Die erste Einführung in die Einführung in die Angewandte Bioinformatik 15.04.2010

Prof. Dr. Sven Rahmann

Team

Prof. Dr. Sven Rahmann (Vorlesung)
Dipl.-Inform. Marcel Martin (Übungen)

Zeit Do 12-14; Übungen um 14, 15, 16, 17 Uhr
Ort Vorlesung in der Chemie HS 3; Übungen in OH14, U04 (Keller)

Alle Informationen

Webseite: http://ls11-www.cs.tu-dortmund.de/teaching/angebio
kürzer: http://tiny.cc/xbdrv

Sprechstunde von Prof. Rahmann

Mo 16-17 in OH14, R214
Bitte möglichst per e-mail anmelden, sonst evtl. sehr lange Wartezeiten!
Sven.Rahmann /at/ tu-dortmund.de

Übungsbetrieb

In dieser Woche (heute) keine Übungen!
Übungen ab nächster Woche.

Anmeldung per email an Marcel Martin.
Anleitung siehe Aushänge
oder http://ls11-www.cs.tu-dortmund.de/teaching/angebio.

Anmeldung so bald wie möglich.
Frist: nächster Mittwoch, 21.04.2010, 12 Uhr

Die Einteilung in Gruppen erfahren Sie Mittwoch Abend auf der Webseite und nächste Woche in der Vorlesung.
Es stehen 4 Gruppen zur Verfügung (Do 14/15/16/17 Uhr).

Ziele der Vorlesung und Übungen

Wissen über

- Aufgaben,
- Methoden,
- Beschränkungen der Bioinformatik

Anwendungskenntnisse zu

- bioinformatischen Datenbanken
- im WWW verfügbarer Software

Einführung in Methoden der Informatik (sehr knapp)
Ziel dabei: Wie kann ich mit Informatikern reden,
biologische und chemische Fragestellungen/Probleme formulieren?

Prüfungsleistung

- Klausur und praktische Prüfung (am Übungs-PC)
  am Do 22.07.2010 in der jeweiligen Übungsgruppe

- Nehmen Sie unbedingt regelmäßig an den Übungen teil !

- Normale Übungsaufgaben enthalten zahlreiche Hinweise.
- Es genügt nicht, diese „abzuarbeiten“ !
- Sie sollen verstehen, was Sie tun, wenn Sie eine Aufgabe lösen !!
- Manche Aufgaben enthalten keine Hinweise, simulieren Klausurbedingungen.
- Diese geben Ihnen eine Rückmeldung zu Ihrem Kenntnisstand.

Literatur (Empfehlungen)

P.M. Selzer, R. J. Marhöfer, A. Rohwer (2004)
Angewandte Bioinformatik – Eine Einführung
Springer-Verlag

Jean-Michel Claverie, Cedric Notredame (2006)
Bioinformatics for Dummies, 2. Auflage
Wiley & Sons

Nello Christianini and Matthew W. Hahn (2007)
Introduction to Computational Genomics – a Case Studies Approach
Cambridge University Press

D.W. Mount (2004)
Bioinformatics: Sequence and Genome Analysis, 2. Auflage
Cold Spring Harbor Laboratory Press

Biologie: bio = Leben, logos = Wissenschaft
Biologie: Wissenschaft des Lebens
Biologie früher: Katalogisieren von Lebensformen
Biologie heute: molekular geprägt (seit der Entdeckung der DNA)
Basis der modernen Biologie: Chemie

Informatik: Wissenschaft der systematischen Verarbeitung von Information
Information: Ordnung, Struktur, Abweichung vom Zufall

Wie passt das zusammen?
Entstehung von Leben = Bildung von Ordnung / Strukturen [?]

Z.B. sind Zellen vor allem damit beschäftigt, die innere Ordnung zu erhalten.
Lebewesen bleiben am Leben, weil sie sich von ihrer Umwelt abgrenzen.

Andere Bio-*x* - Wissenschaften

Bio-x bedeutet eins von zwei Dingen:
[Hier ist *x* ein Platzhalter (Variable)]

1. Wissenschaft *x* leistet Beitrag zum besseren Verständnis der Biologie
2. Biologie inspiriert neue Forschungsrichtungen in Wissenschaft *x*

Beispiele:

- Biochemie
- Biophysik
- Biotechnologie
- Biomathematik
- Bioinformatik

Bioinformatik ist ein weites Feld...

