BLAST

- Acronym for Basic local alignment search tool
- BLAST is the most important tool for comparing nucleotide and protein sequences
- Input sequence (query) vs database
- Can be used online (NCBI BLAST) or locally on your own PC
- Database: publicly available (UniprotKB, NCBI Genomes, etc.) or custom (e.g. local database)

Some background

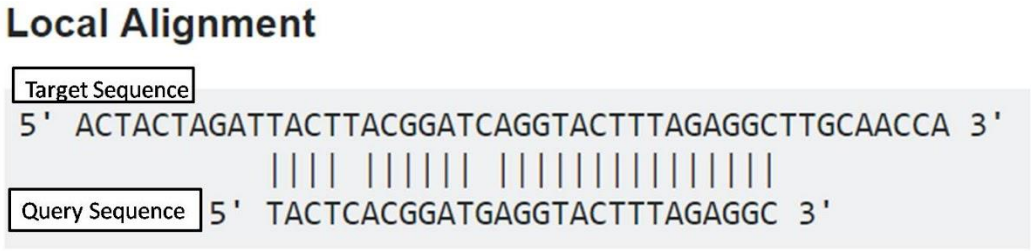
- Based on the di Smith-Waterman algorithm, which is used for **local alignment**
- **Highly sensitive**, so it is optimal for detecting small regions of similarity in squences with low homology
- Encoutered a great success in bioinformatics with **FASTA** and **BLAST** applications (FASTA was another tool for sequence similarity searches whcih we will not study in detail)
- **Based on similarity matrices (PAM o BLOSUM)** which assign positive scores to identical or similar amino acids and negative scores to dissimilar amino acids and gaps

Some basic concepts for bioinformatics noobs

Local vs global alignment

An example can be seen below… which one is the «best» alignemnt?

We need to thing about what we are mostly interested into



Similarity matrix

Two-enry matrix with positive and negative scores calculated based on the similarity among amino acids (i.e. charge, hydrophobicity, etc.)

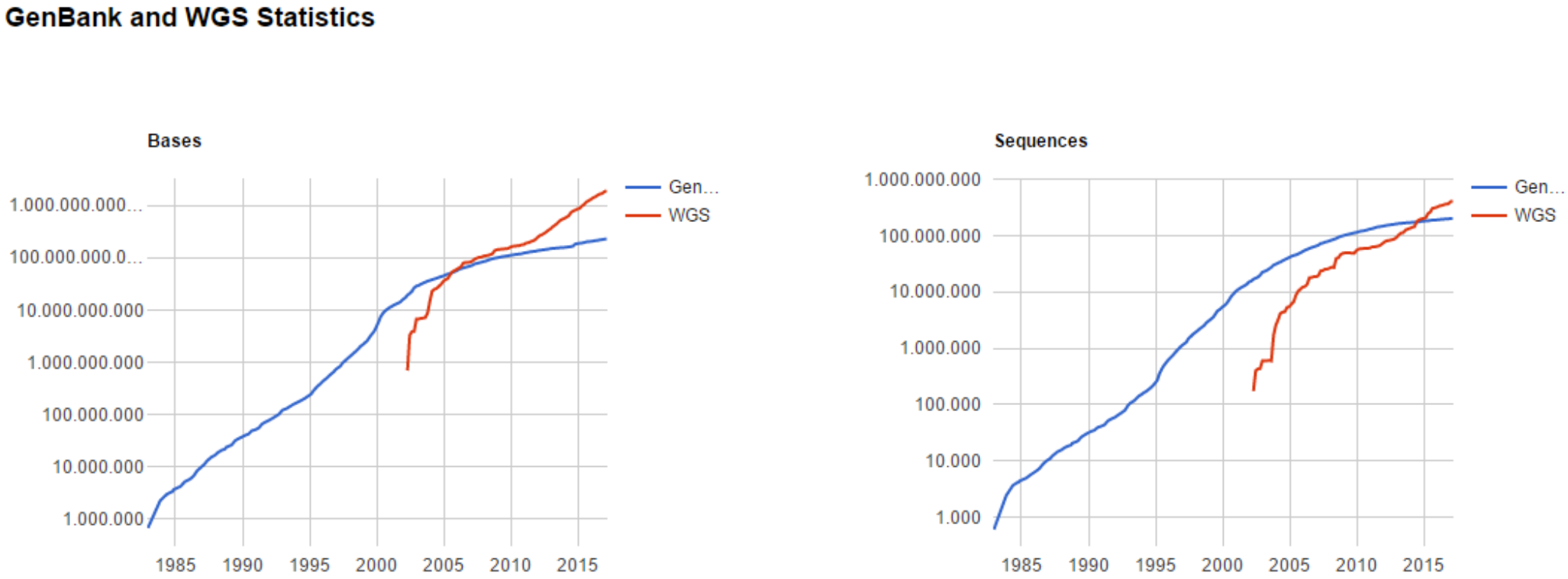
Can be of two types, PAM or BLOSUM, and they have a numeric index (e.g. PAM250)

