Data Mining in Bioinformatics

Erwin M. Bakker
LIACS
Leiden University
Overview

• Introduction
• Bioinformatics: Data Analyses Data Modeling
• DIAL CMSB
• Phenotype Genotype Integration
• CYTTRON
• Sub-Graph Mining
• Conclusion
Leiden - Delft
CS Bioinformatics track
www.delftleiden.nl/BIO

• Organized by:
  – LIACS Leiden University
  – EEMCS, the faculty of Electrical Engineering, Mathematics and Computer Science of the Delft University of Technology

• co-operation with three of the four centres of excellence of the *Nationaal Regie Orgaan Genomics*:
  – the Kluyver Centre for Genomics of Industrial Fermentation in Delft
  – the Cancer Genomics Consortium (of which DUT is a member)
  – the Centre for Medical Systems Biology in Leiden.

• **Focus on: Data Analyses and Data Modeling**
Data Analyses and Data Modeling

- **Zebra Fish Atlas** (dr F. Verbeek)
- **Applied optimization techniques**: EA, GA, NN, etc. (prof T. Bäck)
- **Content Based Indexing and Retrieval** (dr M.S. Lew)
- **Integrating Protein Databases**: Collecting and Analyzing Natural Variants in G Protein-Coupled Receptors (drs. M. van Iterson, drs J. Kazius (LACDR))
- **Mining Phenotype Genotype Data** (drs. F. Colas, LUMC)
- **Data Mining, VL-e** (prof J. Kok), Etc.
Data Mining

Data Mining’ and ‘Knowledge Discovery in Databases’ (KDD) are used interchangeably

– The process of **discovery of interesting, meaningful and actionable** patterns hidden in **large amounts** of data

• Multidisciplinary field originating from artificial intelligence, pattern recognition, statistics, machine learning, bioinformatics, econometrics,
Data Mining in Bioinformatics

• Problem:
  – Leukemia (different types of Leukemia cells look very similar)
  – Given data for a number of samples (patients), can we
    • Accurately diagnose the disease?
    • Predict outcome for given treatment?
    • Recommend best treatment?

• Solution
  – Data mining on micro-array data
CGH-DB

Center for Medical Systems Biology (www.cmsb.nl)
Data Integration and Logistics (DIAL)

CGH-DB: a CGH Database

- Consolidation of Experimental Data
- Integration of CGH data with:
  - Other CGH Experiments
  - Genome Databases
  - Expression Databases
  - Phenotype data
  - Etc.
- Publication, validation, repetition, etc.
Groups Involved

- **Micro Array Core Facility, VUMC**: Bauke Ylstra, José Luis Costa, Anders Svensson, Paul vden IJssel, Mark van de Wiel, Sjoerd Vosse
- **Center for Human and Clinical Genetics, LUMC**: Judith Boer, Peter Taschner, and others
- **Department of Molecular Cell Biology, Laboratory for Cytochemistry and Cytometry**: Karoly Szuhai
- **Leiden Institute of Advanced Computer Science, LIACS**: Joost Kok, Floris Sicking, Erwin Bakker, Sven Groot, Michiel Ranshuysen, Harmen vder Spek
CGH-DB Goals

• A **Secure**, **Reliable**, and **Scalable** database/data management solution for storing the vast amounts of experimental micro array comparative genomic hybridization (CGH) **data** and **images** from the different CMSB research groups.

• **Data Consolidation:** through **standard control** mechanisms for data quality, data preprocessing, data referencing (BAC), and **meta data** (CGH MIAME), it is ensured that the stored data represent the original experimental data in an accurate and highly accessible way.

• **Data Integration:** the applied standards for normalization, smoothing, (BAC) referencing, and MIAME CGH annotation must support multiple experiment integration over various platforms, and a controlled interface with further analysis and visualization tools.
Comparative Genomic Hybridization (CGH)

Mantripragada et al, Trends in Genetics 2003
At BAC, or Oligo positions:

- Normal
- Gains
- Losses
Micro Array CGH Data Flow I
Micro Array CGH Data Flow II

Data Management

MIA ME CGH

BAC Oligo

Experiments

Processing
Norm., Smoothing, etc.

Data Mining

View

Selection

View

TIGR MeV

Statistics

See CGH

Visualization, Reporting, Analysis, ...

Ensemble

Rosetta Resolver
Multiple Experiments: Viewing, Analyses
Chromosome 9
1 - 138,429,268

Detailed view

Use Ensembl to...
- Run a BLAST search
- Search Ensembl
- Data mining [BioMart]
- Upload your own data
- Download data
- Export data

Docs and downloads
Visual Feedback: normalization; smoothing; BAC reference file version; ...
Integration of different experiments: BAC, Oligo, Bluefuse, Imagene, etc.
CGH Databases

• Data Explosion
  – BAC 3500 data points
  – Oligo’s 20000 to 60000 data points 1000 experiments/year (currently)
  – 200k and 500k expected in the near future
  – Within some years: 5M data points ‘routine’ diagnosis
  – 200MB - 1GB Images

• Storage and Computational Requirements
Integration of Genomic Data
• Micro Array Expression Data mRNA levels, …
• Human Genome, Chimp, Rhesus, Mouse, etc.
• Semantic integration

• Scale up of routine analysis
• Scale up of research analysis over integrated data sets
• Data mining for hidden relations
• …
Phenotype Genotype Integration

• Genotype data
  – Annotated genome databases
  – CGH Database
  – Expression databases
  – Etc.

