

Einführung in die Bioinformatik

SS 2013

1. Was ist Bioinformatik?

Kay Nieselt

Integrative Transkriptomik
Zentrum für Bioinformatik Tübingen
Kay.Nieselt@uni-tuebingen.de

Ablauf und Formales



- Ringvorlesung mit 5 Dozenten
 - Dr. Kay Nieselt (ZBIT/WSI)
 - Prof. Dr. Daniel Huson (ZBIT/WSI)
 - Prof. Dr. Oliver Kohlbacher (ZBIT/WSI)
 - Prof. Dr. Karsten Borgwardt (MPI/WSI)
 - Prof. Dr. Andrei Lupas (MPI)
- Webseite mit pdfs zu Vorlesungen, Übungen etc. unter:
<http://ab.inf.uni-tuebingen.de/teaching/ss2013/ebi/>

Modulbedingungen



- Um das Modul zu bestehen, muss eine am Ende des Semesters stattfindende Klausur mitgeschrieben und bestanden werden (Note $\geq 4,0$).
- Am Ende jeder Vorlesung werden Übungsblätter zum Selbststudium ausgeteilt, deren Bearbeitung und Abgabe *freiwillig* aber *empfohlen* ist.
- Antworten an *Benjamin Schubert*, schubert@informatik.uni-tuebingen.de schicken

Themenübersicht



Termin	Thema	Dozent
18.4.	Was ist Bioinformatik?	Nieselt
25.4.	Von der DNA zur Datenbank - Sequenzierung, Assemblierung	Huson
2.5.	Designerdrogen - Wirkstoffe aus dem Rechner	Kohlbacher
16.5.	Darwins Erben - Phylogenie und Bäume	Huson
6.6.	Gut vernetzt hält besser - Analyse biologischer Netzwerke	Nieselt
13.6.	It's hip to Chip - von Microarrays zu personalisierter Medizin	Nieselt
20.6.	Molekulare Maschinen - Proteinstrukturen und ihre Funktion	Kohlbacher
27.6.	Impfen gegen Krebs - Bioinformatik im Impfstoffentwurf	Kohlbacher
4.7.	Aus einer Hand voll Erde - Metagenomik	Huson
11.7.	Lernen mit Kernen - Maschinelles Lernen in der Bioinformatik	Borgwardt
18.7.	Die Sprache der Proteine - Evolution konservierter Proteinstrukturen	Lupas
25.7.	Klausur	-

Was ist Bioinformatik?



Wir beginnen mit Nachschlagen:

“Die Bioinformatik ist eine interdisziplinäre Wissenschaft, die Probleme aus den *Lebenswissenschaften* mit theoretischen computergestützten Methoden löst.” (Wikipedia)

Was ist Bioinformatik?



“**Bioinformatics** is the field of science in which biology, computer science, and information technology merge to form a single discipline.” (NCBI)

Bioinformatik beschäftigt sich insb. mit der Anwendung von computergestützten Methoden, um Information, die mit Biomolekülen assoziiert ist, zu analysieren.

Aufgaben der Bioinformatik



Große technologische Fortschritte auf dem Gebiet der Molekularbiologie => biologische Information exponentiell gewachsen.

- Entwicklung und Implementierung von Werkzeugen, die den effizienten Zugang, Gebrauch und das Management von biologischen Daten / Informationen ermöglichen.
- Entwicklung **effizienter Algorithmen**, mathematischer Methoden und Statistiken.

Ursprung der Bioinformatik



- Die Begründerin der Bioinformatik ist Margaret Dayhoff.
- Sie hat 1965 eine Sammlung von Proteinsequenzen, auch bekannt als *Atlas of Protein Sequence and Structure* veröffentlicht.
- Zusammen mit Richard Eck hat sie den ersten phylogenetischen Baum mit molekularen Sequenzen veröffentlicht.



