Knowledge-Oriented Analysis of Mycroarray Data

Avoiding Paralysis of Analysis: Building an Intellectual Prosthesis

I. Jurisica
Goals

- Parallel analysis of gene expressions
  - Improved understanding of tumorigenesis
  - Tumor classification
- Individualized medicine
  - Improved diagnosis, prognostics, treatment planning & adjustment
  - Targetted therapy & drug design/use
  - Informed patient
Problems

- Multi-dimensionality
  - many degrees of freedom, few datapoints
- Noise
  - Imprecision, variation
  - Low number of repeats
- Non-independence
- Non-linearity
- DBs change
- Integration of results with other DBs & multiple experiments
Intellectual Prosthesis

- Finding appropriate model to support reasoning

- Exceptions
- Evolution

- More Knowledge
- More Data

- Fixed
- Parametric
- Nonparametric with Processing
- Nonparametric
Analysis

- **Clustering** organizes observations into groups by max. iner-cluster and min. inter-cluster similarity
- **Classification/prediction** assigns an observation to a class (finite/infinite)
- **Comparison** describes the item by comparing it to other items
- **Summarization** describes common characteristics of a subset
- **Discrimination** describes minimum features needed to differentiate among classes
- **Association** finds common occurrence of observations
Paralysis

- Source
  - too slow to search the problem space
  - not enough data/processing time available for a system to generate a NP model
  - lack of domain knowledge
  - too much data (including noise) from HTP (high dimensionality)

- A solution
  - HTP & computation
  - Generate - analyze - reduce - test - validate
HTP

- Modified CBR approach
  - symbolic similarity
  - lazy learning combined with
    - clustering & classification
    - summarization
- Analysis-based research
  - DNA microarray analysis
  - annotation
Model-Building Solutions

- Eager approach
  1. analyze data
  2. create a model
  3. use the model

- Lazy approach - data-driven model
  1. incrementally accumulate data
  2. incrementally analyze & evolve

- Generate - analyze - reduce - test - validate
Analyzing and Using MA Data

- Problems
  - Knowledge of classes
  - Providing parameters
  - Clinical attributes as measures of "meaningfulness"
  - Scalability
  - Annotating and explaining results
  - Quality assurance
  - Integratability
Discovery Algorithms

http://cmgm.stanford.edu/pbrown/
Case-Based Reasoning

1. Diagnosis
2. Prognosis
3. Treatment plan

General Demographics & Medical History
Clinical Presentation & Prognostic Factors
Surgical Details
Pathology Staging
Clinical Staging
Research Protocol
Follow-up
Age
Dates
Hematology
Biochemistry
19.2k expression profiles, ....

DIMACS'01
I. Jurisica
Case-Based Reasoning

- DSS
  - Cases represent experiential knowledge
  - Cases are patterns: context, problem, solution
  - Symbolic similarity - context-based
  - Retrieval - k-NN with context and structure
  - Anytime algorithm
- KM for evolving domains
  - Documenting, analyzing, transferring & sharing experience
  - Classification, prediction, guidance in hypothesis discovery
  - Clustering, summarization
  - Acquire now, process later
Patient Information Management

- we need detailed disease classification
- we need markers to improve diagnosis, prognosis and treatment planning
- we need new and systematic methods
CBR for DNA Micro Arrays

- Gene expression signature
- Find patients with similar signature
  - \(k\)-NN approach - without prior domain knowledge
- Provide diagnosis, prognosis & treatment by analogy
- Apply \textit{Explain} function for marker & cancer subtype summarization
Advantage of CBR

- Supports reasoning, not just analysis
- Measure of similarity is based on gene expression profile
- Does not require prior knowledge
- Supports evolution & is more flexible
- Handles inconsistencies
  - Inconsistencies get resolved at run-time with contextual information
  - CBR can be used to find inconsistencies
- Supports discovery & validation
Outliers

- Represent change and deviation
  - data outside of normal region of input
    - unusual but correct
    - unusual & incorrect
  - for numeric attributes
    - detect with histogram
      - remove with threshold filter
    - identify by calculating the mean & stdev
      - remove by specifying "window", e.g., 2 standard deviations from the mean
**KD and CBR**

- Organize genes into groups
- Organize attribute values into taxonomies
Context Relaxation
Patient-Patient Similarity
Open Source BIOdb

