Concepts of homology, orthology, paralogy, synteny and how these concepts are used in Insyght

(Version 1.0.0, October 2015)

Insyght is developed by the Maiage lab at INRA.

Citation:
Lacroix T., Loux V., Gendrault A., Hoebeke M., and Gibrat J-F. (2014) Insyght: navigating amongst abundant homologues, syntenies and gene functional annotations in bacteria, it's that symbol! Nucleic Acids Res. 42(21):e162. PMID: 25249626

Content:

1/ The concepts of homology, orthology and paralogy

The concept of homology, although essential in biology, is often misinterpreted or misused (Fitch WM, Trends genet, 2000). Two genes are homologous if they descent from a common evolutionary ancestor. Two types of homologs can be distinguished: orthologs and paralogs (see the figure below). Orthologs result from a speciation process and paralogs from a duplication event. One can also consider xenologs that result from horizontal gene transfers between organisms. Horizontal gene transfers are mostly observed in prokaryotes. Duplication events, gene losses and horizontal gene transfers being superimposed over the speciation process make the analysis of genomic data more complex.

Homologous proteins (the products of the homologous genes) share common properties, which result from their common origin. The common protein ancestor had a particular sequence, three-dimensional structure and function. Its descendent in modern organisms may have kept, in spite of multiple mutations, insertions and deletions that occurred over time, similar sequences. However, sequence similarity is not the best-conserved property of homologous proteins. Numerous examples of homologous proteins with less than 20% sequence identity after alignment are known. On the other hand, the 3D structures of the descendent proteins are relatively well conserved. The last property, on which is based the so-called homology-based annotation, concerns the conservation of the function of the common ancestor. The central issue that annotators have to solve is the following: do homologous proteins have conserved the ancestor function or do their functions have evolved? Function evolution may result in a change of specificity or regio-specificity or even in a more drastic change of the biochemical function.

It is important to discriminate orthologs from paralogs in the list of homologs provided by sequence comparison methods. It is generally assumed that when a duplication event occurs, one of the copies keeps the initial function whereas the other one is free to adopt a new function. Therefore, it is essential to unravel the evolutionary relationships between homologs to correctly assign the function. Biologists often consider that orthologs exhibit the same function in their respective genome. As shown on the figure, this is not correct since pairs {A1 - B1} and {A1 - B2} are both orthologous but gene B1 has kept the function of the common ancestor while gene B2 has not.

