Literature Databases

Introduction to Bioinformatics
Dortmund, 16.-20.07.2007

Lectures:
Sven Rahmann

Exercises:
Udo Feldkamp, Michael Wurst
Overview

1. Databases
2. Publications in Science
3. PubMed and Entrez (literature DB)
4. ISI Web of Science (citation DB)
Data and Information

• Data :=
  everything that can be collected, stored, and read again.
  Data, by itself, is uninterpreted.

• Information, Knowledge :=
  facts derived from data, interpretation of data
Database

- DB := structured collection of data
- customer DBs
- address DBs
- movie DBs
- literature DBs (PubMed)
- sequence DBs for DNA, RNA, proteins
- protein structure DBs (PDB)
- pathway DBs
DBMS – Database Management System

- DBMS := piece of software to access, modify, delete, add data in a database, may have a graphical user interface (GUI)

- MS Access
- OpenOffice.org Base
- mySQL
- Entrez at NCBI (“meta access”)
Database System

- Database(s) (DB(s))
  + Database Management System (DBMS)
  = Database System (DBS)
Database vs. Flat File

- Even a simple text file can be a “database”
  - tab- or newline-delimited lists of numbers or words

- Flat file
  - requires no special software to access or modify (just a plain text editor)
  - access (searching) is often inefficient
  - integrity and security are difficult to maintain

- Database
  - requires a DBMS
  - allows indexing for efficient complex searches
  - DBMS maintains integrity and security
Publications

• New research results are communicated to the scientific community via
  – conference proceedings ("conference papers")
  – research articles in journals ("journal papers")
  – review articles in journals
  – chapters in edited books
  – student theses
  – textbooks
Who's paying?

- Publications are good for the whole community, even the whole world in the long run
- (Patents are good for one company)
- Who's paying for publications? Two models:
  - classic: the reader (journal subscription)
  - new: the author (open access model) sometimes via institutional flat rate
- Who's making money?
  - the publisher (not the scientist or funding agency)
Quality in Publications

- Publishing research results means progress
- But what if something wrong gets published?
- Quality assurance: peer review
- Other scientists reads an comment on articles before they are published (anonymously)
- Important part of the job
- Not fail-safe, but no one has a better idea
Not all papers are equal

- Write-only papers
  - no one ever reads them
  - waste of resources and money
  - they still add to the publication list of the author
- Small advances
  - frequent, low-risk
- Big advances
- Papers opening a new research field
  - quite rare, high risk
Literature Search

- Important before starting a new project
- Avoids duplicating existing work
- If articles cannot be found (efficiently), they become write-only papers
- We need DBs of research articles
MEDLINE, PubMed, Entrez

- **MEDLINE** := database of biomedical journal citations and abstracts, ca. 5,000 journals; is the largest component of
- **PubMed** := freely accessible online system containing MEDLINE, allows advanced queries
- **PubMedCentral** := free digital archive of biomedical and life sciences journal literature.
- **Entrez** := links PubMed to other databases; common interface
Who's behind it?

- National Institute of Health (NIH)
  - founded 1887, located in Bethesda, Maryland
  - focal point for medical research in the U.S.
- U.S. National Library of Medicine (NLM)
  - world's largest biomedical library
- National Center for Biotechnology Information (NCBI)
  - creates public databases
  - conducts research in computational biology
  - develops software tools for analyzing genome data
  - disseminates biomedical information
NCBI Website

Searching PubMed


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.
Simple Search

• Simply enter key words into the “for” field
• Default:
  – key words are connected by AND,
  – refer to author OR title OR abstract ..., 
  – case insensitive;
  – quotes fix the order of terms
• Examples:
  – Rahmann Microarray
  – Rahmann “Microarray Design”
• Use the 'Help' button!
1. Baumbach J, Brinkrolf K, Czaja LF, Rahmann S, Tauch A.
   PMID: 16478536 [PubMed - as supplied by publisher]

2. Schliep A, Torney DC, Rahmann S.
   Group testing with DNA chips: generating designs and decoding experiments.
   PMID: 16452782 [PubMed - indexed for MEDLINE]