Bioinformatik, Systembiologie: Modewörter der letzten 20 / 5 Jahre

Bioinformatik = mehrere Disziplinen (von theoretisch bis angewandt):

- Biomathematik
- Theoretische Biologie
- (Theoretische) Ökologie
- Biostatistik
- Sequenzanalyse
- „Computational biology“
- Bioinformatik (im engeren Sinn)
- Systembiologie
- Computational *omics: genomics, transcriptomics, proteomics, ...
  [*omics: Untersuchung der Gesamtheit von *; der Stern * ist ein Platzhalter]
- Angewandte = praktische Bioinformatik

Bioinformatiker und Anwender

Bioinformatiker

Person, die Modelle, Methoden und Programme
aus der Informatik und Mathematik
entwickelt und anwendet,
um Fragestellungen aus den molekularen Lebenswissenschaften zu lösen.

Bioinformatik-Anwender

wie oben, ohne „Modelle“ und „entwickelt“.

Wichtig

In dieser Vorlesung werden Sie zu einem (gut informierten) Anwender.
Sie lernen **nicht**, formale Modelle zu entwickeln.
Sie lernen auch nicht programmieren (außer ein wenig R).

Tipp: Besuchen Sie später evtl. einen Programmierkurs.

Informatik in der Biologie

Hauptgrund: Hochdurchsatztechnologien, große Datenmengen
(z.B. DNA-Sequenzierung, Massenspektrometrie, Mikroskopie-Aufnahmen)

Verwaltung von großen Datenmengen

Datenbanken (schneller Zugriff, Auffindbarkeit von Informationen)
Ausfall-tolerante Systeme (z.B. auch bei Festplattencrash)

Analyse von großen Datenmengen

z.B. Genomassemblierung, Identifikation von Metaboliten,
hierzu braucht man effiziente Algorithmen und gute Hardware!

Design von Experimenten

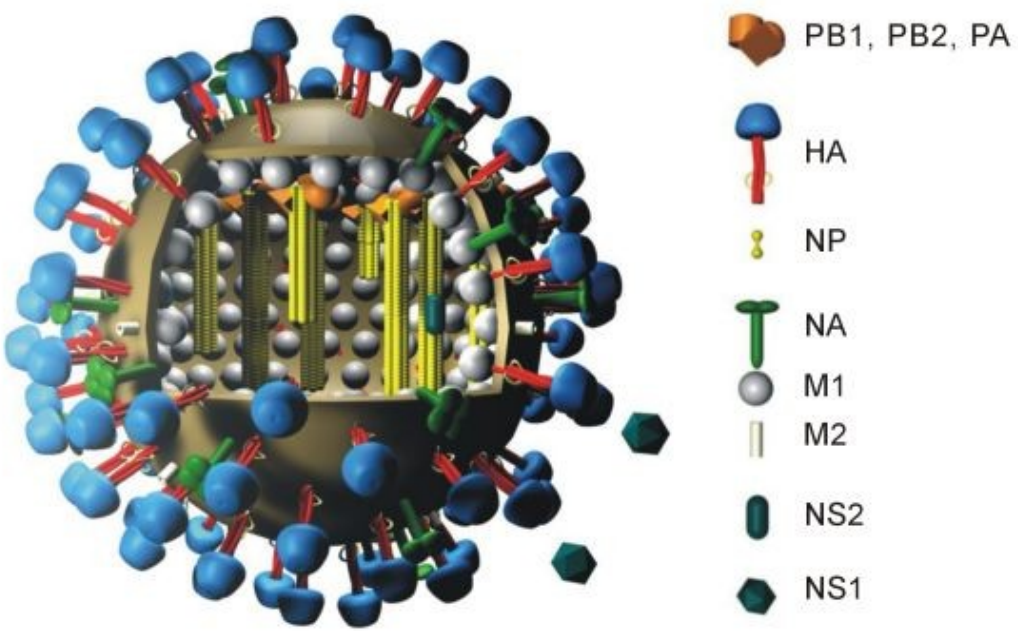
Maximum an neuen Informationen mit möglichst wenig Aufwand?
Welche Untersuchung ist am Erfolg versprechendsten?

Simulation

Vermeidung von (teuren) Experimenten – Vorhersage am Computer
(z.B. Wirkungsweise eines neuen Medikaments anhand molekularer Dynamik)

Ein Beispiel: Sequenzanalyse des Influenza-A Virus H1N1

- Hämagglutinin (HA oder H),
- Neuraminidase (NA oder N),
- Nukleoprotein (NP),
- Matrixproteine (M1) und (M2),
- Polymerase Proteine (PB1), (PB2) und (PA),
- Nichtstrukturproteine (NS1) und (NS2).



