Algorithmen der Bioinformatik I
WS 2017/2018

Burkhard Morgenstern
Peter Meinicke

Dept. Bioinformatics
Institute of Microbiology and Genetics (IMG)
University of Göttingen

October 23, 2017

Organisatorisches

- Praktische Übungen:
  Ca. 1 Woche statt Vorlesung Programmieraufgabe,
  Implementierung eines Algorithmus aus der Vorlesung.

- Prüfung:
  - Mündliche Prüfung (ca. 20 min)
  - Kurzes Testat zur Programmieraufgabe (ca. 5 min)

  Zeit nach Vereinbarung innerhalb des Prüfungszeitraums

Organisatorisches

Bitte alle in *StudIP* für die Vorlesung eintragen!

ABER: Emails bitte besser direkt an P. Meinicke und mich:

`pmeinic@gwdg.de , bmorgen@gwdg.de`

Was ist Bioinformatik?

Was ist Bioinformatik?

Die *Fachgruppe Bioinformatik (FaBi)*

`http://bioinformatik.de/`

Was ist Bioinformatik?

Die Fachgruppe für Bioinformatik (FaBI) hat folgende Definition von Bioinformatik als Grundlage ihrer Arbeit definiert:

"Die Bioinformatik ist eine interdisziplinäre Wissenschaft. Unter Bioinformatik verstehen wir die Erforschung, Entwicklung und Anwendung computergestützter Methoden zur Beantwortung molekularbiologischer und biomedizinischer Fragestellungen. Im Fokus stehen Modelle und Algorithmen für Daten auf molekularer und zellbiologischer Ebene, beispielsweise für

- *Genome und Gene,*
- *Gen- und Proteinexpression und -regulation,*
- *metabolische und regulatorische Pfade und Netzwerke,*
- *Strukturen von Biomakromolekülen, insb. DNA, RNA und Proteine,*
- *molekulare Interaktionen zwischen Biomakromolekülen untereinander und zwischen Biomakromolekülen und weiteren Substanzen wie beispielsweise Substraten, Transmittern, Botenstoffen und Inhibitoren sowie*
- *die molekulare Charakterisierung von Ökosystemen."*

Was ist Bioinformatik? Definition der *FaBi*

Was ist Bioinformatik?

(A) Grundlegende Gebiete

- Datenbanken
- Vergleichende Analyse von DNA-, RNA- und Proteinsequenzen
- Struktur von Makromolekülen
- Rekonstruktion von Stammbäumen (Phylogenie)

Was ist Bioinformatik?

(B) Genomanalyse

- Assemblierung von Genomen
- Genvorhersage
- Vorhersage von nicht-kodierenden funktionellen Elementen im Genom
- Vergleichende Analyse von Genomen
- *Next generation sequencing (NGS)* Datenanalyse, read mapping
- Metagenomics

Was ist Bioinformatik?

(C) ‘Systembiologie’:

- Gen-Expressionsdaten:
  Welche Gene wann ‘exprimiert’?
- Metabolische Netzwerke
- Regulatorische Netzwerke

Was ist Bioinformatik?