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

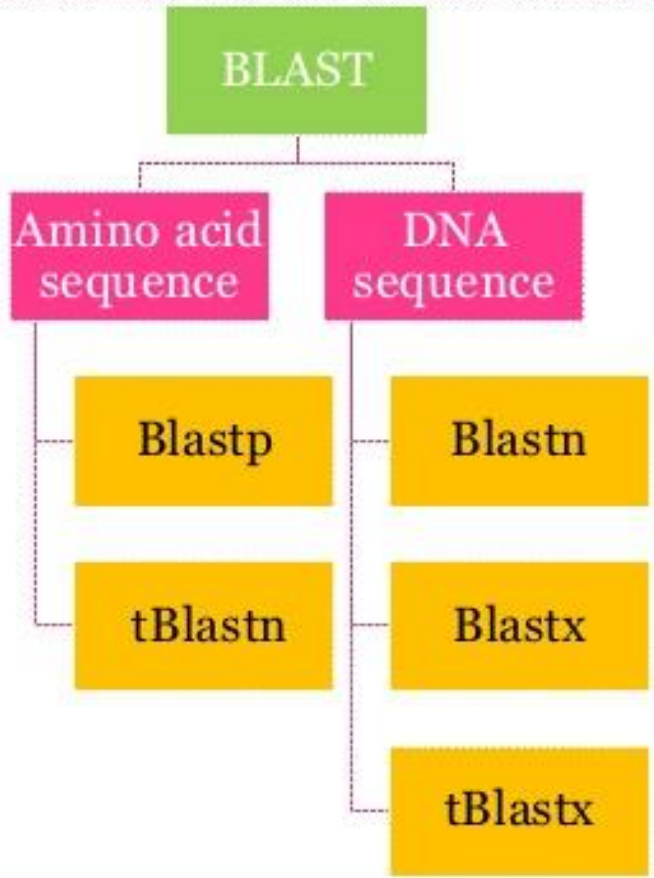
An heuristic approach to sequence alignment

Sequence similarity searches are carried out in enormous databases… we simply can’t expect to evaluate all alignemnt possibilities

N.B. the graphs below are in Log10 scale!!!



BLAST Types



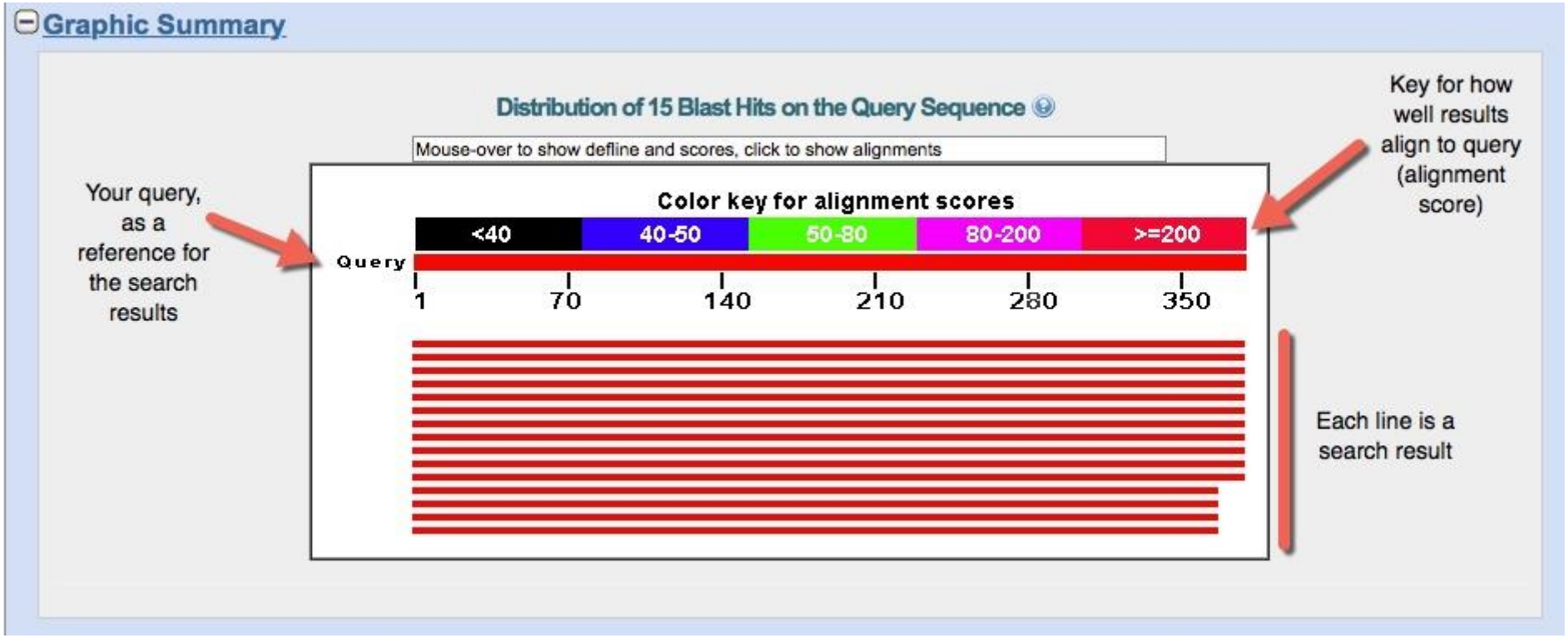
- Blastp : compares protein query against proteins sequence database.
- tBlastn : compares protein query against the all six reading frames of a translated nucleotide sequence database.
- Blastn : compares nucleotide query against nucleotide sequence database.
- Blastx : compares six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.
- tBlastx : compares nucleotide query against translated nucleotide sequence database.