M. Dayhoff, 1980, Quelle: Tochter

Themen der Bioinformatik

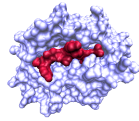


- Rekonstruktion unbekannter Genome
Assemblierung
- Vergleich von (DNS/Aminosäure-)Sequenzen
Suche nach Mustern / Motiven
- Wie sind Sequenzen entstanden:
Evolution von Arten
- Welche Abschnitte im Genom kodieren für Gene?
Maschinelles Lernen

Themen der Bioinformatik



- Strukturen von Molekülen
 - Protein-Strukturvorhersage
 - RNA-Strukturvorhersage
- Metabolische und regulatorische Stoffwechselwege
- Genexpression
 - Microarrays bzw. DNA-Chips
- Datenaufbereitung und Speicherung in biologischen Datenbanken



Biologische Objekte

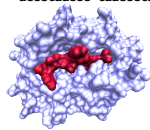


Sequenzen

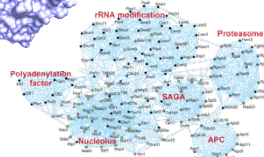
```

taacctaac octaaccta accctaacc taacctaac
taacctaac octaaccta accctaacc taacctaac
aacctaacc octaaccta accctaacc taacctaac
aacctaacc octaaccta accctaacc taacctaac
taacctaac octaaccta accctaacc taacctaac
octaaccta accctaacc taacctaac octaaccta
octaaccta accctaacc taacctaac octaaccta
octaaccta accctaacc taacctaac octaaccta
octaaccta accctaacc taacctaac octaaccta
octaaccta accctaacc taacctaac octaaccta
octaaccta accctaacc taacctaac octaaccta
    
```

Strukturen



Netzwerke



Biologische Sequenzen



- Primärsequenzen werden als Zeichenketten über einem gegebenen Alphabet (z.B. dem Alphabet der Nukleotide $\Sigma = \{A, G, C, T\}$ oder dem Alphabet der Aminosäuren $\Sigma = \{A, R, N, D, C, Q, \dots\}$) definiert
- Beispiel Nukleotidsequenz
ATGCAGGACGACGGATAACGATGACAG
- Beispiel Aminosäuresequenz
MQDDGQRGQTVVALKCCLLKARDDG

Mutationen



- Fehler während der Replikation
- Lokale Veränderungen der Primärsequenz der DNS
 - Substitutionen (Austausch einer Base durch eine andere)
 - Insertionen / Deletionen einer Base
- Umordnung ganzer Segmente entlang eines Chromosoms oder Austausch zwischen Chromosomen

Variabilität



- Mutationen sind die Quelle phänotypischer Vielfalt, die die Grundlage der natürlichen Selektion darstellt.
- Veränderte/neue Spezies
- Untersuchung von Mutationen führt zu einem besseren Verständnis des evolutionären Prozesses.
- Vergleich menschlicher Genome:
1000 Genomes project

Paarweises Alignment



Gegeben seien zwei biologische Sequenzen.
Aufgabe: berechnen Sie die Ähnlichkeit der beiden Sequenzen

Biologische Annahme:

Ähnliche Gene / Proteine haben auch ähnliche Funktionen

Effiziente Algorithmen



- “Ein **Algorithmus** ist eine aus endlich vielen Schritten bestehende eindeutige Handlungsvorschrift zur Lösung eines Problems oder einer Klasse von Problemen.” (Quelle: Wikipedia)
- **Effizient** bedeutet: „es darf nicht zu lange dauern“, selbst wenn es sich um ein großes Problem handelt.

Sequenz-Ähnlichkeit

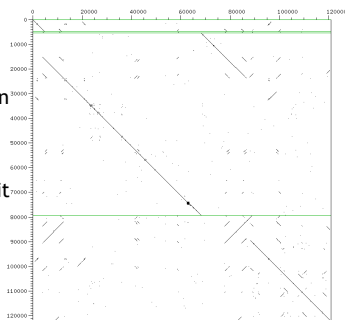


Frage:

Wie kann man feststellen wie ähnlich sich zwei Sequenzen sind?