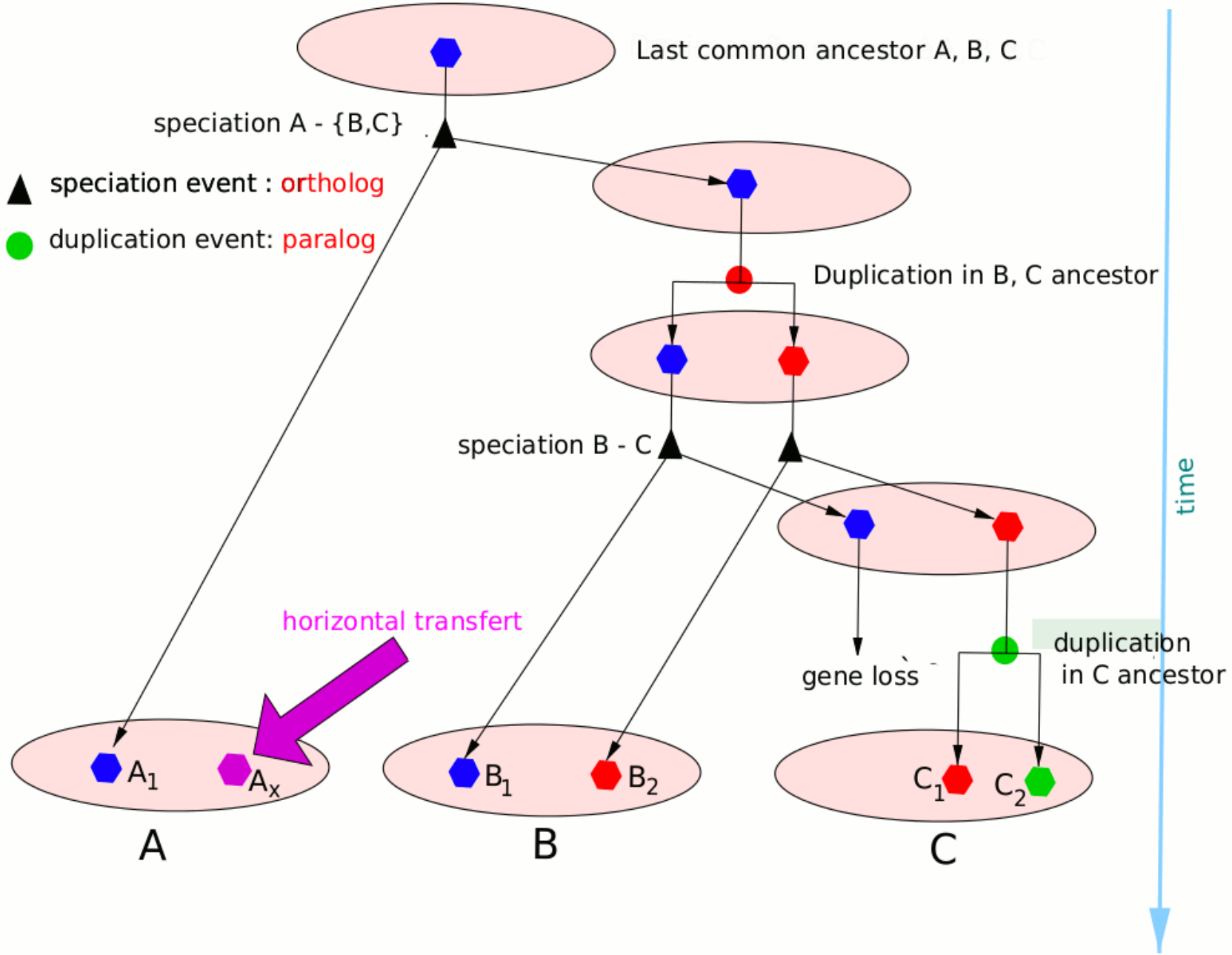
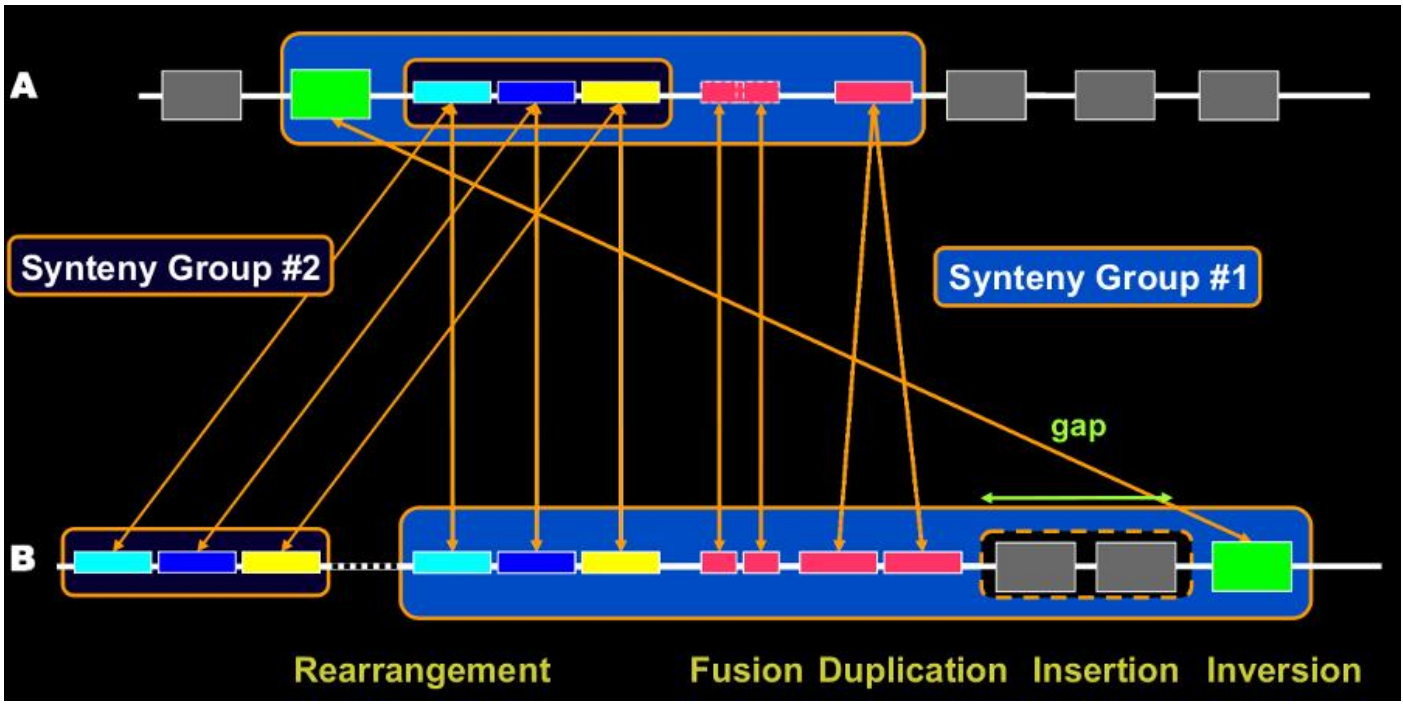