BLAST parameters

- Depend on the type of BLAST!
- We can limit the database to taxonomical entries (e.g. Metazoa, Bacteria, Homo sapiens, Primates, etc.)
- Word size (W): «word» use to nucleate regions of similarity: high W -> higher speed, lower sensitivity; low W: slower speed, higher sentitivity
- Substitution matrix: PAM/BLOSUM. In BLASTn match/mismatch
- Gap cost: penalties applied to gap existence and extension

How to intepret BLAST results?

- All the parameters mentioned above are combined to give you an e-value: describes the number of hits one can "expect" to see just by chance when searching a database of a particular size



How to intepret BLAST results?

- Results will be ordered from the «best» result to the «worst» result, i.e. from the most the least significant hit
- Basically, the significance of each hit is evaluated by the alignment score of the alignment between your **query** sequence (input) and the **subject** sequence (the hit found in the database)
- **The lower the e-value, the better the result**, i.e. highly significant hits will tend to have an-value = 0 or very close to 0
- The e-value depends on the score of the alignement, determined by the **similarity between the query and the subject** (calculated with a PAM/BLOSUM matrix) and by the **length of the alignment**

BLAST in genomic studies

- Annotation of intron/exon regions: mRNAs (query) vs genome (database)
- Functional annotation of genes (by similarity/orthologies): similar genes in different species are likely to encode proteins with similar/identical functions
- Gene prediction: annotation of genes by similarity. E.g. The position and intron/exon organization can be inferred by BLASTing human mRNAs vs the orangutan genome
- Different applications: metagenomics (OTUs assignment), comparative genomics/phylogenomics and many others!

Be careful with your interpretation! SIMILARITYvs HOMOLOGY

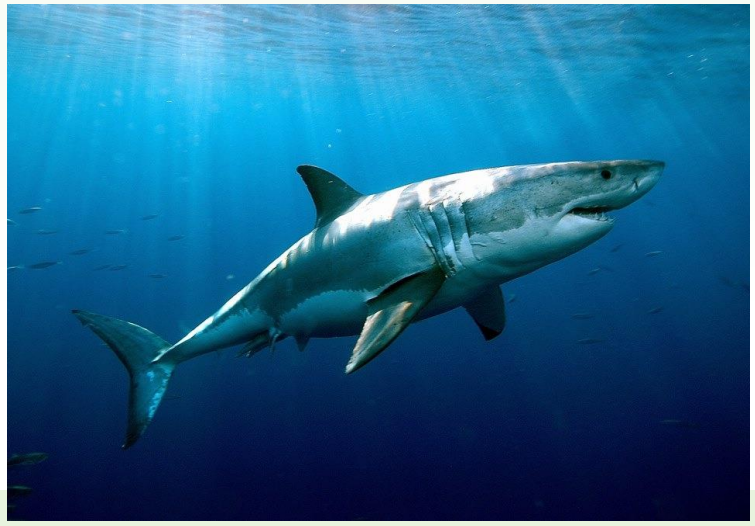
- We need to pay attention to the different meaning of these two words in a biological context. They are not synonymous and considering them as such would be a mistake