• Phenotype data (Multimodal)
  – Blood samples
  – Weight, height, fat %, fat type, etc.
  – Echo, CT, MRI scans
  – Photographs
  – Etc.
Longevity Studies at LUMC
Group headed by E. Slagboom (LUMC)

Current data mining studies by Fabrice Colas (LIACS)

- Mining genetic data sets
- 1-, 2-, and 3-itemsets (frequent item sets)
- Solving the problems in reasonable time was only possible using parallel computing (DAS3)
Towards a Classification of Osteo Arthritis subtypes in Subjects with Symptomatic OA at Multiple Joint Sites.
F. Colas et al NBIC-ISNB2007

GARP study of OA (Osteo Arthritis) subtypes
• Identifying genetic factors
• Assist in development of new treatments
• Genetic etiology is difficult because of the clinical heterogeneity of the disease
• Identification of homogeneous subgroups of OA
• Identify and characterize potentially new disease subtypes using machine learning techniques
• Parallel Computation (DAS3)
Content Based Indexing and Retrieval Techniques

- Image Databases
- Speech Databases
- Video Databases
- Multimodal Databases
- Face recognition, bimodal emotion recognition (N. Sebe, UVA), Semantic Audio Indexing, etc.
Within the CYTTRON project various modes of imaging biological structures and processes will be integrated in a common visualization platform.

The success of the integration and use of the bio-image data strongly relies on new bio-image processing techniques and searching methods.

The research focus is on new visual search tools for bio-image queries, handling multi dimensional image data sets.
CYTTRON

- Different Bio-Imaging Techniques:
  - Light Microscopy
  - MRM
  - Confocal laser Microscopy
  - EM, Cryo, 3D EM
  - NMR
  - Crystallography
  - Etc.
Example: White blood cell
CYTTRON

• Integration
  – Different modalities
  – 2D, 3D, Noisy, Model, random projections
  – Poor annotation
• Database design
• Content Based Searching Algorithms
• Feature Based Annotation
• Automatic Learning: relevance feedback, training sets, etc.
• Computational needs …
Content-based image retrieval

• Searching for images based on content only, using an image as a query

• Using text search for images requires every image to be annotated. This has some disadvantages:
  – Annotating images is time-consuming
  – Annotation can be incomplete
  – Annotation can be almost impossible
Image annotation difficulties

• How would you describe these images?
Basic CBIR paradigm

- Describe a specific visual property (feature) of an image as a vector
  - RGB Histograms
  - Local Binary Patterns
  - Etc.
- Extract features for all database images
- Extract features from query image
- Calculate distance between query image and all database images
- Rank images by distance

Average color

(23, 37, 241)
Relevance Feedback

Iterative search process:
• Search for query image
• Ask user to evaluate the search results
• Use feedback to adjust the query
• Repeat process until user is satisfied
Sub-image search

• Let the user select one or more parts of the query image
• For each database image, calculate the number of sub-images matching (are close to) the selected parts
• Rank results based on number of matching sub-images
Automatic Registration of Microtubule Images
Feiyang Yu, Ard Oerlemans
Erwin M. Bakker and Michael S. Lew

(Artificial images. The original images could not be used due to copyright.)

Microtubule ‘Movie’
Challenges

CYTTRON
- Large number of images
- Insufficient or no annotation
- Multiresolution images (different scales)
- Images made by different types of imaging devices

LML Projects
- High performance feature space computation and indexing (Images, Video’s; batch usage)
- Interactive robust content based indexing techniques: emotion recognition, object recognition, who is talking, what is audible, etc.
  - can be batch usage, but optimally we would like real time usage of DAS3 (!?)
Proteins: structure is function

- 1D and 2D structure computable from models, 3D structure difficult to predict
- Protein sequences => molecular description => structural encoding in graphs
- Existing protein databases can be encoded as graphs
- New sequences can then be encoded as graphs and used to search the graph database
- Mine the graph database => frequent patterns => see if these frequent patterns indicated groups of proteins with the same functionality
GASTON
S. Nijssen, J.Kok ’04

• www.liacs.nl/~snijssen/gaston/iccs.html

• Applications:
  – Molecular databases
  – Protein databases
  – Acces-patterns
  – Web-links
  – Etc.
Frequent Pattern Trees

• Develop new parallel versions for frequent item set mining

• Currently research on Closed and Constrained Frequent Item Set mining
  – Biological Semantics
  – Biological Relevance
  – Evaluation experiments will be run on DAS3
Conclusion

• Data mining in Bioinformatics offer many challenging tasks in which DAS3 plays an essential role:
  – research on novel scalable high performance segmentation of high dimensional and high volume feature spaces.
  – Development and evaluation of novel high performance techniques for data mining
  – research on novel scalable data(base) structures for efficient data querying, analysis and mining of high volume data sets