

**Einführung in die Angewandte Bioinformatik:
Datenbanken, Publizieren, ISI, PubMed
30.04.2009**

Prof. Dr. Sven Rahmann

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

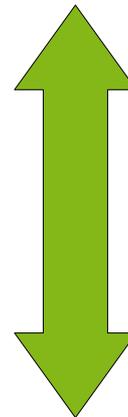
Wissenschaftliches Publizieren und Zitations-Datenbanken

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

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

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

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://scientific.thomson.com/free/essays/journalcitationreports/impactfactor/>
Literaturliste jeder(!) Veröffentlichung durchgehen,
in Bezug zu vergangenen Veröffentlichungen setzen
Manuelle Nachbearbeitung; teuer!

Datenbank dieser Verknüpfungen:

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

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!]

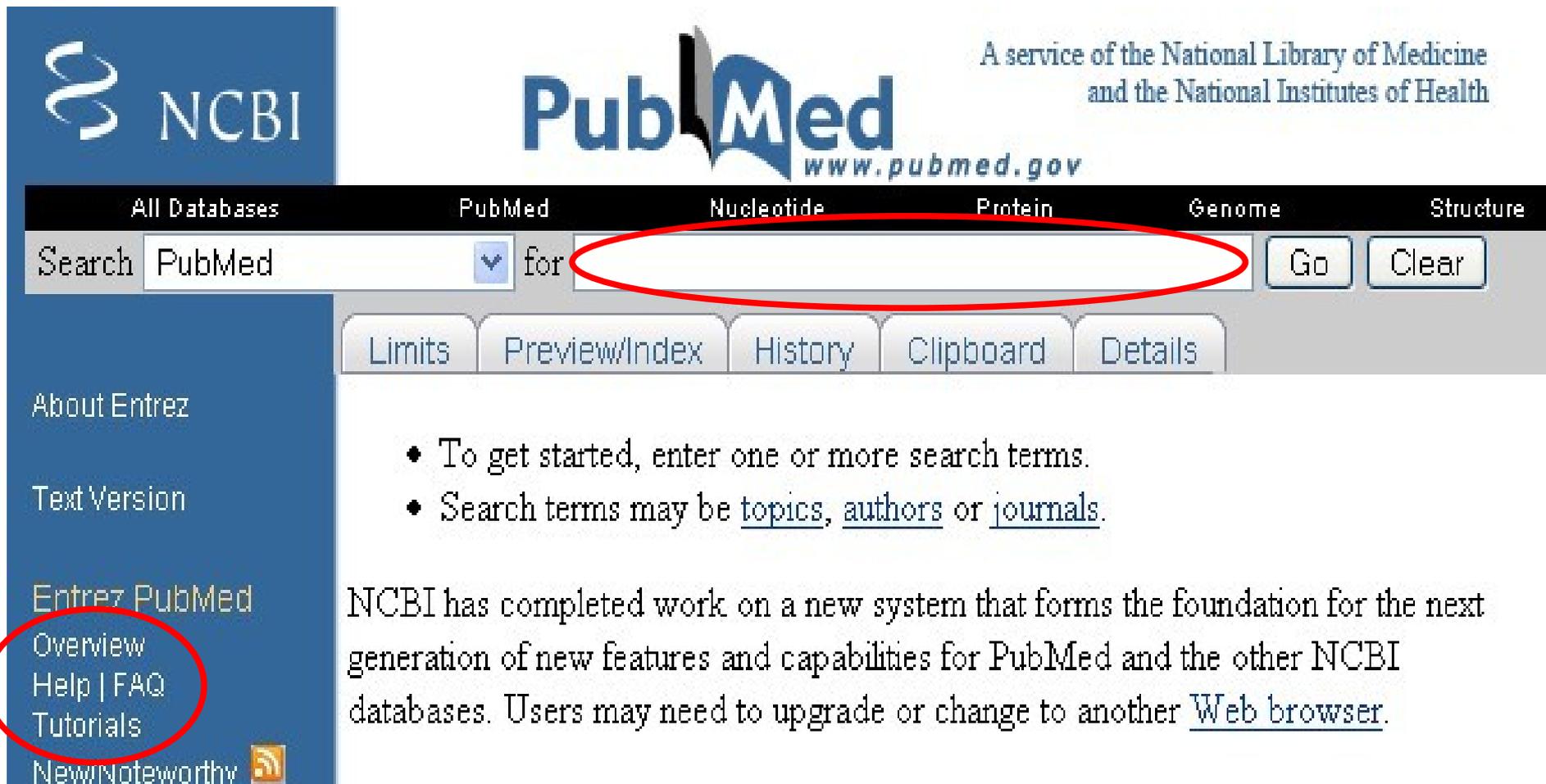
PubMedCentral :=

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 MEDLINE mit PubMed über Entrez unter
<http://www.ncbi.nlm.nih.gov/sites/entrez/>
(Was sind NCBI, NLM, NIH?)



