Einführung in die Angewandte Bioinformatik: Shell, Datenbanken, Publizieren, Impact-Faktoren 22.04.2010

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 zur Vorlesung (NEU!)

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

Fortsetzung

Shell-Befehle

Datei- und Verzeichnisbefehle

`cp Datei1 Datei2` – kopiert Datei1 nach Datei2
`cp Dateiliste Verzeichnis` – kopiert alle Dateien der Liste ins Verzeichnis
`mv Datei1 Datei2` – wie `cp`, aber löscht Original (move, Umbenennung)
`rm Dateiliste` – löscht alle in der Liste angegebenen Dateien
Warnung! Löschen mit `rm` oder `rmdir` kann nicht rückgängig gemacht werden!

Verwendung von wildcards * und ?

Häufig darf man statt einer einzelnen Datei eine Liste von Dateien angeben.
Statt diese explizit aufzulisten, kann man * und ? verwenden:

- * steht für eine beliebige Folge von Zeichen
- ? steht für ein beliebiges Zeichen

z.B. `rm test*.fasta`
löscht alle Dateien, deren Name mit test anfängt und deren Endung .fasta ist.

Vorsicht: Niemals `rm *` ausprobieren!
Löscht alles im aktuellen Verzeichnis!

Untersuchung von Dateien

- `cat Dateiliste`
  zeigt nacheinander die Inhalte aller Dateien an

- `head -N 17 Dateiliste`
  zeigt jeweils die ersten 17 Zeilen an

- `tail -N 17 Dateiliste`
  zeigt jeweils die letzten 17 Zeilen an

- `more Dateiliste`:
  zeigt nacheinander die Inhalte aller Dateien an,
  wartet auf Tastendruck wenn der Bildschirm voll ist

Untersuchung von Dateien – grep und wc

grep (general regular expression matcher)

`grep Muster Dateiliste`
sucht nach Muster in allen Dateien,
gibt alle Zeilen aus, in denen das Muster auftritt.
Beispiel: `grep Meier telefonbuch.txt`
(Datei ist Telefonbuch, ein Eintrag mit Name + Telefonnummer pro Zeile).
Sucht alle Einträge mit Namen „Meier“.
Das Muster kann viel komplizierter sein (reguläre Ausdrücke), z.B.
`grep M..er telefonbuch.txt`
Hier steht der Punkt für ein beliebiges Zeichen, man findet Meier, Mayer, Maler, ...

wc (word counter)

`wc Dateiliste`
gibt für jede Datei 3 Zahlen aus: Anzahl Zeilen, Wörter, Zeichen

Ein- und Ausgabeumleitung, Pipes (<, >, |)

Ausgabeumleitung >

Das Zeichen > leitet die Ausgabe eines Befehls in eine Datei um.
Achtung: Wenn die Datei schon existiert, wird sie überschrieben!

```
grep Meier telefonbuch.txt > meiers.txt
```

Eingabeumleitung <

Analog kann man die Eingabe zu einem Programm
aus einer Datei (statt z.B. von der Tastatur) beziehen.

Pipe |

Will man die Ausgabe eines Programms als Eingabe eines anderen verwenden,
kann man die Pipe (Rohr) | benutzen:

```
ls | wc
```

Zählt, wie viele Dateien im aktuellen Verzeichnis sind.
Die Ausgabe von `ls` (Verzeichnisinhalt) wird als Eingabe für `wc` verwendet.
Die Ausgabe von `wc` erscheint im Shell-Fenster.

Neues Thema

Datenbanken

Daten, Wissen, Datenbanken

Daten (Singular: Datum) :=
alles, das gesammelt, gespeichert und wieder gelesen werden kann.
Daten für sich sind nicht interpretiert, haben keine Bedeutung.

Wissen :=
interpretierte Daten, aus Daten abgeleitete Fakten

Beispiel: 108
Die Ziffernfolge 1-0-8 kann man als die Zahl 108 interpretieren,
diese wiederum als Hausnummer, oder als den 1. August, oder als Geldbetrag, ...

Datenbank (DB) :=
strukturierte Sammlung von Daten

Hinweis zur Notation
Das Zeichen := bedeutet „ist definiert als“, d.h., Sie lesen eine Definition.
Der Doppelpunkt steht auf der Seite des zu definierenden Begriffs.

Inhalt von Datenbanken

Datenbanken enthalten
alles, das gesammelt, gespeichert und wieder gelesen werden kann.
Daten für sich sind nicht interpretiert, haben keine Bedeutung.

Die Datenbank ist nur der „Container“,
der die Daten in strukturierter Form enthält,
so dass sie wiedergefunden werden **können**.

Das ist nutzlos,
solange es keine Möglichkeit gibt, auf die Daten zuzugreifen,
Daten hinzuzufügen, zu löschen, sie zu verändern, etc...

