Einführung in die Angewandte Bioinformatik: PubMed und Einführung in R 29.04.2010

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

Neues Thema

Literatur-Datenbanken

Literatursuche

Wichtige Frage, bevor man ein Forschungsprojekt beginnt:

„Was gibt es schon?“

Um dies herauszufinden, benutzt man Literaturdatenbanken.

Früher:
spezielle Review-Zeitschriften mit Zusammenfassungen anderer Artikel

Datenbanken und Systeme zur biomedizinischen Literatur

MEDLINE :=
öffentliche Datenbank mit (im weitesten Sinne biomedizinischen) Artikeln, Querverweisen, Zusammenfassungen, ca. 5000 verschiedene Zeitschriften

PubMed :=
frei zugreifbares online-System, das die MEDLINE-Datenbank enthält und komplexe Abfragen erlaubt [damit befassen wir uns jetzt!]

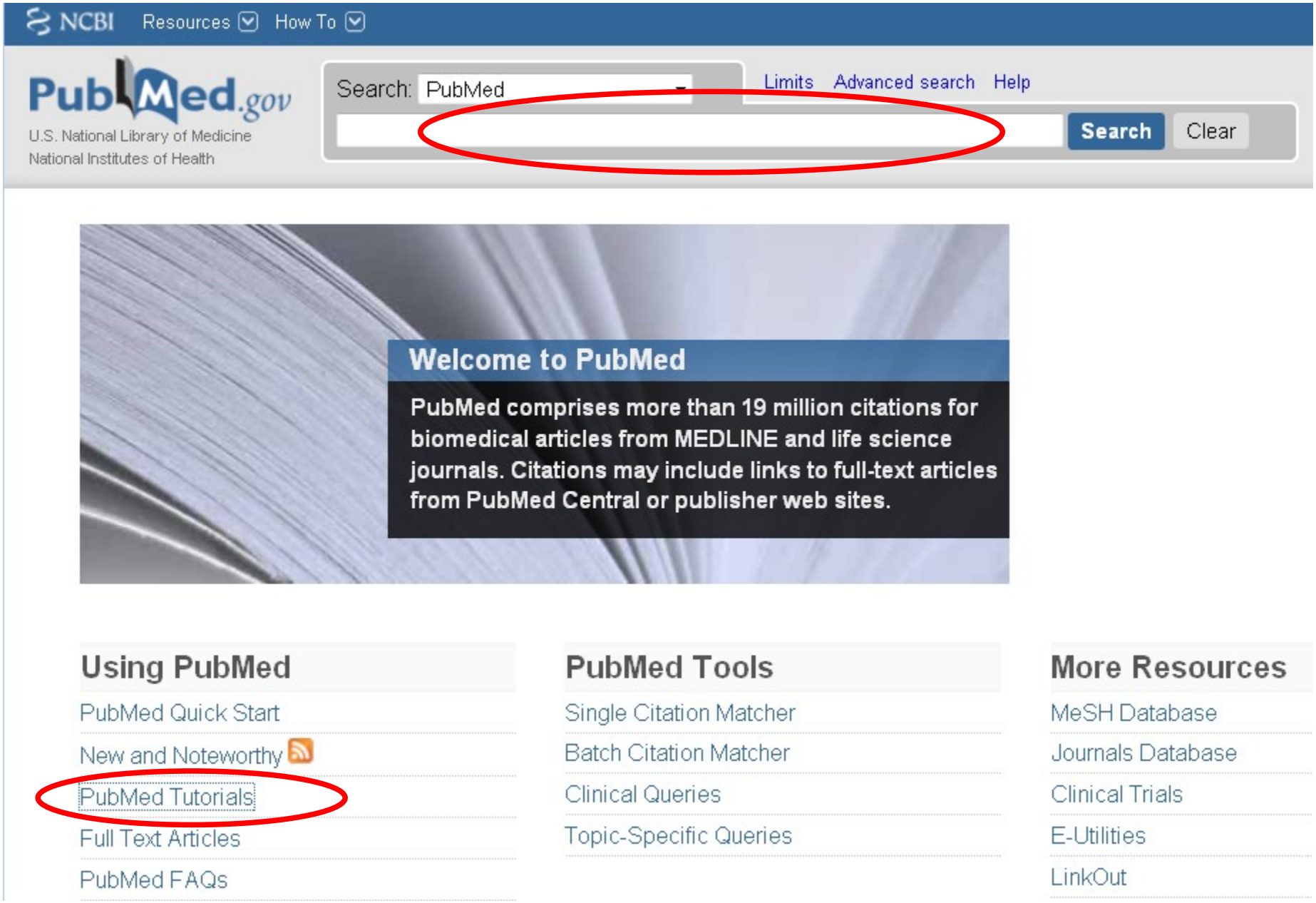
PubMedCentral :=
frei zugängliches digitales Archiv von Artikeln aus den Lebenswissenschaften

