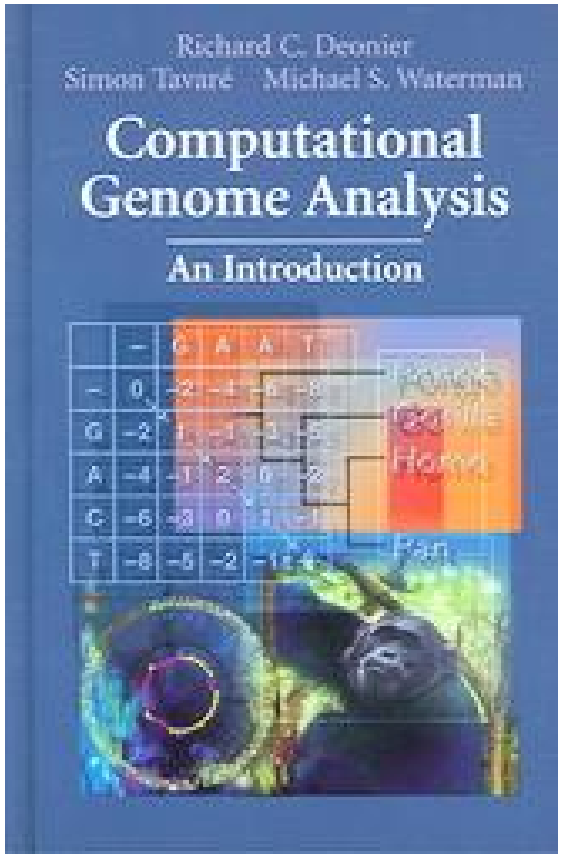
Course contents (18.9.)

- Biological background (book chapter 1)
- Probability calculus (chapters 2 and 3)
- ***Sequence alignment (chapter 6)***
  - This week (18.9. and 21.9.)
- Rapid alignment methods: FASTA and BLAST (chapter 7)
  - Next week (25.9. and 28.9.)
- Phylogenetic trees (chapter 12)
- Expression data analysis (chapter 11)



Sequence Alignment (chapter 6)

- *The biological problem*
- Global alignment
- Local alignment
- Multiple alignment

Background: comparative genomics

- Basic question in biology: *what properties are shared among organisms?*
- Genome sequencing allows comparison of organisms at DNA and protein levels
- Comparisons can be used to
  - Find evolutionary relationships between organisms
  - Identify functionally conserved sequences
  - Identify corresponding genes in human and model organisms: develop models for human diseases

Homologs

- Two genes g_B and g_C evolved from the same ancestor gene g_A are called *homologs*
- Homologs usually exhibit conserved functions
- Close evolutionary relationship => expect a high number of homologs

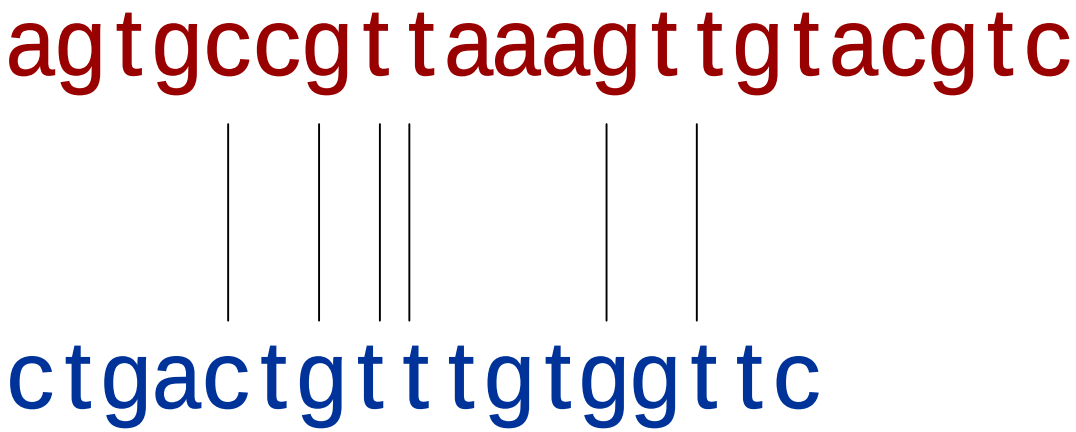
g_A = agtgtccgttaagtgcgttc

g_B = agtgccgttaaagttgtacgtc

g_C = ctgactgtttgtggttc

Sequence similarity

- Intuitively, similarity of two sequences refers to the degree of match between corresponding positions in sequence



- What about sequences that differ in length?

