Methods in Bioinformatics, 1-BIN-301/2-AIN-501

Lecturers:

Broňa Brejová, M-163, brejova@fmph.uniba.sk

Tomáš Vinař, M-163, vinar@fmph.uniba.sk

Web: http://compbio.fmph.uniba.sk/vyuka/mbi/

Announcements, Q&A: Microsoft Teams code h2w2fxb
(guests please write e-mail)

Literature:

I-INF-D-23: Durbin, Eddy, Krogh, Mitchison: Biological sequence analysis. Cambridge University Press 1998.

I-INF-Z-2: Zvelebil, Baum: Understanding Bioinformatics. Taylor&Francis 2008.

Some lecture notes (in Slovak), notes and videos on the web page

Times and Lecture Rooms

- Lecture Thu 15:40-17:10 lecture hall C
- Tutorials (CS) Thu 14:00-15:30 lecture hall C
- Tutorials (Bio) Thu 17:20-18:50
  lecture hall C and computer room M-217

we plan to record/stream lectures and CS tutorials

"CS": students of computer science, bioinformatics, applied informatics; please enrol under 1-BIN-301 code

"Bio": students from the Faculty of Natural Sciences, students of biomedical physics; please enrol under 2-AIN-501 code

others: contact us

Course Goals

- **Everyone:** Overview of basic methods for analysis of biological sequences and other data sets in molecular biology
- **CS:** Algorithms and data structures, machine learning, probability. How to develop mathematical abstractions for real-world problems.
- **Bio:** Mathematical models at the core of popular bioinformatics tools, how to use tools, interpretation of their results.
- **Everyone:** Experience with an interdisciplinary collaboration.

Grading

3 homework assignments 30% (10% each)
Journal club 10%
Quizzes 10% (1 point each week)
Final exam 50%

(no quizzes for English speaking guests)

Final grade: A: 90+, B: 80+, C: 70+, D: 60+, E: 50+
At least 50% of the final exam is required

- Two versions of questions: bio and CS
- Journal club: read a research paper, write summary in a group (optional presentations for bonus points)
- You are allowed 2 double sided A4 pages as a cheat sheet on the exam
- DO NOT COPY, DO NOT CHEAT!

What to expect from lectures and tutorials

Typical lecture

- Biological introduction to a problem
- Formulation/abstraction as a computer science problem
- Algorithm idea for the problem solution(s)

Typical tutorial

- CS: algorithmic details and extensions, background biological knowledge
- Bio: applications to concrete data sets, what do various parameters mean and how to set them, background computer science knowledge,

Weekly quizzes

- Cca 5 short questions concerning last week's lectures and tutorials
- Due on Wednesday 10pm
- Moodle link on the web page
- Goal: review basic concepts from the lecture and the tutorial
- **First quiz already this week**

Example from our research

common marmoset, Callithrix jacchus, 250g, 18cm

Genome sequenced in 2007

(Washington University St. Louis a Baylor College of Medicine, USA)

Analysis published in 2014

IGF1R: Insulin-like growth factor 1 receptor

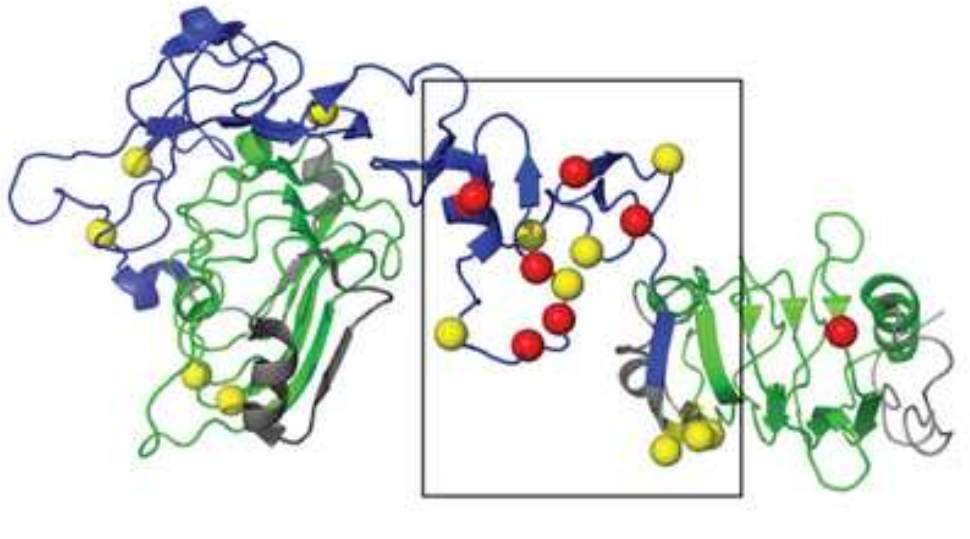
Protein passes through cytoplasmic membrane on the cell surface