Entrez :=
System, das eine gemeinsame Oberfläche und Abfragesystem für PubMed und andere Datenbanken bietet

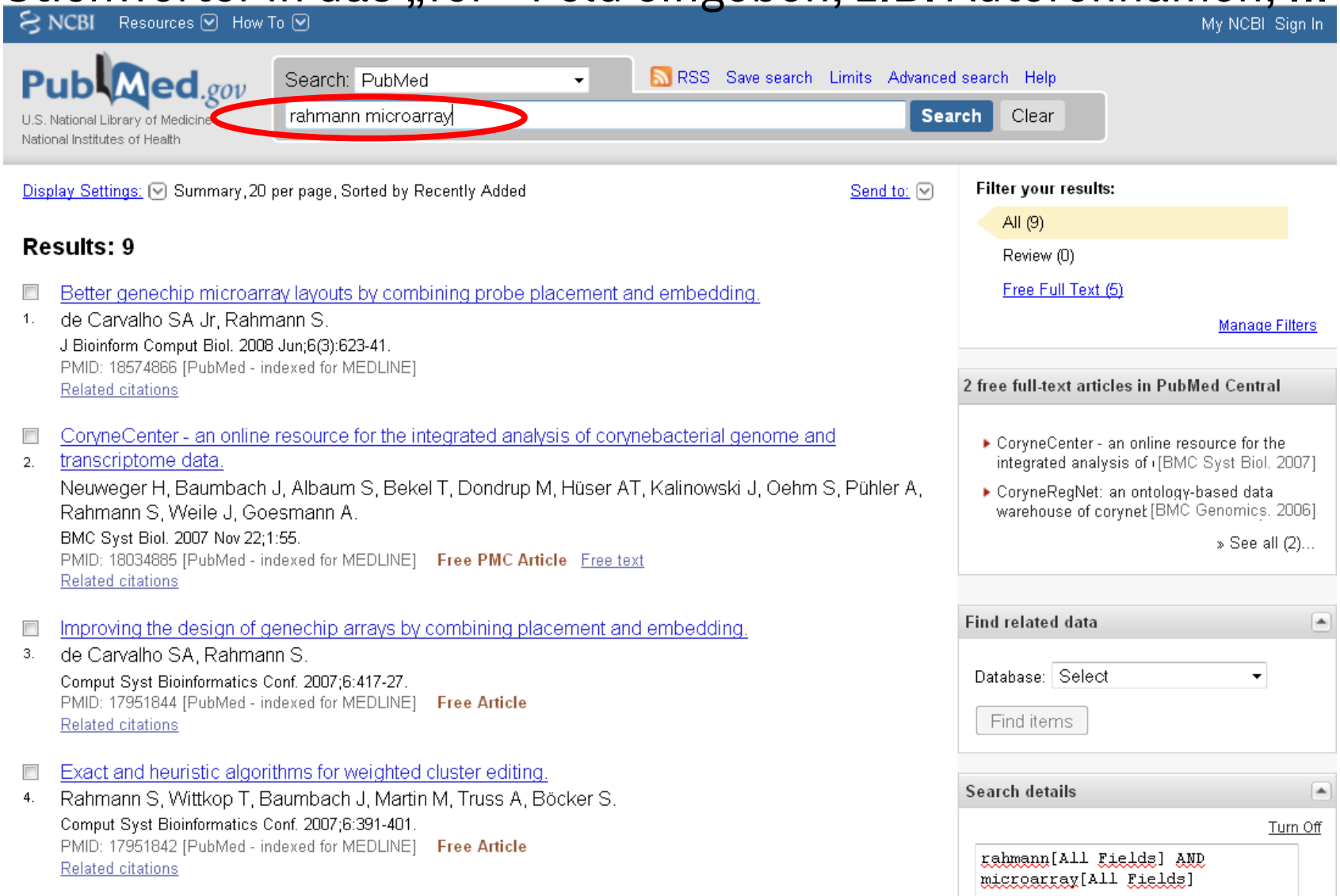
Zugriff auf die MEDLINE-Datenbank mit PubMed