DB / DBMS / DBS

Datenbank-Managementsystem (DBMS) :=
Software, die den Zugriff auf eine Datenbank erlaubt –
bietet überhaupt erst die Möglichkeit, eine Datenbank zu nutzen.
Kann shell-ähnlich sein (man muss Befehle eintippen)
Kann browser-ähnlich sein (graphische Benutzeroberfläche, GUI)
Beispiele: MS-Access, OpenOffice.org Base, mySQL, ...

Datenbanksystem (DBS) :=
DB + DBMS (taucht im Grunde immer zusammen auf,
man kann aber dieselbe DB möglicherweise mit verschiedenen DBMS bearbeiten)

Beispiele für Datenbanken

Allgemeine Beispiele

- Kundendatenbanken (besitzt jede Firma)
- Adressdatenbanken (hat mein email-Programm; auch mein Adressbuch ist eine)
- Filmdatenbanken (z.B. imdb.com)
- Literaturdatenbanken (amazon braucht so etwas für alle Arten von Literatur)

Beispiele, die uns in dieser Vorlesung interessieren

- wissenschaftliche Literaturdatenbanken (z.B. PubMed – später)
- biologische Sequenzdatenbanken
  - für DNA / speziell für RNA / UniProt für Proteinsequenzen
- Proteinstrukturdatenbanken, z.B. PDB
- Datenbanken zu Reaktionsnetzwerken
  - metabolische Netze
  - Protein-Reaktionsnetzwerke / Signaltransduktionsnetzwerke

Textdateien („flat files“) als Datenbanken

Textdateien („flat files“)

Eine Textdatei ist eine Datei, in der die gespeicherte Information direkt, ohne Formatierung oder Meta-Informationen steht.
Beispiel: FASTA-Dateien,
Gegenbeispiel: Word (.doc)-Dateien mit Formatierung und Autoren-Information.

Historische Bedeutung von Textdateien als Datenbanken

Datenbank + DBMS: Infrastruktur nötig; Aufwand
1970er: Erste DNA-, Protein-Sequenzen (wenige!)
einfach und übersichtlich in Textdatei zu speichern
gut austauschbar
„historisch gewachsen“; heute noch immer Verwendung von Textdateien!

Vor- und Nachteile von „flat files“ als Datenbanken

Vor- und Nachteile von Textdateien

+ Man kann sie mit jedem Text-Editor lesen und bearbeiten
  (Shell-Befehle: `cat`, `more`; keine spezielle Software!)
- Suche oft ineffizient
- Datenintegrität und Datensicherheit werden nicht unterstützt

Vor- und Nachteile „echter“ Datenbanken

- Man benötigt ein DBMS zum Zugriff.
  Wer das „richtige“ DBMS nicht hat, kann auf die Informationen nicht zugreifen.
+ Erlaubt effiziente Suchanfragen (z.B. durch Erstellen eines Index)
+ Das DBMS kümmert sich um Integrität und Zugriffsrechte.

Neues Thema

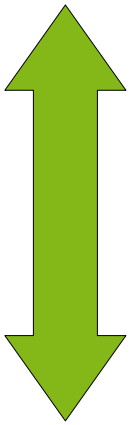
Wissenschaftliches Publizieren und Zitations-Datenbanken

Grundlagen wissenschaftlicher Veröffentlichungen

Wissenschaftler werden aus Steuergeldern bezahlt.
Forschungsergebnisse sollen der Allgemeinheit zur Verfügung stehen.
Es gibt mehrere Möglichkeiten, Ergebnisse zu veröffentlichen.

Arten der Veröffentlichung

- Konferenzbeiträge
- Forschungsartikel in Fachzeitschriften
- Übersichtsartikel in Fachzeitschriften
- Buchkapitel
- studentische Abschlussarbeiten
- Lehrbücher



Neue Resultate
schnelle Veröffentlichung

ältere Resultate
langsame Veröffentlichung

Fachzeitschriften

JournalSeek-Datenbank: http://journalseek.net/

- Welche Fachzeitschriften gibt es in welchem Bereich?
- Was bedeuten die Abkürzungen der Zeitschriften-Titel?
- ISSN (International Standard Serial Number) identifiziert Zeitschrift eindeutig.

Z.B. i.d. Bioinformatik:
J Comput Biol = Journal of Computational Biology hat ISSN 1066-5277

Wer trägt die Kosten wissenschaftlicher Veröffentlichungen?

2 Modelle:

Klassisch:
Der Leser (durch Bestellen der Zeitschrift)

Neuer („open access“):
Der Autor bezahlt den Verlag für die Druckkosten,
häufig gibt es „flat rates“ für eine ganze Universität oder Institute.

Der Verlag, nicht der Forscher, verdient Geld mit Publikationen.

Qualitätssicherung wissenschaftlicher Veröffentlichungen

Grundprinzip

Veröffentlichung von Forschungsergebnissen bedeutet Fortschritt:
Andere machen weiter, wo einem selbst die Ideen ausgegangen sind.
Schlecht, wenn (viel) Unsinn veröffentlicht wird.
Generell sollte gelten: Was gedruckt wird, stimmt.

