Introduction to Bioinformatics

Background concepts for sequence analysis

Jacques van Helden
jvanheld@tagc.univ-mrs.fr
Université de Marseille-Méditerrannée
Lab Technological Advances for Genomics and CLinics (TAGC).
http://www.bigre.ulb.ac.be/Users/jvanheld/

Contents

- Evolutionary models
  - Mutations, gene duplicactions, species divergence
  - Homology, orthology, paralogy, etc.
- Pairwise sequence alignment
  - Dot plots (dottup, dotmatcher)
  - Substitution matrices
  - Gapless alignment
  - Alignment with gaps
    - Global alignment (Needleman-Wunsch)
    - Local alignment (Smith-Waterman)
  - Matching a sequence against a database (Fasta, BLAST)
- Multiple sequence alignment (ClustalX)
- Matching motifs against sequences

Bioinformatics

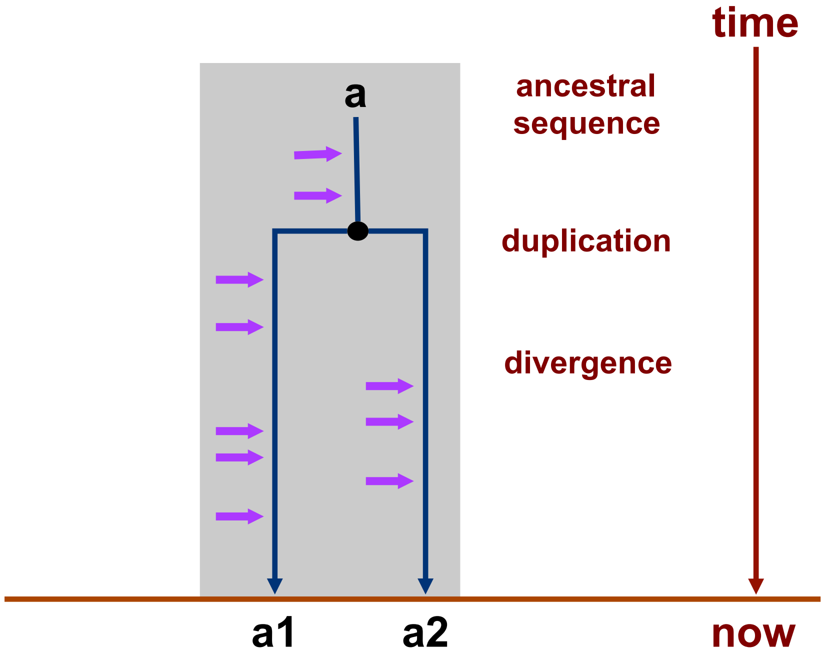
Ana, homo, ortho, para and other logies

Jacques.van.Helden@ulb.ac.be
Université Libre de Bruxelles, Belgique
Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe)
http://www.bigre.ulb.ac.be/