Zugriff auf MEDLINE mit PubMed über Entrez unter
http://www.ncbi.nlm.nih.gov/sites/entrez/
http://www.ncbi.nlm.hih.gov/pubmed

Was sind NCBI, NLM, NIH ?
NCBI = National Center for Biotechnology Information
NLM = National Library of Medicine
NIH = National Institute of Health
(Bethesda, MD, USA)


NCBI
Resources
How To
PubMed.gov
U.S. National Library of Medicine
National Institutes of Health
Search: PubMed
Limits
Advanced search
Help
Search
Clear
Welcome to PubMed
PubMed comprises more than 19 million citations for biomedical articles from MEDLINE and life science journals. Citations may include links to full-text articles from PubMed Central or publisher web sites.
Using PubMed
PubMed Quick Start
New and Noteworthy
PubMed Tutorials
Full Text Articles
PubMed FAQs
PubMed Tools
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
Topic-Specific Queries
More Resources
MeSH Database
Journals Database
Clinical Trials
E-Utilities
LinkOut

Einfache PubMed-Suche:
Stichwörter in das „for“-Feld eingeben, z.B. Autorennamen, ...



PubMed-Suchfunktionen

Ähnlich wie bei Internet-Suchmaschinen:

- Alle eingegebenen Wörter müssen vorkommen
- Reihenfolge der Wörter spielt keine Rolle
- Anführungszeichen legen Zusammenhang und Reihenfolge der Wörter fest
- Man kann generell nach Autor, Titel, Wörtern im abstract (Zusammenfassung), Jahreszahlen, ..., suchen.
- Groß - und Kleinschreibung spielt keine Rolle.

Beispiele

- Rahmann Microarray
  es wird eine Liste der passenden Artikel angezeigt
- Rahmann “Microarray Design”
  bei nur einem Treffer werden direkt mehr Details gezeigt

PubMed.gov
U.S. National Library of Medicine
National Institutes of Health

Search: PubMed | RSS | Save search | Limits | Advanced search | Help

rahmann "microarray design" | Search | Clear

Display Settings: Abstract

Send to:

Proc IEEE Comput Soc Bioinform Conf. 2003;2:84-91.

Group testing with DNA chips: generating designs and decoding experiments.

Schliep A, Torney DC, Rahmann S.

Department of Computational Molecular Biology, Max-Planck-Institute for Molecular Genetics, Inestrasse 63-73, D-14195 Berlin , Germany.
Alexander.Schliep@molgen.mpg.de

Abstract

DNA microarrays are a valuable tool for massively parallel DNA-DNA hybridization experiments. Currently, most applications rely on the existence of sequence-specific oligonucleotide probes. In large families of closely related target sequences, such as different virus subtypes, the high degree of similarity often makes it impossible to find a unique probe for every target. Fortunately, this is unnecessary. We propose a microarray design methodology based on a group testing approach. While probes might bind to multiple targets simultaneously, a properly chosen probe set can still unambiguously distinguish the presence of one target set from the presence of a different target set. Our method is the first one that explicitly takes cross-hybridization and experimental errors into account while accommodating several targets. The approach consists of three steps: (1) Pre-selection of probe candidates, (2) Generation of a suitable group testing design, and (3) Decoding of hybridization results to infer presence or absence of individual targets. Our results show that this approach is very promising, even for challenging data sets and experimental error rates of up to 5%. On a data set of 28S rDNA sequences we were able to identify 660 sequences, a substantial improvement over a prior approach using unique probes which only identified 408 sequences.

PMID: 16452782 [PubMed - indexed for MEDLINE]

Publication Types, MeSH Terms, Substances

Publication Types:

Evaluation Studies
Validation Studies

MeSH Terms:

Algorithms*
Base Sequence
DNA Probes/genetics*

Eingabe eindeutiger Suchwörter:
direkt zum Abstract.