Last common ancestor A, B, C
speciation A - {B,C}
speciation event : ortholog
duplication event: paralog
Duplication in B, C ancestor
speciation B - C
horizontal transfert
gene loss
duplication
in C ancestor
time
A1
Ax
B1
B2
C1
C2
A
B
C

Figure 1: The figure present a hypothetical evolutionary tree that traces the evolutionary history of a gene (the blue hexagon – we assume that gene functions are characterized by their color) present in the last common ancestor of three contemporary organisms labeled A, B and C. Triangles represent speciation events, the first one marking the separation between the A and {B-C} lineages, the second one between the B and C lineages (all intermediary ancestors are not shown in the figure). Colored circles represent duplication events. For instance, the red circle represents the duplication of the blue gene in an ancestor of B and C. The green circle represents the duplication of the red gene in an ancestor of C. The descendent of the blue gene was lost in the lineage leading to organism C. To determine whether two genes are orthologs or paralogs, one has only to move up from the gene locations in the tree until a junction is met. If this junction corresponds to a speciation event the two genes are orthologs, if it corresponds to a duplication events the two genes are paralogs. Therefore paralogs are not limited to homologous genes within the same genome, as often thought. For instance, gene pairs {B1 - C1} and {B1 - C2} are paralogs. By contrast, pairs {B2 - C2} and {B2 - C1} are both orthologs. However, C1 has retained the same function as B1, unlike C2. A1 is ortholog to genes B1, C1, B2 and C2. Notice that the function of A1 has been retained in organism B (by B1) but has been lost in organism C. Ax is a xenolog, that is, a horizontally transferred gene. If the evolutionary history of a gene involves many intermixed speciation and duplication events, together with gene losses in some of the lineages, it can be quite difficult to unravel precisely this evolutionary history.

2/ The concept of synteny

Genomic regions undergo various types of rearrangement at micro and macro scales due to different evolutionary processes. This leads to translocations, duplication, fusion, fission, loss or inversion (El-Mabrouk et al, *Methods Mol Biol*, 2012). Those events participate in conferring the uniqueness of each species or individuals (Dietrich et al, *Science*, 2004 ; Dujon et al, *Nature*, 2004). From a multi-species comparison perspective, each genome can be seen as a succession of regions that are either distinctive or conserved at various degrees. Conserved synteny (or shared synteny) refers to the co-localization of homologous loci across different species:



[1] Vallenet et al, Nucleic Acids Res, 2006.

Together with sequence similarity, gene neighbourhood conservation and phylogenetic profiles provide important clues to identify orthologous genes or infer gene functions (Huynen et al, *Genome Res*, 2000; Zheng et al, *Bioinformatics*, 2005). Conservation in the ordering of genes can help in assigning functions for a train of genes at once or providing clues for hypothetical proteins (Doerks et al, *Nucleic Acids Res*, 2004; Zdobnov et al, *Nucleic Acids Res*, 2005). Moreover, shared synteny may indicate a relationship between gene products such as protein-protein interaction (Dandekar et al, *Trends Biochem Sci*, 1998) or functional coupling (Overbeek et al, *Proc Natl Acad Sci U S A*, 1999; Tamames et al, *J Mol Evol*, 1997). Transcriptional activity has also been correlated to conserved synteny in expression pattern and transcriptional regulation studies (Rodelsperger et al, *Nucleic Acids Res*, 2011; Roy et al, *Nature*, 2002).