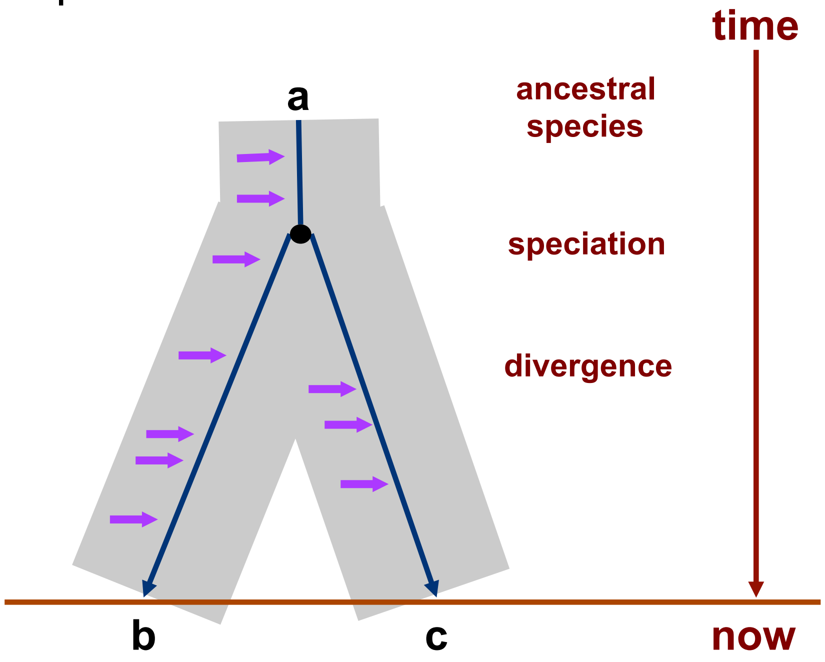
Evolutionary scenario

- We have two sequences, and we suspect that they diverge from some common ancestor (either by duplication, or by speciation).
- Mutational events occur during their evolution
  - substitutions
  - deletions
  - insertions
- Pairwise alignment aims at
  - detecting the regions of similarity between the two sequences
  - inferring the mutational events which occurred from the common ancestor

Gene duplication



Speciation

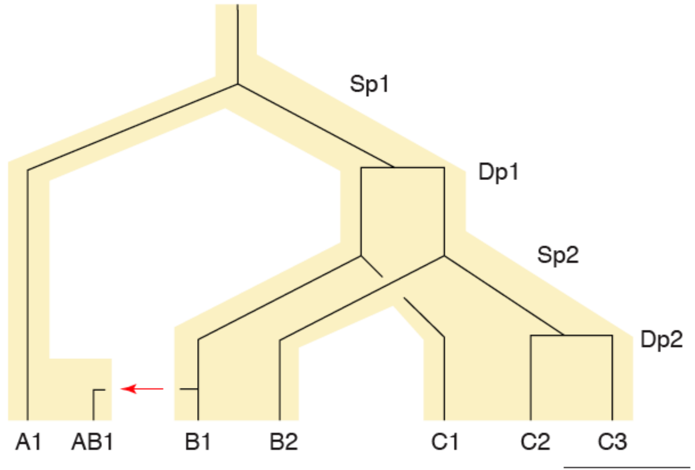


Similarity and homology

- The similarity between two sequences can be interpreted in two alternative ways :
    - ***Homology***: the two sequences diverged from a ***common ancestor***.
    - ***Convergent evolution***: the similar residues appeared ***independently*** in the two sequences, possibly under some selective pressure.
- Inference
    - In order to claim that two sequences are homologous, we should be able to trace their history back to their common ancestor.
    - Since we cannot access the sequence of all the ancestors of two sequences, this is not feasible.
    - The claim that two sequences are homolog thus results from an inference, based on some evolutionary scenario (rate of mutation, level of similarity, …).
    - The inference of homology is always attached to some ***risk of false positive***. Evolutionary models allow to estimate this risk, as we shall see.
- Homology is a Boolean relationship (***true or false*)*:*** two sequences are homolog, or they are not.
    - It is thus **incorrect** to speak about “percent of homology”.
    - The correct formulation is that we can infer (with a measurable risk of error) that two sequences are homolog, because they share some percentage of identity or similarity.

Concept definitions from Fitch (2000)

- Discussion about definitions of the paper
  - Fitch, W. M. (2000). Homology a personal view on some of the problems. Trends Genet 16, 227-31.
- **Homology**
  - Owen (1843). « the same organ under every variety of form and function ».
  - Fitch (2000). Homology is the relationship of any two characters that have descended, usually with divergence, from a common ancestral character.
    - Note: “character” can be a phenotypic trait, or a site at a given position of a protein, or a whole gene, ...
  - Molecular application: two genes are homologous if diverge from a common ancestral gene.
- **Analogy:** relationship of two characters that have developed convergently from unrelated ancestor.
- **Cenancestor**: the most recent common ancestor of the taxa under consideration
- **Orthology:** relationship of any two homologous characters whose common ancestor lies in the cenancestor of the taxa from which the two sequences were obtained.
- **Paralogy:** Relationship of two characters arising from a duplication of the gene for that character.
- **Xenology:** relationship of any two characters whose history, since their common ancestor, involves interspecies (horizontal) transfer of the genetic material for at least one of those characters.