Related articles

- Optimal robust non-unique probe selection using Integer Linear Programmin [Bioinformatics. 2004]
- Decoding non-unique oligonucleotide hybridization experiments [Bioinformatics. 2006]
- Fast and sensitive probe selection for DNA c [Proc IEEE Comput Soc Bioinform Conf. 2003]
- Review Sequencing by hybridization (SBH): advantages [Adv Biochem Eng Biotechnol. 2002]
- Review Conjugated polyelectrolytes for label-free DNA microar [Trends Biotechnol. 2008]

» See reviews... | » See all...

Cited by 2 PubMed Central articles

- Ultraspecific probes for high throughput HLA typing. [BMC Genomics. 2009]
- In silico microarray probe design for diagnosis of multiple pathogens. [BMC Genomics. 2008]

All links from this record

- Related Citations
- Cited in PMC

Search details

Turn Off

```
rahmann[All Fields] AND
"microarray design"[All Fields]
```

Feldrestriktionen

Problem

Die Suchanfrage **Down Syndrome** findet so ziemlich alles:
Artikel über das Down-Syndrom, Artikel von Dr. Down über irgendein Syndrom, ...

Lösung

Man legt fest, in welchem Feld (Autor, Titel, ...) man jeweils suchen will.

[PMID]	PubMed ID; eindeutige Nummer, die einem Artikel zugeordnet ist
[TI]	Titel
[AB]	Abstract, Zusammenfassung
[AD]	Adresse des Instituts des publizierenden Autors
[FAU]	Full author name, Voller Name des Autors
[AU]	Autor (Nachname + Initialen)
[SO]	(Abkürzung des Zeitschrift-Namens)

Feldrestriktionen - Beispiele

Rahmann S [AU]
: alle Artikel von allen Leuten, die S. Rahmann heißen

„Down Syndrome“ [TI]
: zusammenhängender Begriff im Titel

down [AU] AND syndrome [TI]
: Artikel von Dr. Down über Syndrome

12345678 [PMID]
: das durch diese Nummer eindeutig identifizierte paper

Tipp bei der Literatursuche:
PMIDs von relevanten Artikeln notieren

Boole'sche Verknüpfungen

Hintergrund des Namens:

George Boole, engl. Logiker, 1815 – 1864

Operatoren

AND (und, Standard, alle Terme müssen vorkommen)
OR (oder, ein Term muss vorkommen)
NOT (nicht, der Term darf nicht im genannten Zusammenhang vorkommen)

