Data Bases and Data Mining

Universiteit Leiden

liacs medialab

Erwin M. Bakker
LIACS, LML
Leiden University

Databases and Data Mining Organization

- Materials:
  - J. Han and M. Kamber. *Data Mining, Concepts and Techniques.* (2nd Edition) Morgan Kaufmann, 2006
  - Presentations and papers
- Grading
  - Assignments and class participation
- Website of the course
  - www.liacs.nl/~erwin/dbdm2009
- Assistant: Hossein Rahmani

Databases and Data Mining Overview

- Some related LIACS research
- The evolution of database system technology
  - Integrated Data and Information Systems
  - Semantics and Ontologies
- Data Preprocessing
- Data Warehouse, Data Cubes, OLAP
- Grand Challenges and State of the Art

Databases and Data Mining Overview

- Introduction and Overview Data Mining
  - Frequent Item Sets
- Advanced Topics
  - Mining Data Streams and Time Series Data
  - Mining Sequence Patterns in Transactional Data
  - Mining Biological Sequence Patterns
  - Graph Mining
  - Social Network Graphs
  - Multi-relational Data Mining

Databases and Data Mining Overview

- Advanced Topics (continued)
  - Multidimensional Analysis and Descriptive Mining of Complex Data Objects
  - Spatial Data Mining
  - Multimedia Data Mining
  - Text Mining
  - Mining the World Wide Web
- Applications and Trends in Data Mining

LIACS Projects

Universiteit Leiden

liacs medialab

TU Delft Technische Universiteit Delft

Centre for Medical Systems Biology

- Bioinformatics: Data Analyses & Data Modeling
- DIAL CMSB 1 & 2
- Phenotype Genotype Integration
- CYTTRON
- Sub-Graph Mining
- GRID Computing
- Sensor Networks

Cyttron

DAS-3

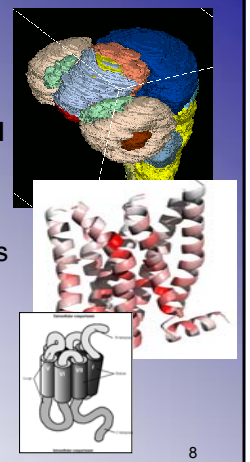
Leiden - Delft CS Bioinformatics track

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 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

Centre for Medical Systems Biology

- CMSB (Centre for Medical Systems Biology, www.cmsb.nl) is one of the Genomics Centre of Excellence
  - Genomics for identifying hidden connections between diseases: Improving diagnosis, treatment and prevention of common diseases *such as Alzheimer's, cardiovascular disease, diabetes and rheumatism.*
- DIAL (Data Integration, Analysis and Logistics) The project works on the data of the experimental projects of the CMSB.

Centre for Medical Systems Biology

Six Projects

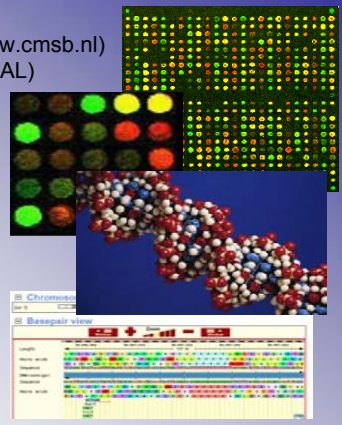
- Epidemiology: cohorts & genotyping
- Systems Biology: transcriptomics/arraying proteomics metabolomics
- Technology: magnetic resonance microscopy and others imaging molecular interactions
- Model Systems: animal models (mouse, zebra fish etc).
- Clinical Applications: translation (cells, vaccines, viral, pharmaceutical)
- DIAL: Data Integration, Analysis and Logistics

DIAL Example Study Groups

- RotterdamStudy (ERGO: Hofman, van Duijn, a.o.)
  - population-based cohort study of 12,000 subjects aged 55+ years. Patients have been followed for over 10 years now.
- Grip Cohort Study (Rotterdam: van Duijn, Oostra)
  - population-based cohort study of 3 generation families (2500 subjects). They are screened for the presence of multiple diseases.
- Netherlands Twin Register (Boomsma VU Amsterdam)
  - number of twins (60,670) and siblings (3,175)

DIAL CMSB: 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.

Centre for Medical Systems Biology

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, Antanas Kaziliünas

Universiteit Leiden

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.

DIAL CGH Database Platform

How?: Security, Reliability, Availability, Scalability

Hardware:
- Alcolu Intel SR2500 2U Server
- 2x Dual Core Xeon 5160 3.7GHz, 4x2GB
- RAID5, 6x250GB Hot Spare
- UPS, NAS Backup Storage (Attached)
- Currently: mirrored 2x Quad Core 8Tbyte systems (geographically separated)

Operating System and Database:
- Unix (Linux, Solaris)
- Database (Oracle, Postgres)
- Web-server (Apache, near future: https, SSL login)
- Applications: PERL modules

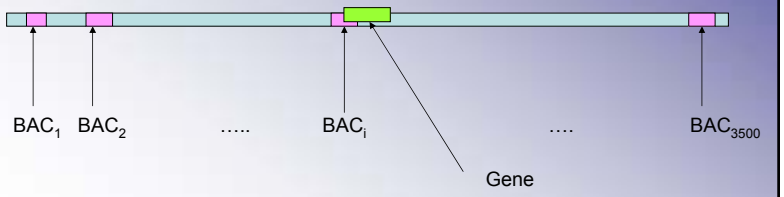
BAC's and OLIGO's

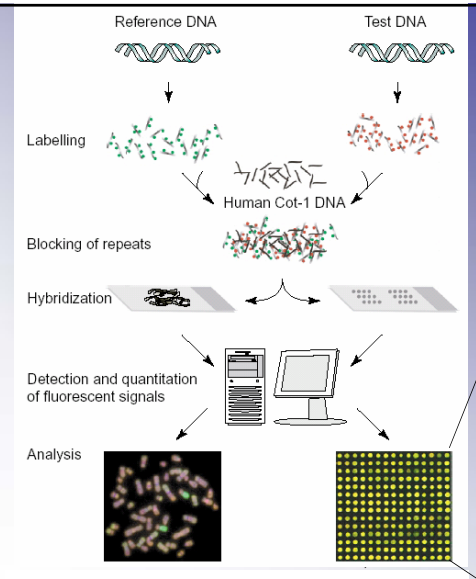
- DNA: a long polymer of simple units called nucleotides, with a backbone made of sugars. Attached to each sugar is one of four types of molecules called bases (A, C, T, G).
- Bacterial Artificial Chromosome (BAC), typically relatively short sequences of up to 200k bases.
- Oligonucleotides (OLIGO's) are short sequences of nucleotides (RNA or DNA), typically with <=20 bases. Automated synthesizers allow the synthesis of oligonucleotides up to 160 to 200 bases.
- SNP: single nucleotide polymorphism a variation in the DNA of length 1 nucleotide, i.e., some people will have …CGG**T**AAC…, whereas others will have …CGG**C**AAC… in their DNA

