Introduction to Computational Biology

Review of the material
For the test

Bartek Wilczyński

May 26th 2020

DNA sequences and graphs

- Biological sequences, DNA, RNA, proteins,alphabets, trascription DNA→RNA, translation RNA→protein, base complementarity, DNA replication, genetic code
- K-mers, k-mer spectrum, sequencing by hybridization, deBruijn graphs, Hamiltonian and Eulerian path approaches
- Microarrays, unique array probe design, DNA melting temperature

Biological sequence evolution

- The hypothesis of the „tree of life” with all DNA originating from the same sequence
- Species vs. Organisms and species evolution vs. DNA evolution
- DNA replication, mutations, selection on the phenotype
- Parsimonious evolutionary models, ancestral sequence, time-reversibility
- Markov models of DNA evolution(JC69, K80, F81), parameters (rate matrix), estimation of evolutionary divergence time, parameter estimation
- Protein substitution matrices, PAM, BLOSUM, log-odds vs multiplicative models

Pairwise sequence comparison

- Hamming distance, types of mutations (silent or coding)
- Types of errors during replication (base errors, non-homologous recombination, etc) and resulting mutations
- Edit distance and editing scenarios – definition and calculation
- Sequence alignment – definition, relation to the edit distance, local and global variant
- Dynamic programming approaches to computing alignments, local and global
- Fixed and affine gap penalties

Phylogenetic tree reconstruction

- Distance vs similarity matrices
- Phylogentic trees: binary and star-like, rooted and unrooted
- Molecular clock hypothesis, ultrametric trees
- UPGMA algorithm for reconstruction of ultrametric trees
- Neighbor joining for reconstruction of trees from distance matrices

Multiple sequence alignment

- Ambiguity of ancestral sequence in multiple pairwise alignments
- Multiple sequence alignment (MSA) definition, sequence profiles
- The sum-of-pair metric for MSAs, naive dynamic-programming approach, np-completeness of the MSA problem
- Progressive alignment idea, profile alignment, CLUSTAL
- Improvements to the basic progressive alignment – MUSCLE and T-COFFEE

Markov models

- Markov model definition, transition and emission matrices
- Gaussian HMM models
- Viterbi and Baum Welch algorithms
- Markov models of higher order, CpG islands model
- HMMer – Markov models for protein domains
- Gene structure modeling: Variable Order Markov Models and Interpolated Markov Models

Searching for similar sequences

- Examples of highly similar gene sequences in very diverged species
- Biological sequence databases (Genbank) – their role and work principle
- Searching for related sequence problem: short query, long DB variant
- Heuristic approaches, identifying short identity hits - FASTA
- BLAST algorithm: the basic ideas behind the method, statistical model for evaluation (e-values, extreme value distribution)

Gene homology and gene functions

- Duplications and speciacion – basic events in gene evolution
- Homologs: Paralogs, orthologs, ohnologs and xenologs
- Faster divergence of paralogs and neo-functionalization
- Different types of gene families with respect to paralog count
- Bi-directional blast hits and clusters of orthologous genes
- Gene Ontology
- Statistical tests for functional overrepresentation: Fisher’s exact test and GSEA

Tree reconciliation

- Gene and species trees, evolutionary scenarios, reconciliation, gene losses
- Reconciliation costs: deep coalescence, duplication, losses
- LCA mapping and its properties with respect to cost functions
- Optimal reconciliation in the duplication-loss case
- Horizontal gene transfer idea and evolutionary networks (trees with transfers)
- Examples of applying reconciliation to gene families

DNA sequence motifs

- The role of non-coding DNA sequence
- DNA binding sites and gap-less alignments
- Count-, frequency- and log-odds matrices
- Sequence information content, entropy and motif logos
- Motif databases: JASPAR and TRANSFAC
- Motif seqrching problem: consensus method, Gibbs sampling and EM approaches

NGS read mapping

- Comparison of Sanger and NGS sequencing
- The problem of mapping milions of short sequences on a long genome
- Suffix trees, suffix arrays and Burrows-Wheeler transform
- FM-index for identification of short matches in low memory
- Errors and SNPs in sequencing, read splitting vs. back-tracking
- Even faster methods for read quantification: STAR and Kallisto

Genome assembly and metagenomics

- Different approaches of shotgun-sequencing vs traditional sequencing (Celera vs. Human genome consortium)
- Genome assembly using deBruijn Graphs, bubble and tip removal
- Metagenomic sequencing – the idea and associatied computational problems
- Metagenomic sequencing applications: sea, gut, soil metagenomes, faecal transplants