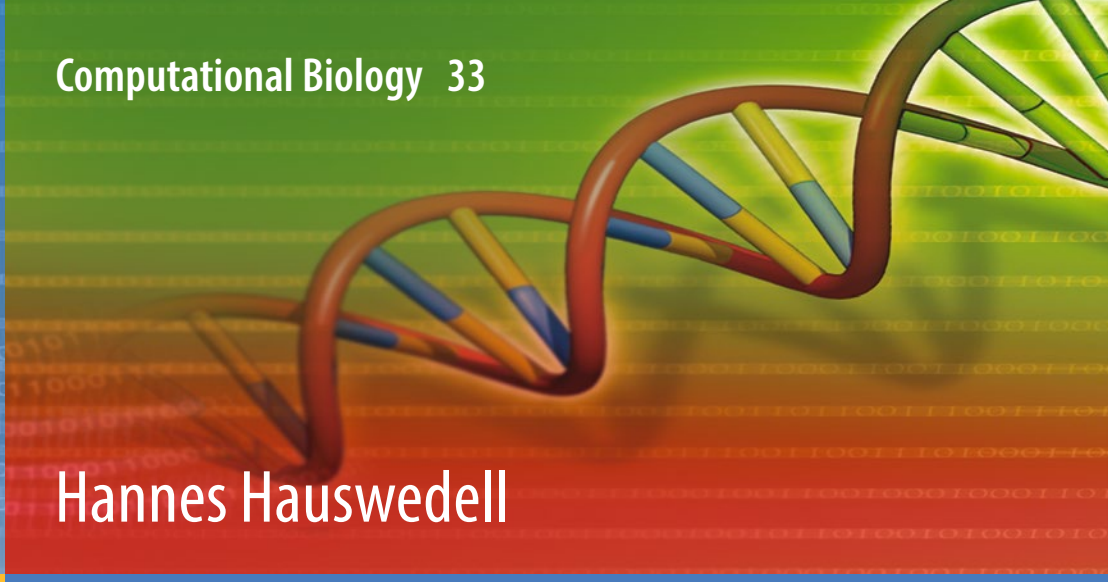
Computational Biology 33



Hannes Hauswedell

Sequence Analysis and Modern C++

The Creation of the SeqAn3 Bioinformatics Library

ISCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

Springer

Computational Biology

Volume 33

Endorsed by the *International Society for Computational Biology,* the *Computational Biology* series publishes the very latest, high-quality research devoted to specific issues in computer-assisted analysis of biological data. The main emphasis is on current scientific developments and innovative techniques in computational biology (bioinformatics), bringing to light methods from mathematics, statistics and computer science that directly address biological problems currently under investigation.

The series offers publications that present the state-of-the-art regarding the problems in question; show computational biology/bioinformatics methods at work; and finally discuss anticipated demands regarding developments in future methodology. Titles can range from focused monographs, to undergraduate and graduate textbooks, and professional text/reference works.

More information about this series at https://link.springer.com/bookseries/5769

Hannes Hauswedell

Sequence Analysis and Modern C++

The Creation of the SeqAn3 Bioinformatics Library

Springer

ISCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

Hannes Hauswedell
Reykjavik
Iceland

ISSN 1568-2684 ISSN 2662-2432 (electronic)
Computational Biology
ISBN 978-3-030-90989-5 ISBN 978-3-030-90990-1 (eBook)
https://doi.org/10.1007/978-3-030-90990-1

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This is a book about software engineering, bioinformatics, the C++ programming language and the SeqAn library. In the broadest sense, it will help the reader create better, faster and more reliable software by deepening their understanding of available tools, language features, techniques and design patterns.

Every developer who previously worked with C++ will enjoy the in-depth chapter on important changes in the language from C++ 11 up to and including C++ 20. In contrast to many resources on Modern C++ that present new features only in small isolated examples, this book represents a more holistic approach: readers will understand the relevance of new features and how they interact in the context of a large software project and not just within a "toy example". Previous experience in creating software with C++ is highly recommended to fully appreciate these aspects.