BAC's and OLIGO's

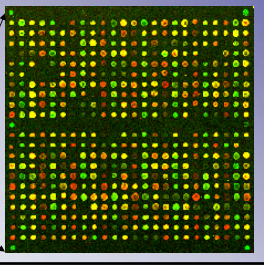
- Site specific hybridization of control and sample DNA or cDNA to target DNA (BAC, or OLIGO's)

Complete Genome, all chromosomes consecutively:





Comparative Genomic Hybridization (CGH)

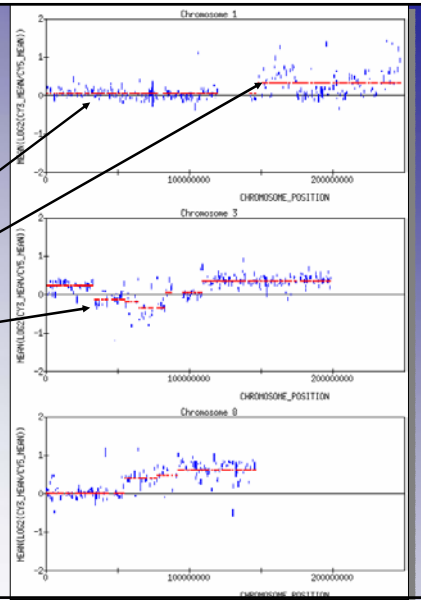


Mantripragada *et a,l Trends in Genetics* 2003

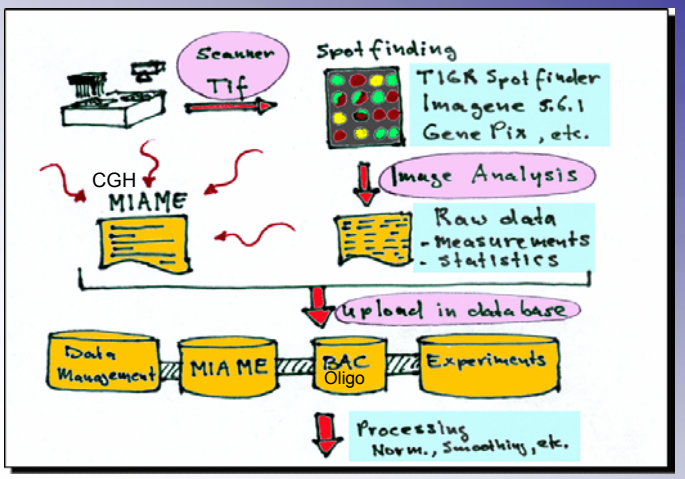
At BAC, or
Oligo positions:

- Normal
- Gains
- Losses

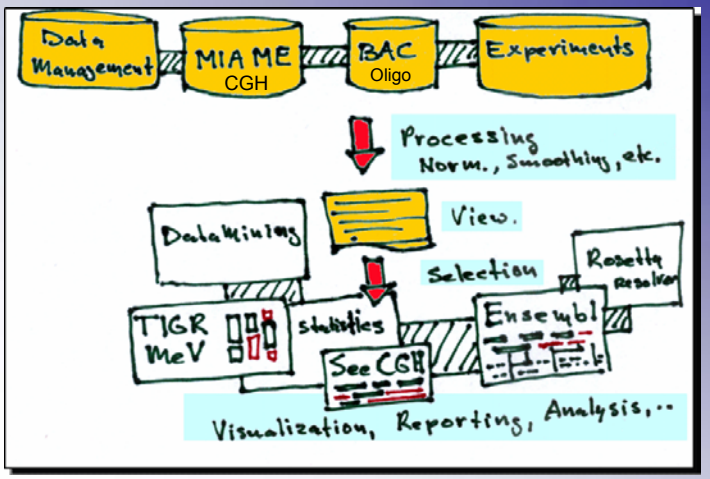
Log-ratios = $\log(I_{Red}/I_{green})$, where
I_{red}= measured red channel intensity
I_{green}=measured green channel intensity