SIMILARITY: does not imply any hypothesis concerning the reasons behind similarity itself

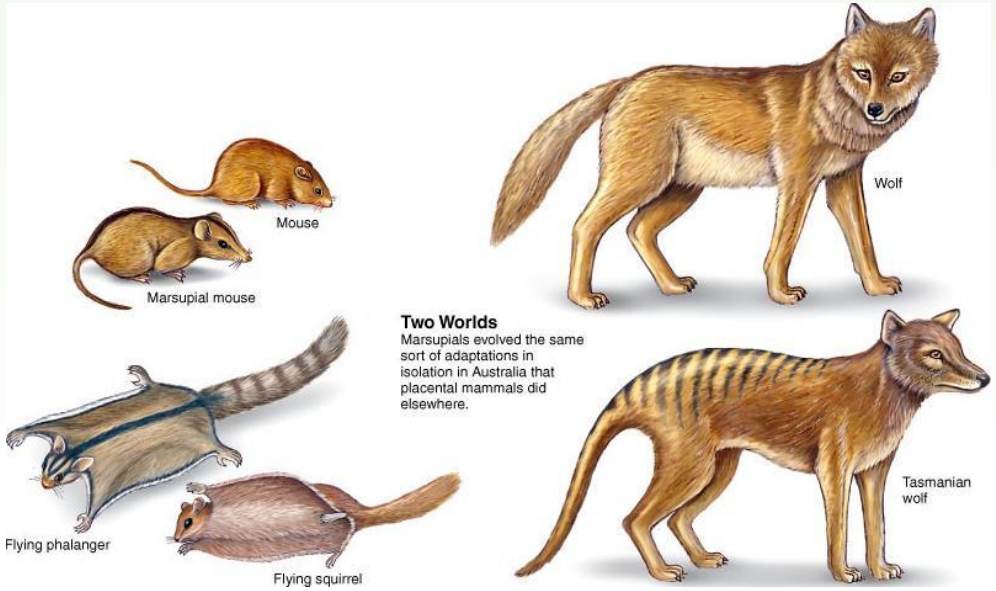
HOMOLOGY: two sequences are homologous if the share the same phylogenetic ancestry

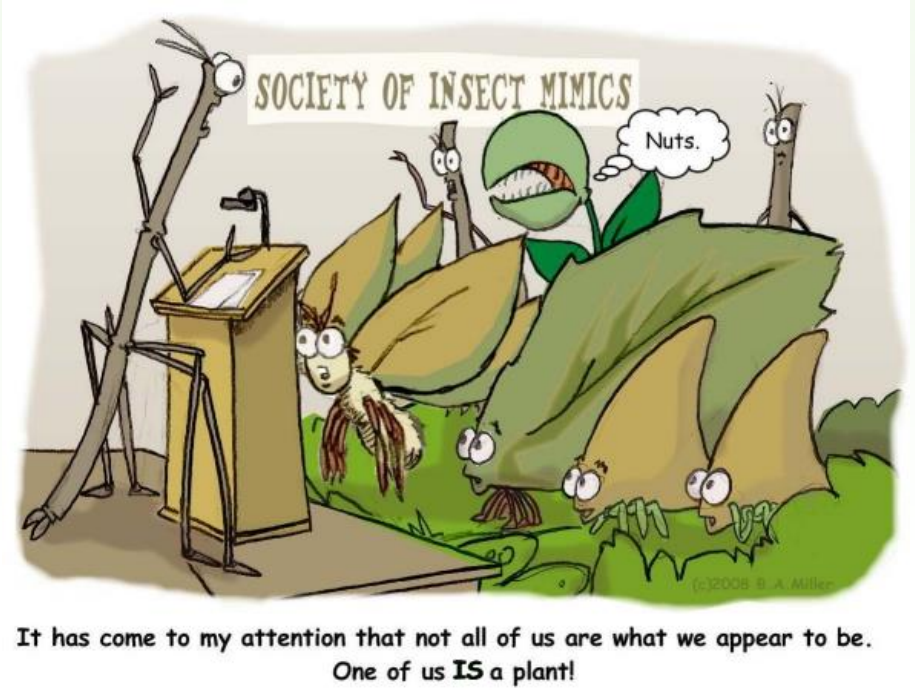
- Biological similarity is often due ti homology, but it can also occurr by chance or be linked to **adaptive convergence**

