Bioinformatics

Introduction: biology, genes

Martin Saturka

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

EBI version 0.5

Bioinformatics background

methods for dealing with huge amounts of biological data
experimental data: sequences, microarrays, structures

Distinctions

- theoretical bioinformatics
  - model and algorithm development
- technical bioinformatics
  - databases and computation systems
- computational biology
  - practical usage of bioinformatics tools

connections: data-mining, medical informatics, text parsing

Course topics

1. background knowledge
   - introduction into molecular biology
   - graphs and Hidden Markov Models - **essential!**
2. DNA sequences - genomes
   - fragment assembly, exact matching
   - approximate matching, heuristic algorithms
3. gene expression - microarrays
   - linear methods, factor analysis, scaling
   - clustering, nets, data-mining
4. structures and databases
   - structural biology, structure prediction
   - protein, RNA structures, genomes with genes
5. computational biology
   - task examples, gene ontology
   - bioinformatics software tools

Biology

Never say always.

- DNA and stability - (retro)transposomes
- horizontal gene transfer
- RNA viruses, viroids, virusoids, RNAzymes.

Easy to make misunderstandings

- accidental shapes, events.
- (non-existent) extinction periods
- DNA junks - necessary or trash?

Physics

It stands always.

- Relevant parts:
  - thermodynamics - what is possible
  - kinetic theory - how fast it is
- Mathematical point of view:
  - nonlinear dynamics
    - description of open systems
  - game theory
    - linear approximation near equilibria

Molecular interactions

- energy minimization:
  - hydrogen bonds, electrostatic, van der Waals interactions
- entropy maximization:
  - amount of accessible states

hydrophobic effect

- preferred interactions water-water
- count of water shell configurations

Enzymatic kinetics

Description based on graphs with steady state assumptions.

- Michaelis-Menten equation

$$E + S \underset{\longleftarrow k_{-1}}{\overset{k_1 \longrightarrow}{}} ES \quad k_2 \longrightarrow \quad E + P$$

$$v = v_{max} \frac{[S]}{[K_M]+[S]} \qquad v_{max} = k_2 * [S] \qquad K_M = (k_{-1} + k_2)/k_1$$

dual role of enzyme-substarte affinity

- higher affinity $\rightarrow$ faster ES formation
- lower affinity $\rightarrow$ faster P release

receptor-ligand binding alike

Central dogma

The central dogma of molecular biology:

$$\circlearrowleft \text{DNA} \rightarrow \text{RNA} \rightarrow \text{Proteins}$$

- DNA
  - **stable, data carrier, replication**
  - very weak possible enzymatic functions
- RNA
  - less stable, can be a data carrier as well
  - **transfer: transcription, translation**
  - substantial enzymatic functions - translation by rRNA
  - gene expression regulation
- Proteins
  - **structural, enzymatic functions**

DNA data

DNA: semi-conservative replication

- stem cells divisions
  - asymmetric strand distribution

high inter-species genome similarities

- why: computation like usage
  - how to do something, how to interact
  - no plans of final structures, organs

highly unrelated expression profiles
even for many conservative ORFs

NA molecules

Nucleic acids: DNA, RNA

- DNA double helix
  - two complementary strands
  - antiparallel directions
- mutation and repair
  - zygote divisions a critical stage
  - expressed genes: both mutations and repair
  - aging as a defence against cancer?
- RNA
  - data carrier for some viruses and alike
  - various structures, enzymatic functions

Chromosomes

several levels of packing

tight chromosomal structures formed during M phases

- components
  - DNA: centromeres, telomeres - repetitions
  - Proteins: histons, transcription factors
- chromatin
  - homochromatin - accessible to expression
  - hetrerochromatin - tightly packed

Terminology

DNA strands

- Chromosomes:
  - $+$ (plus) strand
  - $-$ (minus) strand

  Every position according to the plus strands!