3. Rahmann S.
   Rapid large-scale oligonucleotide selection for microarrays.
   PMID: 15838123 [PubMed - indexed for MEDLINE]

4. Rahmann S.
   Fast large scale oligonucleotide selection using the longest common factor approach.
   PMID: 15290776 [PubMed - indexed for MEDLINE]
Search Results

• Note: If there is exactly one hit, you usually see immediately the abstract.
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, Inestrass 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 challenging data sets and experimental error rates of up to 5%. On a data set of 285 rDNA sequences we were able to identify 660 sequences, a substantial improvement over a prior approach using unique probes which only identified 408 sequences.
Structure of MEDLINE Entries

Field name: PMID (unique PubMed ID)
Value: a unique number identifying the article

PMID- 16452782
OWN - NLM
STAT- MEDLINE
DA - 20060202
DCOM- 20060808
PUBM- Print
IS - 1555-3930 (Print)
VI - 2
DP - 2003
TI - Group testing with DNA chips: generating designs and decoding experiments.
PG - 84-91
AB - 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 283 rDNA sequences we were able to identify 660 sequences, a substantial improvement over a prior approach using unique probes which only identified 408 sequences.

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

• Down Syndrome:
  – finds, e.g., articles by Dr Down on any syndrome
• “Down Syndrome”
  – using quotes is much better
• Down [ti]
  – restricts to articles with 'Down' in the title
• Down [au]
  – articles whose author is Dr Down
• [pmid]: unique PubMed identifier
  – always remember it for any interesting article
Fields

- PMID: PubMed ID
- TI: Title
- AB: Abstract
- AD: Address
- FAU: Full author name
- AU: author (Last name + Initials)
- SO: Journal abbreviation
- ...

...
Boolean Operations

• AND (default)
• OR
• NOT
• Example
  – Find microarray experts in Dortmund
    microarray [ti] AND Dortmund [ad]
  – Find (many) microarray papers not by Rahmann
    microarray [ab] NOT Rahmann S [au]
Using the “Limits” Tab

- Convenient alternative to specify restrictions
A Few Hints

- Use quotes
- Add initials to author names
- Write down the PMID
- If you find too many results, restrict your search
- If you find too few (or no) results, broaden your search.
- (Use synonyms)

Problem: The same “idea” or “concept” is often represented by several different terms.
MeSH Terms

- MeSH = Medical Subject Headings
- MeSH is a controlled vocabulary used for indexing articles for MEDLINE/PubMed.
- MeSH terminology provides a consistent way to retrieve information that may use different terminology for the same concepts.
Example: Microarray

1: Microarray Analysis
The simultaneous analysis, on a microchip, of multiple samples or targets arranged in an array format.
Year introduced: 2005

2: Protein Array Analysis
Ligand-binding assays that measure protein-protein, protein-small molecule, or protein-nucleic acid interactions using a very large set of capturing molecules, i.e., those attached separately on a solid support, to measure the presence or interaction of target molecules in the 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 attached to a solid support, to determine sequence or to detect variations in a gene sequence or expression or for gene mapping.
Year introduced: 1999
MeSH Terms in a PubMed Search

- Directly enter a [MH] field
- Or “build” the query interactively using the GUI

![PubMed Search Interface]

- Search: PubMed
- Query: "oligonucleotide array sequence analysis" [mh]
- Results: 5 papers
  1. Schliep A, Rahmann S.
  - Decoding non-unique oligonucleotide hybridization experiments of targets related by a phylogenetic tree.
  - PMID: 16873503 [PubMed - indexed for MEDLINE]
ISI Web of Science

- Commercial (Thomson Scientific) citation database
  http://isiknowledge.com/
- Access to journal articles in the sciences, social sciences and arts and humanities.
- Web of Science also provides a unique search method, cited reference searching.
- Access to the Science Citation Index® (1900-present)
Explanations

- New research builds on past research.
- Newer papers cite older papers they refer to.
- Interesting question: How influential is a paper (how often cited)?
- [The papers of] which journals are influential? => “Impact factor”
- Science Citation Index provides this information
- A fair basis for evaluating researchers?