After binding to growth hormones IGF1, IGF2 signals into the cell

Functions related to the cell growth and division,

organism growth, cancer

```
human     RDFCANILSAESSDSEGFVIHDGECMQECPSGFIRNGSQSMYCIPCEGPCPKVC-EEEKKTK
chimp     RDFCANILSAESSDSEGFVIHDGECMQECPSGFIRNGSQSMYCIPCEGPCPKVC-EEEKKTK
orang     RDFCANILSAESSDSEGFVIHDGECMQECPSGFIRNGSQSMYCIPCEGPCPKVC-EEEKKTK
macaque   RDFCANILSAESSDSEGFVIHDGECMQECPSGFIRNGSQSMYCIPCEGPCPKVC-EEEKKTK
marmoset  RQFCASIVSSENSENNKFVIHDGECMQDCPSGFIRDTTHSMQCIPCKGPCPKVC-D-EQMAK
mouse     RDFCANIPNAESSDSDGFVIHDDECMQECPSGFIRNSTQSMYCIPCEGPCPKVCGDEEKKTK
rat       RDFCANIPNAESSDSDGFVIHDGECMQECPSGFIRNSTQSMYCIPCEGPCPKVCGDEEKKTK
dog       RDFCANIPSAESSDSEGFVIHDGECMQECPSGFIRNGSQSMYCIPCEGPCPKVC-EEEKKTK
```



What bioinformatics tools were needed for this research?

1. Assemble genome from sequencing reads
2. Find sequence similarities to other genomes
3. Find genes coding for proteins
4. Find genes under positive selection
5. Determine structure and function of the proteins

