Information Systems in the Life Sciences (ISiLS)

Fons J. Verbeek
Imaging & BioInformatics, LIACS

Lectures & Supervision

- Dr. Erwin Bakker
- Dr. Nies Huijsmans
- Dr. Fons Verbeek (coordination)

Unraveling the complexity of live, the ultimate challenge

Contents

- Introduction
- Organization of the Course
- Data
- Informatics
- Processes
- Contents
- Questions

Organization (1)

- Seminarium
  - Limited number of participants
  - Taking the course is participating in the course
  - Attending the course
- Introductory Lectures (2 series)
  - Bakker
  - Huijsmans
  - Verbeek

Organization (2)

- Paper presentations
  - Schedule
  - Groups, depending on # people attending
  - Deadline
- Paper writing
  - End of the course
  - Discuss with course administration
  - Subjects equally divided over participants
  - Deadline

Information Systems in LS

- Fons Verbeek
  - From signals to systems
  - Virtual Cell
  - Virtual organism
- Nies Huijsmans
  - Molecular modeling, shape
  - Content bases search, image
- Erwin Bakker
  - Data integration MicroArray (DIAL)
  - Gene Browsers
- *Other*

Information Systems in LS

- Closely related to BioInformatics
  - What is BioInformatics?
- Information Retrieval
  - How is the information structured
    - Sequences
    - Graphics
    - Images
    - Other …
  - Learn basic components

BioInformatics

BIO-Informatics ?

$(2b) \vee \neg(2b)$

Bio-Informatics: $(2b) \vee \neg(2b)$

- In the early 1980 Super-Computer scientists (Hwa Lim) realized the potential of combining of biology and computer science: CompBio (“that is not a word …”)
- More whimsical: Bioinformatique
- This changed to: Bio-informatics
- Using - or / was troublesome: Bioinformatics

Interpretations of BioInformatics

- Political interpretation: definition is under debate.
- Narrow interpretation: the information science techniques needed to support genome analysis.
- Broader interpretation: synonymous with computational biology or computational molecular biology.

Definitions of BioInformatics (1)

Bioinformatics:
Research, development or application of computation tools and approaches for expanding the use of biological, medical behavioral or health data, including those to acquire, store, organize, archive, analyze or visualize such data.

after:
NIH Biomedical information Science & Technology Initiative Consortium

Definitions of BioInformatics (2)

Computational Biology:

The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral and social systems.

after:
NIH Biomedical information Science & Technology Initiative Consortium

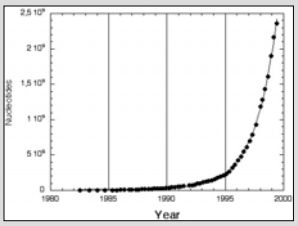
Definitions of BioInformatics (3)

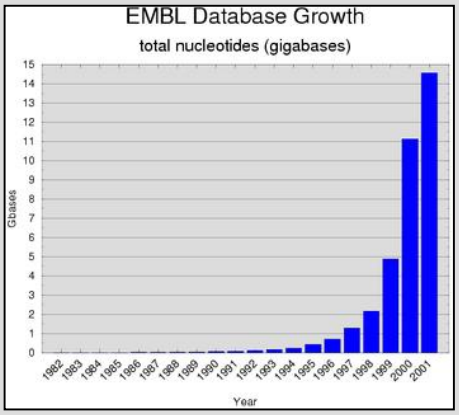
BioInformatics:

- Integration of mathematical, statistical and computer methods to analyze biological, biochemical and biophysical data.
- The science of developing computer **databases** and **algorithms** for the purpose of speeding up biological research (Human Genome Project)

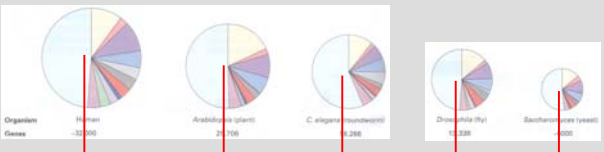
Why the Need? Massive Data Explosion

- Growth of EMBL nucleotide database
- And this is just the Nucleotide database.
- What is a nucleotide





