Introduction to Bioinformatics for Computer Scientists

Lecture 15
Wrap-up and Exam Preparation

Exam

- Don't underestimate the exam!
- In general, if you get equations wrong, that's not a catastrophe
- You should always know and be able to explain how things work in principle though!
- **Register for the exam via the campus system!!!!!!**
- Find a time slot at: https://terminplaner.dfn.de/bocteM4I6nQ32RNj
- Monday Feb 10 slots: only for students that can't make the later spots, justification via email required!
- You can chose to do the exam in English or German

Exam II

- Exam dates:
  - Feb 10
  - April 22, 23, 24
- 20 minutes oral exam
- **Bring along your ID and student card!**
- Room 234

Course Overview

- Biological background knowledge
- Pair-wise alignment
- Sequence Assembly
- Multiple Alignment
- Phylogenetics
- MCMC
- Population genetics

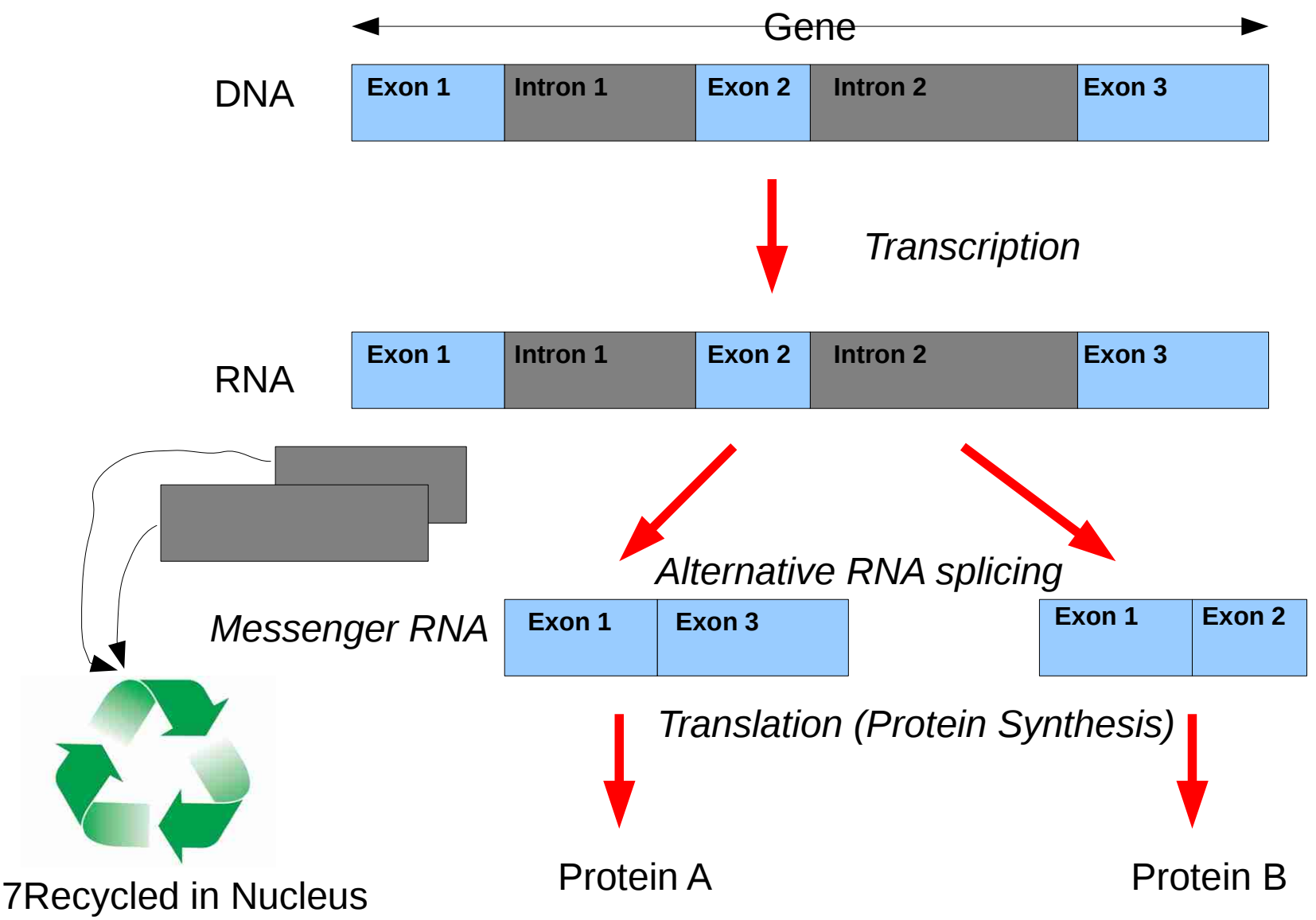
Biological Knowledge

- DNA and AA alphabets
- What are paired-end reads?
- What's a genome?
- Name some model organisms
- Why do we use model organisms?
- Coding versus non-coding sequence data
- What's a transcriptome?
- Is the transcriptome constant or does it change?
- What's a gene?