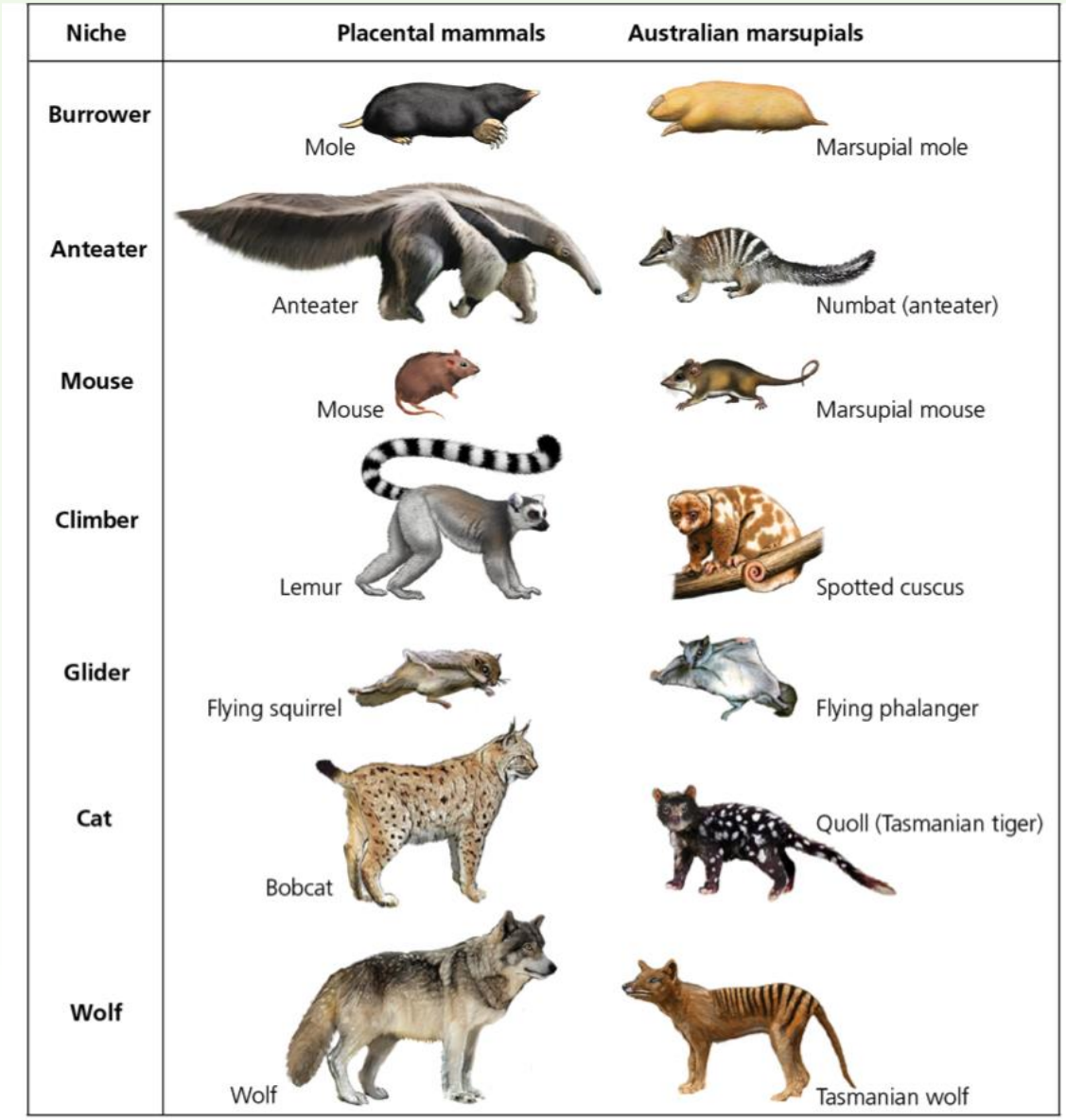
HOMOPLASY: SIMILAR FEATURES WITH INDEPENDENT EVOLUTIONARY ORIGIN (CONVERGENT EVOLUTION)



- It may be easier to approach these concepts by thinking about species, which will be more familiar to students without a molecular biology background
- We need to keep in mind that **sequences evolve in parallel with species** and, although they can be subject to particular selective pressure, they are strictly linked with the species of origin