Necessity of BioInformatics



Metabolism
DNA replication/modification
Transcription/Translation
Intracellular signaling
Cell-cell communication
Protein folding and degradation

Transport
Multifunctional proteins
Cytoskeleton/structure
Defense and immunity
Miscellaneous function
Unknown

Adapted from human genome consortium: Nature 2001

Data in Information Systems

Molecular Basis of Living Systems

- A gene is a unit of information within the chromosome that can be inherited
- Expression of genomic information involves a complex sequence of steps
  - DNA to mRNA (transcription)
  - mRNA to proteins (translation)
- DNA contains only "alphabets"
  - Nucleotides
- The next significant step will come from deep understanding of protein expressions and interactions
  - *Functional* genomics

101

- Information systems & Molecular Biology
  - Requires knowledge of databases
  - Data communication: Internet
  - Data type, i.e. Molecular biology
- Understanding basic Principles of
  - Molecular biology
  - BioChemistry
  - Molecular genetics
  - 101 will be provided (pdf)

Key Components

- Cell
- Cell Nucleus
- Chromosome

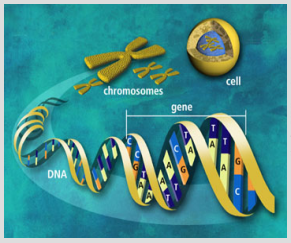
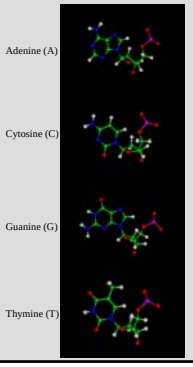


Photo from DOE Human Genome Project, ORNL

- Gene:
  - smallest physical unit of heredity coding: information carrier of a feature
- Let us further decompose a gene …

Building Blocks: Nucleotides

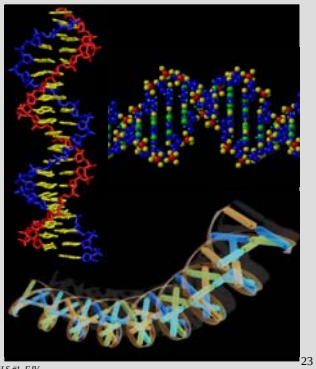
- A nucleotide is the building block of DNA and RNA
  - Nucleotide = bases + sugars + phosphate
- Bases:
  - Adenine
  - Cytosine
  - Guanine
  - Thymine, replaced by Uracil in RNA
- Complementary pairs
  - A complements with T
  - C complements with G



Diagrams from NYU. G=carbon B=nitrogen R=oxygen W=hydrogen P=phosphorus

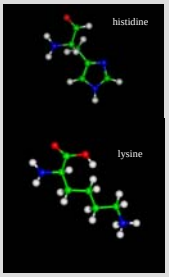
Nucleic Acids

- DNA and RNA
- In eukaryotes, DNA most commonly occurs as a double helix
  - Sugar-phosphate backbone on outside
  - Base paired by hydrogen bonds stacked on the inside
- DNAs are highly stable
  - Dipole-dipole interactions
  - Hydrophobic
- Complementary chains
  - The sequence of bases in one chain determines the sequence in the other chain



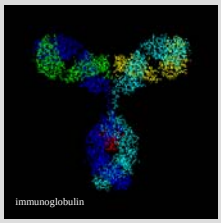
Amino Acids

- Translation
  - Bases of mRNA are translated in groups of 3 (codons)
  - A codon translates into amino acid
    - CAU, CAC -> His
    - AAA, AAG -> Lys
    - Etc.
  - Specific codons for each of the amino-acids
- Proteins are chains of amino acids
  - 20 amino acids
  - 20 letter alphabet



Proteins

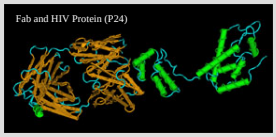
- Proteins are chains of amino acids
  - 20 amino acids,
  - 20 letter alphabet
- Variety of functions
  - Enzymes
  - Membrane receptors
  - Transport (e.g., hemoglobin)
  - Structure (e.g., collagen)
  - Nutrition (e.g., ovalbumin)
  - Immunity (e.g., antibodies)
  - Regulatory

