Data Analysis in the Life Sciences
- The Fog of Data -

Michael R. Berthold
ALTANA-Chair for Bioinformatics & Information Mining
Konstanz University, Germany
Michael.Berthold@uni-konstanz.de

[ some (the cool) pictures thanks to Thomas Exner ]

Overview

• Motivation / Background
  – Data Analysis in the Life Sciences: Understanding Biology
  – Main Focus of Today: Drug Discovery

• Visual Clustering of Screening Data
  – Different Notions of Similarities
  – Visual Clustering Using Neighborgrams

• Finding Interesting Molecular Fragments
  – Mining Graphs
  – The Space of Discriminative Molecular Fragments

• Conclusions
Towards Understanding Biology

Ultimate Goal:

- Complete Understanding (with Model!) of Interactions within human body
  - Gene and Protein functions
  - Protein-protein and -cell interactions
  - Metabolic pathways
- Clear model to understand (and avoid) side effects
- Design of new drugs “in silico” possible
- Personalized drugs for specific geno- and phenotype.

But until then…

Discovering Drugs

**Desired Effect**

First define a worth($)-while disease to tackle
- Often driven by existing ideas
- Nowadays also life style drugs
Discovering Drugs

Desired Effect → Target Identification

Find points to attack (usually proteins) by analyzing
- Gene expression changes
- Changes in proteins
- Information about biological connections (metabolic pathways)
- Biological intuition and expert knowledge

Target Identification → Target Validation

Validate that identified targets have desired effect
- Knockout specific genes, eliminate proteins
- analyze e.g. gene expression for desired effect
- doesn’t work? Try to find new target(s)
Discovering Drugs

- **Desired Effect**
- **Target Identification**
- **Target Validation**
- **Assay Design**

Design an experiment that allows to quickly check if desired action happens
- Needs to allow quick and easy testing of “activity”

---

Discovering Drugs

- **Desired Effect**
- **Target Identification**
- **Target Validation**
- **Assay Design**
- **High Throughput Screening**

Test of hundreds of thousands of candidate molecules from a – usually carefully designed (targeted library) – set of candidates
- Data is generally of rather bad quality
  (many false positives, unknown false negatives)

**Data Exploration**
Discovering Drugs

Desired Effect → Target Identification → Target Validation → Assay Design → High Throughput Screening → Lead Finding

From HTS results, try to identify “lead” structures, i.e. interesting pieces of molecules that may be responsible for the observed effects.

Structured Data Mining

Discovering Drugs

Desired Effect → Target Identification → Target Validation → Assay Design → High Throughput Screening → Lead Finding

Lead Optimization

Optimize active molecules:
- Increase activity
- Eliminate undesired side effects (toxicity!)
- didn’t work? Try to find new lead or new target
Discovering Drugs

1. Desired Effect
2. Target Identification
3. Target Validation
4. Assay Design
5. High Throughput Screening
6. Lead Finding
7. Lead Optimization
8. Animal Testing
9. 1st phase Clinical trials
10. 2nd phase Clinical trials

Check if it really works and has no nasty side effects.
- didn’t work?...

~10-12 years later: $$$$

New Drug
Challenges

- Analysis of truly heterogeneous data
  - Classical Data Analysis: production plants, shopping baskets, … ⇒ few, reasonably self-contained data sources
  - Life Sciences:
    - truly diverse data- and information sources
    - complex data types (molecular structures, sequences,…)
    - many different notions of similarity
- Data of rather different quality
  - noisy, unreliable, faulty, outdated,…
- Complicated Domain
  - Usually: Goal of Analysis also understandable by Non Specialist
  - Life Sciences: Without continuous incorporation of expert feedback (biologist, biochemist) no chance for success.
    - domain knowledge not explicitly known
    - question often unclear / unknown at beginning
    - need for true data exploration cycle incorporating the expert

Overview

- Motivation / Background
  - Data Analysis in the Life Sciences: Understanding Biology
  - Main Focus of Today: Drug Discovery
- Visual Clustering of Screening Data
  - Different Notions of Similarities
  - Visual Clustering Using Neighborgrams
- Finding Interesting Molecular Fragments
  - Mining Graphs
  - The Space of Discriminative Molecular Fragments
- Conclusions
The Context of Today: High Throughput Screening

Rapidly screen 100-thousand’s of candidates.

• Problems
  – Often thousands of actives
  – Data extremely noisy
  (up to 50% false positives, unknown false negatives!)
  – Positives almost always active for different reasons
    ⇒ Separate, diverse clusters!

⇒ Goal:
  ⇒ Find subsets of similar active molecules
    (help user find clusters of actives)

What is “Similarity”?

