Bioinformatics

Computational biology

Martin Saturka

http://www.bioplexity.org/lectures/

EBI version 0.4

Branches of computations in biology

Bioinformatics usage in common biology and bioindustry.

- phylogenesis and health care oriented research.
- approach targeting. experiment driven methods.

Main topics

- computational biology
  - evolution, cladistic trees
  - practice, restriction maps
- biomedical informatics
  - pharmaceutical usage
  - medical informatics
- laboratory
  - image processing, lims
  - DNA computers, robots

Genetics

classical analysis

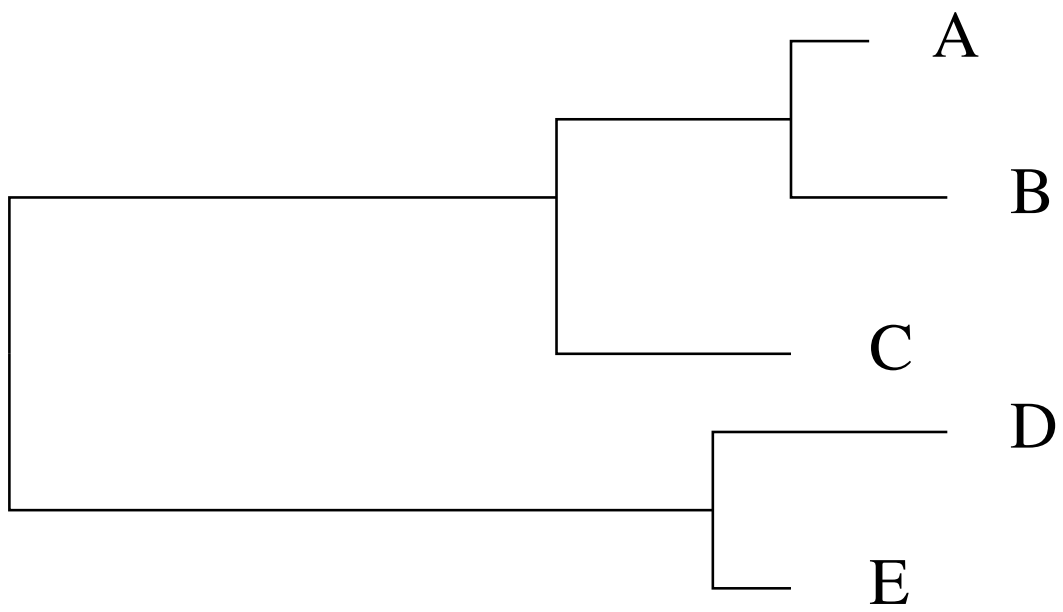
- linkage of traits
  - genetic markers
  - segregation independence
    - OLS, GLM: variance fraction explained
- hereditary diseases
  - cca 4000 genetic disorders known
  - cystic fibrosis - cca 5% in Europe/USA heterozygotes
    - intrigued advantage against cholera toxin, typhoid fever
- heredity
  - paternal: Y chromosome
  - maternal: mitochondria

Cladistics

- phylogenetic trees
  - tree structure
    - rooted vs. unrooted
    - bifurcating, multifurcating
  - distances
    - dendrograms - with distances
    - cladograms - without distances
  - taxonomic units
    - OTU - operational taxonomic units
    - TU, HTU, clades

- description
  - Newick format
    - (((A,B),C),(D,E))
    - (((A:1,B:2):3,C:3):7,(D:3,E:1):9)

Trees



- number of different (bifurcating) phylogenetic trees
  - n-th leaf: $2n - 5$ possibilities
  - unrooted tree: $(2n - 3)!/[2^{n-2} \cdot (n - 2)!]$
  - rooted tree: $(2n - 5)!/[2^{n-3} \cdot (n - 3)!]$
  - root with the help of an outgroup

Computational phylogenetics

evolution tree algorithms

- methods
  - distance matrix methods
    - UPGMA - unweighted pair group method with arithmetic mean
    - ME - minimal evolution
    - NJ - neighbor joining
  - character state methods
    - MP - maximum parsimony
    - ML - maximum likelihood

- tree approach
  - phenograms - without evolution history (UPGMA)
  - phylograms - with evolution history (other methods)

Tree construction - distances

algorithms based on distance evaluations

- UPGMA
  - distance matrix construction (e.g. alignment scores)
  - the nearest pair forms a new TU
    values on the new TU by the arithmetic mean
  - artificial results for different evolution speeds
- ME
  - to minimalize total branch lengths
  - good idea, however hard to compute
- NJ
  - ME approximation, frequently used
  - start with the star tree (i.e. maximal multifurcation)
  - iterate node joining to minimize total branch lengths