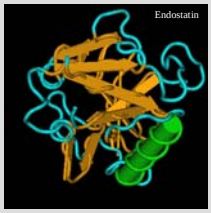


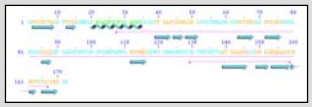
Protein Structure

- Complex structures
  - Primary – sequence
  - Secondary – α-helix and β-sheets
  - Tertiary – folding structures
  - Quaternary – multi-chain (multimeric) arrangements
- Protein structure determines function
  - Where is the active site(s)
  - What is the catalytic strength
  - Interaction in complexes

Examples Protein Structure







Key Principles Molecular Biology

- DNA acts as a template to replicate itself
- DNA is transcribed into RNA
- RNA is translated into a protein
- Sequence and structural homology (similarity) between molecules can be used to infer structural and functional similarity

Fields of Application

- Genomic sequencing
- Comparative genomics
  - Comparisons to find similarities/differences
- Expression quantification
  - Relative abundance of expression during development or disease stages
- Functional genomics
  - Large-scale mapping of gene functions and associations
- Proteomics
  - Catalog of activities that characterize interactions among gene products
- Structural genomics
  - Protein structure mapping and predictions
- Research informatics and data management
  - Experimental data management

Applications of BioInformatics

- Sequence analysis, Alignment
- Comparison BLAST, FASTA
- Molecular modeling, Prediction modeling
- Databases for EST's, Sequences (HGP), Linkage Maps (Syntenies), Physical Maps, Probes, Gene Array data.
- Databases of Gene Expression

Bio Molecular Databases

- Used for 3 major tasks
  - Lookup
    - Is there a gene known for my protein?
    - Is there mutations known causing this disease?
  - Compare
    - Are there sequences available resembling my cloned protein?
    - Are these two sequences similar (to what extent)?
  - Predict
    - Can the active site residues of the this enzyme be predicted?
    - Can a 3D model of this protein be made?
- Answers are not necessarily found in "1" database
- Combined search, Integrate results
  - How can this be realized
  - Interoperable databases

Core Databases

- Sequence search, BLAST:
  - Basic Local Alignment Search Tool
  - http://www.ncbi.nlm.nih.gov
- Protein structure: PHD
  - http://www.embl-heidelberg.de/predictprotein/predictprotein
- Molecular modeling and imaging: RasMol
  - http://www.umass.edu/microbio/rasmol/

Core Databases

- Data repositories
  - GenBank: NCBI Nucleotide database
  - Protein DataBank (PDB): http://www.rcsb.org/pdb
  - Repository for processing and distribution of 3D biological macromolecular structure data
- KEGG
  - Kyoto Encyclopedia of Genes and Genomes
  - From Genes to BioChemical Pathways

Searching in Databases

- Key issue is the different information that is available in the different databases
- Added value is obtained if these databases are transparently accessible
- Information is shared!
- Learning the specific contents of a database
- Ontology's

Searching in Sequences

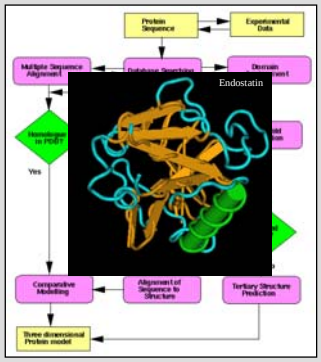
- Complications
  - Sequence DBs contain enormous amounts of nucleotides
  - Query sequence is not exact
  - It is important to find non-exact matches (homologues)
- Techniques
  - Sequence alignments
  - Multiple sequence alignments
  - Sequences of common structure or function

Sequence Alignment

Drosophila "eyeless" (S) gene vs human aniridia (Q)

