

Bioinformatics for biomedicine

Multiple alignments and phylogenetic trees

Lecture 3, 2006-10-03

Per Kraulis

<http://biomedicum.ut.ee/~kraulis>

Course design

1. What is bioinformatics? Basic databases and tools
2. Sequence searches: BLAST, FASTA
3. **Multiple alignments, phylogenetic trees**
4. Protein domains and 3D structure
5. Seminar: Sequence analysis of a favourite gene
6. Gene expression data, methods of analysis
7. Gene and protein annotation, Gene Ontology, pathways
8. Seminar: Further analysis of a favourite gene

Previous lecture: Sequence searches

- Algorithms
 - Complexity
 - Heuristic or rigorous
- Sequence alignment
 - Global or local
 - Needleman-Wunsch, Smith-Waterman
- Sequence search
 - BLAST, FASTA

Task from previous lecture

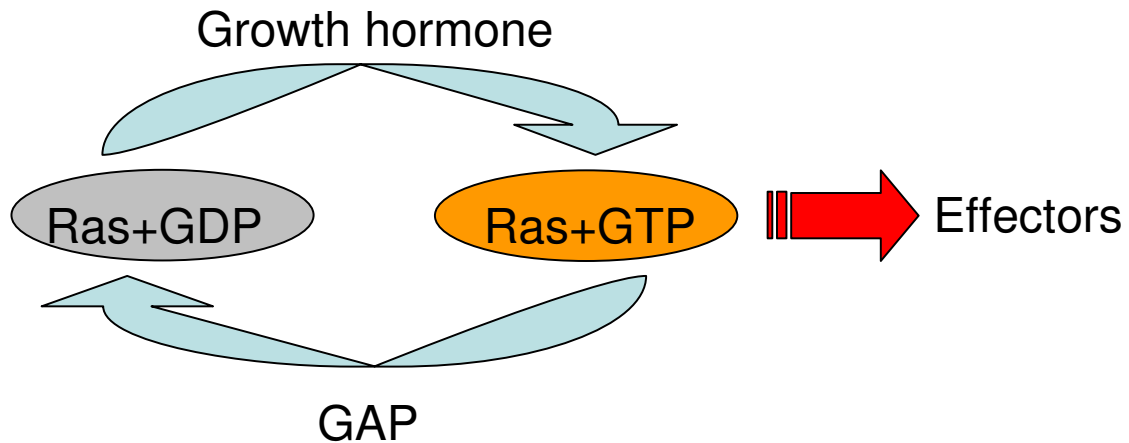
- BLAST for short oligonucleotide
 - Difference in params
 - Why?
 - More sensitive
 - More hits expected
- Inherent limitations
 - Too many matches for short oligonucleotides

Task from previous lecture

- FASTA search using EBI server
- <http://www.ebi.ac.uk/fasta/index.html>
 - Protein: UniProt, UniRef, PDB, patents
 - Nucleotide: EMBL, subdivisions
 - Proteomes or genomes
- Choose proper DB
- Consider choice of parameters

Sequence search example 1

- Example protein: H-Ras p21
 - Signalling protein (growth hormone, etc)
 - Binds GDP or GTP
 - GTPase: GTP \rightarrow GDP (slowly)



Sequence search example 1

- Find proteins similar to H-Ras p21 (signalling GTPase, growth hormone cascade, etc)
- UniProt at EBI:
<http://www.ebi.uniprot.org/>
 - "ras p21" in description lines
 - Not found: but mouse homolog!
 - Use it to find human via Ensembl
 - Ortholog list
 - BLAST search
 - "H-ras" and limit to "Homo sapiens"
 - Combine data sets using "intersection" operation

Homology, orthology, paralogy

- Confusion when applied to sequences
 - Evolutionary biologists upset with bioinfo (mis)use
- Terms describe evolutionary hypothesis
- May correlate with function
 - Not necessary, but probable
- Language note
 - "X is homologue of Y" = "X and Y are homologs"

Homology

- Biological structures (sequences) are alike because of shared ancestry
 - Hypothesis about history
 - Claim: common ancestor
 - Probably indicates similar function

Similarity

- Degree of sequence similarity
- Measurable
 - % Identity (identical residues)
 - % Similarity (conserved residues)
- May indicate homology
 - Depends on statistics
 - Other knowledge or assumptions

Is homologous, or not!