Beispiele

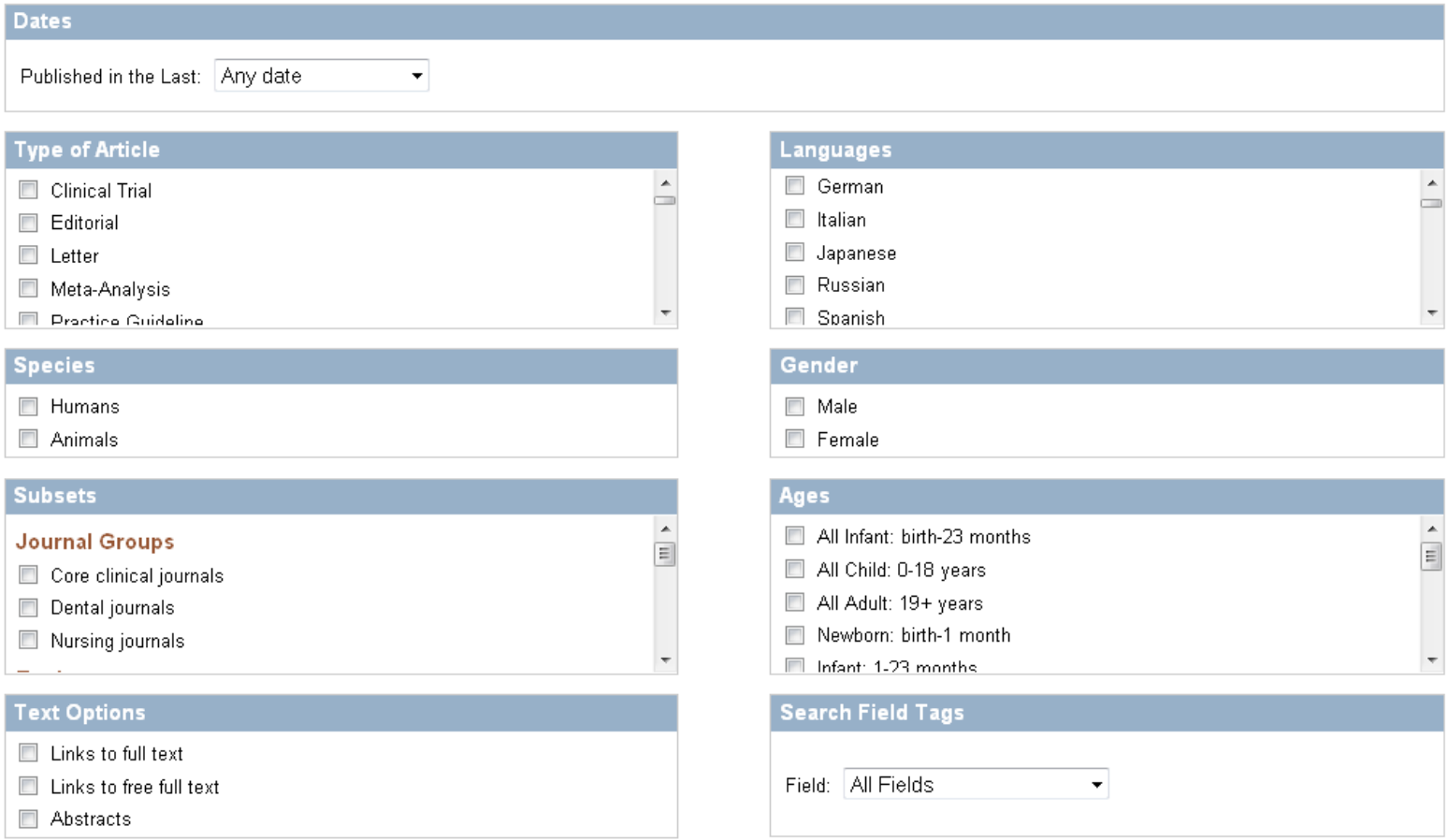
microarray [ti] AND Dortmund [ad]

Finde Microarray-Experten in Dortmund

microarray [ab] NOT Rahmann S [au]

Finde Arbeiten über Microarrays, die nicht von S.Rahmann sind

Limits



Statt die Suchanfrage „per Hand“ einzugeben,
kann man die Einschränkungen bequemer in einer Suchmaske eingeben;
aber: weniger flexibel bei komplexen Anfragen

Tipps beim Suchen

- Anführungszeichen benutzen, wo möglich
- Initialen der Autoren benutzen, wenn bekannt
- PMIDs notieren
- Bei zu vielen Ergebnissen, Suche weiter einschränken
- Bei zu wenig Ergebnissen, Suche erweitern

Problem: Synonyme
Dasselbe Konzept, dieselbe Idee
wird durch verschiedene Wörter oder Wortkombinationen ausgedrückt.
Man müsste alle ausprobieren, um sicher zu sein, alles zu finden!

Lösung
Standardisiertes Vokabular: MeSH-Terme (Medical Subject Headings)

MeSH-Terme

MeSH – Medical Subject Headings :=
standardisiertes und kontrolliertes Vokabular
zur Indizierung von Artikeln in MEDLINE / PubMed

MeSH-Terme erlauben,
auf konsistente Weise Informationen zu Themen zu erhalten,
die sich mit verschiedenen Begriffen beschreiben lassen.

Neues Problem dabei

Woher bekomme ich die richtigen MeSH-Terme zu einem Thema?

