Course objectives

- How do you become a good bioinformatics problem solver?
  - You need to know basic analysis and data mining modes
  - You need to know some important backgrounds of analysis and prediction techniques (e.g. statistical thermodynamics)
  - You need to have knowledge of what has been done and what can be done (and what not)
- Is this enough to become a creative tool developer?
  - Need to like doing it
  - Experience helps

At the end of this course...

- You will have seen a couple of algorithmic examples
- You will have got an idea about methods used in the field
- You will have a firm basis of the physics and thermodynamics behind a lot of processes and methods
- You will have an idea of and some experience as to what it takes to shape a bioinformatics tool

Contents (tentative dates)

<table>
<thead>
<tr>
<th>Date</th>
<th>Lecture Title</th>
<th>Lecturer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 wk 19</td>
<td>Introduction</td>
<td>Jaap Heringa</td>
</tr>
<tr>
<td>2 wk 19</td>
<td>Microarray data analysis</td>
<td>Jaap Heringa</td>
</tr>
<tr>
<td>3 wk 20</td>
<td>Molecular simulations &amp; sampling techniques</td>
<td>Antton Feenstra</td>
</tr>
<tr>
<td>4 wk 21</td>
<td>Introduction to Statistical Thermodynamics I</td>
<td>Antton Feenstra</td>
</tr>
<tr>
<td>5 wk 21</td>
<td>Introduction to Statistical Thermodynamics II</td>
<td>Antton Feenstra</td>
</tr>
<tr>
<td>6 wk 23</td>
<td>Machine learning</td>
<td>Elena Marchiori</td>
</tr>
<tr>
<td>7 wk 23</td>
<td>Clustering algorithms</td>
<td>Bart van Heute</td>
</tr>
<tr>
<td>8 wk 24</td>
<td>Support vector machines and feature selection in bioinformatics</td>
<td>Elena Marchiori</td>
</tr>
<tr>
<td>9 wk 24</td>
<td>Databases and parsing</td>
<td>Sandra Smit</td>
</tr>
<tr>
<td>10 wk 24</td>
<td>Ontologies</td>
<td>Frank van Harmelen</td>
</tr>
<tr>
<td>11 wk 25</td>
<td>Benchmarking, parallelisation &amp; grid computing</td>
<td>Thilo Kielmann</td>
</tr>
<tr>
<td>12 wk 25</td>
<td>Method development I: Protein domain prediction</td>
<td>Jaap Heringa</td>
</tr>
<tr>
<td>13 wk 25</td>
<td>Method development II</td>
<td>Jaap Heringa</td>
</tr>
</tbody>
</table>

Bioinformatics

“Studying informatic processes in biological systems”
(Hogeweg)

“Information technology applied to the management and analysis of biological data”
(Attwood and Parry-Smith)

Applying algorithms and mathematical formalisms to biology (genomics)
This course

• General theory of crucial algorithms (GA, NN, HMM, SVM, etc.)
• Method examples
• Research projects within own group
  – Repeats
  – Domain boundary prediction
• Physical basis of biological processes and of (stochastic) tools

Genomic Data Sources

• DNA/protein sequence
• Expression (microarray)
• Proteome (x-ray, NMR, mass spectrometry, PPI)
• Metabolome
• Physiome (spatial, temporal)

Bioinformatics

Large - external
(integrative)

Science
Planned Science
Population Biology
Sociobiology
Systems Biology
Biology

Human
Cultural Anthropology
Sociology
Psychology
Medicine

Small – internal (individual)

Molecular Biology
Chemistry
Physics

Protein structural data explosion

Protein Data Bank (PDB): 14500 Structures (6 March 2001)
10900 x-ray crystallography, 1810 NMR, 278 theoretical models, others...

Mathematics
Statistics
Computer Science
Informatics

Bioinformatics inspiration and cross-fertilisation

Bioinformatics

Chemistry
Biology
Molecular biology
Mathematics
Statistics
Computer Science
Informatics
Physics
Medicine

Algorithms in bioinformatics

• string algorithms
• dynamic programming
• machine learning (NN, k-NN, SVM, GA, ...)
• Markov chain models
• hidden Markov models
• Markov Chain Monte Carlo (MCMC) algorithms
• stochastic context free grammars
• EM algorithms
• Gibbs sampling
• clustering
• tree algorithms (suffix trees)
• graph algorithms
• text analysis
• hybrid/combinatorial techniques and more…
Joint international programming initiatives

- Bioperl
  http://www.bioperl.org/wiki/Main_Page
  http://bioperl.org/wiki/How_Perl_saved_human_genome

- Biopython
  http://www.biopython.org/

- BioTcl
  http://wiki.tcl.tk/12367

- BioJava
  www.biojava.org/wiki/Main_Page

Integrative bioinformatics @ VU

Studying informational processes at biological system level

- From gene sequence to intercellular processes
- Computers necessary
- We have biology, statistics, computational intelligence (AI), HTC, ...
- VUMC: microarray facility, cancer centre, translational medicine
- Enabling technology: new glue to integrate
- New integrative algorithms