- “80% homologous” is meaningless!
- Should be: “80% similar”
- Or:
 - Closely homologous
 - Distantly homologous
 - Refers to age
 - Probably inferred from similarity

Homology and medicine?

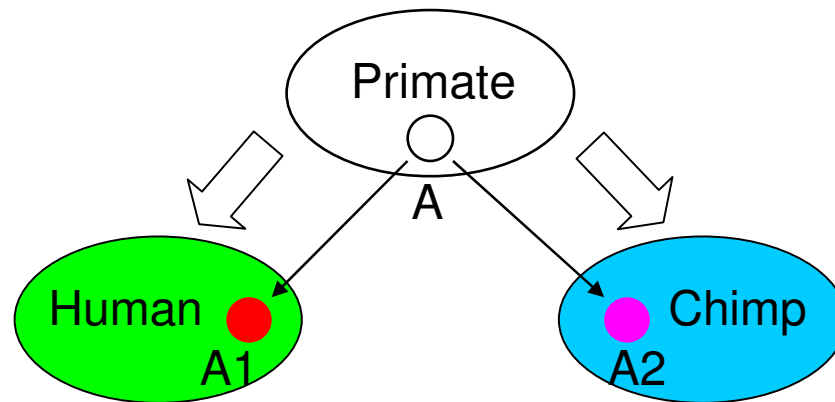
- Model organisms
 - Animals used in research
 - Choice essential for prediction into human
 - Efficacy (does a drug work?)
 - Toxicity (does a drug cause bad effects?)
- How does a drug target work?
 - Human-specific gene, or ancient
 - Pathways, essential component or not
 - Has pathway changed over time?

Orthology and paralogy

- Special cases of homology
- Hypothesis about evolutionary history
- Is hard to determine conclusively
 - Depends on
 - Model of evolution
 - Statistics
 - New evidence may alter interpretation

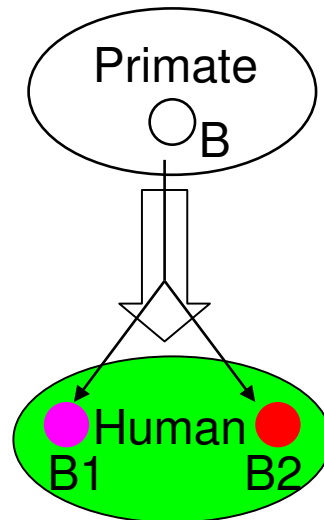
Orthologs

- Two genes are orthologs if they were separated by a speciation event
 - In two different species
 - Typically: same or similar function