Similarity vs homology

- Sequence similarity is not sequence homology
  - If the two sequences g_B and g_C have accumulated enough mutations, the similarity between them is likely to be low

#mutations	
0	agtgtccgttaagtgcgttc
1	agtgtccgttatagtgcgttc
2	agtgtccgcttatagtgcgttc
4	agtgtccgcttaagggcgttc
8	agtgtccgcttcaaggggcgt
16	gggccgttcatgggggt
32	gcagggcgtcactgagggct

#mutations	
64	acagtccgttcgggctattg
128	cagagcactaccgc
256	cacgagtaagatatagct
512	taatcgtgata
1024	acccttatctacttcctggagtt
2048	agcgacctgcccaa
4096	caaac

Homology is more difficult to detect over greater evolutionary distances.

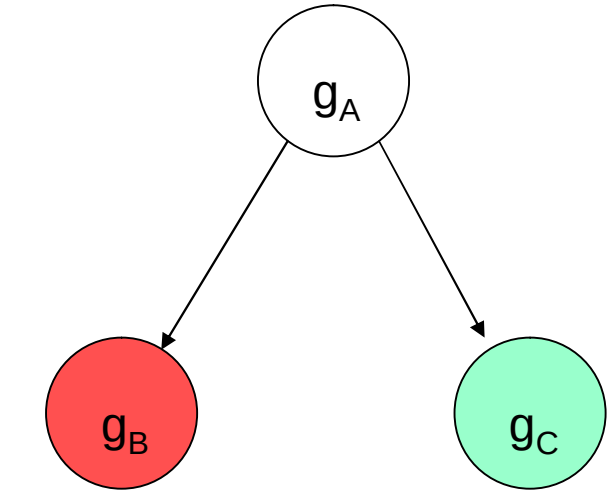
Similarity vs homology (2)

- Sequence similarity can occur by chance
  - *Similarity does not imply homology*

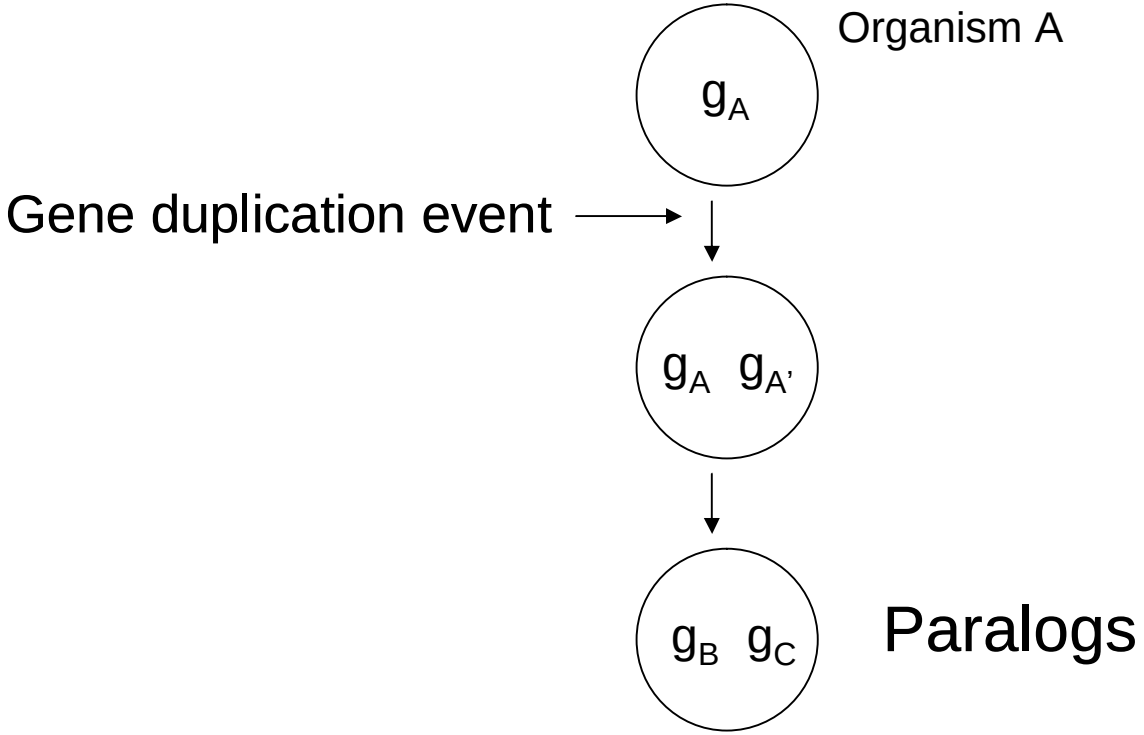
- Consider comparing two short sequences against each other