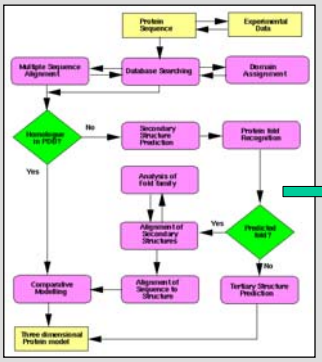
```
pir||A41644 homeotic protein aniridia - human
      Length = 447

 Score = 256 bits (647), Expect = 5e-67
 Identities = 128/146 (87%), Positives = 134/146 (91%), Gaps = 1/146 (0%)

Query: 24 IERLPSLEDMAHKGHSGVNQLGGVFVGGRPLPDSTRQKIVELAHSGARPCDISRILQVSN 83
          I R P+  M + HSGVNQLGGVFV GRPLPDSTRQKIVELAHSGARPCDISRILQVSN
Sbjct: 17 IPRPPARASMQNS-HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSN 75

Query: 84 GCVSKILGRYYETGSIRPRAIGGSKPRVATAEVVSKISQYKRECPSIFAWEIRDRLLQEN 143
          GCVSKILGRYYETGSIRPRAIGGSKPRVAT EVVSKI+QYKRECPSIFAWEIRDRLL E
Sbjct: 76 GCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEG 135

Query: 144 VCTNDNIPSVSSINRVLRNLAAQKEQ 169
           VCTNDNIPSVSSINRVLRNLA++K+Q
Sbjct: 136 VCTNDNIPSVSSINRVLRNLASEKQQ 161
```

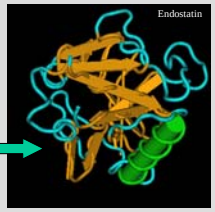
Structure Prediction, Drug design



Adapted from EMBL

Structure Prediction, Drug design





Adapted from EMBL

Key definitions & relations

- Cell % Cell Nucleus
- Cell Nucleus % Chromosome
- Chromosome % Gene
- Gene % DNA

- % = has part
- Or reverse: is part of
- ontology

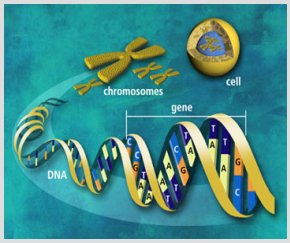


Photo from DOE Human Genome Project, ORNL

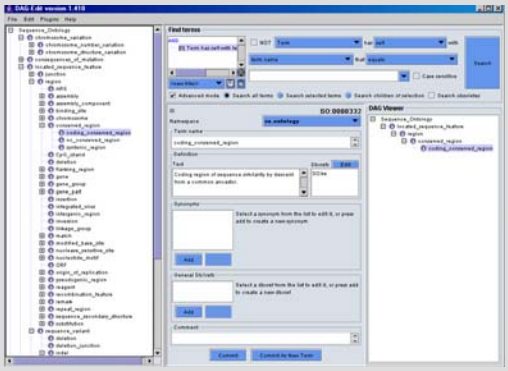
Facts on Ontology

- Share common understanding of a domain
- Make domain knowledge explicit
- There is no explicit method of writing an ontology
  - Depending on the application in mind
  - Obtained through iteration
- Concepts
  - Objects & Relationships in domain of interest
  - Nouns & Verbs in domain to be described
- Biology (Life Sciences)
  - GO-BO initiative coordinated by EBI

Open Biological Ontologies (OBO)

- Produced with DAG edit
  - Directed Acyclic Graph
- Ontologies stored in MySQL database
- Regular updates of the ontologies submitted to the database
  - Sequence ontology
  - Microarray Gene Expression Data (MGED)
  - Generic Model Organism databases

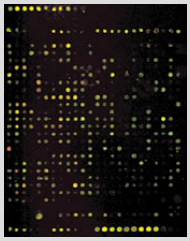
DAG Edit



DNA-chip

- Glass chip consisting of array of spots,
  - each spot 20-100 μm diameter
- Each spot contains a RNA of interest – "probe"
- Fluorescently tagged mRNA samples flow over probes
- Two-color fluorescence can be used to identify:
  - normal from abnormal,
  - over-expression from under-expression, etc.

