu erschliessen. Dieses Kapitel wurde im Oktober 2014 publiziert, als die Epidemie noch in der Ausbreitung begriffen war. Obwohl der kurze Zeitraum, während dessen die Daten gesammelt wurden, es nicht erlaubt hat, zeitliche Veränderungen infolge von Bekämpfungsmassnahmen zu erfassen, konnten wir zeigen, dass verschiedene Modelle der Phylodynamik zu ähnlichen Ergebnissen gelangen. Darüber hinaus sind die Resultate konsistent mit Schätzungen, die anhand von Krankheitsfällen erhalten wurden. Damit bieten sie eine unabhängige Verifizierung dieser Ergebnisse und zeigen, dass Inferenzen mit phylodynamischen Methoden robust und zuverlässig sind.

Kapitel 4 baut auf den Ergebnissen von Kapitel 3 auf, um einen allgemeinen Rahmen für Phylodynamik in Echtzeit auf der Grundlage von Geburts- und Todesprozessen vorzustellen. Zur Illustration dieses Rahmenkonzepts führe ich eine phylodynamische Analyse von EBOV Genomsequenzdaten aus Guinea, Liberia und Sierra Leone durch, die während eines Zeitraums von fast einem ganzen Jahr gesammelt wurden und die intensivste Phase der Epidemie abdecken. Ich zeige, dass ein relativ einfaches Modell genügt, um robuste Aussagen über zeitliche Änderungen epidemiologischer Kennzahlen unter einer Vielzahl verschiedener Annahmen über die a-priori-Verteilung zu treffen. Wichtig dabei ist, dass diese Inferenzen auch robust gegenüber dem Zeitfenster sind, in denen die Daten gesammelt wurden; dies zeigt, dass es möglich ist, dieses Framework zu nutzen, um während eines Ausbruchs zuverlässige Rückmeldung in Echtzeit zu erhalten.

Um zuverlässige Schätzungen zu erhalten, ist es ausserordentlich wichtig, dass das zugrundeliegende Modell, das für den Geburts- und Todesprozess benutzt wird, eine realistische

Näherung der Krankheitsdynamik darstellt. Das letzte Kapitel dieses Teils, Kapitel 5, untersucht die Likelihood-Oberfläche eines realistischeren Modells, das explizit Sättigungseffekte bedingt durch die abnehmende Zahl an suszeptiblen Individuen berücksichtigt. Ich zeige anhand einer Reihe von Simulationen, dass es möglich ist, epidemiologische Kennzahlen des Modells abzuleiten, und dass diese Inferenzen robust gegenüber geringfügigen Verletzungen der Modellannahmen zur Probenahme sind. In Fällen, in denen Stochastizität eine grosse Rolle spielt, verringern Korrelationen zwischen Kennzahlen jedoch die Genauigkeit der Schätzungen, was gute Annahmen über die a-priori-Verteilung unabdingbar macht.

Der zweite Teil meiner Dissertation (Kapitel 6 und 7) befasst sich mit Trajektorien einer adaptierenden Population unter verschiedenen evolutionären Einflüssen und Umwelten wie etwa einer Erregerpopulation unter dem Einfluss verschiedener Medikamente. Die Fitnesslandschaft ist eine Abbildung von Genotypen auf Fitnesswerte, die den Verlauf der Anpassung festlegt, da Selektion Populationen zu Fitnessgipfeln treibt. Während Selektion wirksam ist, hinterlässt sie ein Muster im Genom, das extrahiert werden kann, um die Art der Selektion zu ermitteln, der Loci ausgesetzt sind.

Die hohe Dimensionalität des Genotypenraums macht es unmöglich, umfassend Daten aller Varianten von biologisch sinnvollen Sequenzen zu sammeln. Kapitel 6 untersucht die Eignung von Regressionsmodellen, eine Fitnesslandschaft auf der Grundlage einer lediglich kleinen Auswahl von Erregergenotypen, zu denen Fitnessmessungen vorliegen, zu rekonstruieren. Da wir nicht über die vollständige reale Fitnesslandschaft verfügen, simuliere ich quasi-empirische RNS Fitnesslandschaften mit vergleichbarer Komplexität. Ich zeige, dass es, obwohl es im Allgemeinen unmöglich ist, die Fitnesslandschaft akkurat vollständig zu rekonstruieren, möglich ist, die Landschaft lokal bemerkenswert gut zu beschreiben, vorausgesetzt dass die Probenahme geeignet erfolgt ist.

Obwohl ich mich hauptsächlich auf Viruspopulationen konzentriere, können die Konzepte, die ich in dieser Dissertation benutze, ohne Weiteres für die Evolution von Arten erweitert werden und sind ebenso anwendbar auf Wirtsgenome, die über viele Generationen hinweg evolvieren, um dem Druck von Krankheitserregern zu entkommen. In Kapitel 7 erforsche ich Selektionsabdrücke im Immunsystem von 5 Bienenarten, die auf einem Gradienten in der Intensität sozialer Interaktion von eusozialen Honigbienen über primitiv eusoziale Hummeln bis zu solitär lebenden Blattschneiderbienen reichen. Die Resultate zeigen, dass sich der Selektionsdruck zwischen den Kladen unterscheidet, was darauf hindeuten könnte, dass Erreger in Abhängigkeit der sozialen Struktur der Wirtspopulation unterschiedlichen Selektionsdruck ausüben.

Zum Schluss diskutiere ich in Kapitel 8 die in dieser Dissertation erhaltenen Resultate im Kontext aktueller Entwicklungen des Forschungsgebiets und schlage zukünftige Forschungsschwerpunkte vor.