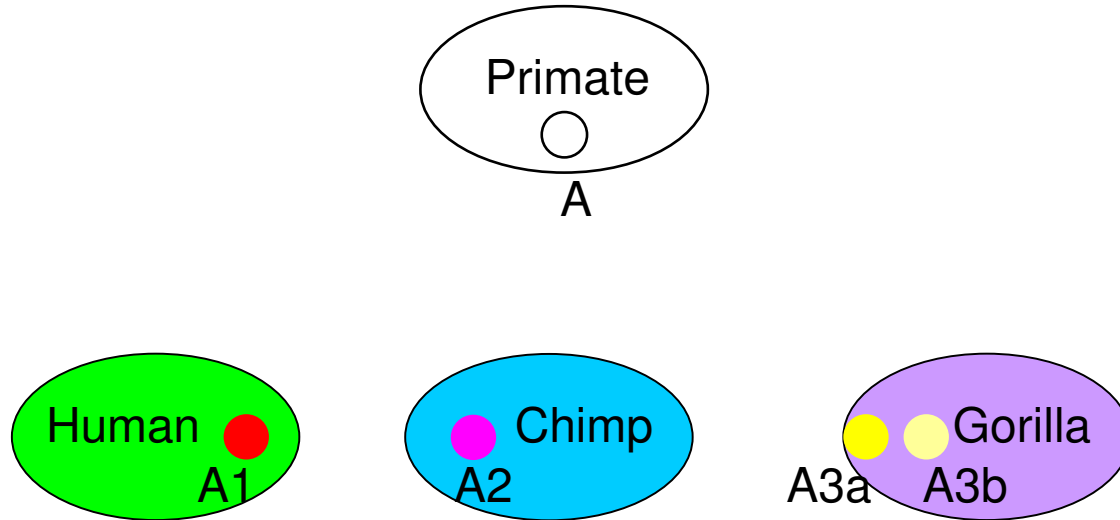


Paralogs

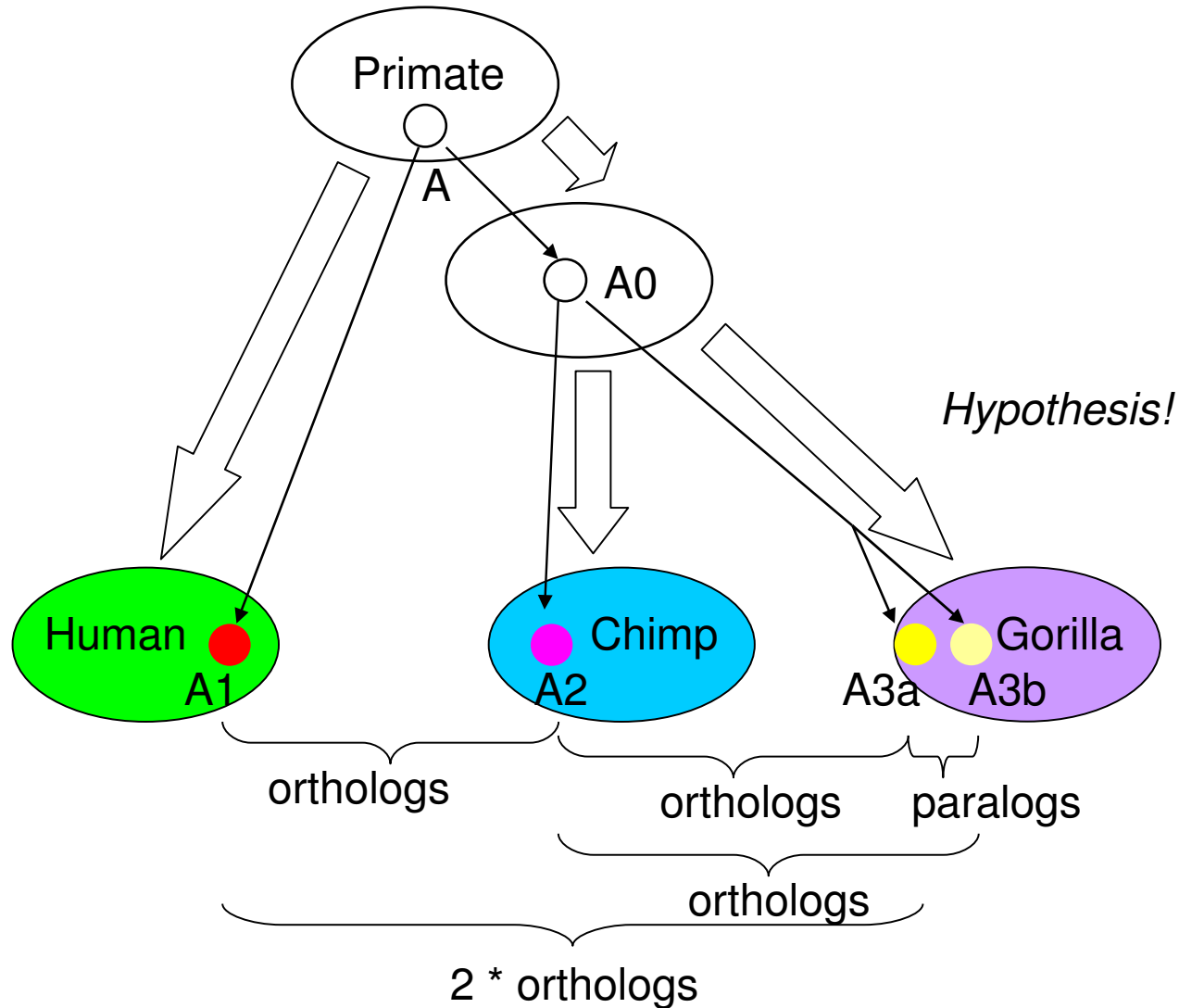
- Two genes are paralogs if they were separated by a gene duplication event
 - In a same or different species
 - Functions may have diverged