Antwort: z. B. visuell (mit einem Dotplot)

Aber kann man es auch berechnen?



Sequenz-Alignment



Berechnung: mit Hilfe eines **paarweisen Alignments** (Ausrichtung zweier Sequenzen)

Alignment: Buchstaben einer gegebenen Sequenz werden den Buchstaben der anderen Sequenz so zugeordnet, dass die Reihenfolge erhalten bleibt und jeder Buchstabe genau einem anderen Buchstaben oder einem Gap (Leerstelle, Lücke) zugeordnet ist.

Sequenz-Alignment



Beispiel:

```
Seq 1 : T A T A - T A C G C T A G C A
Seq 2 : T A T A A T A G G C T - G C A

      ↑      ↑      ↑
    Match Mismatch Gap (Lücke)
```

Obiges Alignment hat 80% Identität, denn 12 von den 15 ausgerichteten Positionen sind identisch.

Sequenz-Alignment



Offensichtlich gibt es viele Möglichkeiten zwei Sequenzen auszurichten (zu alignieren).

Folgendes Alignment hat nur 60% Identität.

```
Seq 1 : - T A T A T A C G C T A G C C
Seq 2 : T A T A A T A G G C T - G C A
```

Sequenz-Alignment



Das optimale globale Alignment Problem:

Gegeben seien zwei Nukleotid- (Protein-) Sequenzen. Gesucht ist das Alignment, das alle Charaktere der Sequenzen beinhaltet und dessen Wert für eine gegebene Bewertungs-Funktion maximal ist.

Bewertungsfunktion



- Wie bewerten man Mutationen, Deletionen / Insertionen?
- Annahme: alle Positionen evolvieren unabhängig voneinander → additive Funktion, Wert des Alignments = Summe der Werte der einzelnen Spalten
- Benötigt wird Bewertung von Match, Mismatch, Gaps

Bewertungsfunktion



Beispiel: Match = +2, Mismatch = -1, Gap = -5

Seq 1 : - T A T A T A C G C T A G C A
Seq 2 : T A T A A T A G G C T - G C A
-5-1-1-1-1+2+2-1+2+2+2-5+2+2+2 = +4

Seq 1 : T A T A - T A C G C T A G C A
Seq 2 : T A T A A T A G G C T - G C A
+2+2+2+2-5+2+2-1+2+2+2-5+2+2+2 = +13

Das zweite Alignment hat also einen höheren Wert. Aber ist es das beste Alignment (optimal)?

Sequenz-Alignment - Optimal



Optimale Lösung:

- Needleman-Wunsch Algorithmus
- Dies ist ein dynamisches Programmierverfahren, eine Technik für das Lösen kombinatorischer Optimierungsprobleme mit den folgenden Eigenschaften:
 - Rekursiv definierter exponentieller Suchraum,
 - Wiederauftretende Matrizen,
 - Zusammensetzung der optimalen Lösung aus den „optimalen“ Lösungen von Teilproblemen
- DPs spielen eine große Rolle in der Bioinformatik

Needleman-Wunsch



- Sei s eine Bewertungsfunktion für Match, Mismatch und d ein Parameter zur Bewertung von Gaps.
- $F(i,j)$ ist dann gleich dem optimalen Wert eines Alignments von $X_i=(x_1,...,x_i)$ und $Y_j=(y_1,...,y_j)$.