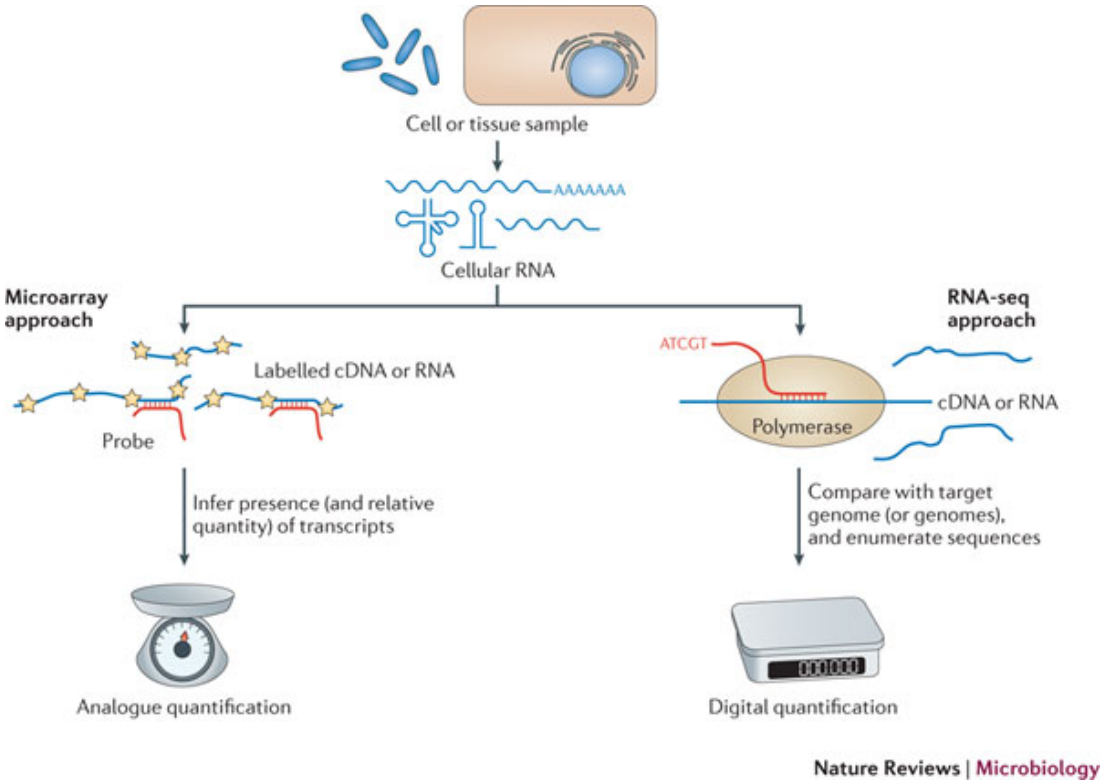

Figure: Genexpressions-Analyse: Vergleich von Mircoarrays zu RNA-Seq (Nature Reviews Microbiology 10, 618-630)

Was ist Bioinformatik?

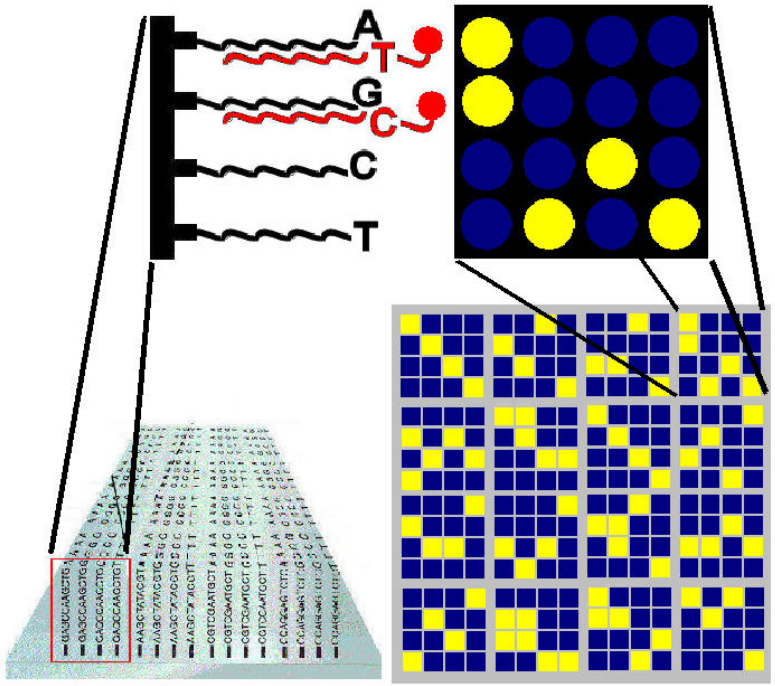

Figure: Microarray (Gene-Chip) Technologie (`https://www.mun.ca/`)

Was ist Bioinformatik?

