GÖTEBORGS UNIVERSITET

INSTITUTIONEN FÖR BIOLOGI OCH MILJÖVETENSKAP

Advancing Evolutionary Biology: Genomics, Bayesian Statistics, and Machine Learning

Tobias Andermann
Institutionen för biologi och miljövetenskap
Naturvetenskapliga fakulteten

Opponent: Dr. Tracy A. Heath
Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa, USA

Examinator: Mari Källersjö
Institutionen för Biologi och Miljövetenskap, Göteborgs universitet

Akademisk avhandling för filosofie doktorsexamen i naturvetenskap, inriktning biologi, som med tillstånd från Naturvetenskapliga fakulteten kommer att offentligt försvaras fredagen den 18 december 2020 kl. 14:00 i Hörsalen, Botanhuset, Institutionen för Biologi och Miljövetenskap, Carl Skottsbergs gata 22B, Göteborg.

ISBN 978-91-8009-136-7 (tryckt)
ISBN 978-91-8009-137-4 (pdf)
Tillgänglig via http://hdl.handle.net/2077/66848

ABSTRACT

During the recent decades the field of evolutionary biology has entered the era of big data, which has transformed the field into an increasingly computational discipline. In this thesis I present novel computational method developments, including their application in empirical case studies. The presented chapters are divided into three fields of computational biology: genomics, Bayesian statistics, and machine learning. While these are not mutually exclusive categories, they do represent different domains of methodological expertise.

Within the field of genomics, I focus on the computational processing and analysis of DNA data produced with target capture, a pre-sequencing enrichment method commonly used in phylogenetic studies. I demonstrate on an empirical case study how common computational processing workflows introduce biases into the phylogenetic results, and I present an improved workflow to address these issues. Next I introduce a novel computational pipeline for the processing of target capture data, intended for general use. In an in-depth review paper on the topic of target capture, I provide general guidelines and considerations for successfully carrying out a target capture project. Within the context of Bayesian statistics, I develop a new computer program to predict future extinctions, which utilizes custom-made Bayesian components. I apply this program in a separate chapter to model future extinctions of mammals, and contrast these predictions with estimates of past extinction rates, produced by a set of different recently developed Bayesian algorithms. Finally, I touch upon newly emerging machine learning algorithms and investigate how these can be improved in their utility for biological problems, particularly by explicitly modeling uncertainty in the predictions made by these models.

The presented empirical results shed new light onto our understanding of the evolutionary dynamics of different organism groups and showcase the utility of the methods and workflows developed in this thesis. To make these methodological advancements accessible for the whole research community, I embed them into well documented open-access programs. This will hopefully foster the use of these methods in future studies, and contribute to more informed decision-making when applying computational methods to a given biological problem.

KEYWORDS

Computational biology, bioinformatics, phylogenetics, neural networks, NGS, target capture, Illumina sequencing, fossils, IUCN conservation status, extinction rates