- Genes:
  - coding strand - the 'same' sequence as mRNA
    - the 'same': transcribed RNA is heavily processed
    - can be on either $+$, $-$ strand of a chromosome
  - template strand - actually transcribed, i.e. complementary

Genes

- Positions:
  - promoter region: ranks up to -1
    - most of binding sites for transcription factors
  - transcribed region: ranks from 1
- Parts:
  - exons - expressed sequences
    - parts of final mRNAs, terminal and coding parts
  - introns - intervening sequences
    - enabling complex protein domains
- Splicing:
  - common for eukaryotes and Archea
  - alternatives - promoters, polyadenylation, introns/exons

Gene expression

- 3D (e.g. of human brain) gene expression maps

Transcription

- DNA $\rightarrow$ RNA
- alphabets: ATCG $\rightarrow$ AUCG

Translation

- RNA $\rightarrow$ Proteins
- triplets of AUCG $\rightarrow$ stopcodons + 20 aminoacyls

Reverse transcription

an inverse process: RNA $\rightarrow$ DNA

RNA types

- mRNA - information carrier

ncRNA - non-coding RNAs

- (transfer) tRNA - helper function
- (ribosomal) rRNA - translation function

- snRNA - splicing, transcription factors, telomeres
- snoRNA - rRNA processing
- (guide) gRNA - mRNA editing

- (micro) miRNA - mRNA inhibition
- (small interfering) siRNAi - RNA interference

Expression regulation

Gene expression is generally not a 'yes'/'no' process.

- DNA structure: regulation by methylation - GC pairs
- level of expression
  - by transcription initiation frequency - transcription factors
- other factors
  - transcription termination
    - prokaryotes
  - alternative splicing
    - eukaryotes
  - mRNA inhibition and/or degradation
    - both natural and therapeutic
  - mRNA editing

Cells

Prokaryota Archea Eukaryota

Cells - basic blocks of living organisms

exceptions: parasites - viruses, virusoids, viroids, what else?

- particular genes expressed on various levels
- physiological states: keeping the homeostatsis

- Eukaryotic cell - cellular membranes:
  - separartion from outer space
  - distinct inner compartments
- Nucleus - chromosome sets: haploid (n), dipoid (2n), etc.
  - pairs of antiparallel DNA molecules
  - many proteins (histons, polymerases, transcription factors),
  - various RNAs

Cell cycle

passing a cell throughout its divison cycle

$$G_1 \rightarrow S \rightarrow G_2 \rightarrow M \rightarrow G_1 \text{ phases}$$

G_0 is the off cycle phase

$\circlearrowleft$ stem cells $\rightarrow$ differentiation

- cell death: apoptosis vs. necrosis
- cancer: two necessary conditions
  - immortal (*stem cells* - each tumor?)
  - out of the contact inhibition

Regulation

Signalling pathways

- receptors - ligands
- autocrine, paracrine, endocrine signalling

Protein phosphorylation

- regulation of enzymes, receptors
- kinases vs. phosphatases

Protein degradation

- ubiquitin proteasome system

Inter-cellular interactions

Extracellular matrix

- 'outer cytoskeleton'
- cell adhesion, interaction mediation

Immune system

- over-feeding - necessary for survival
- low dirt exposure $\rightarrow$ allergy (hypothesis)

Self vs. non-self distinction

- 'basic instinct' of living matter
- markers: saccharide surfaces

Populations

- species
  - evolution dynamics
  - partially understood
- organisms
  - population dynamics
  - deeply understood
- genes: 'selfish gene'
  - competition inside DNA strands
  - competition between organisms

Population dynamics

basic strategies

- r - high growth rate
- K - capacity utilization

logistic growth

$$\dot{x} = r \cdot x \cdot (1 - x/c)$$

$$x(t) = c \cdot \exp rt / (\exp rt + s)$$

Red queen principle

Perpetual competition

”It takes all the running you can do, to keep in the same place.”

- parasites vs. hosts
  - any new attack or defence evokes a counter-action
- trees of tropical forests
  - tree hights are individual drawbecks
  - tree hights are competition necessity

Evolution issues

- Evolution basis

the same basis as the reduction to the molecular level

clashes alleged to religions like flat-earth clashes

- Gene duplications
  - new weak accidental functions of genes
  - subsequent function improvements

- Extinctions - exponential process
  - formerly a wrong cycle proposed based on half-time

Linear games

- Small changes around an equilibrium.
  - suitable linear approximation
  - many natural populations obey it

Zero vs. non-zero games

- Non-zero games
  - possible cooperation
  - targeting win-win strategies
- Zero games
  - attrition wars
  - misunderstanding: mercantelism

Game example

The least loss strategy

- outfit is not worsened by strategy changes of competitors
  - nature tries everything and (immediately) penalizes