trends in Genetics

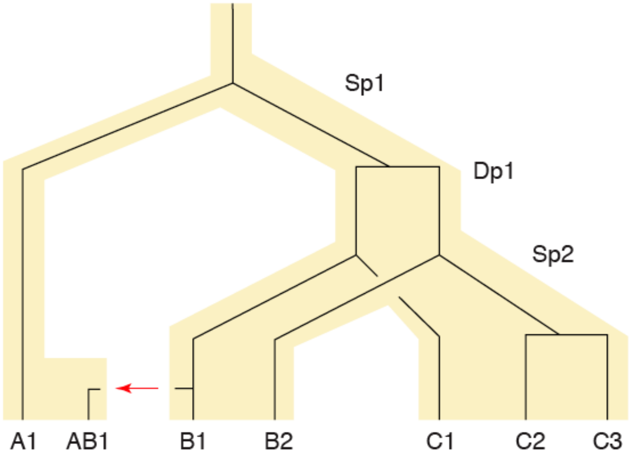
The idealized evolution of a gene (lines) is shown from a common ancestor in an ancestral population (the gray background), descending to three populations labelled A, B and C. There are two speciation events (Sp1 and Sp2), each occurring at the junctions shown as an upside down Y. There are also two gene-duplication events (Dp1 and Dp2), depicted by a horizontal bar. Two genes whose common ancestor resides at a Y junction (speciation) are orthologous. Two genes whose common ancestor resides at a horizontal bar junction (gene duplications) are paralogous. Thus, C2 and C3 are paralogous to each other but are orthologous to B2. Both are paralogous to B1 but orthologous to A1. The red arrow denotes the transfer of the B1 gene from species B to species A. As a result, the AB1 gene is xenologous to all six other genes. All three subtype relationships are reflexive, that is, A1=>B1 implies B1=>A1 where => should be read, for example, as ‘is orthologous to.’ However, the relationships are not transitive. Thus, C2=>A1=>C3 might be true, but it is not necessarily therefore true that C2=>C3, as indeed it is not in the figure if => is read as ‘is orthologous to.’ A different non-transitivity occurs for ‘is paralogous to’ with B2=>C1=>C2.

Analogy
Homology
Paralogy
Xenology or not
(xeonologs from paralogs)
Orthology
Xenology or not

Exercise

- On the basis of Fitch’s definitions (previous slide), qualify the relationships between each pair of genes in the illustrative schema.

❑	P	paralog
❑	O	ortholog
❑	X	xenolog
❑	A	analog



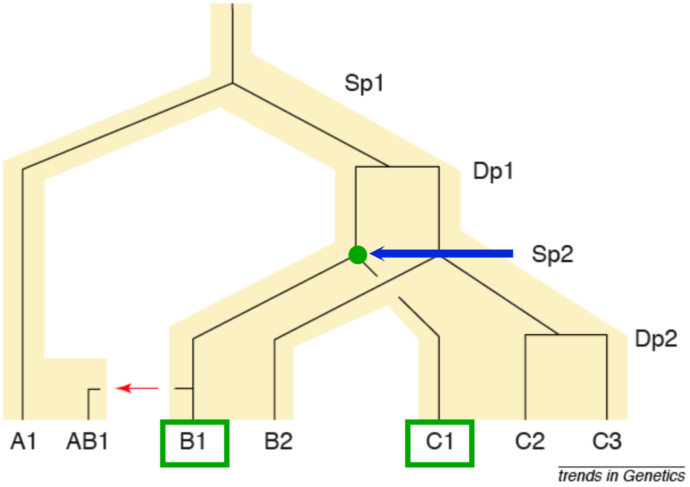
trends in Genetics

	A1	AB1	B1	B2	C1	C2	C3
A1							
AB1							
B1							
B2							
C1							
C2							
C3							

- **Orthologs** can fomally be defined as a pair of genes whose last common ancestor occurred immediately before a speciation event (ex: a_1 and a_2).
- **Paralogs** can fomally be defined as a pair of genes whose last common ancestor occurred immediately before a gene duplication event (ex: b_2 and $b_{2'}$). **Source: Zvelebil & Baum, 2000**

Exercise

- **Example:** B1 versus C1
  - The two sequences (B1 and C1) were obtained from taxa B and C, respectively.
  - The cenancestor (**blue arrow**) is the taxon that preceded the second speciation event (Sp2).
  - The common ancestor gene (**green dot**) coincides with the cenancestor
- -> B1 and C1 are orthologs