• Example: Find similar cars:
  – red Ferrari
  – blue Porsche
  – red Bobby car

• First abstract representation (descriptor) = Speed
  – Group 1: Ferrari & Porsche

• Second abstract representation (descriptor) = Color
  – Group 1: Ferrari & Bobby car

• …and many other notions of similarity can be used

• For molecular structures this is even more complicated!
What is “Similarity”? 

Types of Molecular Similarity

- **Structural similarity:**
  - Same basic layout of overall graph
  - …or at least existence of a common subgraph

- **Geometrical similarity:**
  - Roughly same shape in 3D, independent of exact atom matches
  - Instead of simple shape, also other properties (surface charge…) can be compared

- **Global properties:**
  - Molecular weight
  - Number of hydrogen donors/acceptors…

- **And (mainly used) some of many heuristic approaches:**
  - Fingerprint based
  - 3D grid comparisons (topomers…)

Similarity to Adrenaline: Other Neurotransmitters

Adrenalin

(B)

(D)

N
CH₃

NH₂

CH₃

N

CH₃

OH

CH₃

OH

HO

OH

HO

Similarity: Polarity

#19

#20
Dissimilarity: Hydrophobic / Hydrophilic

Dissimilarity to Adrenaline: Crossing the Blood-Brain Barrier
Dissimilarity to Adrenaline: Crossing the Blood-Brain Barrier

Different Similarities

- Comparing Molecular Structures depends on context:
  - different properties
    - molecules (drugs, proteins, ...) often have more than one function
  - different modes of action
    - the same effect can be caused in various ways
Different Modes of Activity: Example

- Different ways to interact, e.g.
  - drug candidate can block access to receptor site
  - protein can lock into receptor site
  - ...
- Example: Inhibitors of the Trypanothione Reductase

Different Similarities

- Comparing Molecular Structures depends on context:
  - different properties
    - molecules (drugs, proteins, …) often have more than one function
  - different modes of action
    - the same effect can be caused in various ways
- Problems:
  - not a priori known which property / properties matter
    - biology poorly understood
  - not always only one relevant property
    - analysis needs to find patterns in various spaces in parallel…
Motivation: Parallel Universes

• Usually: Data given in a single feature space
  – Mostly high-dimensional and numeric
  – Definition of one, global similarity measure

• Here: Multiple representations for the data:
  Parallel Universes
  – Different similarity measures
  – Only subsets of data will form useful model in each universe
  – Standard learning techniques not applicable

Similarities to other Similarity-Based Methods

• Feature Selection
  – Global selection (entire data set)
  – Global optimization goal (e.g. performance)

• Subgroup Mining
  – Local selection (subset of data)
  – Local groupings of features without semantic

• Feature/Cluster Weighting (e.g. Friedman et al)
  – Individual feature weights for each cluster
  – Interpretation difficult.

• Multi-Instance Learning
  – multiple descriptions per object
  – One similarity function – one or more descriptions must “match”

• Multi View Mining
  – multiple descriptors per object
  – independent hypotheses built on each view
  – ensemble for prediction

• Parallel Universes:
  – Several descriptions / distance metrics exist for objects
  – Parts of Model operate on one (or some) universes only.
Parallel Universes: HTS Example

- Molecular Data Analysis: Descriptors based on
  - Fingerprints
  - Numerical Features derived from 3D shapes, surface charge distribution, etc.
  - Comparisons of chemical graphs
- Different descriptors encode different structural information
- None of them show satisfactory prediction results alone

- One approach: interactive clustering algorithm using Neighborgrams

Parallel Universes – Interactive Clustering

Parallel Universes

- Mining in „parallel universes“ challenging:
  - Helps finding the right description for only a subset of the data
    (David Hand: *find patterns describing anomalous local features…*)
  - Overlap between universes possible and desirable
  - Resulting models may only describe parts of data in each universe
  - User feedback to guide analysis even more important

- Outlook
  - Promising concept to mine heterogeneous data with diverse notions of similarity
    (works well in bioinformatics)
  - Integrates focus of analysis, allows for context switch
    (via interactive Expert query refinement)
  - Presents additional context information about model (pieces).