- Automated annotation
- Schema integration, info validation
- Querying and analysis
- Reasons for local source:
  - certain tasks are more efficient and effective
  - certain tasks become possible
WebOQL

- A system for supporting data restructuring operations
  - to integrate data from different sources (documents, relational tables, hypertexts)
  - to restructure an instance of a given source into an instance of another one
- We used WebOQL to write wrappers for UniGene
  - more generic, dynamic, incremental

http://www.cs.toronto.edu/~weboql
Autoannotations

- Information may not be downloadable
- Information may not be complete

ID=1
TITLE=Hippocampus, Stratagene (cat. __936205)
TISSUE=brain, hippocampus
VECTOR=lambdaZAP-II

Lib.1
Infant, 2 yrs, female
brain, hippocampus
lambdaZAP-II
453 ESTs have been classified, 411 gene sets
Expression Distribution

Thousands

One

Distinct

0

50

100

150

200

250

300

DIMACS'01 I. Jurisica
## Lung

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung</td>
<td>15,410</td>
</tr>
<tr>
<td>Lung-tumor</td>
<td>67</td>
</tr>
<tr>
<td>Lung-tumor &amp; suppressor</td>
<td>26</td>
</tr>
<tr>
<td>Lung-tumor &amp; necrosis</td>
<td>20</td>
</tr>
<tr>
<td>Lung-tumor &amp; antigen</td>
<td>5</td>
</tr>
<tr>
<td>Lung-tumor &amp; susceptibility</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accession</th>
<th>Species</th>
<th>Accession</th>
<th>Description</th>
<th>Score</th>
<th>Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hs.241493</td>
<td>M. musculus</td>
<td>PIR:B47328</td>
<td>B47328 natural killer cell tumor-recognition protein - mouse&quot;</td>
<td>1511</td>
<td>79 %</td>
</tr>
<tr>
<td>Hs.241493</td>
<td>H. sapiens</td>
<td>SP:P30414</td>
<td>NKCR_HUMAN NK-TUMOR RECOGNITION PROTEIN&quot;</td>
<td>1461</td>
<td>100 %</td>
</tr>
<tr>
<td>Hs.19074</td>
<td>H. sapiens</td>
<td>PID:g7212790</td>
<td>large tumor suppressor 2&quot;</td>
<td>1045</td>
<td>100 %</td>
</tr>
<tr>
<td>Hs.48499</td>
<td>H. sapiens</td>
<td>PID:g7144644</td>
<td>AF102177 1 tumor antigen SLP-8p&quot;</td>
<td>965</td>
<td>100 %</td>
</tr>
<tr>
<td>Hs.116875</td>
<td>M. musculus</td>
<td>PID:g7637845</td>
<td>AF172722 1 tumor-rejection antigen SART3&quot;</td>
<td>962</td>
<td>87 %</td>
</tr>
<tr>
<td>Hs.211600</td>
<td>M. musculus</td>
<td>SP:Q60769</td>
<td>TNP3 MOUSE TUMOR NECROSIS FACTOR, ALPHA-INDUCED PROTEIN 3&quot;</td>
<td>789</td>
<td>88 %</td>
</tr>
<tr>
<td>Hs.211600</td>
<td>H. sapiens</td>
<td>SP:P21580</td>
<td>TNP3_HUMAN TUMOR NECROSIS FACTOR, ALPHA-INDUCED PROTEIN 3&quot;</td>
<td>789</td>
<td>100 %</td>
</tr>
</tbody>
</table>

DIMACS'01

I. Jurisica
Conclusions

- Management - representation - reasoning - discovery
  - moving from hypothesis-driven to exploration-driven research (analysis)
  - systematically analyzing the problem space
- HTP
  - automation, systematicity, reproducibility
  - hypothesis search - generation & evaluation
"Most disease processes and treatments are manifested at the protein level"

"Gene-based expression analysis alone will (in certain cases) be totally inadequate for drug discovery"

"Only 2% of diseases are believed to be monogenic - we need to understand protein-protein interactions"

DDT 4(3):129-133, 1999
Thanks

- P. Rogers, M. Sultan
- A. Rehaag, G. Quon
- D. Wigle, O. Huner
- P. Macgregor, M. Albert

- J. Glasgow
- A. Barta
- M. Maziarz
- W. Andreopoulos

http://www.cs.utoronto.ca/~juris