3/ The use of these concepts in Insyght

The best way to disentangle a complex evolutionary history and to precisely assign the function is to make use of phylogeny methods. However, biologists often use, as a proxy for the time-consuming phylogeny methods, the bi-directional best hit (BDBH). Proteins p1 of genome G1 and p2 of genome G2 give rise to a BDBH if p2 is the homolog of p1 with the highest score when comparing p1 with the proteome of G2, and reciprocally, p1 is the homolog of p2 with the highest score when comparing p2 with the proteome of G1. Proteome comparisons are done with a sequence alignment method, in general Blast.

3.1/ Our approach

In Insight, we (knowingly) misuse the term orthology by describing as orthologs two protein coding genes that give rise to a BDBH and whose sequence alignment includes more than 50% of the length of the shortest proteins with an e-value less than 0.01. In fact, we simply mean that the proteins are likely to have kept the same function. Two proteins for which the e-value of the sequence alignment is less than 0.01 and that do not fall into the ortholog category are considered homologous but are not displayed unless they belong to a synteny. Syntenies are delineated with a dynamic programming algorithm that determines the highest scoring paths among all the possible gapped chains of collinear homologues. Figure 2 below displays an example of a scoring matrix:

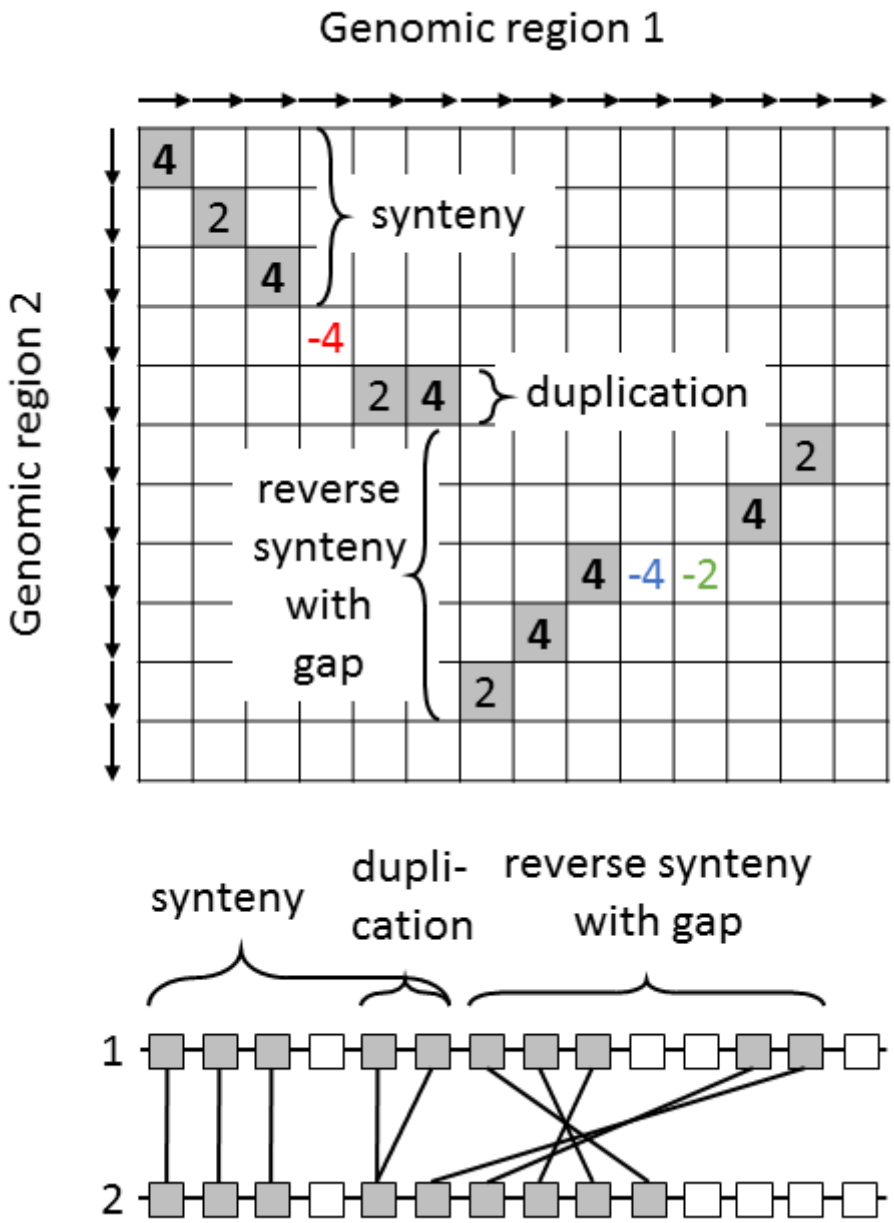


Figure 2: Each cell in the matrix contains a score for a pair of proteins, one from the first genomic segment and the other from the second genomic segment. The scoring scheme that we use by default is:

- Ortholog: **4**
- Homolog: 2
- Mismatch: -4
- Gap creation: -4
- Gap extension: -2
- Unrelated: 0 (displayed as a white cell)

This scoring scheme allows for the insertion of small gaps. Collinear synteny regions appears as diagonals in the matrix. The minimum score to keep a synteny is 4, meaning that we keep syntenies of length 1 corresponding to a single pair of “orthologs”.

3.2/ Discussion and further reading on inference methods

3.2.1/ Warning on annotation errors in current genome files

Regarding the inference of gene functions, the error rate of functional annotations is estimated to lie between 5~40% depending on the annotated genome (Jones et al, BMC Bioinformatics, 2007 ; Poptsova et al, Microbiology, 2010 ; Devos et al, Trends Genet, 2001). Errors are mostly due to the transferring of functional annotations between predicted “homologs” with a low percentage of similarity (ex 30%) or that are missing a domain etc.