- Work in Progress:
  - Fuzzy CU-Means: unsupervised clustering in parallel universes)
Overview

• Motivation / Background
  – Data Analysis in the Life Sciences: Understanding Biology
  – Main Focus of Today: Drug Discovery

• Visual Clustering of Screening Data
  – Different Notions of Similarities
  – Visual Clustering Using Neighborgrams

• Finding Interesting Molecular Fragments
  – Mining Graphs
  – The Space of Discriminative Molecular Fragments

• Conclusions

Remember?: Types of Molecular Similarity

• Structural similarity:
  – Same basic layout of overall graph
  – …or at least existence of a common subgraph

• Geometrical similarity:
  – Roughly same shape in 3D, independent of exact atom matches
  – Instead of simple shape, also other properties (surface charge…) can be compared

• Global properties:
  – Molecular weight
  – Number of hydrogen donors/acceptors…

• And (mainly used) some of many heuristic approaches:
  – Fingerprint based
  – 3D grid comparisons (topomers…)

see also: A. Bender, R. Glen: Molecular similarity: a key technique in molecular informatics, Org. Biomol. Chem., 2:3204-3218, 2004
Motivation for Molecular Fragment Mining

- Goal:
  Find (and describe!) structural groups of molecules that share activity.
- For few molecules, manual inspection is feasible.

Motivation

- Goal:
  Find (and describe!) structural groups of molecules that share activity.
- For few molecules, manual inspection is feasible.
- For more molecules, automated methods are needed…
Motivation

Molecular Fragment Miner (MoFa)

Goal:
- Find Fragments that are discriminative for a class of interest (high activity, good synthesis result, …):
  - Appear often in Positives: freq(high activity)>threshold
  - Appear rarely in Negatives: freq(low activity)<delta

MoFa (and many variations: FSG, gSpan, …):
- Based on Market Basket Analysis (Apriori or Eclat Algorithm)
- Grow Fragment-Candidates from scratch atom-by-atom
- Only report significant and unique fragments

Finding Frequent Molecular Fragments

- More complicated than itemset mining
  - No clear order on items
    - Eliminating duplicates is a problem
      (same fragment can be along different paths)
  - Items can appear more than once (H-O-H)
  - Graph topology crucial – not just co-occurrence in one molecule

- Different types of heuristics exist:
  - MoFa:
    - Uses local order on atom/bond extensions
    - Restricts extensions to recently added atoms
  - MoSS:
    - Uses search tree traversal to generate canonical form
Support Based Pruning

- Support of fragment A:
  \[
  \text{supp}(A) = \text{Frequency of appearance in molecules}
  \]

- Monotone conditions decline with size of fragment:
  
  fragment A is contained in fragment B  
  \[\Rightarrow \text{supp}(A) \geq \text{supp}(B)\]

- If supp(node) in branch is below threshold then all child-nodes will also be below threshold.
Discriminative Fragments

- Just finding frequent fragments usually not interesting
- Find fragments that are
  - frequent in one class of molecules
  - and infrequent in the remainder of molecules
- Discriminative Fragments summarize shared properties.
- Number of actives and inactives (and the ratio) that contain fragment indicates relevance.

Example:
[NCI HIV dataset ~45000 (~400 active) compounds, threshold=15%]

15.08% vs. 0.02%
A few more fragments...

- 5.23% vs. 0.05%
- 4.92 vs. 0.07%
- 9.85% vs. 0.07%
- 10.15% vs. 0.04%

Fragment lattices may still be big...
Example 2:
NCI Cancer Screen (~35000 mol. with several active groups)

Colchicines (antimitotic agents)
Example 2:
NCI Cancer Screen (~35000 mol. with several active groups)

Anthracyclines (i.e., Adriamycin) induce apoptosis in tumor cells.
Example 2:

NCI Cancer Screen (~35000 mol. with several active groups)

Podophyllotoxin derivatives inhibit Tubulin polymerization.
Example 2:
NCI Cancer Screen (~35000 mol. with several active groups)

Camptothecins are Topoisomerase I inhibitors.

Acridines are cytotoxic to tumor cells.
Overview

- Motivation / Background
  - Data Analysis in the Life Sciences: Understanding Biology
  - Main Focus of Today: Drug Discovery

- Visual Clustering of Screening Data
  - Different Notions of Similarities
  - Visual Clustering Using Neighborgrams

- Finding Interesting Molecular Fragments
  - Mining Graphs
  - The Space of Discriminative Molecular Fragments

Conclusions

- Life Science offers lots of challenging data
- Focus is not exclusively on building good predictors
  - Discovered knowledge needs to be meaningful
  - Users want understandable pieces of knowledge ("Information Mining").
  - Integration of expert is crucial: Allow interaction / exploration
  - Two examples discussed today
    - Explorativ tool for potential clusters allows to incorporate expert feedback easily
    - Fragments understandable to chemist
      (Better than e.g. rules/decision trees on mystic attributes)
...Conclusions

- Results only describe one part of the overall picture
  - Successful tools present many different views on the data
- Goal is not outstanding prediction but
  - confirmation of hypotheses,
  - triggering of new ideas, and
  - ...quite often simply: cleanup of data
- No clearly defined target of analysis
  - no real questions asked
  - “find something interesting!” – which is a moving target