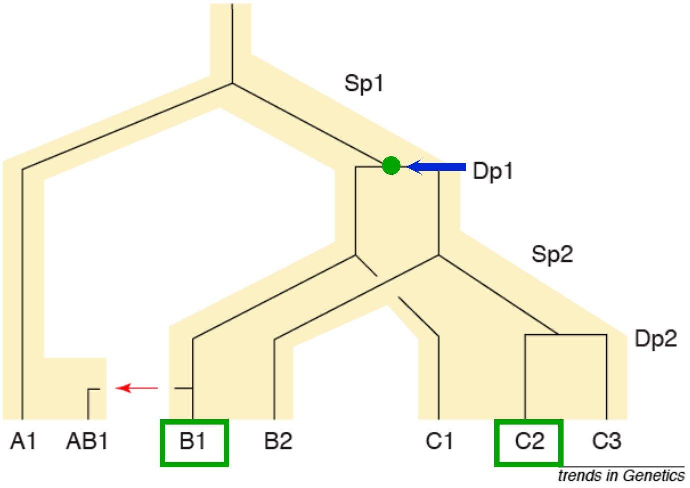
trends in Genetics

	A1	AB1	B1	B2	C1	C2	C3
A1							
AB1							
B1							
B2							
C1			O				
C2							
C3							

- **Orthologs** can fomally be defined as a pair of genes whose **last common ancestor** occurred immediately before a **speciation event**.
- **Paralogs** can fomally be defined as a pair of genes whose last common ancestor occurred immediately before a gene duplication event.
- **Source: Zvelebil & Baum, 2000**

Exercise

- **Example:** B1 versus C2
  - The two sequences (B1 and C2) were obtained from taxa B and C, respectively.
  - The common ancestor gene (**green dot**) is the gene that just preceded the duplication Dp1.
  - This common ancestor is much anterior to the cenancestor (**blue arrow**).
- -> B1 and C2 are paralogs



trends in Genetics

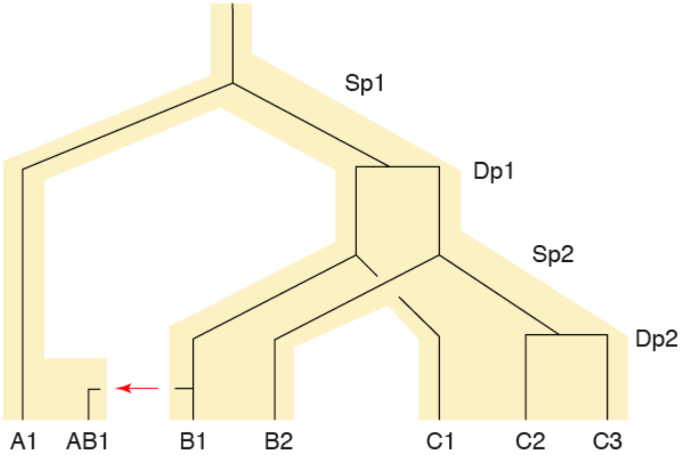
	A1	**AB1**	**B1**	**B2**	**C1**	**C2**	**C3**
A1							
AB1							
B1							
B2							
C1			O				
C2			P				
C3							

- **Orthologs** can fomally be defined as a pair of genes whose last common ancestor occurred immediately before a speciation event.
- **Paralogs** can fomally be defined as a pair of genes whose **last common ancestor** occurred immediately before a **gene duplication event**.
- **Source: Zvelebil & Baum, 2000**

Solution to the exercise

- On the basis of Fitch's definitions (previous slide), qualify the relationships between each pair of genes in the illustrative schema.

❑	P	paralog
❑	O	ortholog
❑	X	xenolog
❑	A	analog



trends in Genetics

The idealized evolution of a gene (lines) is shown from a common ancestor in an ancestral population (the gray background), descending to three populations labelled A, B and C. There are two speciation events (Sp1 and Sp2), each occurring at the junctions shown as an upside down Y. There are also two gene-duplication events (Dp1 and Dp2), depicted by a horizontal bar. Two genes whose common ancestor resides at a Y junction (speciation) are orthologous. Two genes whose common ancestor resides at a horizontal bar junction (gene duplications) are paralogous. Thus, C2 and C3 are paralogous to each other but are orthologous to B2. Both are paralogous to B1 but orthologous to A1. The red arrow denotes the transfer of the B1 gene from species B to species A. As a result, the AB1 gene is xenologous to all six other genes. All three subtype relationships are reflexive, that is, A1=>B1 implies B1=>A1 where => should be read, for example, as 'is orthologous to.' However, the relationships are not transitive. Thus, C2=>A1=>C3 might be true, but it is not necessarily therefore true that C2=>C3, as indeed it is not in the figure if => is read as 'is orthologous to.' A different non-transitivity occurs for 'is paralogous to' with B2=>C1=>C2.