3.2.2/ Protein alignment coverage

Blast is a local alignment algorithm (Altschul et al, J. Mol. Biol, 1990). Using the approach described above (section 3.1/), we observe an alignment coverage of ~90% on average for orthologs. Global alignment algorithms (i.e. Needleman et al, J Mol Biol, 1970) provide an alignment coverage of 100% but have an increased computational cost. Using a global algorithm is not feasible on large dataset (Sonnhammer E.L. et al, Bioinformatics, 2014).

3.2.3/ Comparison of inference methods

The BDBH method we use has been shown to be robust, as highlighted by the following papers that assess different ortholog's inference methods:

- Altenhoff et al, Phylogenetic and Functional Assessment of Orthologs Inference Projects and Methods, PLoS Comput Biol, 2009
- Hulsen et al, Benchmarking ortholog identification methods using functional genomics data, Genome Biol, 2006;
- Chen F et al, Assessing performance of orthology detection strategies applied to eukaryotic genomes, PLoS One, 2007;
- Fulton et al, Improving the specificity of high-throughput ortholog prediction, BMC Bioinformatics 2006;
- Yu et al. QuartetS: a fast and accurate algorithm for large-scale orthology detection; Nucleic Acids Res. 2011.

False positives and false negatives arise from the complexity of the evolutionary mechanisms and are inherent to all predictive computational methods. Those aspects are discussed for example in the following papers:

- Lerat et al, PLoS Biol, 2003
- Koski et al, J Mol Evol, 2001
- Yu et al, Nucleic Acids Res, 2011
- Sonnhammer et al, Bioinformatics, 2014

The main goal of the Insyght method and display tool is to give biologists clues on putative biologically relevant proteins relationships to be further investigated in wet lab experiments.

3.2.4/ The choice for the cutoffs

The cutoff we chose for BDBH orthologs is based on both the e-value (<0.01) and the alignment coverage (>50%). Different types of cutoff are reported in the literature. Another example is the “BLAST Score Ratio Values” (BSRV) which is the ratio bit score / maximal bit score (Lerat et al, PLoS Biol, 2003). Every cutoff-based method has its limitations (i.e. for the BSRV: Gibbons et al, Evaluation of BLAST-based edge-weighting metrics used for homology inference with the Markov Clustering algorithm, BMC Bioinformatics, 2015).