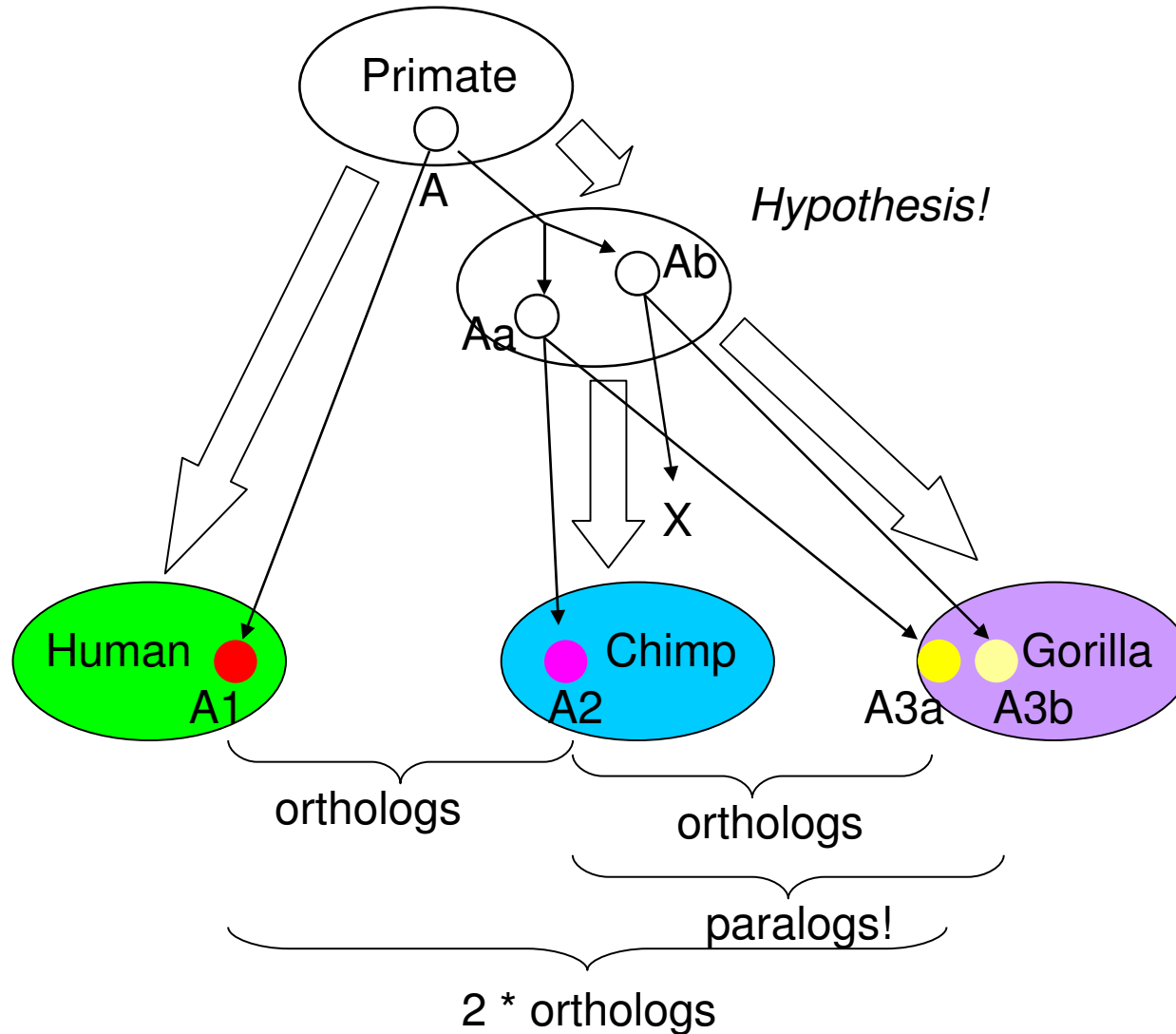
An example scenario



Hypothesis 1: late duplication



Hypothesis 2: early duplication



Does it matter?

- If hypothesis 1 (late duplication):
 - A2 probably same function as A1
 - Chimp is good model organism
 - Gorilla less good
- If hypothesis 2 (early duplication):
 - A2 may have same function as A1
 - But may also have changed!
 - Chimp is probably best, but doubts remain

Deciding orthology or paralogy

- Orthology; simple approach
 - Reciprocal best hits
 - Given gene X in genome A; find in genome B
 - Similarity of X in genome B; best hit B1
 - Similarity of B1 in A; best hit A1
 - If $X=A1$, then B1 is orthologue to X (probably)
- But: is a complex problem
 - Current research
 - More genomes very helpful

Genes vs. fossils

- Fossil and genetics complementary
 - Genes and DNA viewed as fossils
- But: Some controversies and mysteries
 - Disagreement on trees in some cases
 - Paralog/ortholog problem
 - Pax6 gene in embryonic eye development
 - Clearly homologous genes; same ancestor
 - Fossils: Eye arose several time independently
 - Unlikely that same gene used, independently

Multiple alignment (MA)

- >2 sequences aligned to common frame
- Identify
 - Conserved regions
 - Hypervariable regions
 - Insertion/deletion regions
 - Strictly conserved residue positions