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$$

- Initialisierung mit $F(i,0)=-id$, $(F(0,j)=-jd$

Beispiel



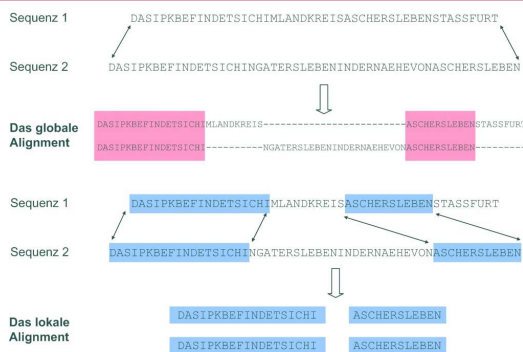
- Beispiel: $X= CACGA$, $Y=CGA$
match= +1, mismatch = -1, $d = 1$

	C	A	C	G	A
C	0	-1	-2	-3	-4
G	-1	1	0	-1	-2
A	-2	0	1	0	-1
	-3	-1	1	1	0

- Alignment:

CACGA
--CGA

Globales vs. lokales Alignment



Problem



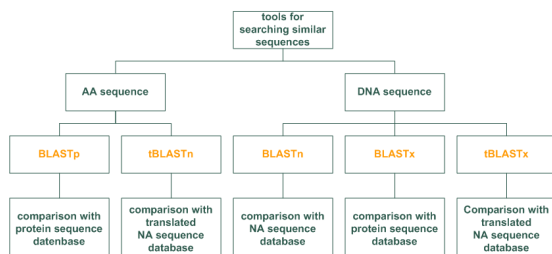
- Needleman-Wunsch und der äquivalenter Algorithmus für lokale Alignments, der Smith-Waterman-Algorithmus berechnen optimale Alignments benötigen aber zu lang (quadratische (in der Länge des Sequenzen) Laufzeit)
- Für Suchen in großen Datenbanken werden schnellere Verfahren benötigt

BLAST-Basic Local Alignment Search Tool



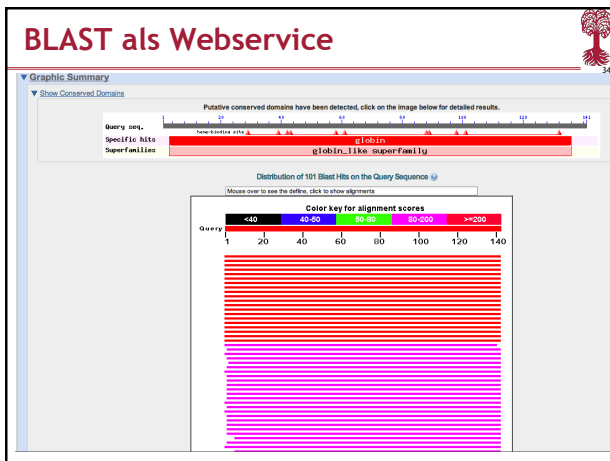
- Ist das häufigst benutzte Programm zur Suche einer Sequenz (Anfrage - *query*) in einer Sequenz-Datenbank (*subjects*, *targets*)
- Es ist ein lokales Alignment-Verfahren
- Es ist eine **Heuristik**: schnell ohne Garantie der Optimalität

Die BLAST-Familie



[illegible]

BLAST als Webservice



BLAST als Webservice

Descriptions

Legend for links to other resources: UniProt, Gene, Structure, Map Viewer, PubChem BioAssay

