Rainer Schmidt / Matthias Brandt
Melina Schellhorn / Georg Füllen

Script zu Teilen der Bioinformatik-Vorlesung des 5. Semesters Medizin

3te Auflage

Wissenschaft & Technik

dissertation.de

http://www.dissertation.de

Rainer Schmidt

Matthias Brandt

Melina Schellhorn

Georg Füllen

Script zu Teilen der Bioinformatik-Vorlesung des 5.Semesters Medizin

3. Auflage

Universität Rostock

Medizinische Fakultät

Institut für Biostatistik und Informatik

in Medizin und Alternsforschung

Ernst-Heydemann-Str. 8

18057 Rostock

Schmidt, Rainer; Brandt, Matthias; Schellhorn, Melina; Füllen, Georg:

Script zu Teilen der Bioinformatik-Vorlesung des 5.Semesters Medizin / Rainer Schmidt; Matthias Brandt; Melina Schellhorn; Georg Füllen. –

ISBN 978-3-86624-628-7

Auflage: 3

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Winter-Industries GmbH, Berlin 2015

Alle Rechte, auch das des auszugsweisen Nachdruckes, der auszugsweisen oder vollständigen Wiedergabe, der Speicherung in Datenverarbeitungsanlagen, auf Datenträgern oder im Internet und der Übersetzung, vorbehalten.

Winter-Industries GmbH

URL: http://www.winter-industries.com

Inhaltsverzeichnis.

1. ALGORITHMEN 7

1.1 Algorithmus-„Bausteine“ 8

1.2 Notation von Algorithmen 11

1.3 Rekursion 13

1.4 Suchen in sortierten Folgen 14
1.4.1 Sequenzielle Suche 14
1.4.2 Binäre Suche (Divide and Conquer) 15

1.5 Entwurfsprinzipien für Algorithmen 18

1.6 Anwendungsbeispiel: Kürzeste Wege 20

1.7 Übungsaufgaben 26

2 SEQUENZ-ALIGNMENT UND PHYLOGENETISCHE BÄUME 27

2.1 Arten des Sequenz-Alignment 28
2.1.1 Paarweises Alignment 28

2.2 Alignment-Bewertung 31
2.2.1 Scoring für Proteine 32

2.3 Methoden des Sequenz-Alignments 35

2.4 Blast 35

2.5 Von Gerüchten zur Rentenpolitik bis zu phylogenetischen Bäumen 55

2.6 Übungsaufgaben 63

3. SINGLE NUCLEOTIDE POLYMORPHISMEN 64

4. BIOMARKER-FINDEN MIT R 69

4.1 Einleitung 69

4.2 Die Golub-Daten 75
4.2.1 Statistische Auswertungen der Golub-Daten, Suche nach Biomarkern (1.Ansatz) 77
4.2.2. Suchen nach Biomarkern in den Golub-Daten (2.Ansatz) 84
4.2.3. Gene Sortieren und Clustern 86
4.2.4 Dynamic Programming 88

4.3 Übungsaufgaben **91**

5. INFORMATIONSBESCHAFFUNG **93**

ANHANG **103**

Lösung der Übungsaufgaben: **103**

Literaturverzeichnis: **105**

Vorwort

Die erste Fassung dieses Skripts wurde von Matthias Brandt (Student im WS 2011/2012) erstellt. Überarbeitet und erweitert wurde es von Dr. Rainer Schmidt und Prof. Dr. Georg Füllen. Im Winter 2011/2012 wurde das Script von Melina Schellhorn (zu *Kürzesten Wegen, Maximum Likelihood und Single Nucleotide Polymorphismen)* erweitert.
Das Skript ist angelehnt an die Vorlesung zur Bioinformatik von Prof. Füllen, die ein Teil der gemeinsam mit Prof. Kundt gehaltenen Vorlesung „Epidemiologie, Medizinische Biometrie und Medizinische Informatik“ ist.