Orthologs and paralogs

- We distinguish between two types of homology
  - Orthologs: homologs from two different species, separated by a *speciation* event
  - Paralogs: homologs within a species, separated by a *gene duplication* event



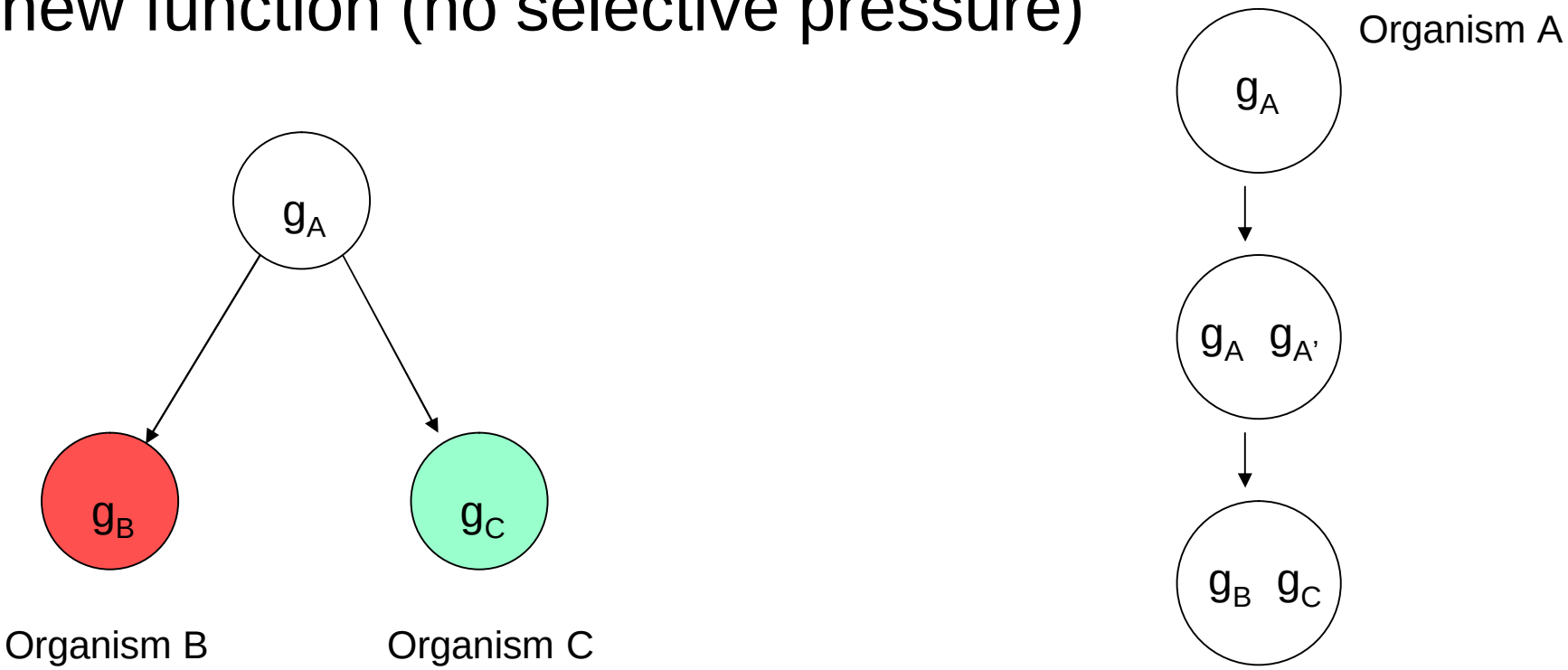
Orthologs



Paralogs

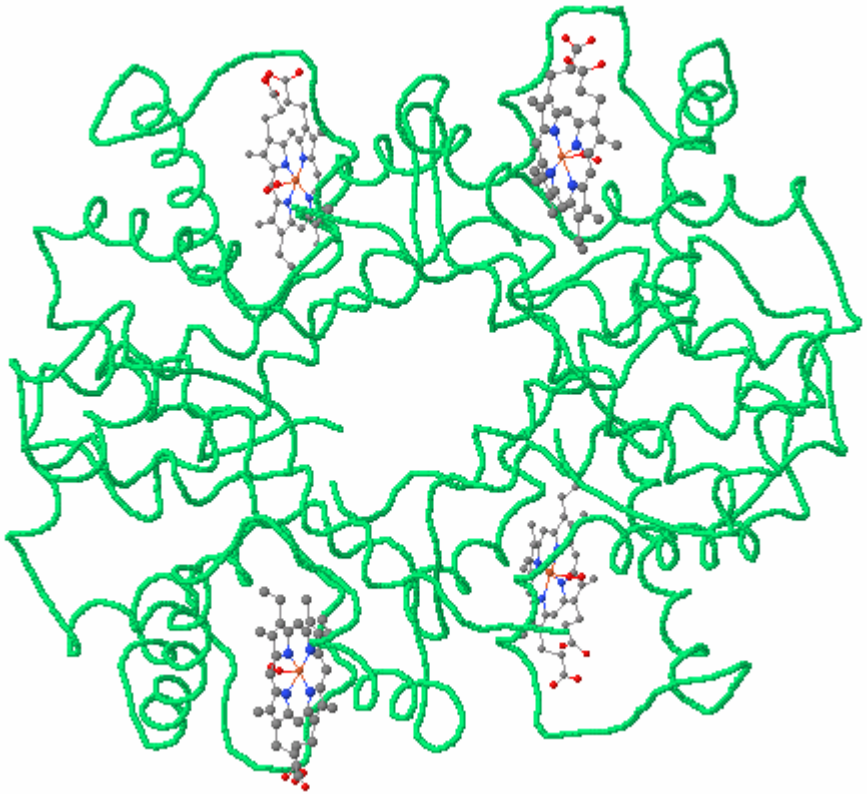
Orthologs and paralogs (2)

- Orthologs typically retain the original function