Biological Knowledge

- What is RNA data?
- What's a Codon?
- 1st & 2nd versus 3rd Codon position
- Synonymous versus non-synonymous substitutions
- Where do genes start and end?
- DNA: what's the *3′* and *5′* end?
- What are the three domains of life?
- What's the difference between Prokaryota and Eukaryota?

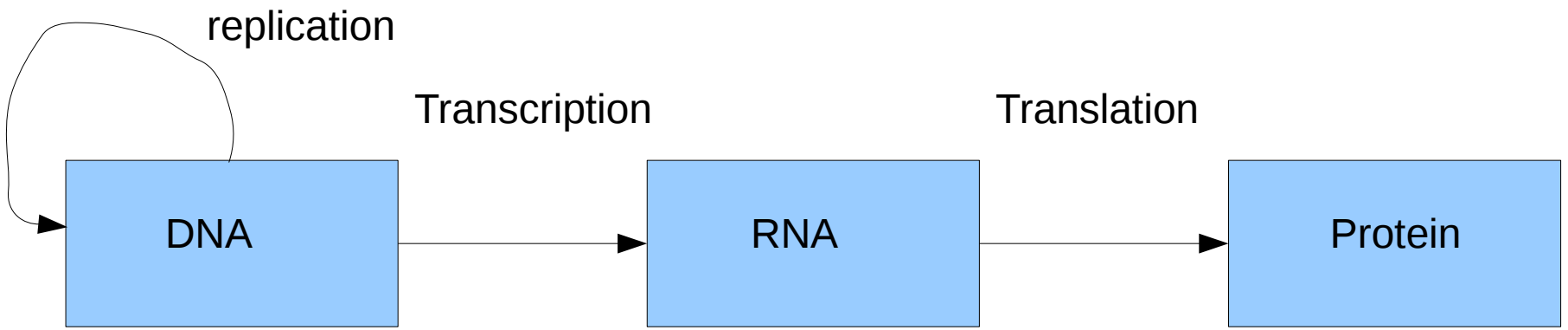
Alternative Splicing



RNA

- Which types of RNA do you know, what do they do?
- Why is RNA interesting for building phylogenies?
- Name some interesting RNA genes!
- Why is the RNA secondary structure interesting for RNA evolution?

Central Dogma of Molecular Biology



Biological Knowledge

- What is a meta-genome?
- What's a chromosome?
- What's a taxonomy?
- What's a phylogeny?
- What is an outgroup?

Pairwise Sequence Alignment

- Name some distances for comparing strings
- What is the difference between local and global pair-wise sequence alignment?
- How is the edit distance defined?
- What's the definition of the Hamming distance?
- How do we define an optimal pair-wise sequence alignment?
- Outline how a pair-wise sequence alignment algorithm that uses dynamic programming works!

Pair-wise Sequence Alignment

- What's the difference between the Needleman-Wunsch and Smith-Waterman algorithms?
- What is their time and space complexity?
- How can we parallelize dynamic programming algorithms?
- What is a substitution matrix?

Blast & Genome assembly

- By reference versus de novo assembly (mapping)
- What is BLAST?
- What is BLAST good for?
- Why not use Smith-Waterman instead?
- How does BLAST work → *seed, extend, evaluate*!
- What is Genome assembly?

De novo Genome Assembly

- How does *de novo* assembly work?
- What is an overlap graph?
- How do we traverse this graph to assemble a genome?
- What is a *de Bruijn* graph?
- What is a *k-mer*?
- How do we traverse a *de Bruijn* graph?

By reference assembly - Mapping

- Which problem are we trying to solve?
- Why not use Blast?
- Why not use pair-wise sequence alignment?
- What techniques can we apply to accelerate mapping?
- How does mapping with hashing work?
- How do we select a *k-mer* representing a read?
- How do we extend the read?
- What's the drawback of the hashing procedure?
- How can this be improved?

Multiple Sequence Alignment

- What is homology?
- How can we assess the quality of an MSA?
- How do we compute the *SP* score?
- What are MSAs good for?
- *Orthology* versus *Paralogy* versus *Homology*?
- What's a gene duplication?
- Does sequence similarity induce homology?

MSA

- Can we build an MSA with an optimal SP score?
- What's the time complexity?
- How does the star alignment heuristic work?
- How is the tree alignment problem defined?

  → can you compute a tree alignment score on a given tree?

  → students always get this wrong … this is not a guide tree method, but an explicit criterion!!!!!
- How do practical approaches for MSA work?
- Describe how progressive MSA methods work in principle
- How does pair-wise profile alignment work?
- What are the shortcomings of progressive alignment methods?
- Solutions to overcome these?
- How do we benchmark MSA programs?