Accession	Description	Max score	Total score	Query coverage	E-value	Links
NP_001000001.1	hemoglobin subunit alpha [Pan troglodytes] >ref NP_001000001.1	254	254	100%	2e-75	U G
XP_002753388.1	PREDICTED: hemoglobin subunit alpha-like [Callithrix jacchus]	225	225	100%	3e-73	U G
NP_001000001.1	PREDICTED: hemoglobin subunit alpha-like [Macaca mulatta]	224	224	100%	3e-73	U G
NP_001000001.1	hemoglobin subunit theta-1 [Macaca mulatta]	223	223	100%	3e-72	U G
NP_001000001.1	hemoglobin subunit alpha [Ptilinopus	222	222	100%	3e-68	U G
NP_001000001.1	hemoglobin alpha, adult chain 2 [Mus musculus] >ref NP_033244.2	221	221	100%	3e-66	U G
NP_001000001.1	hemoglobin subunit alpha [Bos taurus] >ref XP_001788728.1 PREDI	220	220	100%	3e-65	U G
NP_001000001.1	hemoglobin subunit alpha [Equus caballus]	219	219	100%	3e-65	U G
NP_001000001.1	hemoglobin subunit alpha [Equus caballus]	218	218	100%	3e-64	U G
XP_000000001.1	PREDICTED: hemoglobin subunit alpha-like [Alluropoda melanoleuca]	218	218	100%	3e-64	U G
NP_001000001.1	hemoglobin subunit alpha-1/2 [Dryotrogus curvulus]	218	218	100%	3e-64	U G
NP_001000001.1	PREDICTED: hemoglobin subunit alpha-like isoform 1 [Sus scrofa] >	217	217	100%	3e-63	U G
NP_001000001.1	globin, alpha [Rattus norvegicus]	221	221	100%	3e-66	U G
XP_001125120.1	PREDICTED: similar to Hemoglobin subunit alpha (hemoglobin alpha	216	216	100%	3e-65	U G
NP_001000001.1	hemoglobin alpha 2 chain [Rattus norvegicus]	202	202	100%	3e-52	U G
XP_001000001.1	PREDICTED: hemoglobin subunit alpha-A-like [Meleagris gallopavo]	202	202	100%	3e-52	U G
NP_001000001.1	hemoglobin subunit alpha [Monodelphis domestica]	202	202	100%	3e-52	U G
NP_001000001.1	hemoglobin subunit alpha-1/2 [Rattus norvegicus]	202	202	100%	3e-52	U G
NP_001000001.1	hemoglobin subunit alpha-A [Callithrix jacchus]	202	202	100%	3e-51	U G
NP_001000001.1	hemoglobin subunit alpha-A [Taeniopygia guttata]	202	202	100%	3e-51	U G
XP_000000001.1	PREDICTED: hemoglobin subunit alpha-like [Mus musculus] >ref XP_	193	193	99%	3e-48	U G
NP_001000001.1	hemoglobin subunit beta [Taeniopygia guttata]	188	188	99%	3e-46	U G
XP_001000001.1	PREDICTED: hemoglobin subunit alpha-D-like [Meleagris gallopavo]	181	181	100%	3e-45	U G
NP_001000001.1	hemoglobin subunit alpha-S [Meleagris gallopavo]	182	182	99%	3e-45	U G
NP_001000001.1	hemoglobin, theta T1 [Mus musculus]	182	182	99%	3e-45	U G

Bioinformatik-Sequenzdatenbanken

- GenBank am National Center for Biotechnology Information (NCBI), USA; <http://www.ncbi.nlm.nih.gov/>, Suchmaschine ENTREZ
- European Molecular Biology Laboratory (EMBL) am European Bioinformatics Institute (EBI) in Hinxton, England; <http://www.ebi.ac.uk/>
- UniProt (Universal Protein Resource); <http://www.uniprot.org/>
- und viele viele weitere

Fazit



- Während der letzten 3 Jahrzehnte ist bedingt durch bedeutende Fortschritte auf dem Gebiet der Molekularbiologie, verbunden mit den Fortschritten in genomischen Technologien, ein enormer Wachstum der durch die wissenschaftliche Gemeinschaft erzeugten biologischen Information / Daten zu beobachten.
- Diese Explosion genomischer Daten hat damit notwendigerweise Bereitstellung von Datenbanken für das Aufbewahren, Organisieren und Indizieren sowie speziell entwickelten Werkzeugen zur Ansicht und Analyse der Daten geführt.

Referenzen



- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. J Mol Biol. 1990 Oct 5;215(3): 403-10.
- Dayhoff, Eck, Chang, and Sochard. Atlas of Protein Sequence and Structure, 1965
- Eck RV, Dayhoff MO. Science. 1966 Apr 15;152 (3720):363-6.
- Henikoff S, Henikoff JG. Proc Natl Acad Sci USA. 1992 Nov 15;89(22):10915-9.
- Needleman SB, Wunsch CD. J Mol Biol. 1970 Mar;48(3):443-53.