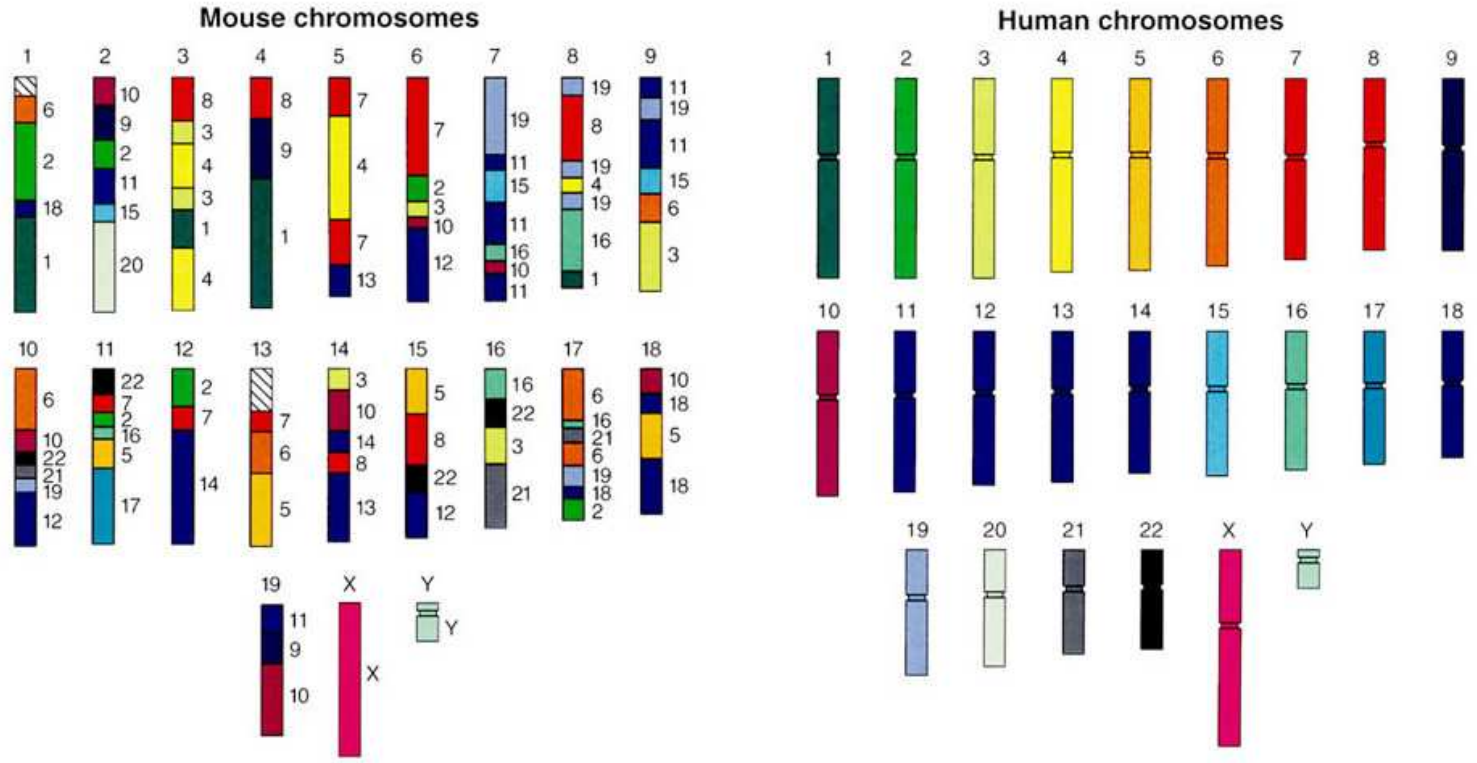
1. Genome assembly

- We can only sequence short fragments of DNA
  (e.g. of length 1000)
- Each place in the genome is sequence multiple times (for marmoset on average 6×)

- We need to "glue" sequencing reads together based on overlaps
- Huge amount of data ⇒ need efficient algorithms

2. Finding similarities to other genomes

For each place in the marmoset genome find corresponding places in other genomes (e.g. human, chimp, mouse, . . .)