Lösung

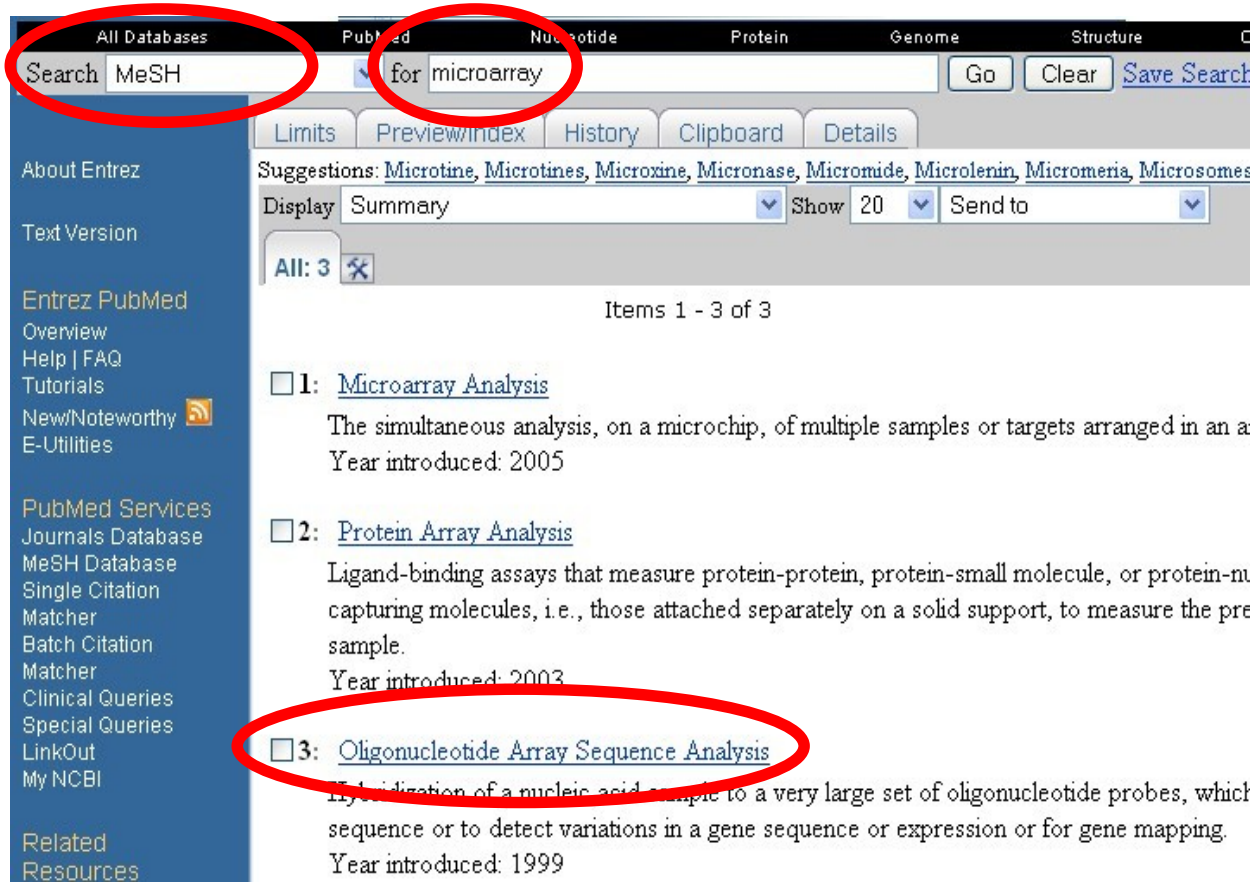
Durchsuche die MeSH-Datenbank,
die wie PubMed über Entrez zugänglich ist

Beispiel zu MeSH-Termen

Suche nach MeSH-Term, der bioinformatische Arbeiten zur Lösung des Microarray-Design-Problems umfasst.

Dann Verwendung in PubMed Suche mit Feldnamen [MH],

rahmann [AU] AND „oligonucleotide array sequence analysis“ [MH]