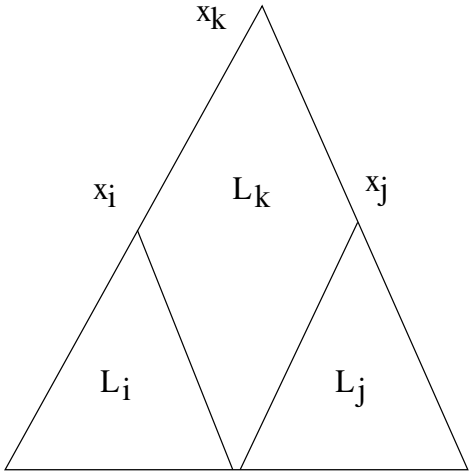
Tree construction - mutations

algorithms based on character mutations

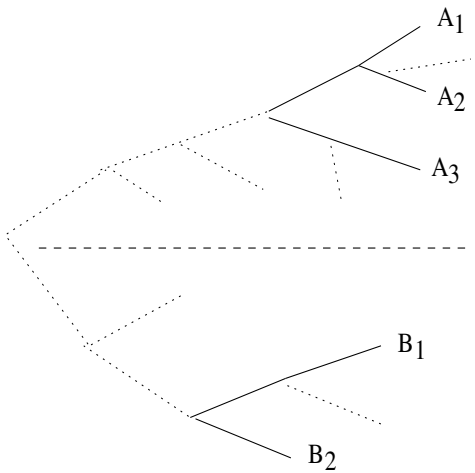
- MP
  - trees with minimal total amount of mutations
  - branching for separate positions
    final tree as a consensus on the positions
  - artificial results on long/fast evolution
    - long branch attraction
      long branches with near-root multiple bifurcation

- ML
  - multiple tree construction
  - probability of the sequence set generation by each tree
  - taking the tree of the maximal probability
  - good idea, however hard to compute
  - used in the case of small data sets

Branches



subtree probability $\Pr(L_k|x_k)$



branch separation

- Likelihood computation
  - $\Pr(L_k|a) = \sum_{b,c}[\Pr(L_i|b)\Pr(a \to b) \cdot \Pr(L_j|c)\Pr(a \to c)]$
  - $\Pr(L_k|x_k) = \sum_a \Pr(L_k|a) \cdot \Pr(a)$

Tree relevance

robustness of the constructed trees

- branch probability
  - bootstrapping on nucleotide positions
  - fraction of trees with a given fork
    - separation of two branches
  - good branches with probabilities above 0.9
- tree adjustment
  - EM - expectation maximization
  - topology changes like simulated annealing and genetic algorithms

practical usage

- daily laboratory tasks
  - restriction sites analysis
    - short palindromic sites serving for fragment gluing
    - databases of restriction sites available
  - primers construction
    - vectors usually constructed with specific primer sequences
  - partial digest analysis
    - had been used in the past for sequencing
    - branch and bound method
      take fragments from the largest
      check whether all the subfragments are present

Targeting

- population research (multiple units)
  - epidemiology, virus spreading
  - ecology systems, pest dynamics
  - evolution exploration
- health care (unique organisms)
  - drug development and testing
  - treatment / survival statistics
  - illness diagnosis and prognosis
- ethics
  - stem cell research, (human) cloning
  - experiments on humans/animals/computers
  - potential arms usage, abuse
  - nature diversity preservation

Targets - sequences

sequences exploration

- what to sequence
  - genomes, coding sequences
  - intra-species, inter-species comparisons
    disease study
- basic methods
  - fragment assembly, sequence localization
  - databases of sequence variations
- subsequent targets
  - gene prediction - comparisons, markers, HMMs
  - gene annotation - product structure and function

Targets - genes

genes - ontology

- gene structure
  - regulatory sequences, promoters
  - exons and introns, splicing sites
- gene comparison
  - evolution: homology, analogy
  - populations: SNPs, CNVs
  - product localization prediction
- gene function
  - coding genes, RNA genes
  - expression experiments
    function prediction

Targets - structures

3D structures

- function guessing
  - domains/folding prediction
    - protein cores, active sites
  - homology to other structures
    - sequence patterns
- drug development
  - protein surfaces
    - ligand targeting, activity modulation
  - protein localization
    - envelope protein search

Targets - expressions

microarray experiments

- gene expression
  - tissue, cell cycle, development experiments
  - illness, drug explorations
  - data preprocessing, normalization
- data exploring
  - data visualization
  - expression alteration statistics
  - illness diagnosis
  - gene characterization

Targets - relations

expression data

- data learning
  - separation into clusters
  - illness diagnosis methods
  - networks of gene expression
- data understanding
  - relation mining
  - characterization features
  - trend prediction

Health care

actual usage in pharmacology

- sequence approaches
  - SNPs, CNVs for liability to disease prediction
  - virus envelope protein search and immunization
  - cellular surface protein glycosylation
- expression approaches
  - gene expression alterations on drug usage
  - cancer diagnosis - appropriate for binary discrimination
  - patient time-series visualization

Medical informatics

patient information

- electronic medical records
  - continuation of health care
  - standards: CEN - EHRcom, HISA; DICOM, HL7, openEHR
  - organizations: EuroRec.org

- physician support systems
  - best practice guidelines / recommendations
  - decission support / expert systems
  - hospital information systems