Vergleich des NS1-Proteins zweier H1N1-Viren

- Idee: Wähle zwei möglichst unterschiedliche Viren aus
  Puerto Rico, 1934 und Taiwan, 2002
  Datenbank: http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/select.cgi
- Sequenzen im FASTA-Format, Länge 693 nt = 231 aa

```
>gb|J02150:27-719| /Human/NS1/H1N1/8/Puerto Rico/1934/// nonstructural protein ns1
ATGGATCCAAACACTGTGTCAAGCTTTCAGGTAGATTGCTTTCTTTGGCATGTCCGCAAACGAGTTGCAG
ACCAAGAACTAGGTGATGCCCCATTCCTTGATCGGCTTCGCCGAGATCAGAAATCCCTAAGAGGAAGGGG
CAGCACTCTTGGTCTGGACATCGAGACAGCCACACGTGCTGGAAAGCAGATAGTGGAGCGGATTCTGAAA
GAAGAATCCGATGAGGCACTTAAAATGACCATGGCCTCTGTACCTGCGTCGCGTTACCTAACCGACATGA
CTCTTGAGGAAATGTCAAGGGAATGGTCCATGCTCATACCCAAGCAGAAAGTGGCAGGCCCTCTTTGTAT
CAGAATGGACCAGGCGATCATGGATAAAAACATCATACTGAAAGCGAACTTCAGTGTGATTTTTGACCGG
CTGGAGACTCTAATATTGCTAAGGGCTTTCACCGAAGAGGGAGCAATTGTTGGCGAAATTTCACCATTGC
CTTCTCTTCCAGGACATACTGCTGAGGATGTCAAAAATGCAGTTGGAGTCCTCATCGGAGGACTTGAATG
GAATGATAACACAGTTCGAGTCTCTGAAACTCTACAGAGATTCGCTTGGAGAAGCAGTAATGAGAATGGG
AGACCTCCACTCACTCCAAAACAGAAACGAGAAATGGCGGGAACAATTAGGTCAGAAGTTTGA
>gb|DQ249269:27-719| /Human/NS1/H1N1/8/Taiwan/2002/// NS1 protein
ATGGATTCCCACACTGTGTCAAGCTTTCAGGTAAACTGCTTCCTTTGGCATGTCCGCAAACAAGTTGCAA
ACCAAGGTCTAGGCGATGCCCCCTTTCTTGATCGGCTTCGCCGAGATCAAAAGTCTCTAAAGGGAAAAGG
CAGCACTCTCGGTCTGAACATCAAAACAGCCACTTGTGTTGGAAAGCAAATAGTAAAGAGGGTTCTGAAA
AAAAAATCCGATGAGGCATTTAAAATGACAATGGCCTCCGCACTTGCTTCGCGGTACCTAACTGACATGA
CTATTGAAAAAATGTCAAGGGACTGGTTCATGCTCATGCCCAAGCAGAAAGTGGCTGGCCCTCTTTGTGT
CAAAATGGACCAGGCGATAATGGATAAGAACATCATACTGAAAGCGAATTTCAGTGTGATCTTTGATCGG
TTGGAGAATCTGACATTACTAAGGGCTTTCACCGAAGAGGGAGCAATTGTTGGCGAAATTTCACCATTGC
CTTCTCTTCCAGGACATACTAATGAGGATGTCAAAAATGCAATTGGGGTCCTCATCGGGGGACTTGAATG
GAATGATAACACAGTTCGAGTCTCTGAAACTCTACAGAGATTCGCTTGGAGAAGCAGTAATGAGACTGGG
GGACCTCCATTCACTCCAACACAGAAACGGAAAATGGCGGGAACAATTAGGTCAGAAGTTTGA
```

Das FASTA-Sequenz-Dateiformat

Biologische Sequenzen (DNA, Proteine)
weden oft im sogenannten FASTA-Format gespeichert.
Dabei können sie in einer „Titelzeile“
mit zusätzlichen Informationen annotiert werden.

Eine Datei kann aus mehreren Sequenzen bestehen.
Jede Titelzeile (header) beginnt dabei immer mit `>`.
Esfolgen Sequenzdaten bis zum nächsten header oder bis zum Dateiende.

Beispiel (2 Sequenzen)

```
>Lieblingssequenz
ACGTTGCA
>andere Sequenz aus dem Internet
AAAAAAAAAA
AAAAAAAAAA
AAAAAAAAAT
```