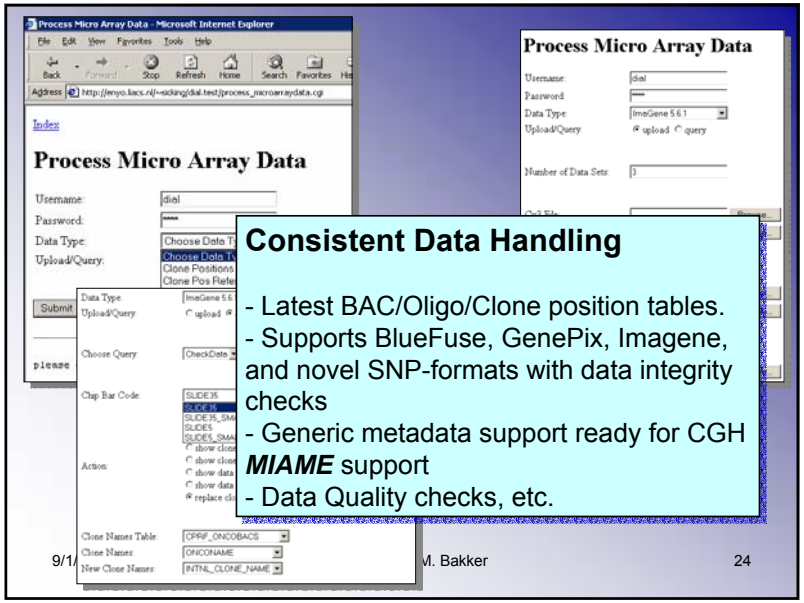


Micro Array CGH Data Flow I



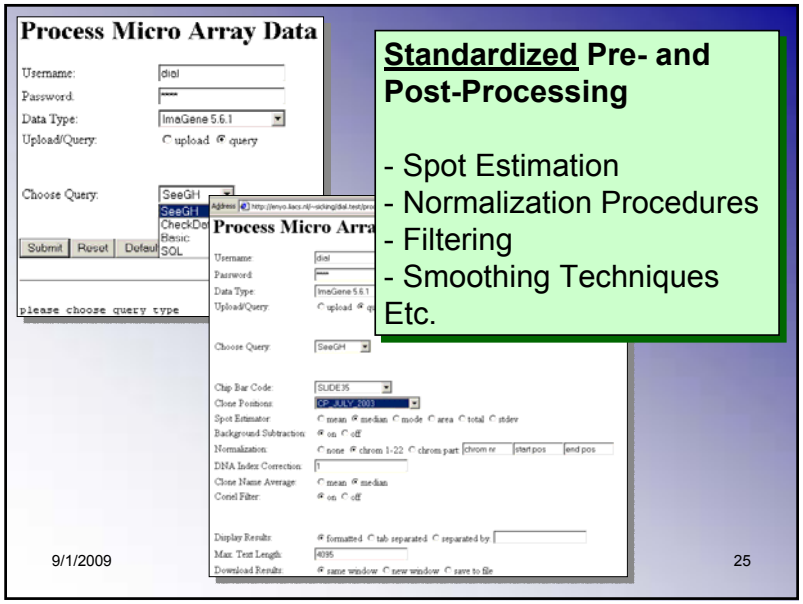
Micro Array CGH Data Flow II

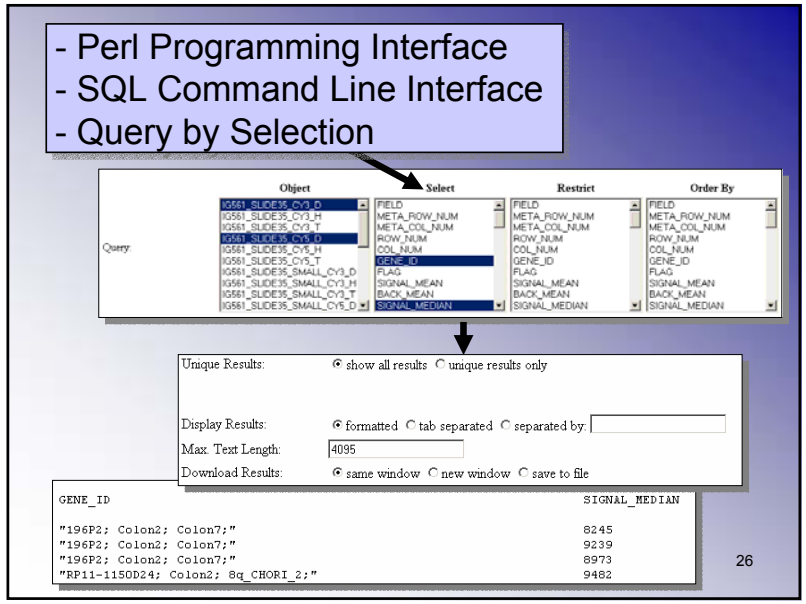


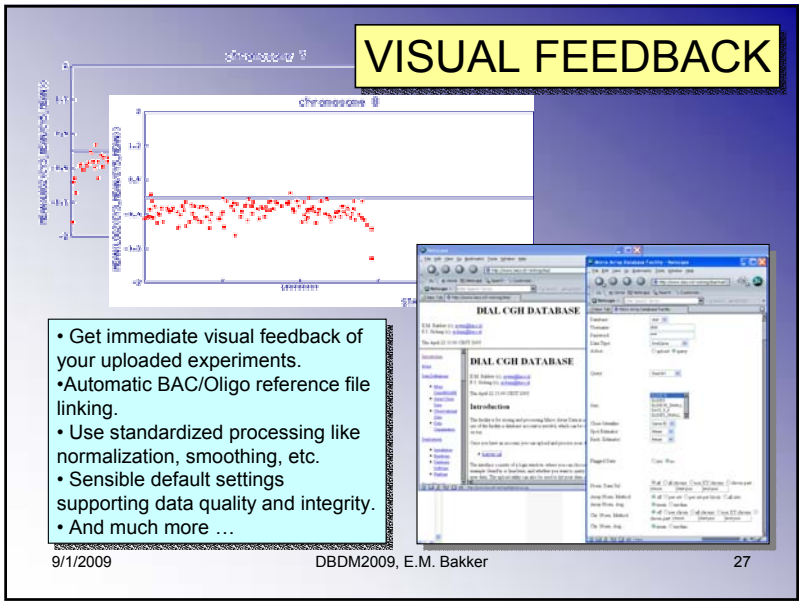


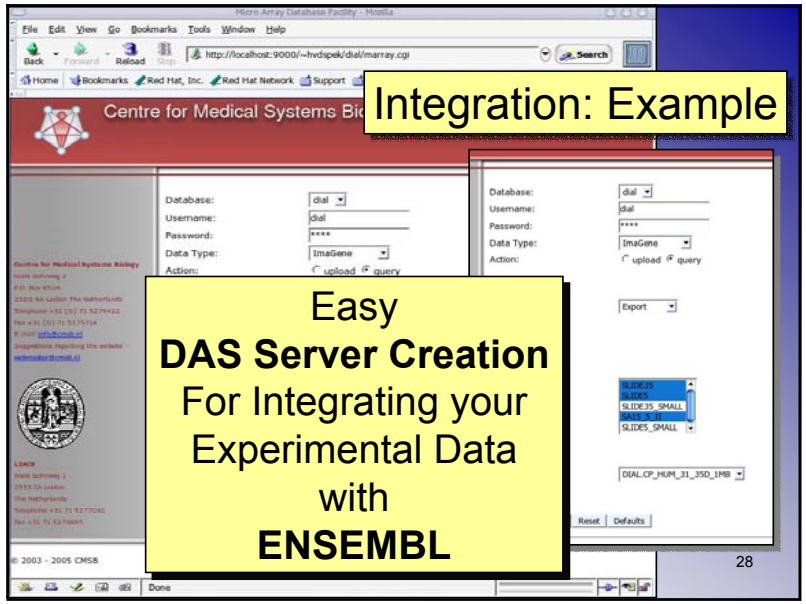
Consistent Data Handling

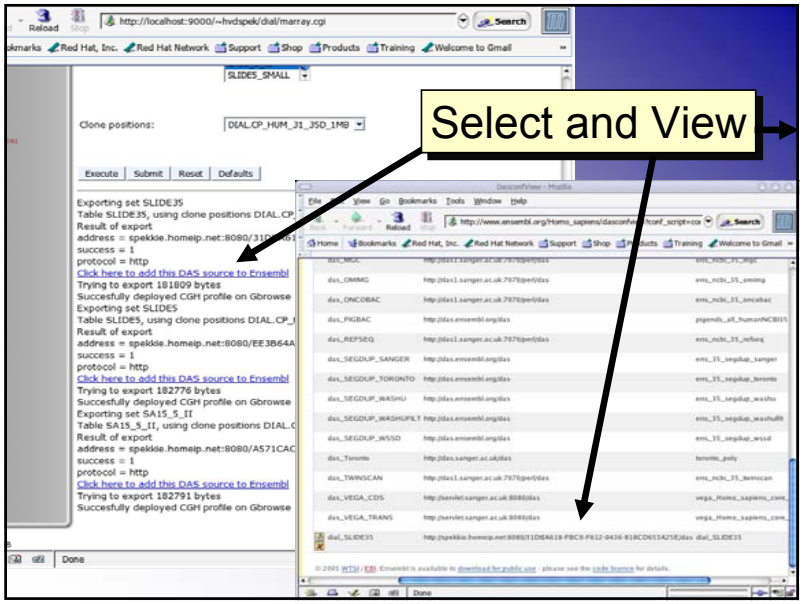
- Latest BAC/Oligo/Clone position tables.
- Supports BlueFuse, GenePix, Imagene, and novel SNP-formats with data integrity checks
- Generic metadata support ready for CGH ***MIAME*** support
- Data Quality checks, etc.