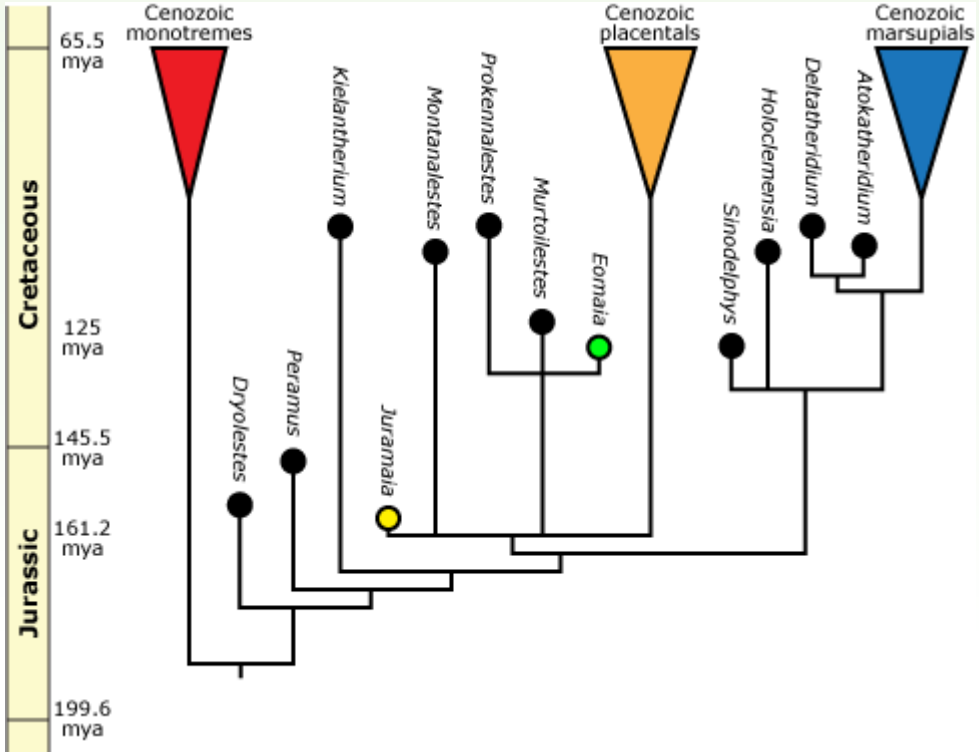


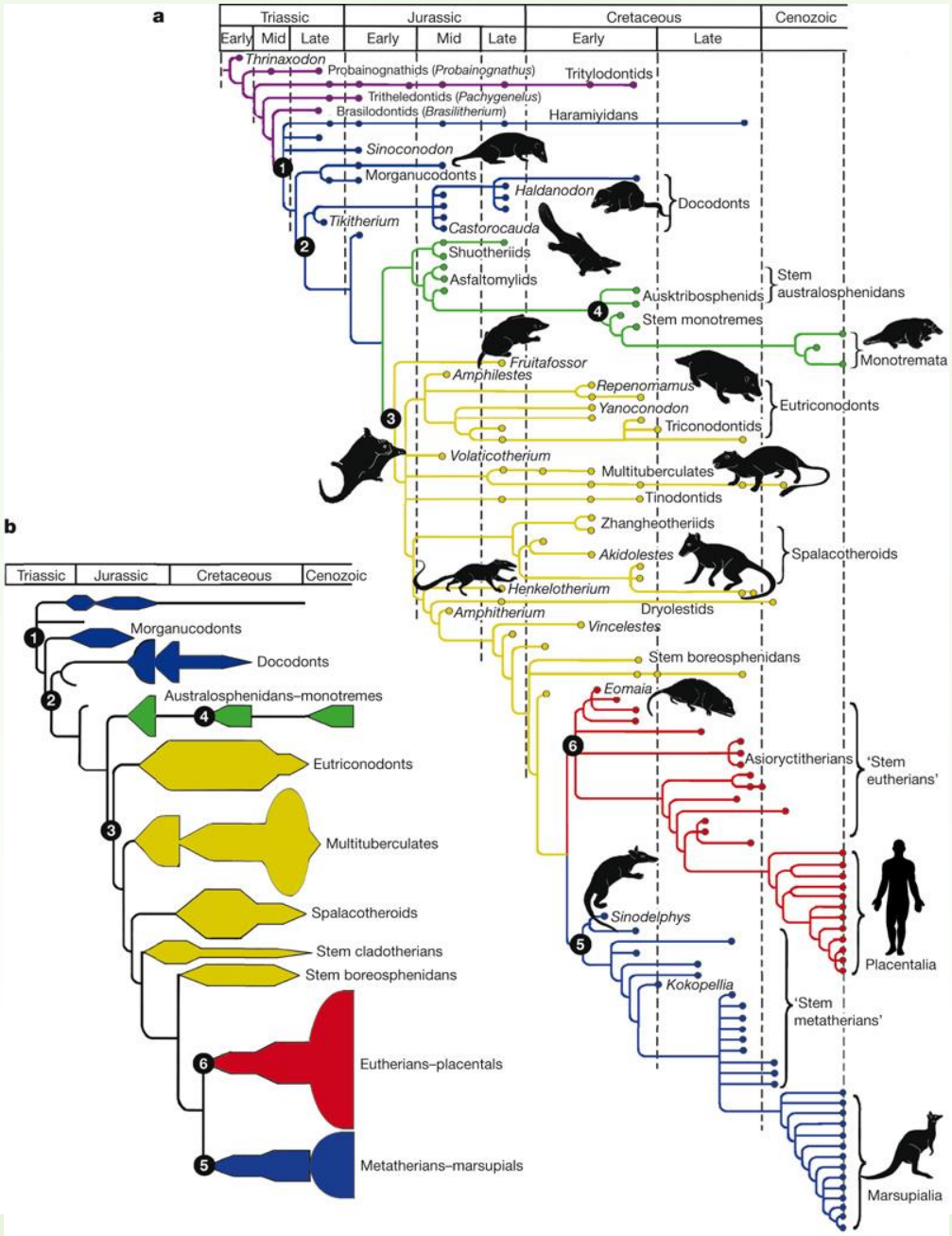
Marsupials represent a great example if we compare them with placental mammals