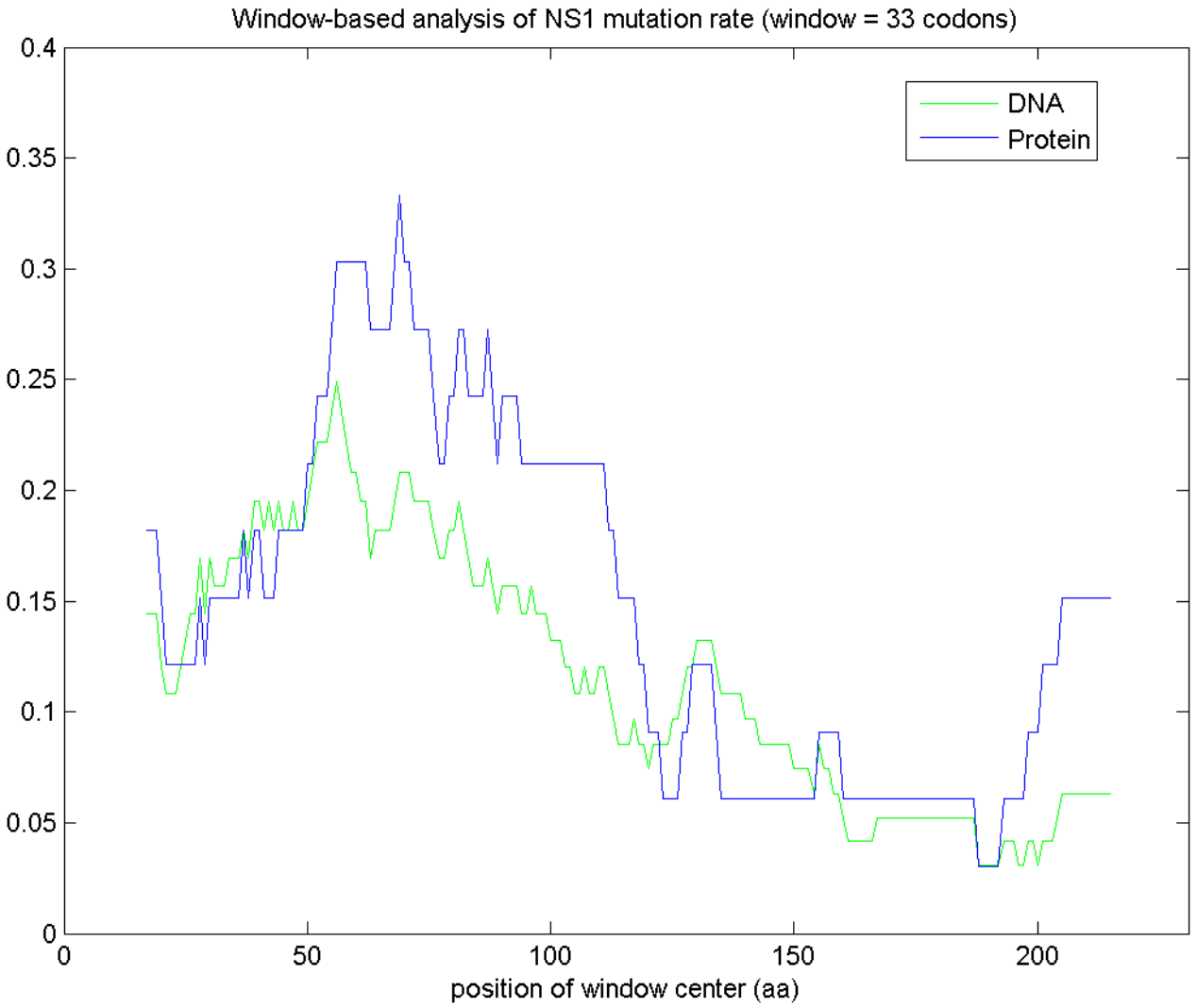
Erinnerung: Genetischer Code (Standard)

```
  AAs  = FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts = ---M---------------M---------------M---------------------------
Base1  = TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
Base2  = TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
Base3  = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

Untersuche die Mutationsrate fensterweise

- Wähle (beliebig) Fensterlänge 99 = 49+1+49 bp (Basenpaare), also 33 = 16+1+16 aa (Aminosäuren)
- Für jeden Fenstermittelpunkt:
  Wie viele nt / aa sind im Fenster unterschiedlich?
- Mutationsrate = Anzahl der Unterschiede / Fensterlänge
- Zeigt, welche Bereiche des Gens sich stärker verändert haben, sowohl auf DNA- als auch auf Proteinebene

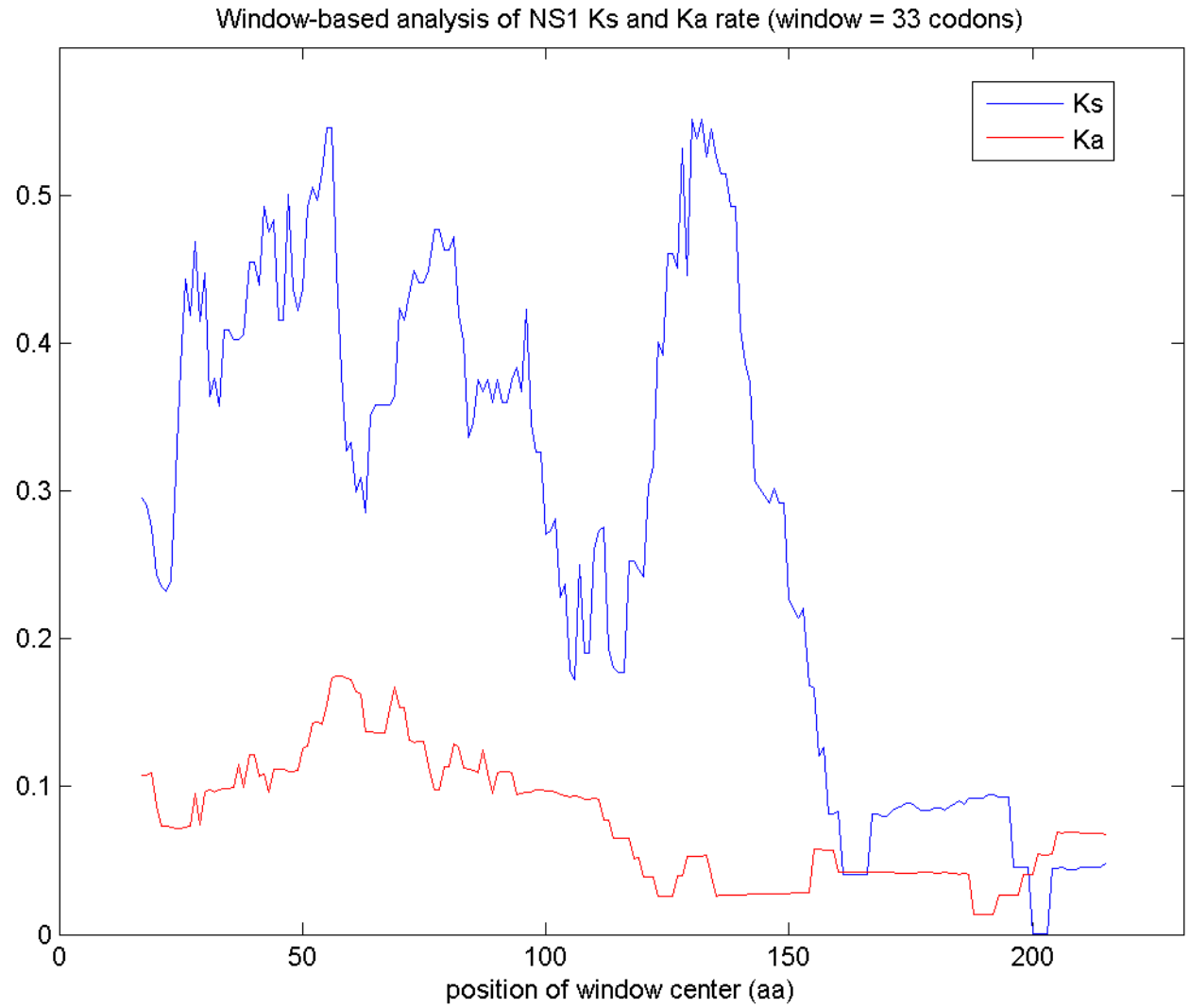
Mutationsrate auf DNA- und Proteinebene



Synonyme und nichtsynonyme Substitutionsrate

- DNA-Codon-Mutation, die die Aminosäure nicht ändert, heisst synonym.
- Betrachte ein Codon (mit seiner Aminosäure)
  Welche Tendenz zu synoymer/nichtsynonymer Mutation ?
  Betrachte die 9 Codons, die durch eine Substitution an einer der drei Stellen entstehen können.
  Berechne die Anzahl synonymer Codons dieser 9.
  Dividiere durch 3, um die „Anzahl synonymer Stellen“ zu erhalten.
- Beispiel: TTA (Leucin): 2 von 9,