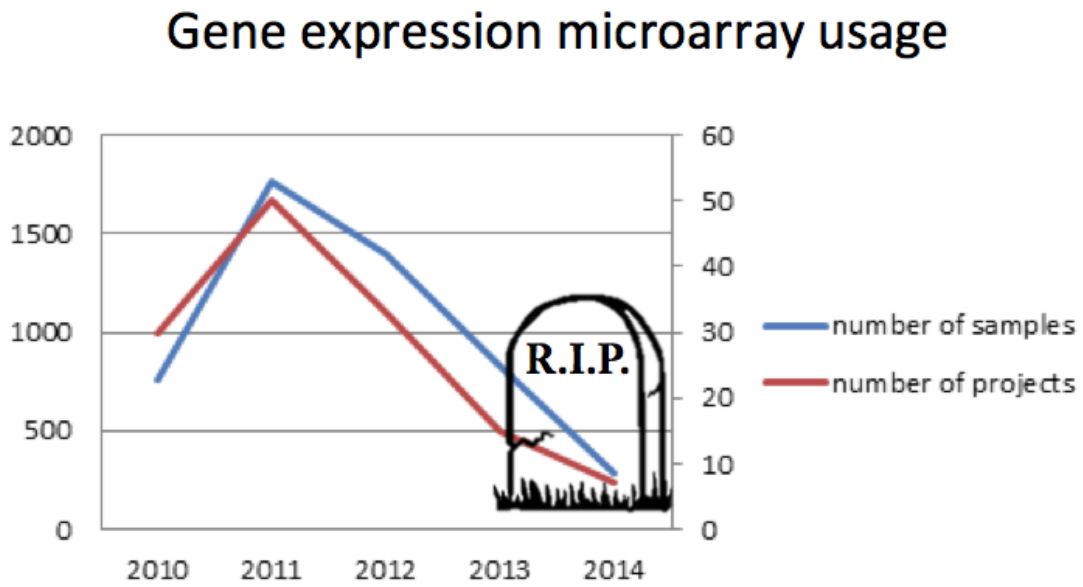

Figure: `http://core-genomics.blogspot.de/`

Was ist Bioinformatik?

Ausgangslage:

- Exponentielles Wachstum der Datenbanken
- Information auf Sequenz-Ebene schnell und billig verfügbar
- *Viel* weniger Information auf Struktur- und Funktions-Ebene vorhanden!

Was ist Bioinformatik?

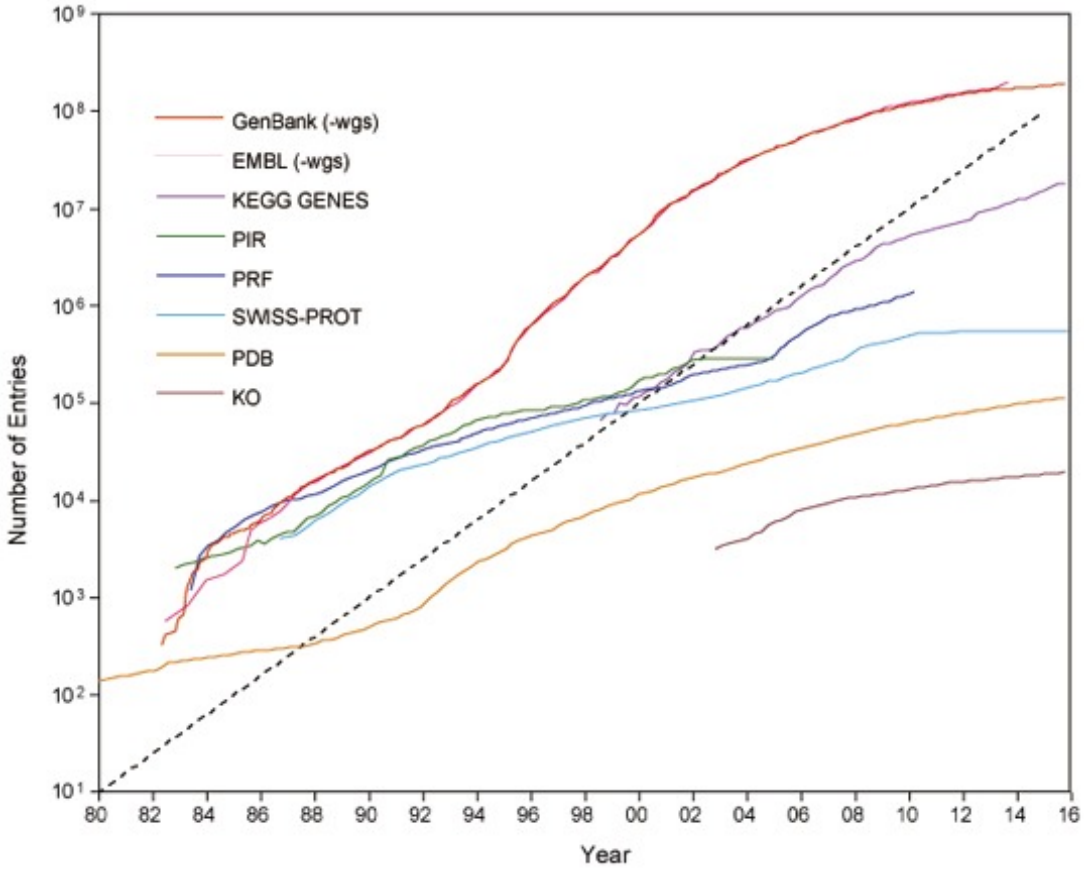

Figure: Wachstum von Sequenz- und Strukturdatenbanken (GenomeNet, Kyoto University)