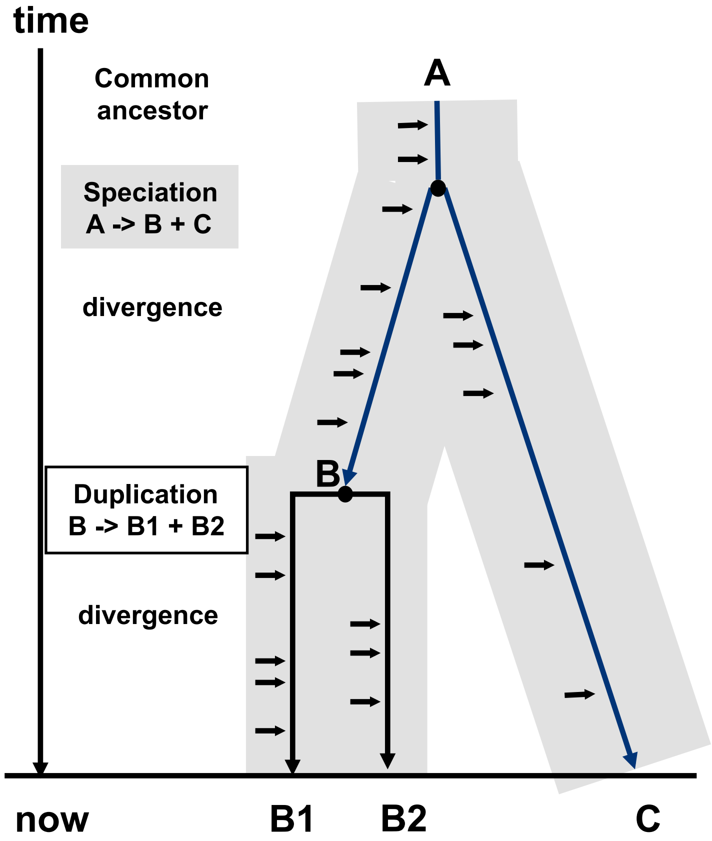
	A1	**AB1**	**B1**	**B2**	**C1**	**C2**	**C3**
A1	I						
AB1	X	I					
B1	O	X	I				
B2	O	X	P	I			
C1	O	X	O	P	I		
C2	O	X	P	O	P	I	
C3	O	X	P	O	P	P	I

How to detect orthology relationships ?

- Bidirectional best hits
  - Concepts
    - Best hit (BH)
    - Reciprocal (RBH) or bidirectional (BBH) best hit.
  - Problem 1: non-reciprocity of the BH relationship, which may result from various effects
    - Multidomain proteins -> non-transitivity of the homology relationship
      - Detection: no paralogy
    - Paralogs in one genome correspond to the same ortholog in the other genome
    - Non-symmetry of the BLAST result
      - Can be circumvented by using dynamical programming (Smith-Waterman)
  - Problem 2: unequivocal but fake reciprocal best hit
    - Duplication followed by a deletion
    - Two paralogs can be BBH, but the true orthologs are not present anymore in the genome (due to duplication).
    - Ex: Hox genes
  - Conceptual problem: intrinsically unable to treat multi-orthology relationships
    - Ex: Fitch figure: B2 is orhtolog to both C2 and C3, but only one of these will be its Best Hit.
- Conclusion: the analysis of BBH is intrinsically unable to reveal the true orthology relationships