  d.h. 2/3 syonyme Stellen, 7/3 nichtsynonyme Stellen
- Für ein Fenster:
  Zähle insgesamt synonyme und nichtsynonyme Stellen durch Summation.
  Zähle insgesamt synonyme und nichtsynonyme Mutationen.
  Verhältnisse heissen K_A (nichtsynonym) und K_S (synonym).
  K_S = synonyme Mutationen pro synonymer Stelle

Ks- und Ka-Rate

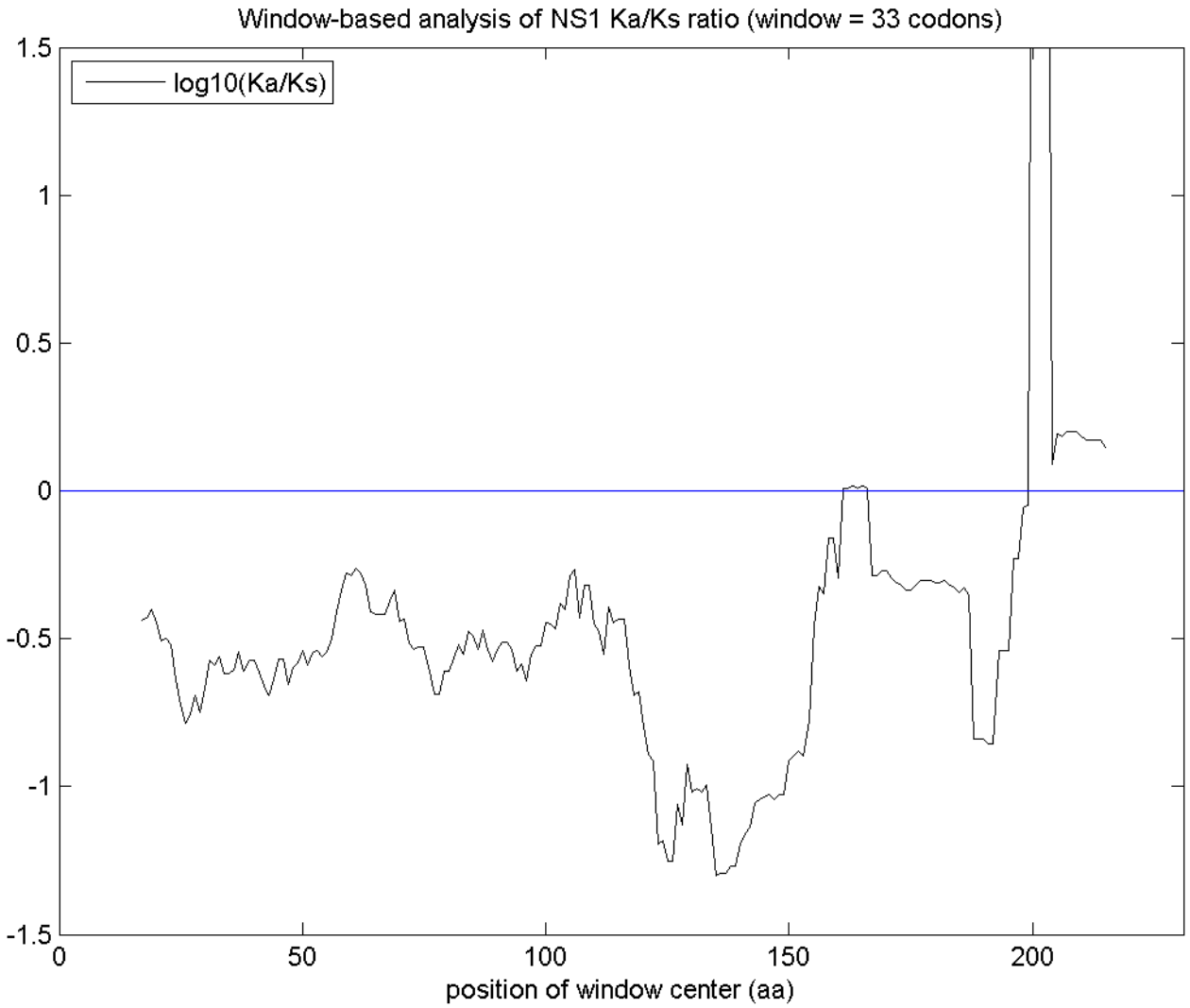


KA/KS - Analyse

- Betrachte die Verhältnisse K_A (nichtsynonym) / K_S (synonym).
- Verhältnis > 1: mehr nichtsynonyme Mutationen pro nichtsynonymer Stelle als synonyme Mutationen pro synonymer Stelle
- Verhältnis < 1: entsprechend

- Was bedeutet das? (hier nur sehr grob)
- Verhältnis < 1 legt den Schluss nahe, dass das Protein unter negativer Selektion steht.
  Veränderungen der aa-Sequenz wirken sich vermutlich direkt negativ auf die Fitness des Proteins aus.

Ka/Ks-Analyse



Themen der Vorlesung

Themen

- Literatursuche mit PubMed
- Die NCBI-Datenbanken und das Entrez-Interface
- Nukleotidsequenz-Datenbanken
- Suche in Sequenzdatenbanken mit BLAST
- Protein-Datenbanken und Werkzeuge zur Proteinanalyse
- Proteinstruktur
- Phylogenetik
- DNA-Microarrays
- Zelluläre Netzwerke

Fähigkeiten

- Typische Arbeitsabläufe der Bioinformatik
- Arbeiten mit der shell unter Linux/Unix
- Einfache statistische Auswertungen mit R
- Benutzung Web-basierter Software

Arbeiten mit Linux / Unix: Begriffserklärungen