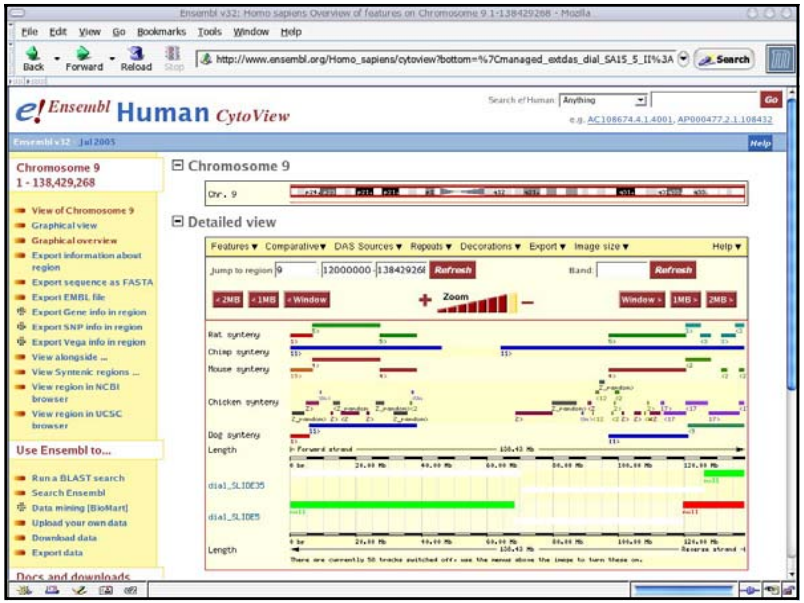

Process Micro Array Data
Username:
Password:
Data Type:
Upload/Query:
Choose Query:
Submit
Reset
please choose query type
Process Micro Array Data
Standardized Pre- and Post-Processing
- Spot Estimation
- Normalization Procedures
- Filtering
- Smoothing Techniques
Etc.

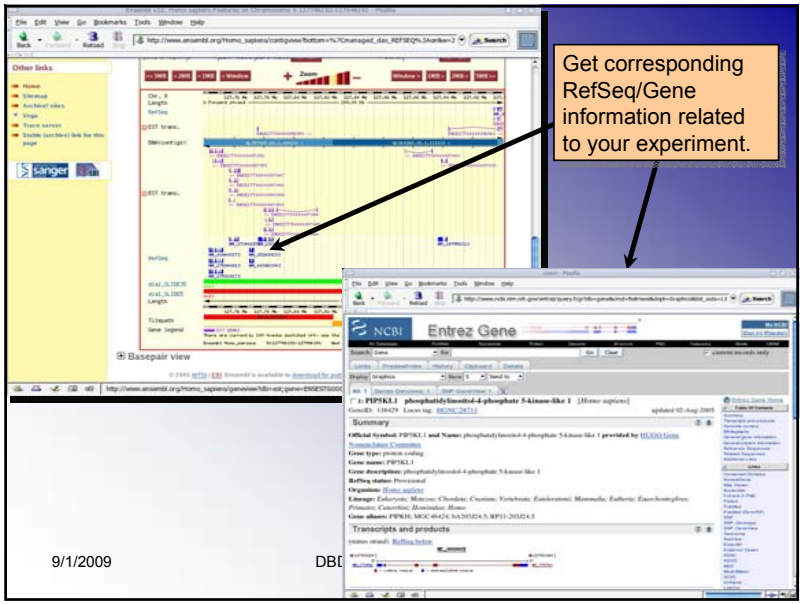

- Perl Programming Interface
- SQL Command Line Interface
- Query by Selection
Object
Select
Restrict
Order By
Query:
Unique Results:
show all results
unique results only
Display Results:
formatted
tab separated
separated by:
Max. Text Length:
4095
Download Results:
same window
new window
save to file
GENE_ID SIGNAL_MEDIAN
"196P2; Colon2; Colon7;" 8245
"196P2; Colon2; Colon7;" 9239
"196P2; Colon2; Colon7;" 8973
"RP11-115OD24; Colon2; 8q_CHORI_2;" 9482

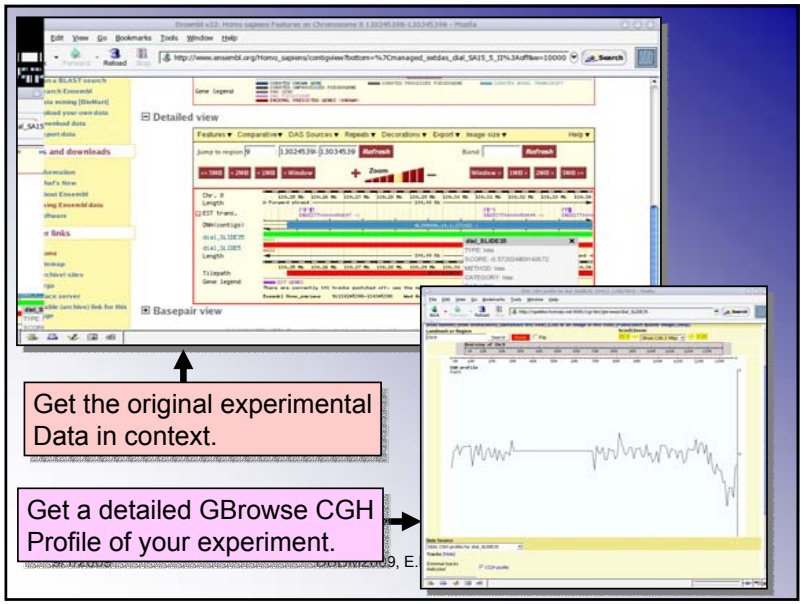

VISUAL FEEDBACK
• Get immediate visual feedback of your uploaded experiments.
•Automatic BAC/Oligo reference file linking.
• Use standardized processing like normalization, smoothing, etc.
• Sensible default settings supporting data quality and integrity.
• And much more …
DIAL CGH DATABASE


Integration: Example
Centre for Medical Systems Biology
Database:
Username:
Password:
Data Type:
Action:
Easy
DAS Server Creation
For Integrating your Experimental Data with
ENSEMBL