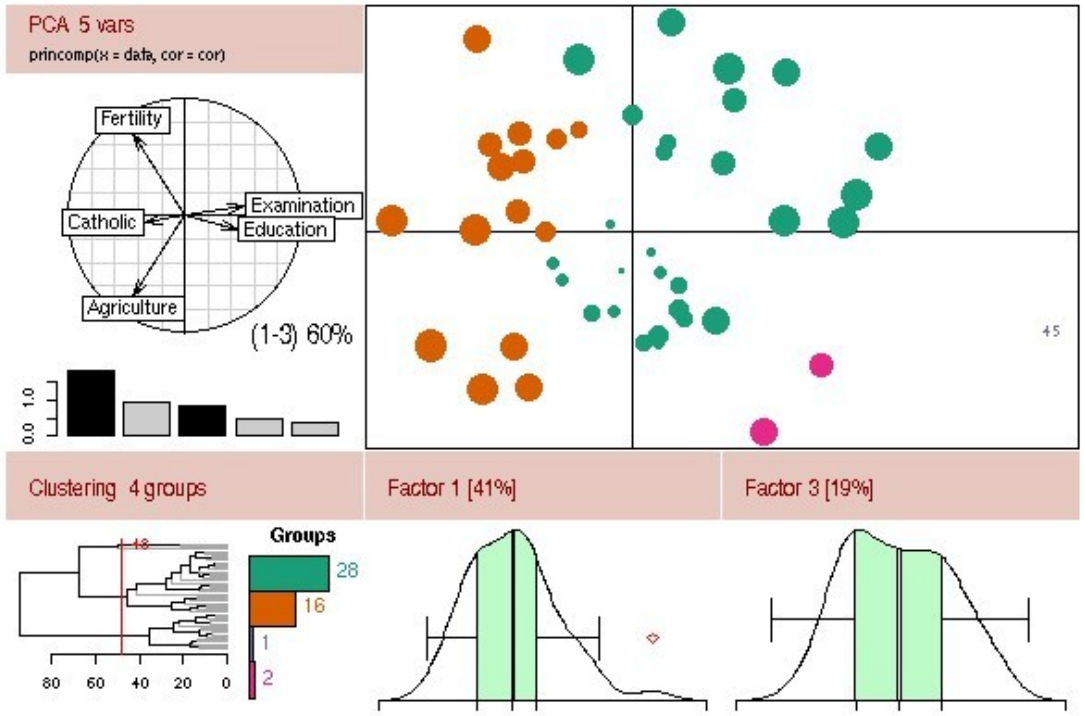


Neues Thema

Statistik mit R

Statistische Datenanalyse mit R: http://www.r-project.org

The R Project for Statistical Computing



About R
What is R?
Contributors
Screenshots
What's new?

Download
CRAN

R Project
Foundation
Members & Donors
Mailing Lists
Bug Tracking
Developer Page
Conferences
Search

Documentation
Manuals
FAQs
Newsletter
Wiki

Getting Started:

- R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred CRAN mirror.
- If you have questions about R like how to download and install the software, or what the license terms are, please read our

Über R

R ist eine freie Software zur statistischen Datenanalyse
(free software – free as in „beer“ and as in „speech“).

Wir besprechen einige statistische Grundlagen,
und wie man R zur einfachen Datenanalyse einsetzt.

Auf den Rechnern im Übungspool ist R vorinstalliert.
Zu Hause laden Sie sich R umsonst unter [http://www.r-project.org] herunter.

Erste Schritte:

- Starten und Verlassen von R
- R als Taschenrechner
- R als Plotter

Starten und Verlassen von R; Hilfe

Starten von R

Im Übungs-Pool: Eingabe des Kommandos „R“ in der Shell
Unter Windows: Herunterladen und Aufrufen des Installers -> Desktop-Icon

Hilfe in R

R funktioniert wie die Shell: Befehle werden nacheinander abgearbeitet
(aber es stehen andere Befehle zur Verfügung als in der Shell!)
Hilfe bekommt man mit `help()` oder einem vorangestellten Fragezeichen `?`

Verlassen von R

Aufruf von `q()`
q ist eine Funktion, daher wird sie mit Klammern aufgerufen.
In den Klammern stehen ggf. Argumente, aber q hat keine.

R als Taschenrechner

Es stehen alle arithmetischen Operationen,
elementaren mathematischen Funktionen,
sowie logischen Operationen zur Verfügung.

Die Eingabe eines Ausdrucks liefert dessen Wert (Ergebnis) zurück.
Ergebnisse können Variablen zugewiesen werden,
um sie für die Dauer der R-Sitzung zu speichern.

```
(5 + 3) * 17 / 2 ** 2
x = (5 + 3) * 17 / 2 ** 2
x
x + 6
?sin # Hilfe zur sin-Funktion
sin(tan(x))
```

Wissenschaftliche Notation von Zahlen

Große und kleine Zahlen werden oft in der Exponentialnotation angegeben.