- Goals: understanding cellular networks in terms of genomes; fighting disease (VUMC)

Bioinformatics @ VU

Progression:

- DNA: gene prediction, predicting regulatory elements, alternative splicing
- mRNA expression
- Proteins: (multiple) sequence alignment, docking, domain prediction, PPI
- Metabolic pathways: metabolic control
- Cell-cell communication

Fold recognition by threading:

THREADER and GenTHREADER

Query sequence

Compatibility scores

Fold 1
Fold 2
Fold 3
Fold N

Polutant recognition by microarray mapping:

Query array

Compatibility scores

Cond. 1
Contaminant 1

Cond. 2
Contaminant 2

Cond. 3
Contaminant 3

Cond. N
Contaminant N

ENFIN WP4

- Functional threading
- From sequence to function
  - Multiple alignment
  - Secondary structure prediction, Solvation prediction, Conservation patterns, Loop enumeration
ENFIN WP4

- Functional threading
- From sequence to function
  - Multiple alignment
  - Secondary structure prediction, Solvation prediction,
    Conservation patterns, Loop enumeration

ENFIN WP5 - BioRange (Anton Feenstra)

- Protein-protein interaction prediction
- Mesoscopic modelling
- Soft-core Molecular Dynamics (MD)
  - Fuzzy residues
  - Fuzzy (surface) locations

Where are important new questions?

New neighbouring disciplines

- Translational Medicine
  A branch of medical research that attempts to more directly connect basic research to patient care
  Translational medicine is gaining importance in the healthcare industry, and in a form whose precise definition is in flux. In particular, including discovery and development, translational medicine typically refers to the "translation" of basic research into real therapies for real patients
  The emphasis is on the linkage between the laboratory and the patient's bedside, without a real disconnect. This is often called the "bench to bedside" definition

- Computational Systems Biology

- Neuro-informatics
  Neuroinformatics combines neuroscience and informatics research to develop and apply the advanced tools and approaches that are essential for major advances in understanding the structure and function of the brain

Translational Medicine

- “From bench to bed side”
- Genomics data to patient data
- Integration
Systems Biology

is the study of the interactions between the components of a biological system, and how these interactions give rise to the function and behaviour of that system (for example, the enzymes and metabolites in a metabolic pathway). The aim is to quantitatively understand the system and to be able to predict the system’s time processes

- the interactions are nonlinear
- the interactions give rise to emergent properties, i.e. properties that cannot be explained by the components in the system

Systems Biology

understanding is often achieved through modeling and simulation of the system’s components and interactions.

Many times, the ‘four Ms’ cycle is adopted:
- Measuring
- Mining
- Modeling
- Manipulating
A system response

Neuroinformatics

• Understanding the human nervous system is one of the greatest challenges of 21st century science.
• Its abilities dwarf any man-made system - perception, decision-making, cognition and reasoning.
• Neuroinformatics spans many scientific disciplines - from molecular biology to anthropology.

Neuroinformatics

• Main research question: How does the brain and nervous system work?
• Main research activity: gathering neuroscience data, knowledge and developing computational models and analytical tools for the integration and analysis of experimental data, leading to improvements in existing theories about the nervous system and brain.
• Results for the clinic: Neuroinformatics provides tools, databases, models, networks technologies and models for clinical and research purposes in the neuroscience community and related fields.
Bioinformatics @ VU

Qualitative challenges:
- High quality alignments (alternative splicing)
- In-silico structural genomics
- In-silico functional genomics: reliable annotation
- Protein-protein interactions.
- Metabolic pathways: assign the edges in the networks
- Fluxomics, quantitative description (through time) of fluxes through metabolic networks
- New algorithms

Bioinformatics @ VU

Quantitative challenges:
- Understanding mRNA expression levels
- Understanding resulting protein activity
- Time dependencies
- Spatial constraints, compartmentalisation
- Are classical differential equation models adequate or do we need more individual modeling (e.g. macromolecular crowding and activity at oligomolecular level)?
- Metabolic pathways: calculate fluxes through time
- Cell-cell communication: tissues, hormones, innervations

Need 'complete' experimental data for good biological model system to learn to integrate

Bioinformatics @ VU

VUMC
- Neuropeptide – addiction
- Oncogenes – disease patterns
- Reumatic diseases

Bioinformatics @ VU

Quantitative challenges:
- How much protein produced from single gene?
- What time dependencies?
- What spatial constraints (compartmentalisation)?
- Metabolic pathways: assign the edges in the networks
- Cell-cell communication: find membrane associated components

Integrative bioinformatics

Data integration

Integrate data sources
Integrate methods
Integrate data through method integration (biological model)
Bioinformatics

“Nothing in Biology makes sense except in the light of evolution” (Theodosius Dobzhansky (1900-1975))

“Nothing in Bioinformatics makes sense except in the light of Biology”