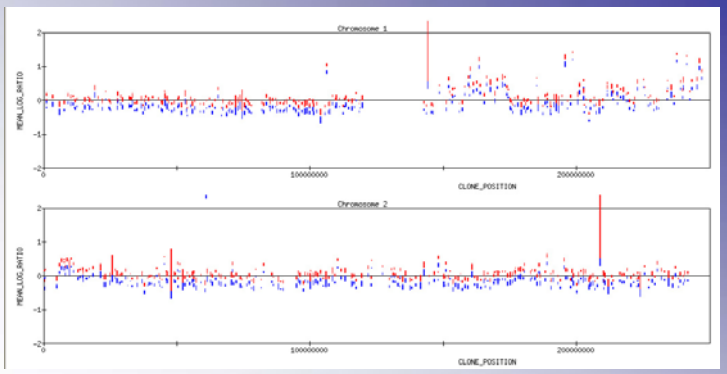

Select and View


Ensembl Human CytoView
Chromosome 9
1 - 138,429,268
Chromosome 9
Detailed view

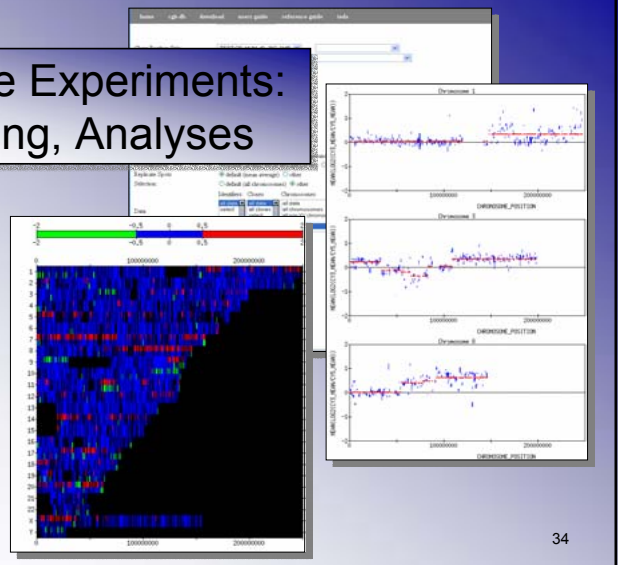

Get corresponding RefSeq/Gene information related to your experiment.
NCBI Entrez Gene


Get the original experimental Data in context.
Get a detailed GBrowse CGH Profile of your experiment.

Visual Feedback: normalization; smoothing;
BAC reference file version; …



Multiple Experiments:
Viewing, Analyses



DIAL Micro Array CGH

MIAME minimum information about a micro array experiment

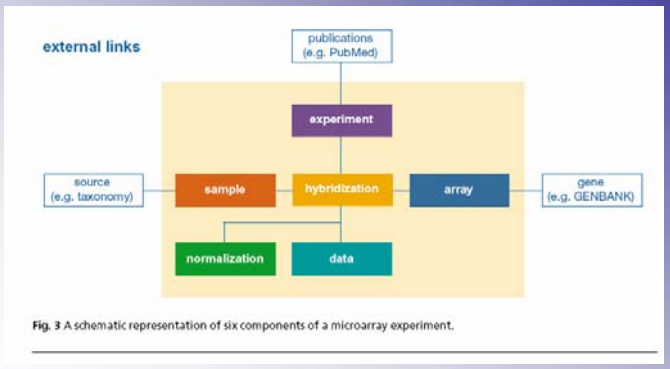
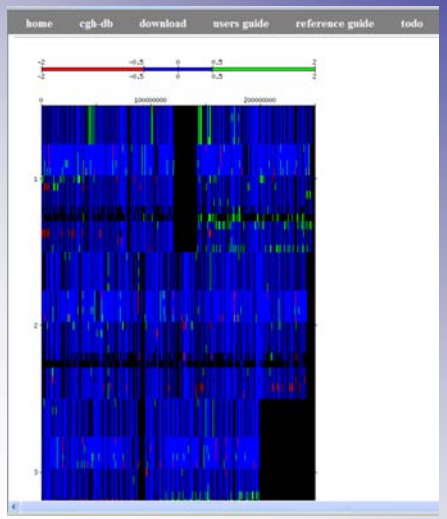


Fig. 3 A schematic representation of six components of a microarray experiment.

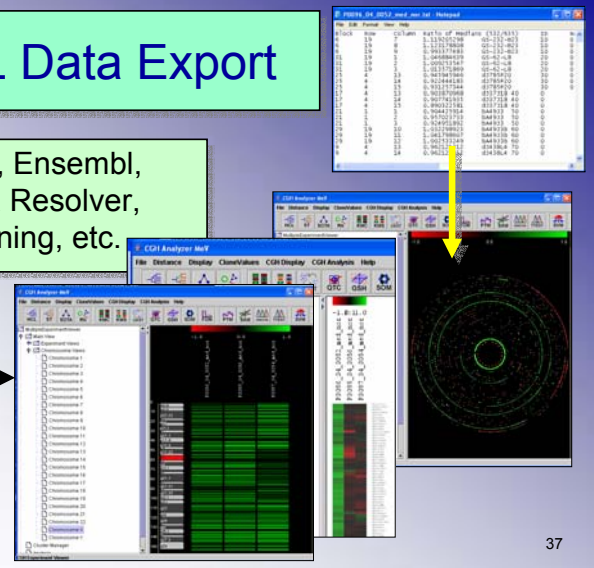


Integration of different experiments: BAC, Oligo, Bluefuse, Imagene, Affymetrix, etc.

DIAL Data Export

- SeeGH, Ensembl, Rosetta Resolver, data mining, etc.

TIGR MeV:



Other Proof of Concepts and Projects

- Interfacing with MySQL data warehouse
- Clustering Module (Python, R)
- Data Mining Algorithms for Multiple CGH Experiments (C++)
- Experimentation with novel CGH Segmentation Methods (Matlab, R)
- Genotype Phenotype Integration using semantic wrappers (Postgres, JAVA)
- Processing pipeline: C#, R
- … YOUR PROJECT … -> erwin@liacs.nl

CGH Databases

- Data Explosion
  - BAC 3500 data points
  - Oligo's 20000 to 60000 data points 1000 experiments/year
  - 200k and 500k currently
  - Soon: 5M data points 'routine' diagnosis
  - 200MB - 1GB Images
- Storage and Computational Requirements

Challenges

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
- …

DIAL CGH Database Key Benefits

- **Consolidation** of the Micro Array Experiment
- **Converging data handling methods** within CMSB => Data Quality and Data Integration
- **Automatic BAC and Oligo referencing** and version management
- **Converging data annotation** within CMSB: MIAME CGH
- **Straightforward Integration:** multi experiment; interfacing for further analyses; export to other databases; Ensembl; Data mining; Publication Export; Your Favorite Analysis Tool, etc.

