BC. CROSS-BORDER BIOINFORMATICS STATE EXAM TOPICS

Bioinformatics expertise

Bioinformatics basics and pairwise sequence alignment

- Provide an overview about the landscape of data bases relevant for BI, including the organizations behind and the methods to generate data.
- What is the general reason behind sequence alignment (pairwise and multiple)?
- How are scoring matrices for proteins constructed to reflect the biological plausibility of alignments? And for DNA?
- Explain the general idea of the Needleman-Wunsch algorithm. What is the difference to Smith-Waterman? When do you apply NW, and when NW?
- Explain the difference between NW or SW to BLAST.
- How could we assess sequence similarity without aligning them?

Multiple sequence alignment and phylogeny

- Can we extend the principle behind SW/NW to multiple sequence alignment?
- Explain a widely used approach to MSA
- How do we score a multiple alignment?
- What are the two major approaches to tree-building (out of a MSA)?
- Sketch ClustalW.

Microarrays

- Explain the working principle of a microarray machine. What is the interface to the BI pipeline?
- Explain two or three ideas for quality control of CEL files
- What are the (four) preprocessing steps? What is the input for the first step and the result from the last one?
- What methods are available to come from a matrix of gene expressions (genes x samples) to a list of „interesting“ genes? When is a gene „interesting“?

Sequencing and genome assembly

- Characterize first, second and third generation sequencing machines in terms of read length, error rate and cost per base.
- Characterize the Oxford Nanopore MinIon machine
- Sketch the procedure that leads from reads via contigs and scaffolds to the final sequence.
- What can we do with sequencers (Whole genome sequencing, …)
- How do you do quality control for reads, and what can you do with low-quality reads?

- Explain the statistics behind coverage and the probability of having a single base not covered by any read. What is the Lander-Waterman equation taking into account? Sketch a typical LW curve.
- What are the major assembly paradigms?
- How are sequencing errors (wrong bases in the reads) treated by Overlap-LayoutConsensus versus DeBruijn graph based algorithms?
- What is used in „scaffolding“?
- Explain Single-Cell Sequencing and the problems associated with it.
- How can you assess the quality of an assembly result?

Structural bioinformatics

- Sketch RNA-seq. To what techonlogy does it compete? Can it compete?
- What „signals“ in a DNA sequence can be used to infer that we have a gene here?
- How can you find isochores (regions of similar GC-content)?
- How do you get the 3D coordinates of the atoms in a protein?
- How do you assess the 3D-similarity of two proteins?
- How could you predict the secondary structure out of a) the 3D-coordinates and b) the sequence alone?
- Explain the four possible approaches to predict the 3D structure of proteins

Biology expertise

Genomics

- Principles of DNA sequencing and its application in genomics
- Sequence databases and the common DNA/AA sequence formats
- Evolution of prokaryotic and eukaryotic genomes
- Genome/transcriptome assembly
- Annotation of prokaryotic and eukaryotic genomes

Diversity of Life

- Give us a BRIEF outline of phylogenetic system of living things, as it is gradually being revealed in current decades, much owing to molecular and phylogenetic methods. Pay particular attention to a) mutual positions of Archaebacteria, Eubacteria and Eucaryota; b) system of five eucaryotic realms; c) position of vascular plants, fungi and metazoan animals within the system of five realms.
- Brief summary of modern phylogenetic methods used for classifying living things. Why was phylogenetic methodology "revolutionary" compared to earlier approaches? How and why is the methodology useful in biogeography, ecology and conservation?

- Discuss the "dynamic" (or "disturbance dynamics) paradigm, currently prevailing in ecology and conservation. Could you speculate on scientific issues in ecology, which might be resolved using bioinformatic methods?
- Discuss the environs of the two cities where you study - České Budějovice and Linz - from the perspectives od biogeography and biodiversity conservation. In which phytogeograohy / zoogeography realms and provinces are the cities located; what biomes and habitats can be found there; which important nature conservation localities are located nearby; why are these localities important regionally, continentally and globally.

Molecular Biology & Genetics

- DNA structure, DNA packaging, DNA replication and the genetic code.
- Transcription (prokaryotic vs. eukaryotic mRNAs) – eukaryotic mRNA processing.
- Translation (genetic code & tRNAs), ribosome cycle.
- Mechanisms of gene expression control (e.g. prokaryotic gene operons, localised mRNAs, transcription factors, epigenetics/ imprinting/DNA methylation/post-translational histone modifications, insulators, RNA editing, riboswitches, RNAi)
- What is recombinant DNA? Cloning and techniques involved (e.g. plasmids, cDNA, PCR, restriction enzymes, DNA sequencing etc.)

Informatics Expertise

Data Structures & Algorithms

- Algorithms: Time and Space complexity, example of algorithms and their complexity
- Basic Data Structures: (array, linked-list, stack, queue), implementations and usage
- Sorting algorithms (selection sort, insertion sort, bubble sport, quicksort

Python Basics

- Python data types
- Python modules and functions
- Python error handling
- Python object oriented programming

Applied Python

- Itertools – basic principles, examples of usage
- Numpy – basic principles, examples of usage
- Pandas – basic principles, examples of usage
- Matplotlib – basic principles, examples of usage

Biopython

- Basic sequence analysis, sequence formats
- Analyzing data in VCF
- Working with BAM/SAM-formatted files
- Accessing GenBank and NCBI databases
- Finding a protein in multiple databases
- Biopython's PDB module for working with the Protein Data Bank

Parallel Programming

- MPI parallelization: communication of threads, memory, execution (advantages and disadvantages)
- OpenMP parallelization: communication of threads, memory, execution (advantages and disadvantages)
- Strategies in parallel programming: input and output operations, splitting of the work load (especially in cycles), efficiency (timing, number of threads)
- Pointers and references, relations between arrays and pointers, pointer arithmetic.