- $107 = 1.07 * 10^{2} = 1.07E+02$
- $0.334 = 3.34 * 10^{-1} = 3.34E-01$

Das E steht für Exponent (zur Basis 10), es sollte ein großes E sein;
leider sieht man bisweilen auch ein kleines e.

Dieses E oder e hat nichts mit der Zahl e (etwa 2.71) zu tun!

Logik in R

Logik in R ist 3-wertig:
- wahr (T, TRUE)
- falsch (F, FALSE)
- nicht entscheidbar / fehlende Daten (NA)

```
x = 35                  # Wert von x setzen
x == 34                 # Test auf Gleichheit zu 34: FALSE
x > 40                  # FALSE
(x > 30) & (x < 40)     # TRUE, Ver-und-ung zweier Tests
(0 / 0) > 2             # NA, da 0/0 nicht definiert
```

Aus Berechnungen, in die NA hineingesteckt wird, kommt wieder NA heraus.
Vergleich mit NA: nicht `==`, sondern Funktion `is.na()` verwenden!

```
y = ((0 / 0) > 2)    # y ist jetzt NA
y == NA              # liefert nicht TRUE, sondern wieder NA
is.na(y)             # liefert TRUE
```

NA und NaN

NA

logischer Wert neben TRUE und FALSE
steht für „not available“
benutzt für nicht entscheidbare Tests und fehlende Daten

NaN

„numerischer“ Wert, der aber keine Zahl repräsentiert
steht für „not a number“
benutzt für nicht definierte Ergebnisse von Rechnungen, wie 0/0

Zusammenhang

Wenn man NaN auf etwas testet, ist das Ergebnis NA.
(Nicht entscheidbar, ob NaN beispielsweise > 17 ist.)

Vektoren

Häufig soll dieselbe Operation auf mehrere Werte angewendet werden.
Dies kann man erreichen, wenn die Daten in einem Vektor gespeichert sind.
Erzeugen eines Vektors mit c() durch Aufzählung oder mit „:“ (von-bis).

```
y = c(2,3,5,7,11,NA,13)
z = 1:10

y * 2
z + 3
```

Man kann zwei Vektoren element-weise miteinander verknüpfen.
Dabei wird der kürzere solange wiederholt, bis die Länge des längeren erreicht ist.
Man erhält eine Warnung, wenn die Längen keine Vielfachen voneinander sind.

```
y + z
```

Vektoren mit seq()

Regelmäßig strukturierte Vektoren erzeugt man mit der seq()-Funkton durch Angabe von Startwert, Endwert, Schrittweite („by“):

```
x = seq(0, 10, by=0.5) # erzeugt 0, 0.5, 1, 1.5, ..., 9.5, 10
x = seq(0, 10, 0.5)    # dasselbe

y = sin(x)             # Elementweise Sinus-Funktion anwenden
```

Einfache Plots

Die graphischen Fähigkeiten von R sind sehr mächtig.
Beispielsweise kann man leicht einfache Funktionsplots erstellen,
z.B. von $y = \sin(x^2)$ auf dem Intervall [-5, 5]:

```
x = seq(-5, 5, by=1/16) # kurze Schrittweite für schönen plot
y = sin(x**2)
plot(x,y)
```

Der resultierende Punktplot sieht nicht sehr übersichtlich aus.
Besser, wir hätten eine durchgehende Linie statt einzelner Punkte.

```
plot(x,y, type="l")
plot(x,y, type="o")  # Punkte und Linien übereinander
```

Oder in Blau mit schönem Titel:

```
plot(x,y, type="o", col="blue", main="Parabel")
```

Noch einfachere Plots

Eben sind wir so vorgegangen:
- Erst Definition von Stützstellen auf der x-Achse (Vektor `x`)
- Berechnung einer beliebigen Funktion an diesen Stellen (Vektor `y`)
- Plot von `y` gegen `x` mit `plot(x,y)`.

Ist die Funktion „eingebaut“ (wie sin, cos, exp, log, ...),
hat also einen Namen, geht es noch einfacher:
`plot(sin, -5, 5)`
plottet die Funktion sin zwischen -5 und 5.

Die andere Methode ist aber universeller einsetzbar.