Der erste Teil der Vorlesung basiert teilweise auf dem Buch „Algorithmen und Datenstrukturen - Eine Einführung mit Java“ von Saake/Sattler und auf Materialien von R. König aus Heidelberg (mit freundlicher Genehmigung).
Der vierte Teil des Skripts „Biomarker-Finden mit R“ basiert teilweise auf dem Buch „Applied Statistics for Bioinformatics using R“ von Wim P.Krijnen. In diesem Teil erfolgt auch eine Einführung in die Programmiersprache „R“, mit der nicht nur viele Rechnungen im Bereich der Bioinformatik durchgeführt werden, sondern auch die typischen statistischen Untersuchungen für medizinische Promotionen im Bereich Biostatistik/Biometrie möglich sind.

Auf den folgenden Seiten soll es um die Bioinformatik gehen. Angelehnt an die Vorlesung haben wir uns mit den Folien (siehe: http://www.ibima.med.uni-rostock.de/IBIMA/ unter „Teaching1“) auseinandergesetzt und versucht, jeweils zu erklären, worum es geht. Somit sollte es mit diesem Script möglich sein, den hier behandelten Teil der Vorlesung eigenständig nachzuvollziehen. Doch vorweg erst einmal die Antwort auf die zwei wichtigsten Fragen zu diesem Thema:

Was ist Bioinformatik?

Bioinformatik ist ein Teilbereich der Informatik, welcher mit Hilfe von Programmen Probleme und Fragestellungen der Biologie und Medizin zu lösen versucht.

Was geht mich das an?

Im Rahmen heutiger Fragestellungen, bei denen es nicht mehr nur darum geht, ob z. B. ein Medikament wirkt oder nicht, sondern auch danach gefragt wird, wo es wirkt (Rezeptoren, Carrier, …) und wie genau es dort angreift, kommt man an der Visualisierung und Berechnung derartiger Zusammenhänge einfach nicht mehr vorbei. In vielen Fällen sind die Datenmengen, die mit modernen Geräten gewonnen werden, so groß, dass sie ohne vorherige Sortierung und Organisierung gar nicht greifbar (und somit auswertbar) sind. Jede Information ist immer nur soviel wert wie der Zusammenhang, in den sie gesetzt wird.
Genau hier greift die Bioinformatik an und versucht mit Hilfe von Berechnungen Ordnung ins Chaos zu bringen. Dass uns dies noch nicht täglich in der Klinik begegnen wird ist klar. Wenn es jedoch darum geht, Forschung zu betreiben oder zu verstehen, woher die Daten von Genanalysen stammen und was damit gemacht werden kann, ist es recht hilfreich, einen Einblick bekommen zu haben.

In diesem Sinne hoffen wir, dass wir all jenen weiterhelfen können, die Probleme mit dem Themengebiet haben!

Beschrieben werden hier die Inhalte der 1.Vorlesung sowie des 1. Seminars, und die Themen ‚Maximum Likelihood Phylogenie', ‚Blast' und ‚SNPs'.

Am Ende der Kapitel befinden sich Übungsaufgaben. Hierbei handelt es sich um Multiple-Choice-Aufgaben, bei denen jeweils nur eine Antwort richtig ist. Die Lösungen zu den Übungsaufgaben befinden sich im Anhang.

Der Inhalt wurde grundsätzlich auf Korrektheit geprüft. Es kann jedoch hier und da Fehler geben, genauso wie in den Folien zur Vorlesung. Wir bitten um Hinweise.

Für die 3. Auflage wurde Kapitel 5 zur Informationsverarbeitung neu eingefügt.

Prof. Georg Füllen: fuellen@uni-rostock.de

Dr. Rainer Schmidt: rainer.schmidt@uni-rostock.de

ISBN 978-3-86624-628-7

(D) 19,95 EUR

dissertation.de

http://www.dissertation.de