MicroArray



- DNA chip a.k.a.
  - = MicroArray
  - = DNA Array
- Miniaturization
- Temporal-Information
- Little Space Information
- Lots of genes tested at the same time
- Renders a pattern of gene expression

Applications of MicroArrays

- Genomics
  - Fundamental research
  - Systems biology
- Toxico genomics
  - Field of functional genomics focusing on environmental health
  - Samples are taken from
    - environment or
    - from a food production process
- Food genomics

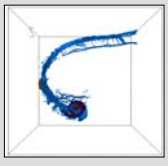
Spatio-Temporal Frameworks

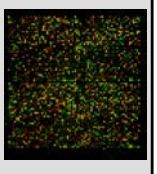
- Lots of data are generated which are not stored in one single repository
- Link repositories, create conditions to make that feasible
- Model system specific
  - Fast (short generation time)
  - Slow, mammalian, close to human genome (rat, mouse)
- Gene-expression can be applied on
  - Micro-arrays
  - In situ (in vivo), whole mount
  - Different model systems
- Relate gene expression to a model system
- Relate model systems

Patterns of Gene Expression

translation

topro anti-tubulin



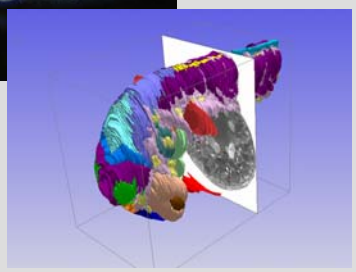


transcription

wnt1

3D Atlas Zebrafish Development

48 hrs.pf.
3.1 mm

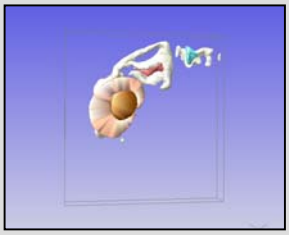


200 Mb image data,

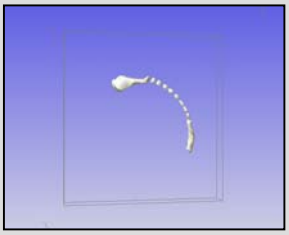
< 2 Mb annotations

"FISH" Gene Expression Patterns

- CLSM images
- Imaging protocol for processing

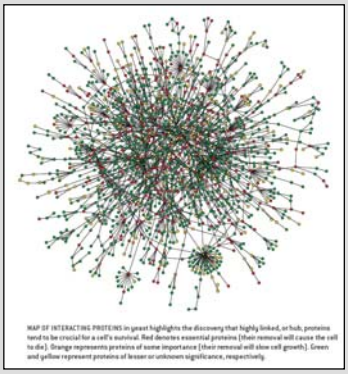


wnt1 (24 hrs pf.)



myoD (24 hrs pf.)

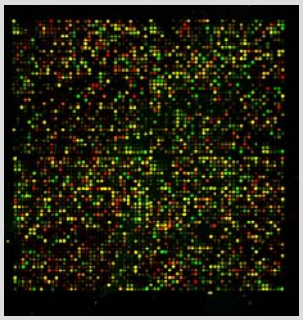
Integration of Information



Finding new genes

- Genetic code is an alphabet (4 letters)
- Genome is a string "…ATTGCGTA …" (very long)
- Looking for meaning in that string
  - Codon: group of 3 coding for an amino acid
- Scanning for Open Reading Frames (ORF)
  - Start codon
  - Stop codon
  - Complication: Intron (non-coding) and Exon (coding)
- Algorithmic approaches, all available data i.e.:
  - Molecular Biology,
  - Model Fitting to intron-exon boundaries,
  - Similarity to other organisms