Hardware

Oberbegriff für die maschinentechnische Ausrüstung eines Systems - („kann man anfassen“).

Software

Zusammenfassender Begriff für Programme und Daten in Computern - Dazu gehören Betriebssysteme und Anwendungsprogramme.

Betriebssystem (operating system, OS)

Software, die die Verwendung (den Betrieb) eines Computers ermöglicht.
Verwaltet Betriebsmittel wie Speicher, Ein- und Ausgabegeräte.
Steuert die Ausführung von Programmen.
Beispiele: Linux, Unix, Solaris, Windows, Mac OS, ...

Begriffserklärungen

Browser (to browse: durchstöbern)
Programm, um Informationen übersichtlich zu betrachten, z.B.
- Dateibrowser zeigt Dateien in einem Verzeichnis an.
- Webbrowser zeigt Seiten aus dem Internet an.

Account, home, login, password
Solaris (das Betriebssystem der Firma Sun) und Linux: Mehrbenutzersysteme.
Jeder hat seinen eigenen Arbeitsbereich (home-Verzeichnis).
Zugriffsbeschränkung durch eigene Benutzerkennung (account),
dazu geheimes Passwort.

Verzeichnis (directory, Ordner)
„Behälter“ für verschiedene Dateien.
Man benutzt eine Hierarchie von Verzeichnissen, um Ordnung zu halten.
So entsteht ein Verzeichnisbaum.

Begriffserklärungen

Arbeitsumgebung und Fenster-Manager (z.B. KDE mit KWin)

- Grundlegende Funktionen für die Arbeit mit dem Computer
- Grundlegende Fensterfunktionen (Minimieren, Vergrößern, Schließen)