Conclusions - Orthology versus paralogy

- The shaded tree represents the history of the species, the thin black tree the history of the sequences.
- Homlogy can be subdivided in two subtypes, depending on the species/ sequence history.
  - **Orthologs** are sequences whose **last common ancestor** occurred immediately before a **speciation** event.
  - **Paralogs** are sequences whose **last common ancestor** occurred immediately before a **duplication** event.
    (Fitch, 1970; Zvelebil & Baum, 2000)
- Example:
  - B and C are orthologs, because their last common ancestor lies just before the speciation
    A -> B + C
  - B1 and B2 are paralogs because the first event that follows their last common ancestor (B) is the duplication
    B -> B1 + B2
- Beware ! These definitions are often misunderstood, even in some textbooks.
- Contrarily to a strong belief, orthology can be a 1 to N relationship.
  - B1 and C are orthologs, because the first event after their last common ancestor (A) was the speciation A -> B + C
  - B2 and C are orthologs because the first event after their last common ancestor (A) was the speciation A -> B + C
- The strategy to **search reciprocal best hits** (**RBH**) is thus a simplification that misses many true orthologs (it is essentially justified by pragmatic reasons).
- The orthology relationship is ***reciprocal*** but ***not transitive***.
  - C <-[orthologous]-> B1
  - C <-[orthologous]-> B2
  - B1 <-[paralogous]-> B2
- The commonly used concept of **clusters of orthologs is thus an aberration**.



How to circumvent the weaknesses of RBH ?

- Solutions to the problems with RBH
  - Domain analysis: analyze the location of the hits in the alignments
    - Resolves the problems of gene fusion (two different fragments of a protein in genome A correspond to 2 distinct proteins of genome B)
  - Analysis of the evolutionary history : full phylogenetic inference + reconciliation of the sequence tree and the species tree
    - Resolves the cases of multiple orthology relationships (n to n)
    - Does not resolve the problems of differential deletions after regional duplications
  - Solving the problem of regional duplications followed by differential deletion
    - Analysis of synteny: neighbourhood relationships between genes across genomes
    - Analysis of pseudo-genes: allows to infer the presence of a putative gene in the common ancestor
    - This is OK when the duplication affects a regions sufficiently large to encompass multiple genes.
- These solutions require a case-by-case analysis -> this is not what you will find in the large-scale databases.

- Resources:
  - EnsEMBL database
  - SPRING database

Criteria for genome-wise detection of orthologs

- Criterion for detecting paralogy
  - Two genes from a given species (e.g. C) are more similar to each other than to their best hit in genome B.
- Pairs of orthologous genes
  - BeT (Best-scoring BLAST hit)
    - Insufficient to infer orthology
  - Bidirectional best hit (BBH)
    - Better approximation
    - Discuss the problem of gene loss
- Clusters of orthologous genes (COGs)
  - Triangular definition of COGs (Tatusov, 1997)
  - KOG: euKaryotic Orthologous Groups
    - Question: is there any interest of defining a new term for eukaryotes ?
- To discuss
  - theoretical weakness of the COG concept, since orthology is NOT a transitive relationship.
  - Pragmatic value of the concept