NCBI

PubMed www.pubmed.gov

A service of the National Library of Medicine
and the National Institutes of Health

All Databases PubMed Nucleotide Protein Genome Structure

Search PubMed for

Limits Preview/Index History Clipboard Details

- To get started, enter one or more search terms.
- Search terms may be topics, authors or journals.

NCBI has completed work on a new system that forms the foundation for the next generation of new features and capabilities for PubMed and the other NCBI databases. Users may need to upgrade or change to another Web browser.

About Entrez
Text Version
Entrez PubMed
Overview
Help | FAQ
Tutorials
New/Noteworthy

Einfache PubMed-Suche

<http://www.ncbi.nlm.nih.gov/sites/entrez/>

Stichwörter
in das „for“-Feld
eingeben, z.B.
Autorennamen,
Titel-Teile

The screenshot shows the PubMed search interface. At the top, there are navigation tabs for PubMed, Nucleotide, Protein, Genome, Structure, OMIM, PMC, Journals, and Books. The search bar contains the text "for rahmann microarray" and is highlighted with a red oval. To the right of the search bar are buttons for "Go", "Clear", and "Save Search". Below the search bar are tabs for "Limits", "Preview/Index", "History", "Clipboard", and "Details". The "Display" dropdown is set to "Summary", "Show" is set to "20", and "Sort by" is set to "Relevance". The search results show 4 items. The first item is "Baumbach J, Brinkrolf K, Czaja LF, Rahmann S, Tauch A." with a "Related Articles, L" link. The second item is "CoryneRegNet: An ontology-based data warehouse of corynebacterial transcription factors and regulatory networks." with a "BMC Genomics. 2006 Feb 14;7(1):24 [Epub ahead of print]" citation and "PMID: 16478536 [PubMed - as supplied by publisher]". The third item is "Schliep A, Torney DC, Rahmann S." with a "Group testing with DNA chips: generating designs and decoding experiments." citation and "PMID: 16452782 [PubMed - indexed for MEDLINE]". The fourth item is "Rahmann S." with a "Rapid large-scale oligonucleotide selection for microarrays." citation and "PMID: 15838123 [PubMed - indexed for MEDLINE]". The fifth item is "Rahmann S." with a "Fast large scale oligonucleotide selection using the longest common factor approach." citation and "PMID: 15290776 [PubMed - indexed for MEDLINE]".

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

Zugriff auf weiterführende Informationen

Display-Auswahlbox: „Abstract Plus“; sowie „Related Links“ rechts

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals

Search PubMed for rahmann "microarray design" Go Clear Save Search

Limits Preview/Index History Clipboard Details

Display AbstractPlus Show 20 Sort by Send to

All: 1 Review: 0

1: 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

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

Related Links

- ▶ Decoding non-unique oligonucleotide hybridization experiments of targets related by a [Bioinformat
- ▶ Fast and sensitive probe selection for DNA chip jumps in match [Proc IEEE Comput Soc Bioinform Co
- ▶ Selecting signature oligonucleotides to identify organisms using DNA arrays. [Bioinformat
- ▶ Fast and accurate probe selection algorithm for genomes. [Proc IEEE Comput Soc Bioinform Co
- ▶ Comprehensive aligned sequence construction automated design of effective probe [Bioinformat
- ▶ See all Related Articles...

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

Search PubMed for [] Go Clear

Limits Preview/Index History Clipboard Details

Limit your search by any of the following criteria.

Search by Author Add Author CLEAR

Search by Journal Add Journal CLEAR

Full Text, Free Full Text, and Abstracts CLEAR

Links to full text Links to free full text Abstracts

Dates CLEAR

Published in the Last: Any date

Added to PubMed in the Last: Any date

Humans or Animals CLEAR

Humans Animals

Gender CLEAR

Male Female

Languages CLEAR

English French

Subsets CLEAR

Journal Groups

About Entrez
Text Version
Entrez PubMed
Overview
Help | FAQ
Tutorials
New/Noteworthy
E-Utilities
PubMed Services
Journals Database
MeSH Database
Single Citation
Matcher
Batch Citation
Matcher
Clinical Queries
Special Queries
LinkOut
My NCBI
Related Resources
Order Documents
NLM Mobile
NLM Catalog
NLM Gateway
TOXNET

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]

The screenshot shows the Entrez PubMed search interface. The search bar contains the text "Search MeSH" and "for microarray". The search results are displayed in a list format. The first three results are:

- 1: [Microarray Analysis](#)
The simultaneous analysis, on a microchip, of multiple samples or targets arranged in an array.
Year introduced: 2005
- 2: [Protein Array Analysis](#)
Ligand-binding assays that measure protein-protein, protein-small molecule, or protein-nucleic acid interactions by capturing molecules, i.e., those attached separately on a solid support, to measure the presence of a target in a sample.
Year introduced: 2003
- 3: [Oligonucleotide Array Sequence Analysis](#)
Hybridization of a nucleic acid sample to a very large set of oligonucleotide probes, which are known or unknown in sequence or to detect variations in a gene sequence or expression or for gene mapping.
Year introduced: 1999

The search bar and the third result are circled in red in the original image.