- In paralogs, one copy is free to mutate and acquire new function (no selective pressure)



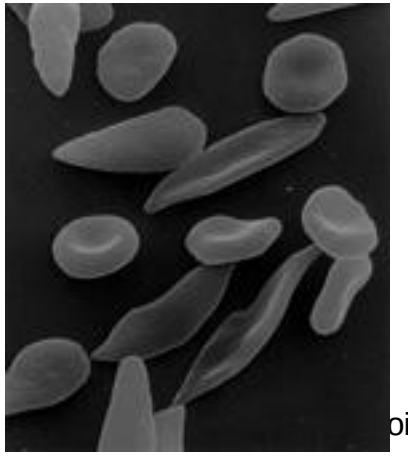
Paralogy example: hemoglobin

- Hemoglobin is a protein complex which transports oxygen
- In humans, hemoglobin consists of four protein subunits and four non-protein heme groups



Hemoglobin A, www.rcsb.org/pdb/explore.do?structureId=1GZX

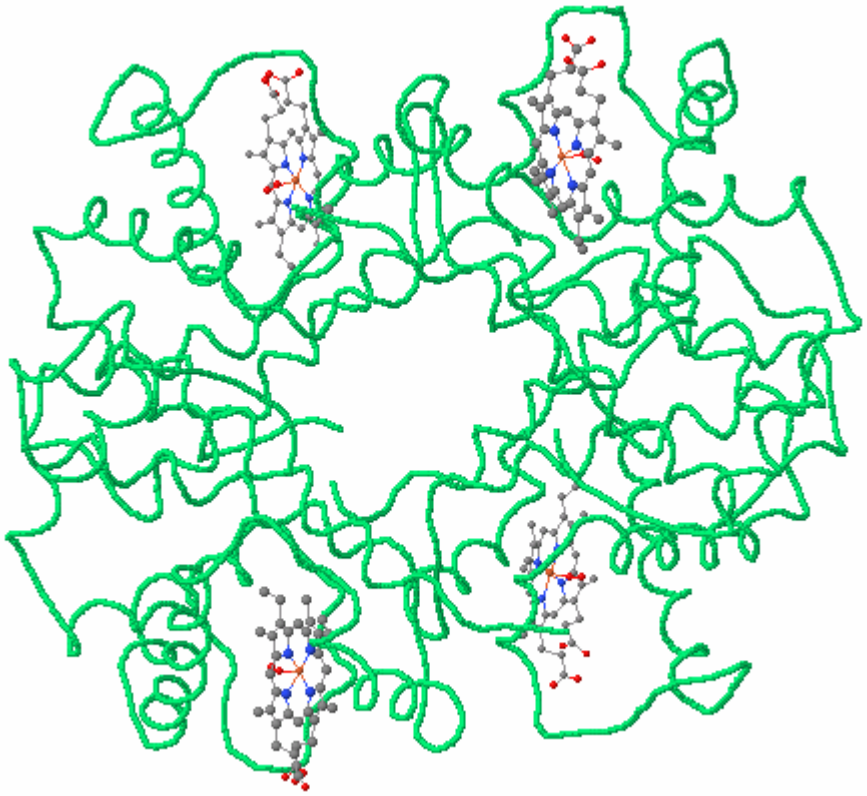
Sickle cell diseases are caused by mutations in hemoglobin genes