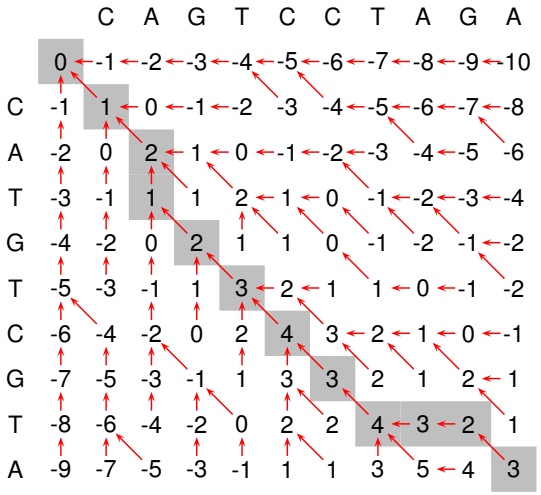


2. Finding similarities to other genomes

- We are looking for similarities between DNA sequences

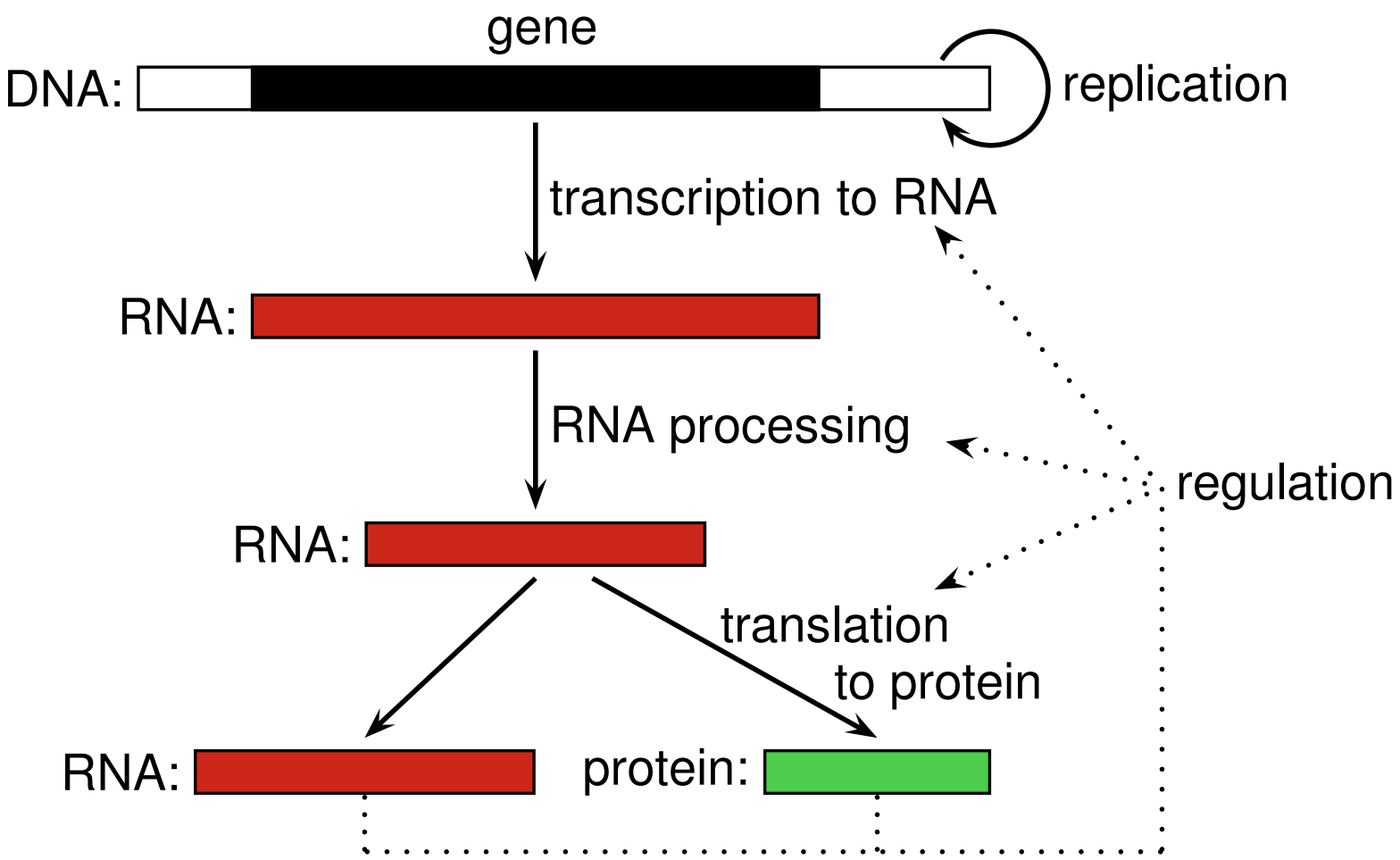
```
  Human AGTGGCTGCCAGGCTG---GGATGCTGAGGCCTTGTTTGCAGGGA
 Rhesus AGTGGCTGCCAGGCTG---GGTTGCTGAGGCCTTGTTTGCCGGGA
 Mouse GGTGGCTGCCGGGCTG---GGTGGCTGAGGCCTTGTTGGTGGGGT
   Dog AGTGGCTGCCCGGCTG---GGTGGCTGAGGCCTTATTTGCAGGGA
Chicken AGTGGCTGCCAGTCTGCGCCGTGGCCGACGTCTTGCTCGGGGGAA
```

- Basic technique used here is called **dynamic programming** which can decompose a large problem into many smaller (and easier) ones



- The table is very large, in practice many improvements and heuristics to make this practical

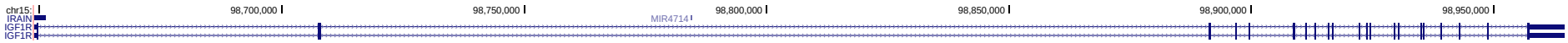
3. Finding genes coding for proteins



Which parts of the sequence genome code for proteins

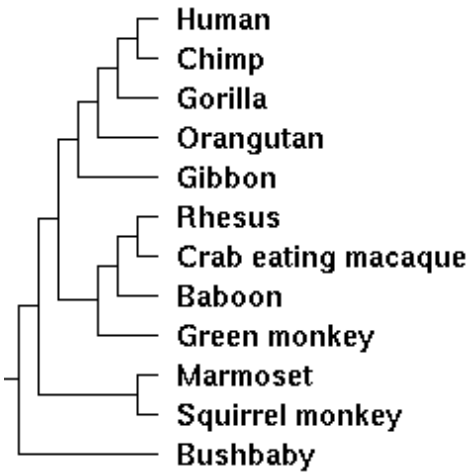
3. Finding genes coding for proteins

- Needle in a haystack: only 1% of human genome codes for proteins
- Code for a single protein is broken into many short parts (exons)
- IGF1R covers 315 569nt, but only 4101nt in 21 exons code for the protein



- Take known genes, collect various statistics
  find other regions of the genome with a similar statistical profile

4. Search for genes under positive selection



- Study of evolutionary processes
- Mutations in DNA over time are subject to natural selection
- Most of random changes in a protein are harmful, thus segments encoding proteins typically mutate very slowly

4. Search for genes under positive selection

- Sometimes a beneficial mutation is discovered, followed by a surge of other mutations optimizing the new function $\rightarrow$ positive selection

5. Determining function and structure of proteins

- After steps 1-4, we have a list of 37 genes under positive selection in the marmoset genome
- What is their function? Any of them related to marmoset size?
- What is the shape of the protein, where are the position under positive selection located?
- Protein structure (shape) can be determined experimentally expensive and time consuming, instead 3D structure predictions