Although all these lifeforms have been developed morphologically after the split between marsupials and placentals, we can notice extraordinary morphological (and behavioral) similarity between species

The classification of these animals based on simple «similarity» criteria would produce trivial errors as it would probably not identify marsupials as descending from a common ancestor

La radiazione delle varie forme morfologiche dei marsupiali e dei vertebrati è avvenuta indipendentemente. Sono pressioni selettive comuni ed ambienti simili ad aver portato a similaritù notevoli

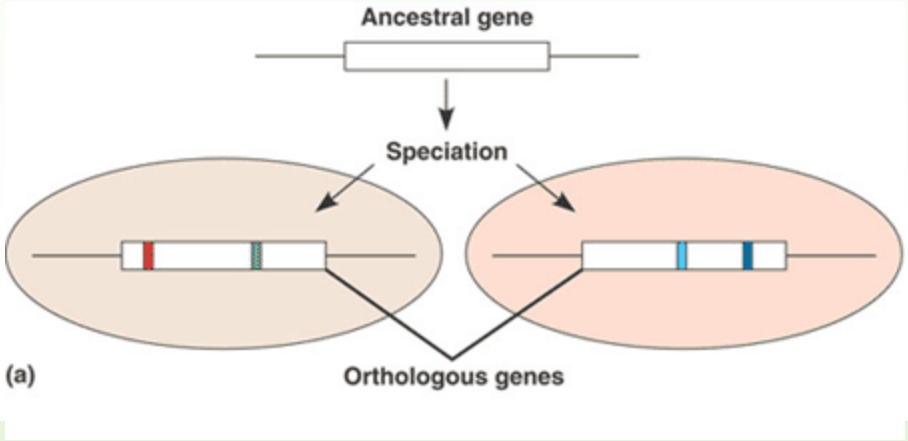




QUANDO DUE SEQUENZE SONO SIMILI, OMOLOGHE, ORTOLOGHE O PARALOGHE?

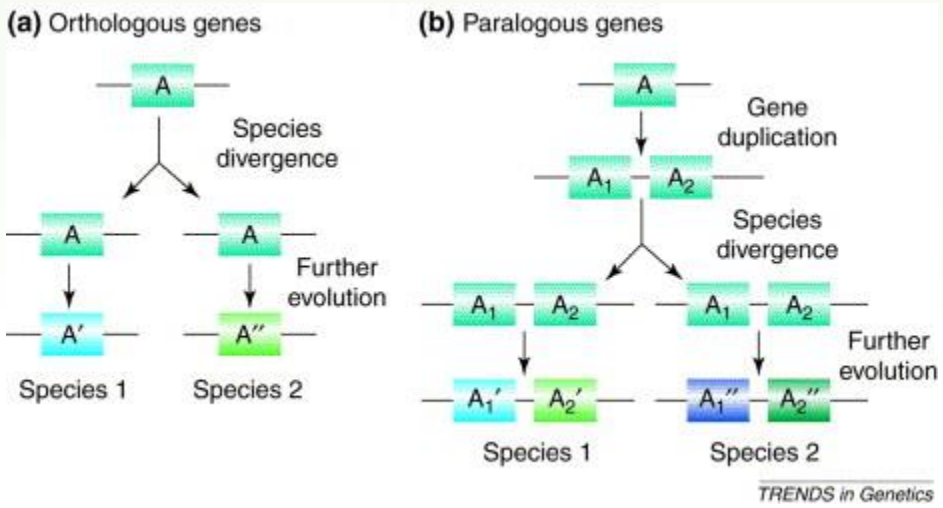
- **Nel trattare le sequenze è sempre più corretto utilizzare il termine similarità, in quanto è sempre possibile stabilire quanto due sequenze siano simili, mentre non sempre si può decidere se la similarità sia dovuta ad omologia, a convergenza adattativa, oppure al caso**