SeqAn3 is a new, re-designed software library. The conception and implementation process is detailed in this book, including a critical reflection on the previous versions of the library. This is particularly helpful to readers who are about to create a large software project themselves, or who are planning a major overhaul of an existing library or framework. While the focus of the book is clearly on software development and design, it also touches on various organisational and administrative aspects like licensing, dependency management and quality control.

The field that SeqAn3 provides solutions for is sequence analysis or, in a broader sense, bioinformatics. Readers working in this domain will recognise many of the discussed problems. However, almost all content is useful to software engineers in general and research software engineers in particular; no background in biology or previous experience with the SeqAn library is required.

This book is based on a dissertation, so the general style is more reminiscent of a "story" than might be typical for a computer science book. Some readers will enjoy reading it cover to cover while others will want to jump to sections of interest directly. The original preface of the dissertation is given on the following page as

the acknowledgements section. In addition to the persons mentioned there, I would like to thank Martin Vingron who was part of my defence committee and suggested this book project. I would also like to thank Susan Evans and the team at Springer Nature for helping it become reality.

Reykjavik, Iceland Hannes Hauswedell

The original version of this book was revised: Revised book has been uploaded to Springerlink. The correction to this book is available at https://doi.org/10.1007/978-3-030-90990-1_13.

Acknowledgements

The SeqAn library is a very active project with a long history. Over the last more than 10 years, it has had different core developers and many people who contributed features and fixes. Although SeqAn3 contains almost no code from SeqAn1/2, the experience of working on and with previous versions was invaluable in the development of SeqAn3. I feel that it is therefore only proper to mention Andreas Gogol-Döring, David Weese, Enrico Siragasu and Manuel Holtgrewe at this point, all of whom contributed significantly to SeqAn1/2. Of course Knut Reinert has always guided and does until today lead the project. His experience is the main pillar of its continued success.

This thesis introduces a new and radically different version of the SeqAn library. The scope of this project is huge, and it certainly would not have been possible to create the library single-handedly in this time. I do, however, credit myself with its inception, the vision behind the project and the endurance to pursue a complete rewrite of the library when most people called it infeasible. The design process, the overarching goals and the technical decisions are overwhelmingly my work—that is the foundation of this thesis. On the practical side, I have also written and changed more code than the next most important contributors combined, but I want to state clearly that relevant parts of SeqAn3 have also been implemented by people other than myself.

René Rahn has shared the responsibility of leading the project with me on a social and administrative level. Since the early beginnings of SeqAn3, I relied strongly on his counsel. Later, we assembled the SeqAn *core team* to discuss design and strategy matters on a regular basis. This included Svenja Mehringer, Marcel Ehrhardt and Enrico Seiler. All members of the core team have left their mark in some way on the library, and I am confident that SeqAn3 is in good hands after I leave the project.

I would like to thank everyone who contributed to SeqAn3, but more generally I want to also thank everyone for the great time at Freie Universität and the unforgettable SeqAn retreats! Special thanks go to Sara Hetzel and Felix Heeger who provided very helpful comments on a draft of this dissertation. Sara will also continue work on Lambda, an application presented later in this thesis.

On a professional and personal level, my sincere gratitude goes to Knut Reinert who has been my mentor now for so many years. None of this would have been possible without him. I would also like to express my sincere gratitude to Stefan Kurtz who agreed to co-supervise this (quite comprehensive) thesis although we had not worked together previously.

Attending the meetings of and contributing to the ISO C++ committee has had the most profound influence on my understanding of C++ and has thus helped greatly with creating SeqAn3. I would like to thank Fabio Fracassi and Nico Josuttis from the DIN Arbeitskreis Programmiersprachen as well as Corentin Jabot and JeanHeyd Meneide for helping me find my way around WG21.

Before working at Freie Universität, my studies were funded through a stipend of the Max-Planck-Gesellschaft. I additionally received a fellowship by the Hans-Böckler-Stiftung which allowed me to attend various extracurricular activities, for which I am very grateful.

Finally, I would like to thank my parents for supporting me during my youth and my early university studies. I am privileged to have had access to computers as a child and to grow up in an environment that fostered my curiosity in science and technology. I am grateful for the support of my friends and especially Romy and Betti. I look forward to spending more time with everyone again!

Contents