Introduction to Bioinformatics

Dortmund, 16.-20.07.2007

Lectures:
Sven Rahmann

Exercises:
Udo Feldkamp, Michael Wurst

Goals of this course

- Learn about
  - Software tools
  - Databases
  - Methods (Algorithms) in bioinformatics
- Know what's out there (too much!)
- Acquire first experience with a selection of standard tools

Overview

- Monday: lectures and exercises
  - What is bioinformatics?
  - Literature databases
  - Bioinformatics databases
  - Sequence analysis
    - pairwise sequence alignment (theory)

Overview

- Tuesday: all-lecture day
  - Sequence Analysis (cont'd)
    - sequence database search (BLAST, BLAT)
    - multiple sequence alignment (CLUSTAL)
    - protein domain analysis (HMMs)
  - Phylogenetics
  - Protein structure
  - Transcriptomics: Gene expression analysis

Overview

- Wednesday: all-exercise day

Overview

- Thursday: lectures and exercises
  - Networks
  - Systems biology

Overview

- Friday: exam day
  - oral exams in small groups
  - questions
  - practical exercises at the computer

What is Bioinformatics ?

What is Bioinformatics?

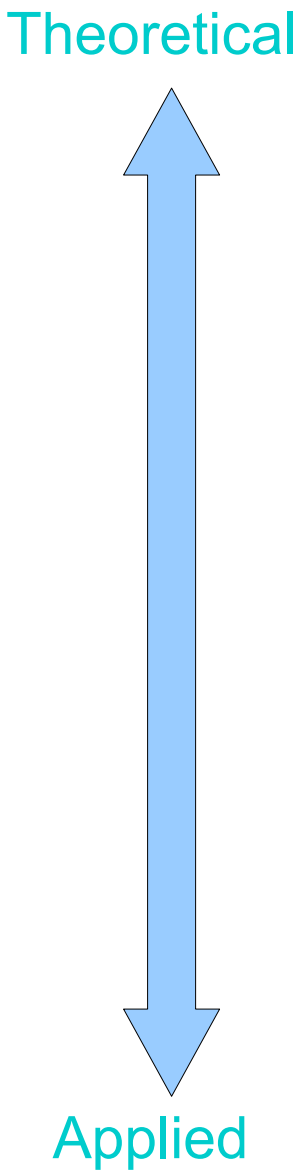
- Biology: bio = life, logos = science
- Earlier centuries: cataloging life forms
- Today: molecular biology (discovery of DNA)
- Basis of modern molecular biology: chemistry
- Life = islands of order or information in chaos
- Information = deviation from randomness
- Informatics: information processing
- Bio-informatics: natural combination

Bio-?

- Biology
- Bio-? :=
  - Science ? helps to understand biology
  - Biology inspires new research directions in ?
- Biochemistry
- Biophysics
- Biotechnology
- Biomathematics
- Bioinformatics

Bioinformatics – a wide field

- Biomathematics
- Theoretical biology
- Ecology
- Biostatistics
- Sequence analysis
- Computational biology
- Bioinformatics
- Systems biology
- Computational *-omics
  - genomics, transcriptomics, proteomics, ...
- Applied or practical bioinformatics



Definition (for this course)

- **Bioinformatician** :=
  person who uses models, methods, programs from computer science and mathematics to solve problems arising in the molecular life sciences

- **Bioinformatics user** :=
  person who uses bioinformatics software

Informatics in Biology

- Management of large amounts of data
  - Databases, Data warehouses
  - Laboratory Information Management Systems (LIMS)
- Analysis of large amounts of data
  - efficient algorithms
  - fast computers and other hardware
- Experiment design
  - most new knowledge at lowest cost
- Simulations
  - avoid expensive lab work altogether

Contrast: DNA Computing

- Bioinformatics is **not** DNA computing.
- DNA computing :=
  Using DNA to solve computational problems

- DNA is an information-storing molecule and can "react" to changes in the environment:
  It can be used as a computational device.
- Adleman (1994) solved the 7-point Hamiltonian path problem with DNA molecules:
  "Molecular Computation Of Solutions To Combinatorial Problems". Science 266(11): 1021–1024

Know your Bioinformatician

- Theoretician?
- Modeler?
- Software Engineer?
- Programmer? Language?
- Database developer?
- Biologist with computer training?
- Lab experience?
- ...

About myself

- Diploma in mathematics (applied probability) (statistics of sequences)
- PhD in bioinformatics (efficient algorithms for oligo microarray design)
- Research group leader computational methods for emerging technologies (in the life sciences)
- Main job:
  - “Extract” computational essence or model from a real-world problem
  - Develop methods for solving it
  - translate back results

How I like to work

- Learn about an interesting problem
  - by chance, or by actively seeking a new one
- Gather information about the problem
  - talk to people, read review papers, who else?
- Wait for new clever ideas ...
- Try out (and frequently modify) these ideas
- Turn ideas into a software product
- Write the publication

Example: Microarray Design

- Microarrays contain 100 000s of DNA probes
- For gene expression analysis, probes must be transcript-specific (otherwise: crosshybridization)
- How to select probes for large arrays efficiently?

Example: Microarray Design

- Modeling: How to measure cross-hyb. risk?
  - binding energy?
  - percent identity between probe and transcript?
  - longest common substring (perfect match)?
- Algorithmics: Which of these allows fast algorithms? Which data structures are needes?
  - Fast LCS computation using enhanced suffix arrays
- Software:
  - input/output format?
  - language, operating system? (PERL, Java vs. C)

Recommended Reading

- JM Claverie and C Notredame:
  Bioinformatics for Dummies, 2nd ed. (2006)
  Wiley
- DW Mount:
  Bioinformatics: Sequence and Genome Analysis, 2nd ed. (2004)
  Cold Spring Harbor Laboratory Press

A Few Web Resources

- A lot of material and software in bioinformatics is freely available on the WWW.
- Good starting points:
  - NCBI: http://www.ncbi.nlm.nih.gov/
    (US National Center for Biotechnology Information)
  - Journal NAR (Nucleic Acids Research) at http://nar.oxfordjournals.org publishes
    - database issue
    - web server issue

    see DB list at http://www3.oup.co.uk/nar/database/c/
  - BiBiServ (Bielefeld Bioinformatics Server): http://bibiserv.techfak.uni-bielefeld.de/