Example multiple alignment

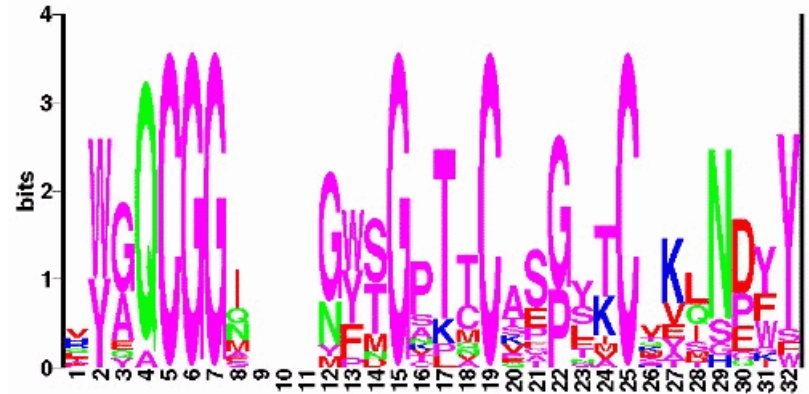
	1	2	3
	45678901...	234567890123456789012	
GUX1_TRIRE/481-509	HYQCGGI...	GYSGPTVCASGTTCCQVLNPYY	
GUN1_TRIRE/427-455	HWQCGGI...	GYSGCKTCTSGTTCCQYSNDYY	
GUX1_PHACH/484-512	QWQCGGI...	GYTGSTTCASPYTCHVLNPYY	
GUN2_TRIRE/25-53	VWQCGGI...	GWSGPTNCAPGSACSTLNPYY	
GUX2_TRIRE/30-58	VWQCGGQ...	NWSGPTCCASGSTCVYSNDYY	
GUN5_TRIRE/209-237	LYQCGGA...	GWTGPTTCQAPGTCKVQNQWY	
GUNF_FUSOX/21-49	IWQCGGN...	GWTGATTCASGLKCEKINDWY	
GUX3_AGABI/24-52	VWQCGGN...	GWTGPTTCASGSTCVKQNDFY	
GUX1_PENJA/505-533	DWAQCGGN...	GWTGPTTCVSPYTCTKQNDWY	
GUXC_FUSOX/482-510	QWQCGGQ...	NYSGPTTCKSPFTCKKINDFY	
GUX1_HUMGR/493-521	RWQCGGI...	GFTGPTQCEEPYICTKLNDWY	
GUX1_NEUCR/484-512	HWAQCGGI...	GFSGPTTCPEPYTCAKDHDY	
PSBP_PORPU/26-54	LYEQCGGI...	GFDGVTCCSEGLMCMKMPYY	
GUNB_FUSOX/29-57	VWAQCGGQ...	NWSGTPCCTSGNKCVKLNDFY	
PSBP_PORPU/69-97	PYGQCGGM...	NYSGKTMCSPGFKCVLNEFF	
GUNK_FUSOX/339-370	AYYQCGGSKSAYPNGNLACATGSKCVKQNEY		
PSBP_PORPU/172-200	RYAQCGGM...	GYMGSTMCVGGYKCMASEGS	
PSBP_PORPU/128-156	EYAACGGE...	MFMGAKCKFGLVCYETSGKW	
consensus	...	QCGG.....G...C.....C.....	

How to view MA?

- Difficult to get overview
- Many different aspects
 - Consensus/conservation
 - Physical properties
- Advise:
 - Biology: what is (un)expected?
 - Explore different approaches
 - No single view is best

Sequence logos

- Highlight conservation
 - Large letter
- Suppress variability
 - Small letters
- Best for short sequences
 - But any number



<http://weblgo.berkeley.edu/>

Inferences from MA

- Functionally essential regions
- Catalytic site
- Structurally important regions
 - But: Function or structure? Hard to tell
- Hypervariable regions not functional
 - Not under evolutionary pressure
 - Or: strong pressure to change!

Problems with MA

- Sequence numbering
 - Often a source of confusion
 - Anyway: gaps cause out-of-register problem
- Regions may not really be aligned
 - Would be left out in local alignment
 - Often shown aligned in MA; confusion!
- Insertions/deletions make overview hard
 - Distantly related sequences problematic

How create MA?

- Given $n > 2$ sequences: How align?
- Naïve approach
 - Align first 2, then add each sequence
 - Problematic
 - The first two given more weight?
 - Order of input important: unacceptable
 - Hard to handle gaps

MA rigorous approach

- Needleman-Wunsch / Smith-Waterman can be generalized for many sequences
- But: $O(2^n)$ in both memory and time!
 - Impossible for more than 8 seq or so
 - Not used in practice
 - Heuristic method required

MA heuristic methods

- Many different approaches
- One simple idea:
 - Find pair of most similar sequences $S1, S2$
 - Align these into $A1$
 - Find next most similar sequence $S3$
 - Align $A1$ and $S3$ into $A2$
 - Continue until finished

Available services

- <http://www.ebi.ac.uk/Tools/sequence.html>
 - ClustalW (Higgins *et al* 1994)
 - T-COFFEE (Notredame *et al* 2000)
 - MUSCLE (Edgar 2004)
- Several others available
 - Kalign (Lassmann & Sonnhammer 2005)
<http://msa.cgb.ki.se/cgi-bin/msa.cgi>

Example: ras proteins

- FASTA input file available at course site
- <http://www.ebi.ac.uk/Tools/sequence.html>
 - ClustalW, MUSCLE
- <http://msa.cgb.ki.se/cgi-bin/msa.cgi>
 - Kalign