When is a gene expressed



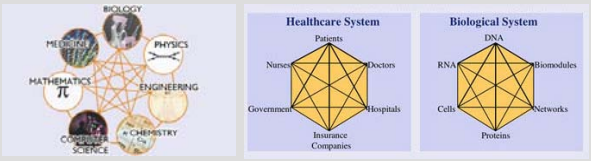
- MicroArray (DNA Chip)
- Large collection of genes in 20-100 µm spots
- Two samples
- Different conditions
- Lots of data
  - Co-expression
  - Level of expression
  - Relations between genes
- Cluster analysis
  - Algorithmic approach
- Database
  - Storage, Retrieval
  - Integration with other data

Where is a gene expressed

- MicroArrays (and other tools): temporal expression
- Microscopy of whole organism with *in situ hybridisation*: spatial expression (3D)
- Microscopy of whole organism with *immuno-histochemistry*: spatial expression (3D)
- Combine different images
  - Built a framework to look for expression
  - Learn about the genetic networks underpinning a function

Systems Biology

- Molecular components of a system
- Understand interactions at system leven
- Integration of different data
- Integration of different disciplines

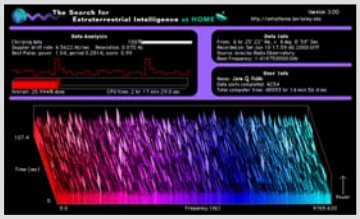


Data Deluge

Solutions to Data Deluge

- GRID
  - Bringing data to computer power, Sharing data
  - High energy physics (CERN)
- E.g.: SETI
  - Searching for Extra-Terrestrial Intelligence
  - Distributing data to grid of computers
  - Berkeley University CA, USA
- E.g. compound screening (chemistry)
  - Looking for right molecule as cancer drug
  - Lifesaver project, Oxford university, UK
- E.g. VL-e
  - Using the GRID for a Virtual Lab (environment)

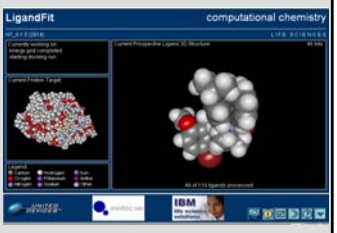
Sharing data & Computing power



- Search for Extraterrestrial Intelligence
- program that downloads and analyzes radio telescope data

Screensaver

- Oxford sends molecules and protein targets to UD.com
- UD.com Global MetaProcessor Grid sends tasks to members
- Screensaver processes the tasks and returns the results to central UD.com server
- Computers are typically idle 90% of the time.

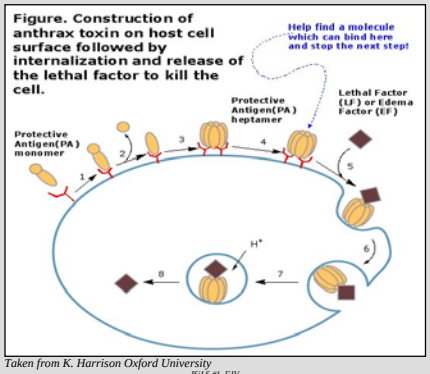


LifeSaver: Computational Chemistry

Two approaches used in the project

- THINK – Keith Davis
  - Cancer and Anthrax targets
- LigandFit – Accelrys Inc.
  - Cancer and Smallpox targets

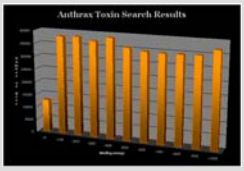
LifeSaver & Anthrax

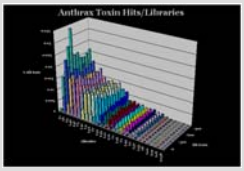


Taken from K. Harrison Oxford University

Results THINK: Anthrax

- Four Weeks
- 376, 064 molecules as hits from the 3.5 billion screened.





The Questions in ISiLS

The Questions

- What are good information systems
- Why is it a good information system (HCI)
  - Success (measurable)
  - tools offered
- How to improve the information system
- What extra tools & techniques are
  - Required
  - To be developed

Summary

- Typical issues in BioTech Information systems
- BioInformatics
- Molecular Biology primer
- Databases & Frameworks
- Examples

- Case studies are worked out in this course