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)

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

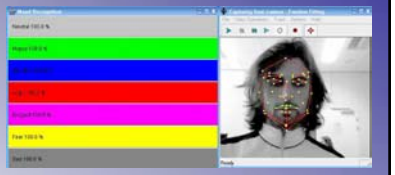
GARP study of OA (Osteo Arthritis) subtypes

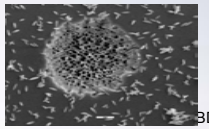
- Identifying genetic factors
- Assist in development of new treatments
- Genetic causes of the disease are difficult to obtain because of the **clinical heterogeneity** of the disease
- Identification of homogeneous subgroups of OA
- Identify and characterize potentially new disease subtypes using machine learning techniques
- Parallel Computation (DAS3)

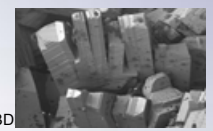
liacs medialab
LEIDEN UNIVERSITY

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.



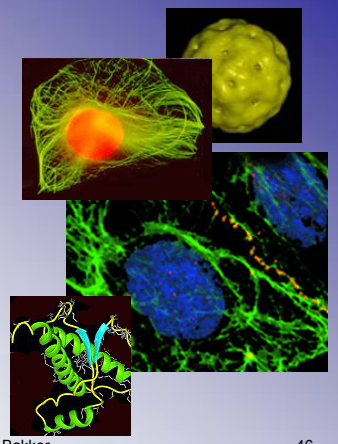




CYTTRON

Headed by prof J.P. Abrahams (LIC), www.cyttron.nl.

- Within the **CYTTRON** project various modes of imaging biological structures and processes had to 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.

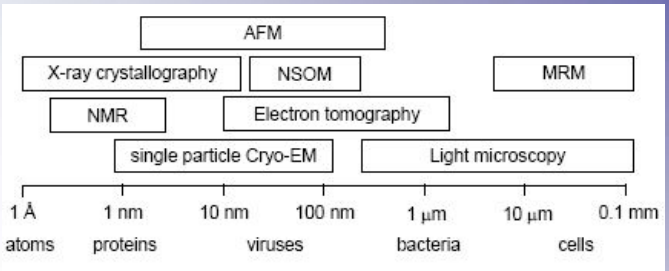


CYTTRON Consortium

- Leiden, Delft, Utrecht, Antwerp and London University, LUMC, Bruker Nonius BV, FEI BV, Key Drug Prototyping BV.
- Headed by Prof J.P. Abrahams (LIC, LU)

CYTTRON

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



Fluorescence Microscopy

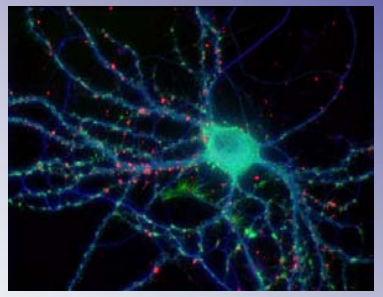


Figure from http://www.wadsworth.org/cores/alm/: A multi-wavelength, three dimensional, wide-field immunofluorescence image of a fixed neuron. The projection was generated using an extended depth of field algorithm. **Cell body** labeled for tubulin is shown in blue, **F-actin** in green, and **presynaptic protein** in Red. Specimen courtesy of Natalie Dowell-Mesfin BMS-PhD student

Fluorescence Microscopy

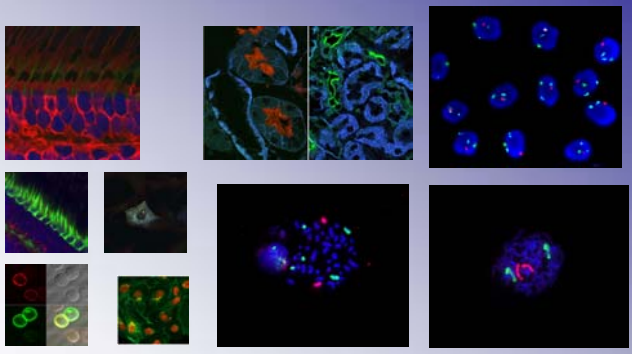


Figure from http://hsc.unm.edu/pathology/microscopy/instru.htm

Confocal Laser Scanning Microscope

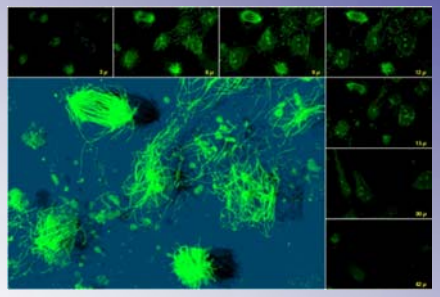
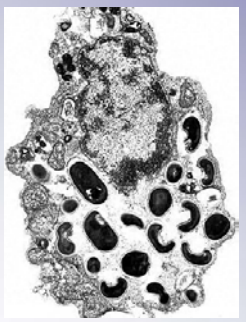
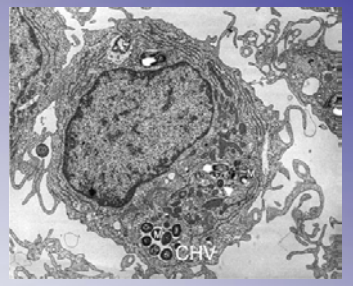


Figure (from http://www.mih.unibas.ch/Booklet/Booklet96/Chapter1/Chapter1.html). *Seven representative optical sections selected from 81 confocal planes (corresponding to a depth of 50 mm) "cut" through a collagen matrix containing growing fibroblasts labeled with fluorescent antibodies to tubulin. Inset, composite shadow-projection image of all 81 confocal sections revealing the spindle apparatus of dividing cells and the regular microtubular network of interphase (i.e., non-dividing) cells*

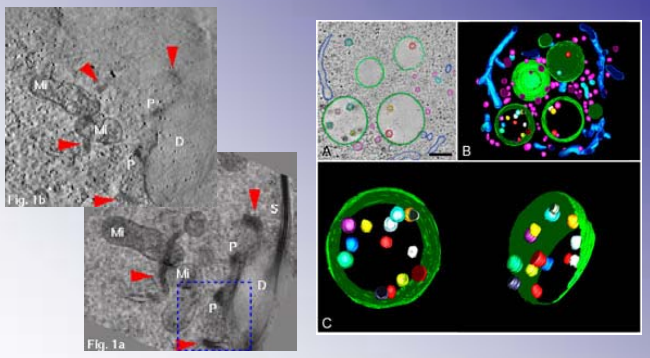
Electron Microscopy





- Some standard (old technique) electron microscopic slides

3D Electron Microscope Electron Tomography

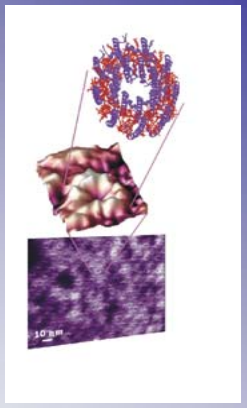


Images from http://www.bio.uu.nl/mcb/3dem/

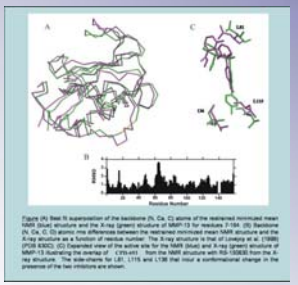
Scanning Probe Microscopy Molecular Imaging

Figure from http://www.physics.leidenuniv.nl/sections/cm/ip/projects/bio-afm/ In a joint project with the Biophysics Department, we are using Scanning Probe Microscopy (SPM) to visualize the molecular and electronic structure of single photosynthetic pigment-protein complexes, of purple bacteria. 2D aggregates of the photosynthetic pigment-protein complexes are prepared for Atomic Force imaging and IV spectroscopy.