	I q	II (1-q)
A p	0.85	0.70
B (1-p)	0.60	0.90

example - virus I/II, vaccine A/B
q - virus type probability
p - vaccine usage fraction

$$E(p,q) = 0.85pq + 0.7p(1-9) + 0.6(1-p)q + 0.9(1-p)(1-q)$$
$$E(p,q) = q(0.45p - 0.3) + 0.9 - 0.2p \rightarrow p' = 2/3$$
$$E(p,q) = p(0.45q - 0.2) + 0.9 - 0.3p \rightarrow q' = 4/9$$
$$E(p',q) = E(p,q') = 0.7667$$

Game strategy

evolutionary stable strategies

$$E(M,P) < E(P,P) \vee [E(M,P) = E(P,P) \wedge E(M,M) < E(P,M)]$$

- P - population, M - mutation
  - qualitative estimation of partial derivatives

- Hawks vs. Doves: population: P = pH + (1-p)D
  - hawks as the mutation, for doves alike

pay-offs	hawk	dove
hawk	-25, -25	50, 0
dove	0, 50	15, 15

hawks: 7 / 12
doves: 5 / 12

$$E(P,P) = -25p^2 + 50p(1-p) + 0(1-p)p + 15(1-p)^2$$
$$E(H,P) = -75p + 50 \quad E(P,H) = -25p \quad E(H,H) = -25$$

Gender games

- to make most offsprings for least energy

males vs. females

- sex/progeny costs
- offspring feeding

- Restraints:
  - females: to force partner to spend energy
  - males: not to take care about other genes

- necessary female cooperation
  - can result in killing a non-cooperating female

Data acquisition

Data types

- sequences: what is it similar to?
  - sequencing fully automated
    - enzymatic polymerization, fluorescence detection
    - alternatives: pyrosequencing, nanopore sequencing, solid-phase sequencing
- gene expressions: what is it coregulated with?
  - acquisition partially automated, progression
    - DNA chips - microarrays: hybridization, fluorescence
    - protein chips, surface plasmon resonance, *in-situ* methods
- structures: how does it look like?
  - methods: RTG diffraction, NMR spectroscopy
  - alternatives: electron microscopy, spectroscopy, AFMs

Molecular biology methods

Enzymatic reactions

- polymerization - PCR, reverse transcription
- restriction endonucleases, ligases, etc.

Genetic material transfer

- vectors: plasmids, viruses, artificial chromosomes

Pairing / binding

- nucleic acids hybridization - blotting
- proteins: antibodies - antigens

Biophysics methods

Spectroscopy

- fluorescence
- NMR, IR, Raman

Microscopy

- electron microscopes
- AFMs

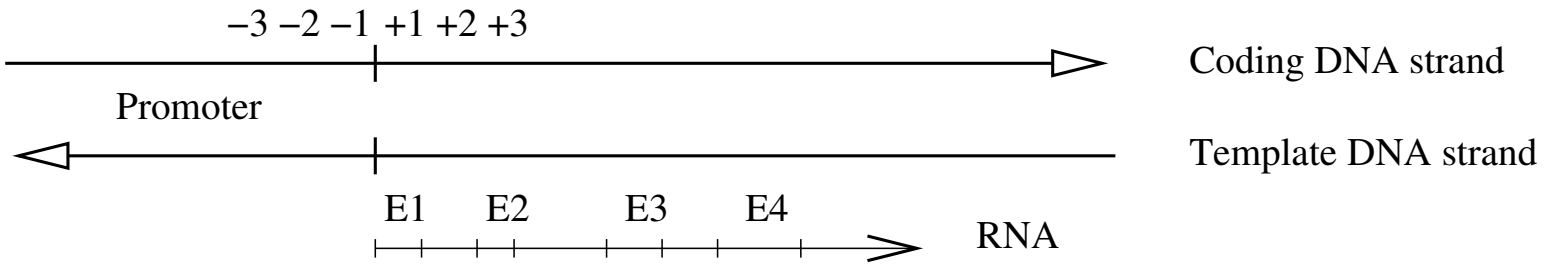
Diffraction

- X-ray crystallography

Items to remember

Nota bene:

- Gene structure
  - promoter, exons, introns
  - positions according to plus strands



- Gene expression
  - transcription, translation
  - regulation on several levels