http://en.wikipedia.org/wiki/Image:Sicklecells.jpg

Paralogy example: hemoglobin

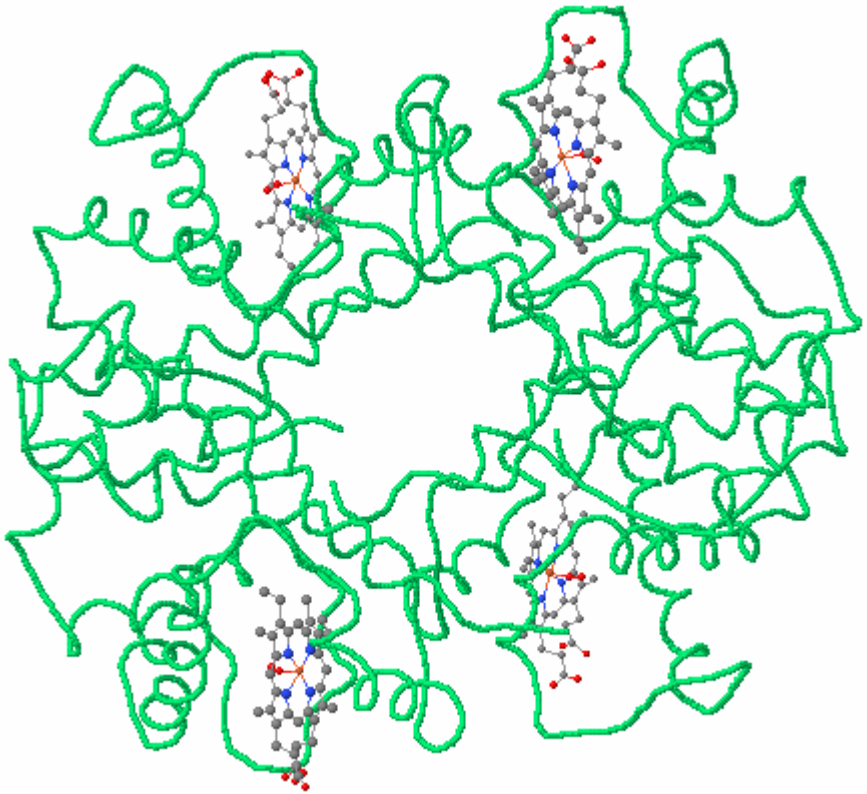
- In adults, three types are normally present
  - Hemoglobin A: 2 alpha and 2 beta subunits
  - Hemoglobin A2: 2 alpha and 2 delta subunits
  - Hemoglobin F: 2 alpha and 2 gamma subunits
- Each type of subunit (alpha, beta, gamma, delta) is encoded by a separate gene



Hemoglobin A, www.rcsb.org/pdb/explore.do?structureId=1GZX

Paralogy example: hemoglobin

- The subunit genes are paralogs of each other, i.e., they have a common ancestor gene
- Demonstration in lecture: hemoglobin human paralogs in NCBI sequence databases
  http://www.ncbi.nlm.nih.gov/sites/entrez?db=Nucleotide
  - Find human hemoglobin alpha, beta, gamma and delta
  - Compare sequences



Hemoglobin A,
www.rcsb.org/pdb/explore.do?structureId=1GZX

Orthology example: insulin

- The genes coding for insulin in human (*Homo sapiens*) and mouse (*Mus musculus*) are orthologs:
  - They have a common ancestor gene in the ancestor species of human and mouse
  - Demonstration in lecture: find insulin orthologs from human and mouse in NCBI sequence databases