- Molecular Imaging: http://www.molec.com/
- Scanning Tunneling Microscopy
- Atomic Force Microscopy
- Scanning Probe Microscopy
- Membrane visualization of living cells



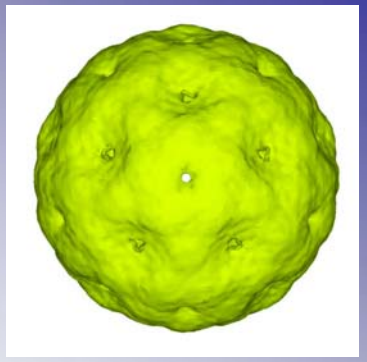
NMR, X-Ray Crystallography



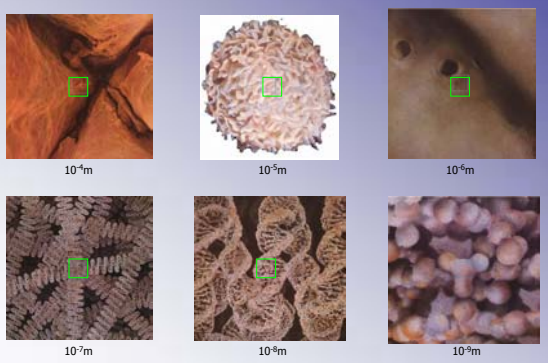
- Structure determination of protein-protein complexes by NMR and X-ray crystallography.

Single Particle Cryo Electron Microscopy

- Reconstruction made by Tyson (reconstruction package).



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 …

Interactive Search in Bio-Image Databases

LIACS Media Lab
Leiden University

Project Background

- Mission: Develop multi-modal (text & image content) search methods for bio-image databases
- Period: 2004 - 2007
- People
  - Ard Oerlemans, PhD candidate
  - Fiona Feiyang Yu, LIACS, PhD candidate
  - Dr. Michael S. Lew, LIACS, supervisor
  - Dr. Erwin M. Bakker, LIACS, supervisor

Image DB Search Background

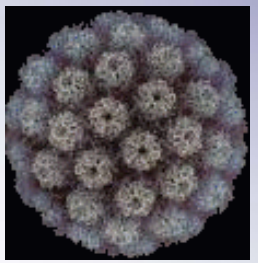
- Principal Investigators for Very Large Image Database projects for
  - Philips Research Eindhoven - Music, Video, Assets
  - IBM
  - Google
  - NWO - 2 grants, Video Databases; Image Databases
  - European Union/Fifth Framework
  - Over 100 peer-reviewed papers in ACM & IEEE, 1995 -
- Organizers for Image DB Research Conferences
  - ACM Multimedia Information Retrieval, ACM Multimedia, IEEE Multimedia, Int. Conf. Image and Video Retrieval, VISUAL, SPIE Storage, IAPR ICPR,...

Introduction

- Problem: the imaging techniques studied in the Cyttron project generate a vast amount of imagery
- **How do we search through these kind of huge databases?**
- Text is useful, but
  - it is not always available - manual annotation
  - it is often fails to capture important pictorial info.

Text is Not Enough

- A picture is worth a thousand words…What words can we use to describe the image structures below?



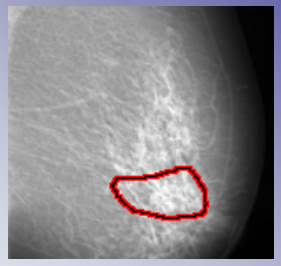
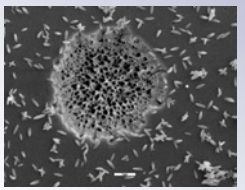


Image annotation difficulties

- How would you describe these images?



Going beyond Google

- Google only searches on text annotation
- We will be searching on both text and the pictorial content of the imagery

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. Disadvantages:
  - Annotating images is time-consuming
  - Annotation can be incomplete
  - Annotation can be almost impossible

Current State of the Art

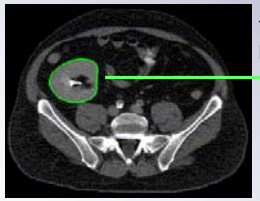
- Most worldwide systems focus on whole-image methods, 1 main object per image
- Current high performance systems focus on 1 particular domain of images - i.e. trademarks, flowers, ...

Basic CBIR Paradigm

- Pre-compute all feature vectors for all images in database
- Calculate feature vectors for query image
- Compare these to the pre-computed feature vectors from the database
- Return the most similar images based on the feature vector distance between query and database

Example

- Given the boundary, convert the interior region to a texture representation such as Linear Binary Patterns
- Quantize the information for efficient searching:



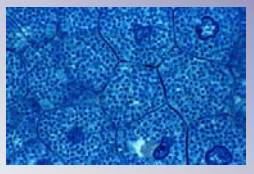
Texture representation:
Linear Binary Patterns

Feature Vector:
F[0…255]

Local Binary Patterns:

```
97   67  20      1 1 0
33   34   5   ->  0 0 1  ->  (110 001 111)
101 123 98      1 1 1
```


Basic CBIR paradigm



Average color → (23, 37, 241)

- Describe a specific visual property (feature) of an image as a vector
  - RGB Historgrams
  - 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

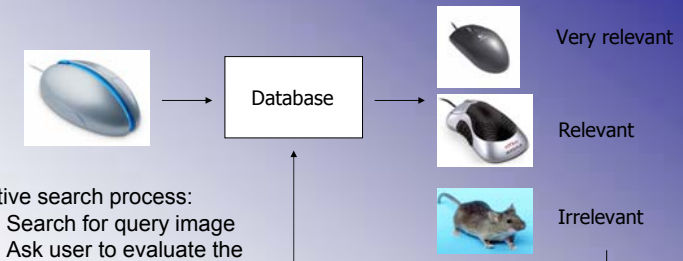
Our Approach

- (1) Go beyond whole-images -> Directly address the subimage problem
- (2) Go beyond single domain -> Integrate automatic machine learning into the search method so that the system can adapt to many domains
- (3) Allow user to interactively improve search results and add domain-specific knowledge

Interactive Search

- Relevance feedback: Based on the initial results, let the user select the most relevant examples and the irrelevant examples. These become positive and negative examples in the learning algorithm.
- Potential: *ability to learn new domains and user-specific queries*.