Themen der Vorlesung

- Paarweises Sequenz-Alignment (Kurzeinführung bzw. Wdh.)
- Multiples Sequenz-Alignment
- Probabilistische Sequenzmodelle (P. Meinicke)
  - Positions-Gewichts-Matrizen
  - Hidden Markov Modelle (HMM)
  - Modelle für Proteinfamilien
- Genvorhersage mit HMMs
- Rekonstruktion phylogenetischer Bäume

Literatur

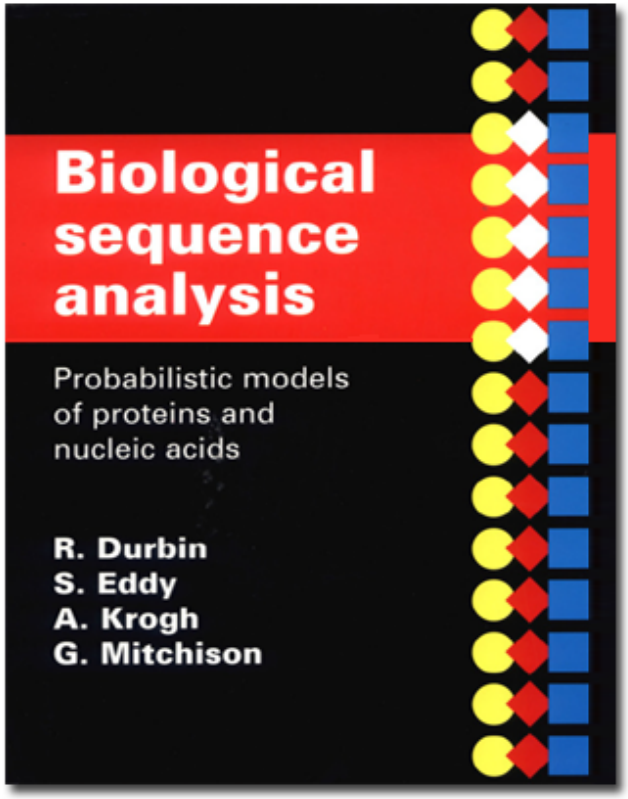

Figure: R. Durbin *et al., Biological Sequence Analysis*

Sequence alignment

```
seq1    W T Y I V M R E A Q E S A Q
seq2    R C L V M R E A Q E W A
seq3    Y I M Q E V Q Q E R A
seq4    A L Y I A M R E V Q Y E S A
```

- Sequence analyis based on *comparison* of sequences
- First step: sequence alignment
  - Assign homologous positions
  - Introduce *gaps* into sequences

Basis of (almost) *all* methods in sequence analysis

Sequence alignment

```
seq1    W T Y I V M R E A Q - E S A Q
seq2    - R C L V M R E A Q - E W A -
seq3    - - Y I - M Q E V Q Q E R A -
seq4    A L Y I A M R E V Q Y E S A -
```

- Sequence analyis based on *comparison* of sequences
- First step: sequence alignment
  - Assign homologous positions
  - Introduce *gaps* into sequences

Basis of (almost) *all* methods in sequence analysis

Sequence alignment

Sequence alignment important for

- Phylogeny
- Genome analysis, gene prediction
- Protein structure
- RNA secondary structure
- Database searching
- *etc.*

$\Rightarrow$ If alignments wrong, *all* results that are based on alignments are wrong!

Sequence alignment

Distinguish:

- *Paarwise* alignment: two sequences aligned
- *Multiple* alignment: more than two sequences aligned

Optimal pairwise alignment

Example (Simple protein alignment)

```
seq1 A L S C V W M I P
seq2 A I S C M I P T
```

Input sequences

Optimal pairwise alignment

Example (Simple protein alignment)

```
seq1 A L S C V W M I P -
seq2 A I S C - - M I P T
```

Possible alignment

Optimal pairwise alignment

Example (Simple protein alignment)

```
seq1 A L S C V W M I P -
seq2 A I S C - - M I P T
```

Alignment as hypothesis about evolution:

- 1 Substitution `L` $\rightarrow$ `I` or `I` $\rightarrow$ `L`
- 2 insertions or deletion events (`V`, `W`, `T`)

Optimal pairwise alignment

Example (Simple protein alignment)

```
seq1 A L S C V W M I P -
seq2 A I S - C M - I P T
```

Alternative alignment (hypothesis):

- 3 substitutions $\texttt{L} \rightarrow \texttt{I}$, $\texttt{V} \rightarrow \texttt{C}$, $\texttt{W} \rightarrow \texttt{M}$
- 3 insertion or deletion events (C,M,T)

Optimal pairwise alignment

Objective function for pairwise protein alignment:

Define score for

- Substitutions: *Score* $s(a, b)$ for every possible substitution of amino acids $a \rightarrow b$
- Insertions/deletions: *penalty* for every *gap* in alignment

Total score of an alignment:

Sum of scores for substitutions und gaps

Optimal pairwise alignment

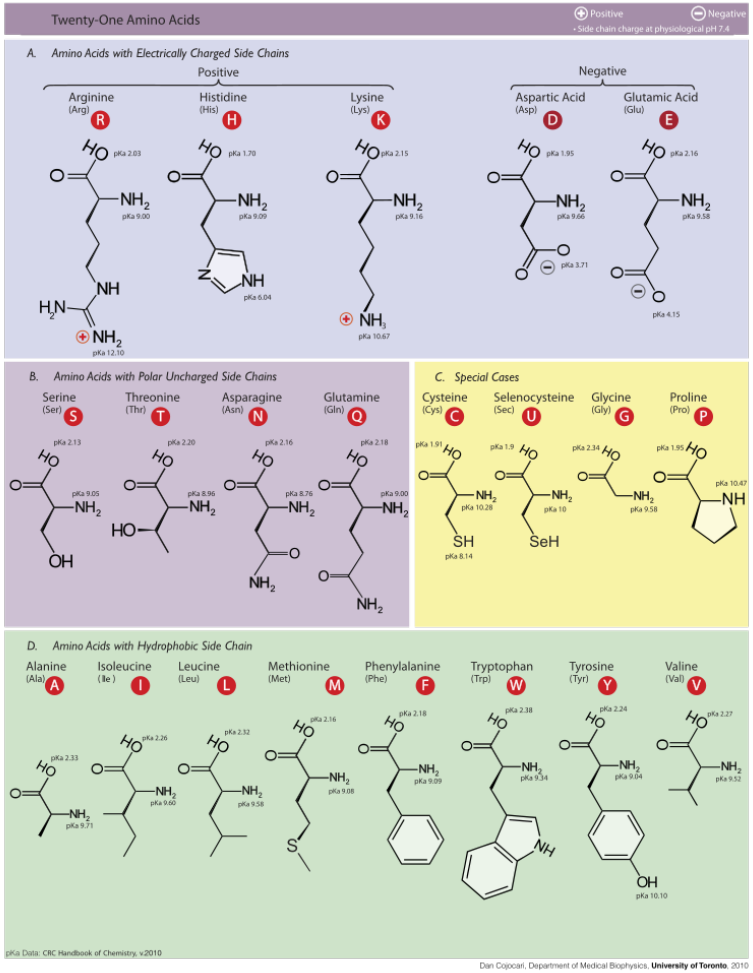

Figure: The 20 amino acides (source: wikipedia)

Optimal pairwise alignment

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W

Figure: The *BLOSUM 62* substitution matrix

Optimal pairwise alignment

- Simples method to score gaps: *linear gap penalty*, i.e. gap of length l gets penalty

$$g \times l$$

for constant g.

For linear gap penalty: every gap symbol '-' receives penalty g.

Optimal pairwise alignment

Find optimal alignment by *dynamic programming (DP)*.

Necessary: *linear gap penalty*!

For sequences X, Y

$$X = X_1 \dots X_m$$

$$Y = Y_1 \dots Y_n$$

of length m and n, respectively, ...

Optimal pairwise alignment

. . . consider sub-problems:

For $0 \leq i \leq m$ and $0 \leq j \leq n$, find optimal alignment of

$$X_1 \ldots X_i$$

$$Y_1 \ldots Y_j$$

('prefixes' of X and Y)

Optimal pairwise alignment

$F(i,j) =$ Score of optimal alignment of prefixes up to i und j

Recursion to calculate $F(i,j)$: consider 3 possibilities for last column of prefix alignment:

- X_i and X_j aligned:

$$\begin{array}{ll} \ldots & X_i \\ \ldots & Y_j \end{array}$$

In this case: $F(i,j) = F(i-1,j-1) + s(X_i, Y_j)$

Optimal pairwise alignment

- Gap in sequence X

$$\begin{array}{ll} \ldots & - \\ \ldots & Y_j \end{array}$$

In this case : $F(i,j) = F(i,j-1) - g$

- Gap in sequence Y

$$\begin{array}{ll} \ldots & X_i \\ \ldots & - \end{array}$$

In this case: $F(i,j) = F(i-1,j) - g$

Optimal pairwise alignment

All together:

$$F(i,j) = \max \begin{cases} F(i-1,j-1) & + & s(X_i, Y_j) \\ F(i-1,j) & - & g \\ F(i,j-1) & - & g \end{cases} \quad (1)$$