LIMS - laboratory information management systems

- LIMS
  - chemicals, biologicals, experiments
  - sample centric, process centric approaches
  - data acquisition and processing
  - results: storage, access, review
  - request and experiment tracking

- literature
  - Pubmed Medline (pubmed.gov)
  - arxiv.org, citebase.org, citeseer.ist.psu.edu, eprints.org
  - text mining

Experiment management

laboratory automation

- bus / protocols
  - parallel: IEEE-488.2 (GPIB), SCPI
  - serial: RS-232/422/423/485, USB, CAN
- real-time OS
  - RTAI.org / Xenomai.org
  - comedi.org, scicos.org
- programming
  - kernel modules / user space, mainly in C
  - RTAI-Lab, graphical interface tool-chain

Image processing

digital signals and images

- microscope imaging
  - CCD cameras, confocal microscopes
  - fluorescent markers - GFP, etc.
  - 2D images, focal depthfs for 3D images
- medical imaging
  - MRI - magnetic resonance imaging
  - PET - positron emission tomography
  - EEG, EKG, radiography
- image transformations
  - fourier transform
  - filtering
  - edge detection

Microarray images

microarray data acquisition

- image processing
  - addressing - center locations
  - segmentation - signal vs. background pixels
    - seeded region growing
  - spot data
    - signal intensities, background, quality
- quality problems
  - spot overlap
  - high background
- data files
  - tiff format, without compression
  - grey/RGB for one/two channels

Systems biology

system function and behavior

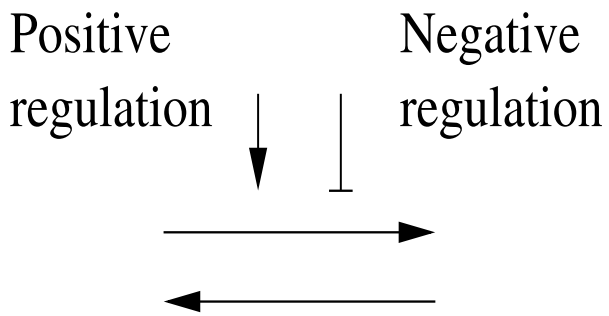
- description
  - dynamical interactions
    - not only statical classification
  - biomolecular processes
  - enzymatic kinetics
  - biochemical pathways

- methods
  - experiments
    - high-throughput methods
  - biochemistry
    - differential equations
  - biology
    - abstract machines

various -omics systems

- interactome
  - fundamental - what interacts with what
  - other: genomics, transcriptomics, metabolomics, etc.

- examples
  - GTPase signal transduction
  - MAPK cascade



Ontogenesis models

standard development

- cell cycle
  - passing through the cell cycle
  - checkpoints - nutrient limitation, DNA damage
    - checking at specific sections of the cell cycle
  - G_0 phase, the non-proliferating phase

- differentiation
  - subsequent cell 'speciation'
    final states non-proliferating
    - for higher multicellular organisms
  - dedifferentiation experiments

Immunology models

non-standard situations

- self-defense
  - physilogical vs. pathological states
  - understanding complex diseases
  - many situations without immediate symptoms
  - modelling long-time evolving

- health-care impact
  - cancer treatment
    - tumor evolution
  - virus infection
    - organism failing

DNA computations

usage of DNA complementarity

- DNA computers
  - hybridization of complementary strands
    - for NP complete problems
    - proof of concept, not a real usage
  - usage of enzymes on DNA
    - endonucleases and ligases for Turing Maching construction
    - currently simple non-TM machine constructed

- DNA nanotechnology
  - hybridization usage for complex 3D object construction
    - cubes of DNA molecules - structure by complementarity, enzymatic fixation
  - DNA robots
    - combination of DNA with other molecular complexes

Adleman experiment

- Hamiltonian path
  - initial data
    - cities: 20-mer oligonucleotides
    - paths between cities: 10+10 20-mer oligonucleotides
  - processing
    - generate random paths - by alignment
    - remove paths with wrong start or end
    - remove paths with not exactly *n* cities
    - remove paths with missing a city
    - any remaining path is a result

- limitations
  - too big amount necessary for large computations
  - accuracy problems for large data sets

NA robots

usage of NA complementarity

- DNA robots
  - mRNA (if present) hybridizes with sequeces on DNA robots
  - ssDNA hairpin $\leftrightarrow$ ssDNA released
    - stochastic changes, probabilities according to levels of the expressed mRNAs
  - proof of concept experiment

- small RNA molecules
  - current the most developed area
  - not a real nano-technology
    - just a modulator of natural cellular processes

Drug and food administration

biotechnology regulation

- therapeutics approval
  - clinical trials
    - cell cultures / animal / humans
    - four trial phases on humans
  - placebo controlled study
  - drug side-effects
  - clinicaltrials.gov, www.fda.gov
- genetic modification
  - GMO - genetically modified organisms
    - random, targeted alterations
    - composition, resistance
  - interference with ecosystems studies

Items to remember

Nota bene:

images, experiments, robots

- Practice
  - cladistic trees
  - sequence features
- Targets
  - sequences, genes
  - gene expression