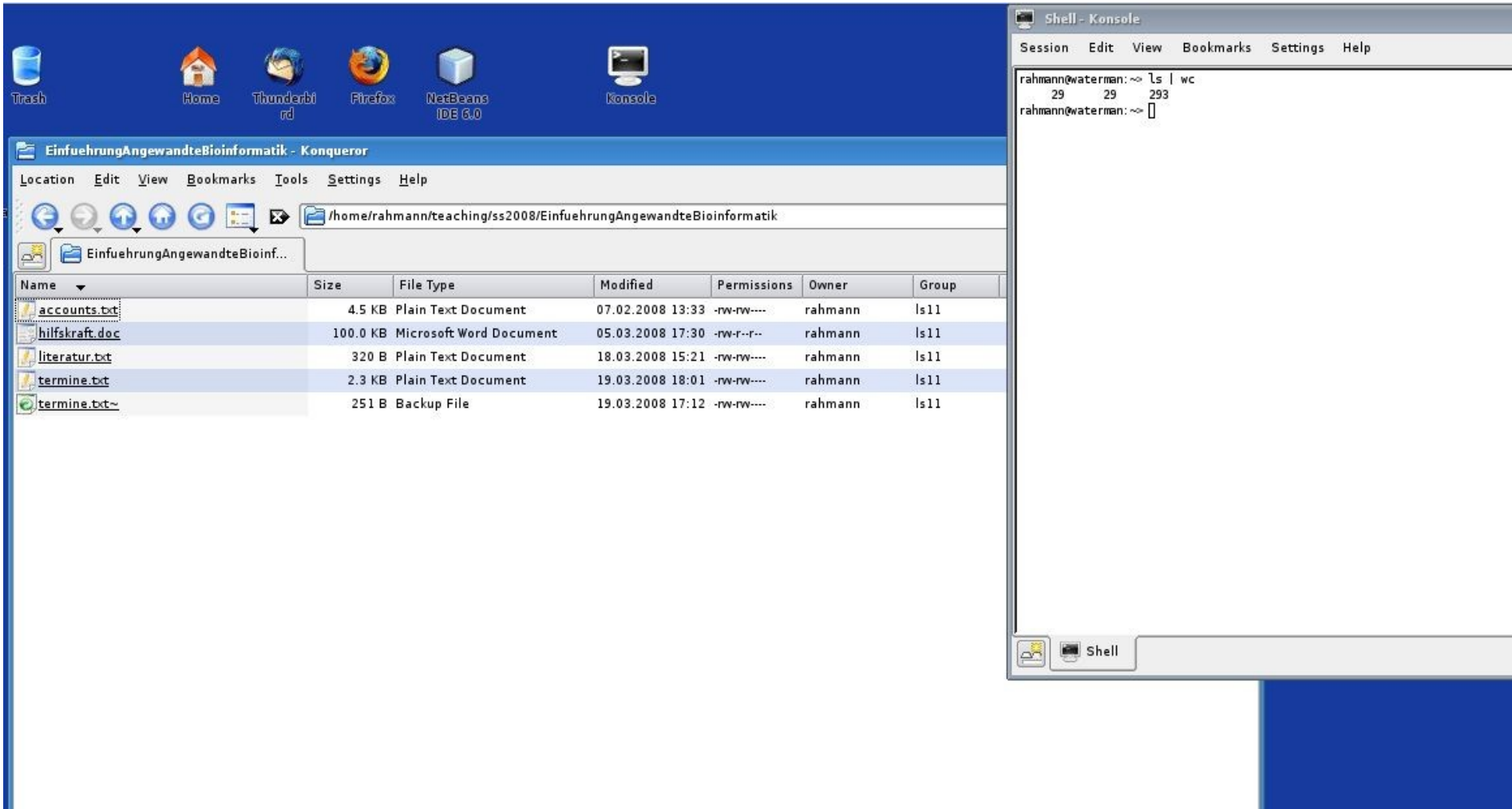
GUI (graphical user interface, graphische Benutzeroberfläche)

erlaubt das Ausführen von Aktionen (z.B. Verzeichnis anlegen)
mit Hilfe von Menüs und Dialog-Fenstern

Shell

Textorientierte Eingabe-Schnittstelle zwischen Computer und Benutzer,
erlaubt die Eingabe von Befehlen (z.B. mkdir zum Anlegen eines Verzeichnisses);
zunächst weniger komfortabel, aber mächtiger als ein GUI.

Beispiel:
KDE-Desktop mit Dateibrowser (Konqueror) und Shell-Fenster



Arbeiten mit der Shell

In einem Shell-Fenster („Konsole“) können Befehle eingegeben werden.

Generelle Befehlssyntax

Befehl Optionen Argumente

- „Befehl“ ist der Name des Befehls
  Beispiel: `ls` listet alle Dateien in einem Verzeichnis auf
- „Optionen“ modifizieren das Verhalten eines Befehls;
  beginnen mit Minus.
  Beispiel zu `ls`: `-a` zeigt alle Dateien (auch sonst versteckte; „all“),
  `-l` zeigt ausführliche Informationen („long“)
- „Argumente“ geben an, worauf der Befehl angewendet werden soll;
  häufig der Name eines Verzeichnisses oder einer Datei (oder mehrere)
  Beispiel: Ein Punkt (.) steht für das aktuelle Verzeichnis.

Gesamtbeispiel: `ls -l -a .`
Kürzer: `ls -la` (Optionen kombiniert; Punkt ist hier Standard, weggelassen)

Verzeichnis-Befehle

- `pwd`: aktuelles Verzeichnis anzeigen
- `mkdir`: neues Verzeichnis anlegen (`mkdir testvz`)