Phylogenetics

- Terminology: *monophyletic, paraphyletic, polyphyletic, patristic distance between taxa*
- Why is an appropriate *outgroup* choice important?
- What is an *ultrametric* tree?
- What are *diversification* rates?
- How do we put real times on a phylogeny?
- What input data can be used to build phylogenies?
- What can we do with phylogenies?
- How many unrooted binary trees exist for *n* taxa?
- How can we come up with this formula? → draw an image

Phylogeny Reconstruction Methods

- Name the two basic classes of reconstruction methods!
  - Distance- versus character-based methods students often seem to be confused by this question
- Name some methods that are NP-hard
- How do Neighbor Joining/UPGMA work **in principle**?

  → run time complexities!!!!
- How does the least-squares algorithm work?

  → suggest a tree search algorithm for the least-squares criterion
- How is the minimum evolution criterion defined?
- How does parsimony work?
- What's the time and space complexity for computing the parsimony score on a tree?
- What's the underlying principle?
- Given a small tree and alignment, calculate the parsimony score!

Search Strategies

- How can we build starting trees?
- How can we change a given comprehensive topology to find a better tree?

Maximum Likelihood

- What's the long branch attraction problem?
- Why shall we model distances as stochastic processes?
- What does a substitution matrix look like?
- What is time-reversibility?
- How do we obtain *P(t)* from *Q*? → eigenvector decomposition, trick with the sum representation of the exponential …
- How does ML work in principle? Remember: **AND** and **OR**
- How does the Felsenstein pruning algorithm work?
- Which parameters do we have (to optimize)?
- What's the time & space complexity for evaluating one tree?
- How can we optimize branch lengths?

ML continued

- How can we model rate heterogeneity among sites?
- How does the Γ model work?
- How are protein substitution models obtained?
- How can we obtain the base frequencies?

Discrete Operations on trees

- What's the phylogenetic bootstrap?
- How does it work?
- How can we map the bootstrap support to the best-known ML tree?
- What is a bipartition?
- What is a trivial bipartition?
- How can we store bipartitions?
- Which flavors of consensus trees do you know?

  → strict, majority, extended majority
- How can we build a strict/majority rule consensus tree?
- Why is the computation of extended majority rule trees NP-hard (intuition!)

Discrete Ops on Trees

- How can we extract and hash the bipartitions of a tree?
- How is the RF-distance defined?
- Given two small trees, compute their RF-distance!

MCMC Methods

- How do they differ from ML methods?
- What are we trying to approximate?
- What are the computational difficulties?
- How do they work in principle?

  → robot metaphor (you can use this to explain everything)
- How do we compute if we want to accept or reject a proposal?
  - Why does this ratio solve a lot of problems?
- Where do we get the priors from?
- How can we summarize samples?
- How does MCMC work in practice for phylogenetics?

MCMC Methods

- What's the difference between the proposal and the target distribution?
- What does the term “good mixing” mean?
- What is the Hastings ratio and why do we need it? → drunk robot
- What is Metropolis-Coupled Markov Chain Monte Carlo?

  → multiple robots on our planet
- What is thinning?
- For DNA under GTR+Gamma what types of proposals do we need?

  → which proposal type would you apply most frequently?
- Can we use MCMC to integrate over different models?

Population genetics

- How can a population evolve?
  → four main evolutionary forces
- Difference: Genotype versus Phenotype?
- Dominant versus Recessive?
- How are alleles inherited?
- How is polymorphism defined?
- What are SNPs?
- Can you describe Hardy's model?
  →assumptions?
  → what's amazing about this model?
- What is the Wright-Fisher Model?
  → how can we simulate a population under it?
  → which assumptions does it make

Population genetics

- What is random genetic drift?
- What is fixation?
- What is Heterozygosity?
- What is positive selection?
- What is mutation-drift balance?
- What do we want to study in population genetics?

  → what's the difference to phylogenetics?

Courses in Summer

Seminar *Hot Topics in Bioinformatics*

- 2 hours per week seminar
- We select interesting Bioinformatics papers and present them

  → select any subject/paper mentioned in the course you find interesting

  → ask us if you are interested in a different topic

  → one of my lab members will help you to understand the paper, prepare the presentation and the report

  → report & presentation language: either English or German, English much preferred though!
- 35 Minute presentation of paper
- Submit a report of **8** pages at the end of the semester
- 3 ECTS

Programming practical

- No practical next summer

Prerequisites for Seminar

- Attended & passed Introduction to Bioinformatics
- To register, write me an email, registration via the campus system will only be possible AFTER you have passed the exam
- Maximum of 10 places available
- Who is interested in the Bioinformatics Seminar?

Thank you

- For the positive course feedback
- For being very concentrated & interested
- For asking a lot of good questions :-)