Qualitätssicherung durch Peer Review

Geschriebene Artikel werden nicht einfach gedruckt.
Andere Fach-Wissenschaftler lesen und kommentieren die Artikel vor Erscheinen.
Wenn alle Fragen und Bedenken ausgeräumt sind, wird der Artikel gedruckt.
„Peer Review“: Kollegen (an anderen Instituten) begutachten die Arbeit.
peer: Gleichrangiger, Gleichgestellter
Problematisch: Kostet viel Zeit und Arbeit, wird nicht bezahlt.
Aber: Ohne peer review bricht guter Wissenschaft das Fundament weg.

Bewertung wissenschaftlicher Veröffentlichungen

Grundidee

Nicht alle Artikel sind gleich wichtig, vor allem auf lange Sicht.
Man zitiert einen Artikel, wenn man sich in seiner Arbeit darauf beruft.
Richtungsweisende Artikel werden häufiger gelesen und häufiger zitiert.

Versuch einer Einteilung

„Write-only“-Artikel:
- werden nicht gelesen
- Verschwendung von Zeit und Geld
- dienen dazu, die eigene Publikationsliste zu verlängern

Kleine Fortschritte, Verbesserungen bestehender Resultate:
- häufigste Art der Publikation

Große Fortschritte, Lösung langer offener Probleme:
- seltener, erzeugen größere Aufmerksamkeit in der Fachwelt

Grundlagen eines neuen Forschungsfeldes:
- sehr sehr selten, Nobelpreisverdächtig

Qualitätsstandards wissenschaftlicher Zeitschriften

Verschiedene Zeitschriften haben verschiedene Qualitätsstandards.

Immer:
- Die Arbeit muss methodisch einwandfrei sein;
- keine offensichtlichen Fehler!

Unterschiede bei:
- Wie groß ist der erzielte Fortschritt?
- Wie aktuell ist das Thema gerade?

Ziel jedes Wissenschaftlers:
- möglichst „hoch“ publizieren („gute“ Zeitschrift)
- eigentlich Unsinn, denn es sollte um den Inhalt des Artikels gehen; Zeitschrift lebt von guten Artikeln, nicht Artikel von guten Zeitschriften (oder doch??)

„Einfluss“ von Zeitschriften

Wie den Einfluss (impact) eines *Artikels* messen?

Übliches Maß: impact factor der *Zeitschrift*!

- Robuster zu berechnen als für jeden Artikel einzeln
- Ungenauer: unbedeutende Artikel können in guten Zeitschriften erscheinen

Zeitschriften achten sehr auf einen hohen impact factor,
und akzeptieren nur Artikel, die diesen vermutlich halten oder übertreffen.

Definition des Impact Factors

$$\frac{\text{Zahl der Zitate im Jahr } x \text{ auf Artikel der Zeitschrift in den Jahren } x\text{-}2 \text{ und } x\text{-}1}{\text{Zahl der Artikel der Zeitschrift in den Jahren } x\text{-}2 \text{ und } x\text{-}1}$$

Kritik

IF misst durchschnittliche Zitierhäufigkeit aller Artikel einer Zeitschrift

- Durchschnitt sagt nichts über individuelle Artikel
- Zitierhäufigkeit sagt etwas über Bekanntheit, nicht Relevanz
- Zeitschrift sagt nichts über individuelle Forscher;
  trotzdem ist kumulativer / durchschnittlicher IF wichtig bei Bewerbungen

Alternative

- Hirsch-Index eines Wissenschaftlers [Übung]

Ermittlung des Impact Factors

Wer macht das mit welchem Aufwand?

Thomson Scientific (früherer Name: Insititue of Scientific Information, ISI)
http://thomsonreuters.com/products_services/science/free/essays/impact_factor/
Literaturliste jeder(!) Veröffentlichung durchgehen,
in Bezug zu vergangenen Veröffentlichungen setzen
Manuelle Nachbearbeitung; teuer!

Datenbank dieser Verknüfpungen:

Science Citation Index Entended (SCIE) im ISI Web of Science / of Knowledge
http://www.isiwebofknowledge.com/

Zukunft:

Automatisierte Suchmaschinen, z.B. Google Scholar (http://scholar.google.com)
Grundlegende Probleme dabei: Korrektheit, Vollständigkeit

Einfluss einzelner Forscher

Wie kann man den „impact“ einer einzelnen Person messen?

Es gibt verschiedene Kennzahlen.
Offensichtlich auch einen großen Markt dafür.

Hirsch-Index

Ein Wissenschaftler hat den Hirsch-Index *h*,
wenn er mindestens *h* Artikel geschrieben hat,
die jeweils mindestens *h*-mal zitiert wurden,
und es kein größeres solches *h* gibt.

(Man kann versuchen, *h* mit google scholar herauszufinden,
Oder mit spezieller Software wie „PublishOrPerish“.)