- strutture o sequenze **ortologhe** in due organismi sono sequenze **omologhe** che sono evolute dalla stessa caratteristica nel loro ultimo antenato comune ma che non necessariamente mantengono la loro funzione ancestrale.



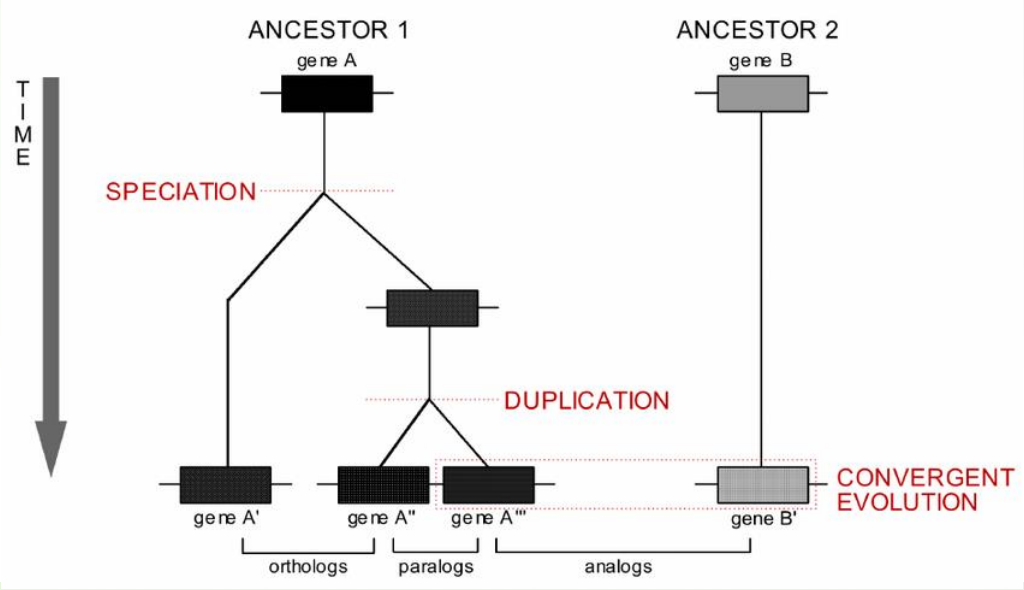
QUANDO DUE SEQUENZE SONO SIMILI, OMOLOGHE, ORTOLOGHE O PARALOGHE?

- sequenze **omologhe** la cui evoluzione riflette invece eventi di duplicazione genica si definiscono **paraloghe**.
- per esempio, la catena alfa dell' emoglobina e' un **paralogo** della catena beta dell' emoglobina e della mioglobina, dal momento che ambedue si sono evolute dallo stesso gene ancestrale attraverso ripetuti eventi di duplicazione genica.



QUANDO DUE SEQUENZE SONO SIMILI, OMOLOGHE, ORTOLOGHE O PARALOGHE?

- Ci possono essere casi più complessi in cui, un pò come per le specie ed i loro caratteri morfologici, si osserva similarità di sequenza senza che ci sia un'origine comune da un unica sequenza ancestrale
- Possiamo in questo caso parlare di sequenze **analoghe**, o molto più semplicemente parlare di **similarità di sequenza senza omologia**



ALCUNI ALTRI IMPORTANTI CONCETTI DA RICORDARE

- Tenete bene a mente che **la similarità di sequenza non nesessariamente si traduce in similarità funzionale**

- Spesso **due sequenze ortologhe svolgono funzioni leggermente diverse in specie diverse**. In caso contrario viene mantenuta anche omologia funzionale

- Spesso **due sequenze paraloghe svolgono funzioni diverse nello stesso organismo**. In caso contrario c'è una ridondanza funzionale.

- Molto spesso **sequenze con nessuna omologia o scarsa similarità svolgono funzioni molto simili se non addirittura identiche**