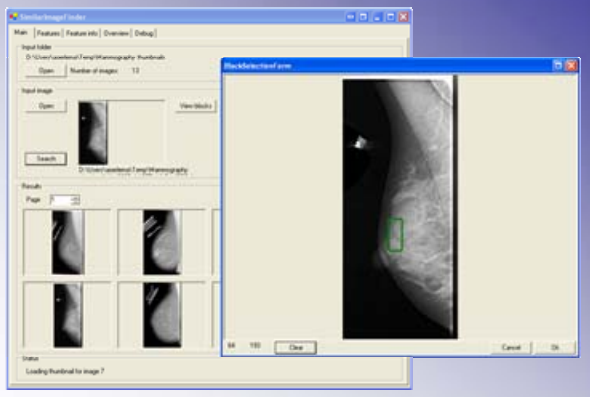
Relevance Feedback



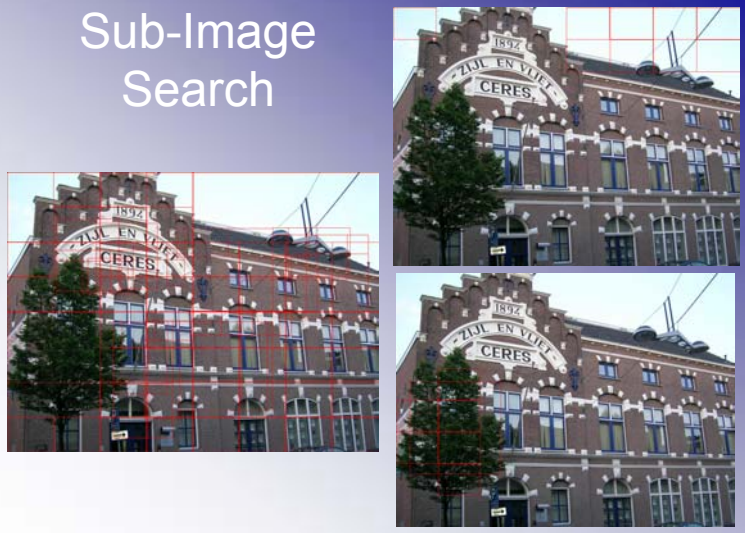
Iterative search process:

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

Example Implementation



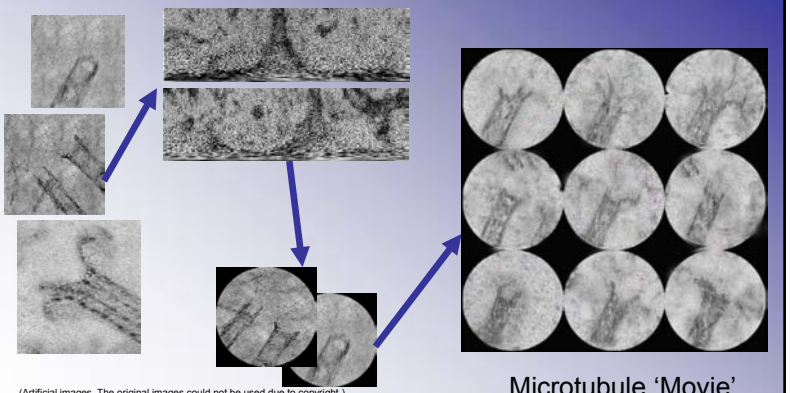
Sub-Image Search



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 Bio-image Searching

- Discover/develop enhanced measures for bio-image similarity
- For example, what features do scientists in biology and chemistry use to decide whether cells are similar? (Very challenging!)
- Sub-image search: develop multi-scale, sub-image search mechanisms for direct usage in the bio-imaging of the cell

Further 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 (!?)

Sub-Graph Mining

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

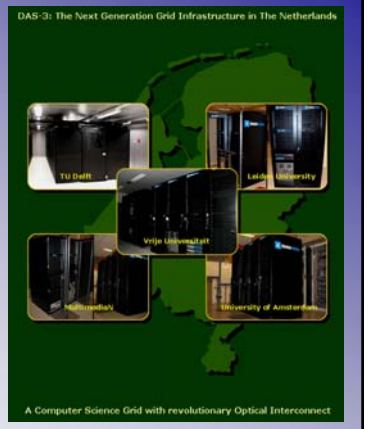
- www.liacs.nl/~snijssen/gaston/iccs.html
- Applications:
  - Molecular databases
  - Protein databases
  - Acces-patterns
  - Web-links
  - Etc.

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 executed on DAS3

DAS3
GRID-Computing

- 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



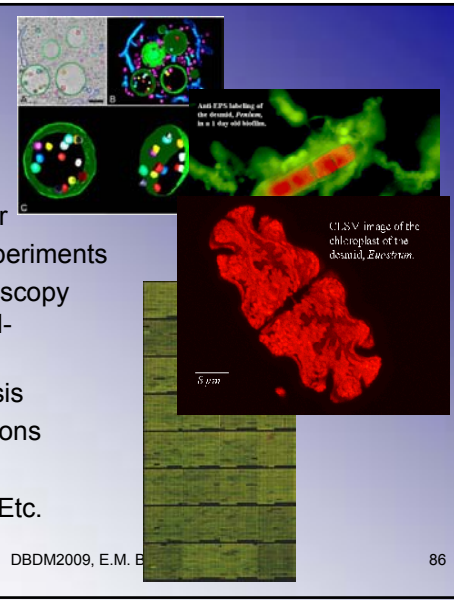
Virtual Laboratory

- Characterizations of E-Science Domains and Application
  - Large amounts of data by simulations or networked instruments
  - Automated experiments
  - Heavily depending on Information and Communication Technology

Virtual Laboratory

- A transparent environment for doing collaborative research using remote and local instrumentation, data, and computational resources, integrating it with existing experimental data libraries
- Scientific Information Management
- Collaborative Experimentation Environments

Examples



- Hadron Collider
- Micro array experiments
- Confocal microscopy experiments (N-dimensional)
- Material Analysis
- Traffic Simulations
- Digital Earth
- DIAL, Cyttron, Etc.

E-Science Domains Common Characteristics

- Diversity of instruments, techniques and information
- Complex experimental procedures, computation intensive in different protocols and techniques
- Large data sizes
- Heterogeneity of the experiments as well as the data
- Collaboration needs due to the expensive and often one of a kind instruments

VLe User Requirements

- Transparent mostly graphical interface to complex instrumentation
- Easy annotation of complex procedures, experiments, data, and analyses