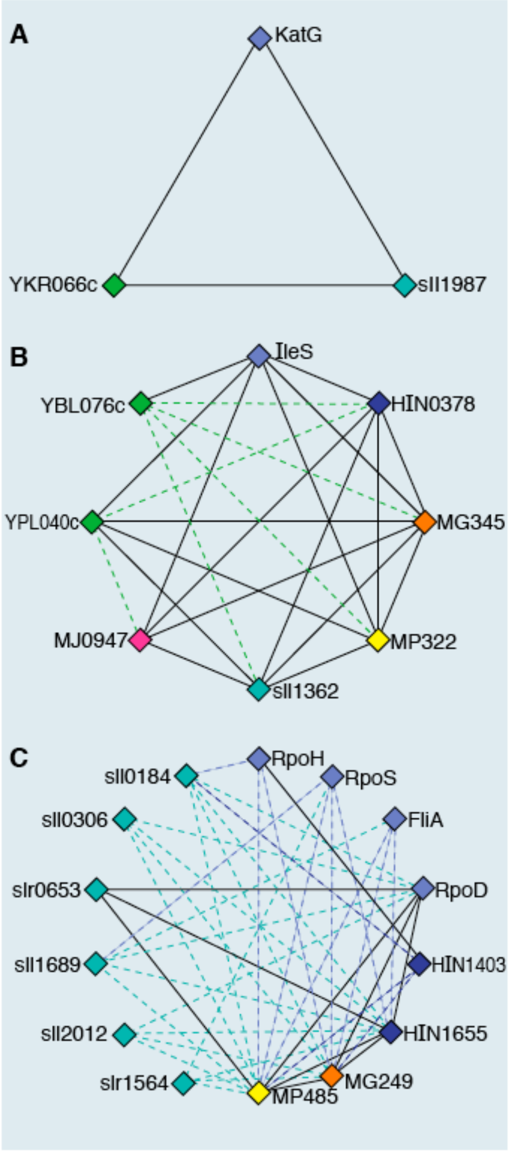


Fig. 1. Examples of COGs. Solid lines show symmetrical BeTs. Broken lines show asymmetrical BeTs, with color corresponding to the species for which the BeT is observed. Genes from the same species are adjacent; otherwise the gene names are positioned arbitrarily. A unique COG ID is indicated in the upper left corner. (**A**) Congruent BeTs form a triangle, the minimal COG. Origin of the proteins: KatG, *E. coli*; sll1987, *Synechocystis* sp.; and YKR066c, *S. cerevisiae*. Note that all the BeTs are symmetrical. (**B**) A simple COG with two yeast paralogs. Origin of the proteins: IleS, *E. coli*; HIN0378, *H. influenzae*; MG345, *M. genitalium*; MP322, *M. pneumoniae*; MJ0947, *M. jannaschii*; and YBL076c and YPL040c, *S. cerevisiae*. Note the adjacent triangles with a common side, for example, IleS-MG345-MJ0947 and sll1362-MG345-MJ1362. YPL040c is the yeast mitochondrial isoleucyl-tRNA synthetase; the bacterial orthologs and that from *M. jannaschii* are the BeTs for this yeast protein, but the reverse is true only of the bacterial proteins (symmetrical BeTs). Conversely, for YBL076c, which is the yeast cytoplasmic isoleucyl-tRNA synthetase, the *M. jannaschii* ortholog is a symmetrical BeT, whereas the bacterial BeTs are asymmetrical. (**C**) A complex COG with multiple paralogs. Origin of the proteins: RpoH, RpoS, RpoD, and FliA, *E. coli*; HIN1403 and HIN1655, *H. influenzae*; MG249, *M. genitalium*; MP485, *M. pneumoniae*; sll0184, sll0306, slr0653, sll1689, sll2012, and slr1564, *Synechocystis* sp. RpoD, HIN1655, slr0653, and MG249 are major sigma factors (σ70), whose function is universal in bacteria; note the fully symmetrical relationships between these proteins. The other proteins are specialized sigma factors whose radiation from the ancestral family apparently was accompanied by modification of the function and involved accelerated evolution; note the asymmetrical BeTs.

Figure from Tatusov, 1997

Example of pairwise alignment

- Example of alignment

```
TTTGCGTT--AAATCGTGTAGCAATTT    s=substitution
s|ss||||ggs||ggggg|||||||s|    g=gap
ATGCCGTTTTTAA-----TAGCAATAT    |=identical residues
```

- Gaps, insertions and deletions
  - Gaps can reflect either an insertion in one of the sequences, or a deletion in the other one.
  - The simple observation of two aligned sequences is insufficient to decide whether a gap results from an insertion or a deletion.
  - The term ***indel*** is sometimes used in this case to designate the evolutionary event.

Global versus local alignment

- Global alignment
  - (e.g. Proteins having a common ancestor and being conserved over their whole sequences)

```
LQGPSKGTGKGS-SRSWDN
|----|--|||---|--|-
LN-ITKSAGKGAIMRLGDA
```

- Local alignment
  - Example: proteins sharing a common domaine

```
LQGPSSKTGKGS-SSRIWDN
      |-|||
LN-ITKKAGKGAIMRLGDA
```

Adapted from Didier Gonze

Some definitions

- Identity
  - The level of identity is a simple calculation of fraction of residues which are identical between the two aligned sequences.
- Similarity
  - Two residues are considered similar if their substitution does not affect the function of the protein.
  - The level of identity is a simple calculation of fraction of residues which are similar between the two aligned sequences.
  - We will see below the criteria to consider that two residues are similar.
- Homology
  - Homology indicates the fact that two sequences diverged from a common ancestor.

Suggested readings

- Homology
  - Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families. Science 278, 631-7.
  - Fitch, W. M. (2000). Homology a personal view on some of the problems. Trends Genet 16, 227-31.
  - Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39, 309-38.
  - Zvelebil, M. J. and Baum, J. O. (2008). Understanding Bioinformatics. Garland Science: New York and London.
- Phylogenetic profiles
  - Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families. Science 278, 631-7.
  - Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96, 4285-8.
- Gene fusion
  - Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. & Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. Science 285, 751-3.