- `cd`: Verzeichnis wechseln (`cd testvz`)
  `cd` ohne Verzeichnisnamen wechselt in Ihr home-Verzeichnis
  Ein Punkt (.) steht für das aktuelle Verzeichnis;
  zwei Punkte (..) für das Verzeichnis darüber.
- `rmdir`: Verzeichnis entfernen, wenn leer (`rmdir testvz`)
  Warnung: Löschen kann nicht rückgängig gemacht werden
- `ls`: Verzeichnisinhalt anzeigen
- `ls -la`: Verzeichnisinhalt detailliert anzeigen

Beispiel: Was passiert jeweils, wenn Sie nacheinander

```
cd; pwd; mkdir uebung; cd uebung; pwd;
cd ..; rmdir uebung; ls; pwd
```

eingeben?

Hinweis zu Übungsaufgaben

Es ist sinnvoll, jeden Übungszettel in einem eigenen Verzeichnis zu bearbeiten, z.B. `uebung1`, `uebung2`, ...
Für die anfallenden Verwaltungsaufgaben
(Anlegen, Umbenennen, Löschen von Verzeichnissen) können Sie:

- einen Dateibrowser (Konqueror) verwenden,
- die Shell (Konsole) verwenden und Befehle eingeben.

Die Wahl steht Ihnen frei.
Versuchen Sie aber, die Shell-Befehle zu üben!

Philosophie der Unix-Befehle

Jeder Befehl führt eine einfache, klar umrissene Aufgabe aus.
Reichhaltige Optionen ermöglichen viele Variationen.
Komplexe Aufgaben werden durch Aneinanderreihung von Befehlen möglich.

Hilfe zu Befehlen und Optionen anzeigen lassen

`man` befehlsname („manual pages“)
z.B. `man ls` zeigt die Wirkung und alle möglichen Optionen von ls

Fortsetzung nächste Woche