Projekt SeqAn

Projektmanagement im Softwarebereich

David Weese and Knut Reinert
Januar 2010

SeqAn - The C++ Sequence Analysis Library

Sequences
- strings
- structured sequences
- gapped sequences
- alterators

Alignments
- alignment data structures
- dynamic Programming
- alignment heuristics
- multidimensional chaining

Indices
- q-gram hashes
- (enhanced) suffix array
- suffix trees
- lazy indices, compress. ind.

Searching
- exact/approximate
- search heuristics
- filtering
- motif search

Graphs
- (structural) align. graphs
- word graphs
- probabilistic automata
- trees

Probabilis.
- profiles, weight matrices
- HMM, SCFG
- p-value computations
-…

Algorithms
- FASTA
- gQUASAR, SWIFT,..
- MEME, PROJECTION,…
- …

Biologicals
- alphabets
- scoring schemes
- file formats
- base pair probabilities

Integration
- using external tools
- STL
- LEDA and Boost graphs
- friend libraries (LISA)

Miscellan.
- allocators
- OS access and support
- helper data structures and algorithms

DNA-Sequenzanalyse

"Eine DNA-Sequenzanalyse ist in der Molekularbiologie und Bioinformatik die automatisierte, computergestützte Bestimmung von charakteristischen Abschnitten, insbesondere Genen, auf einer DNA-Sequenz. Untersucht werden die bei der DNA-Sequenzierung gewonnenen Informationen über die Abfolge und Position der Basenpaare..." [1]

[1] http://de.wikipedia.org/wiki/DNA-Sequenzanalyse (28.01.2010)

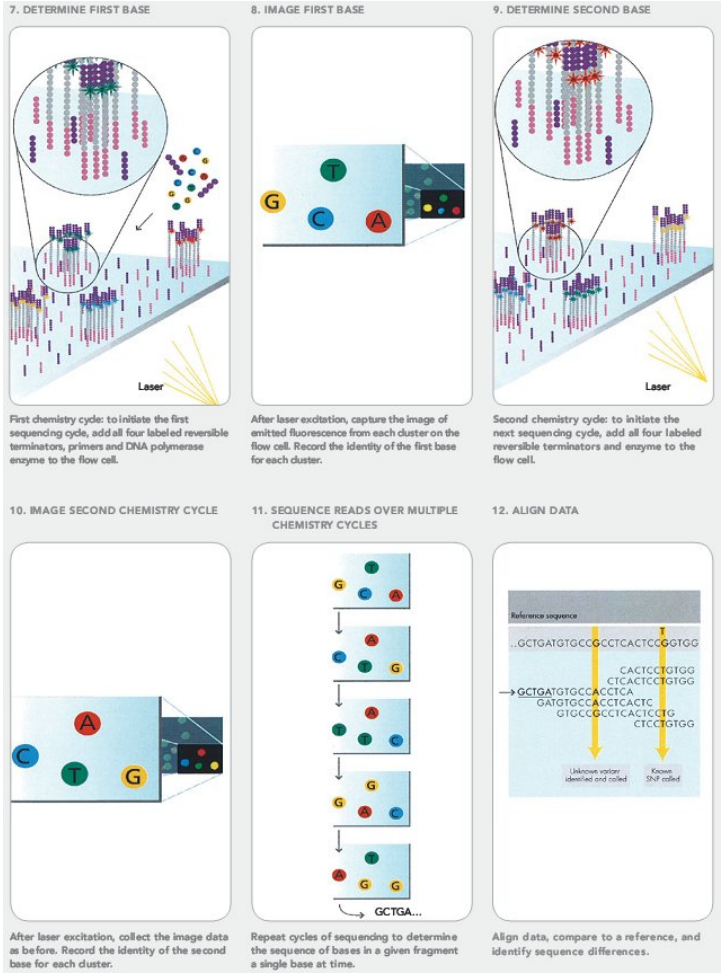
2nd Generation Sequencing

Technologien:

- 454 Sequencing (Roche)
- Solexa Sequencing (Illumina)
- SOLiD Sequencing (Applied Biosystems)

Sequenzierdaten pro Durchlauf [2]:

- 0.8-1.6 Mio. Reads, Länge 350-500 bps (454)
- 10-300 Mio. Reads, Länge 36-100 bps (Solexa)
- 300-1000 Mio. Reads, Länge 20-50 bps (SOLiD)



Solexa Sequencing Technology

[2] Julia Karow, *Survey: Illumina, SOLiD, and 454 Gain Ground in Research Labs; Most Users Mull Additional Purchases*, 2010

Anwendungsgebiete

Genome Assembly

- Resequencing
- De Novo Assembly

RNA-Sequencing

- Gene Expression Analysis
- Gene Annotation

Genome Comparison

- Structural Variations
- Single Nucleotide Polymorphisms
- Copy Number Variations

Metagenomics

Epigenetics

- ChIP Seq
- DNA Methylation

Projektthemen

Teilprojekte:

- Algorithmen zum Alignieren von Reads auf Genome
  - Implementierung
  - Analyse
- Implementierung verschiedener Filteralgorithmen
  - Implementierung
  - Analyse
- Tools zum Nachverarbeiten, Darstellen alignierter Reads
- ...

Anforderungen

Biologie/Chemie

- Grundlegendes Verständnis der Proteinbiosynthese
- DNA

Projektmanagement

- Ausarbeitung und Präsentation eines detaillierten Projektplans
- Umgang mit Werkzeuge zur kollaborativen Softwareerstellung
- Zusammenarbeit in kleineren Projektteams und Entwurf von Schnittstellen

Programmierarbeit

- Programmieren in C++, Templates
- Empfohlene Teilname am Kurs „C++ für Fortgeschrittene“ (6.4. – 9.4.)
- Testen und Dokumentieren der entwickelten Module

Zeitplan

Block 1 (1.4.)
- Vorstellung der Teilprojekte
- Einführung in die Themengebiete, Literatur

Block 2 (12.4. - 16.4.)
- Software Installation
- CMake und Subversion
- SeqAn Tutorials (Sequences, Alignments, Graphs, Indices)
- Zuordnung der Teilprojekte

Recherche und Präsentation des Projektplans (bis 26.4.)
- Vortrag vorbereiten
- Welche Datentypen/Schnittstellen von mir benötigen andere?
- Welche Datentypen/Schnittstellen benötige ich?
- Welche Algorithmen werden benötigt
- Welche davon gibt es schon in SeqAn, welche muss ich neu entwickeln?

Implementierung

Vorstellung der Ergebnisse, **Abschlussbericht** (bis 7.6.)

ENDE