Orthology Inference

Christophe Dessimoz
cdessimoz@inf.ethz.ch

Bioinformatics and Comparative Genome Analysis
Institut Pasteur Paris, July 2010
Outline

• Motivation and Definition
• Orthology Inference
  • Pairwise methods
  • Tree reconciliation methods
• Verification
• Limitations and Future Directions
Motivation and Definition
Definition

DISTINGUISHING HOMOLOGOUS FROM ANALOGOUS PROTEINS

WALTER M. FITCH

Formally

Two homologous genes \((x,y)\) are orthologs if they started diverging through a speciation event.

**Observations**

- relation defined on a *pair* of genes
- non-transitive
Why useful?

- Species Tree Reconstruction

- Functional annotation
  Prevalent model:
  - orthologs share similar function
  - paralogs have different functions

- Physical Mapping between genomes

...
More terminology

in-paralogs( 🐶, 🐾, $S_1$)
out-paralogs( 🐶, 🐾, $S_2$)
Other usages of orthology*

- Genes with the same *function*
  - “isofunctional homologs”
  - “equivalogs”
- Homologs that are in the same genomic context
  - “positional ortholog”

*not recommended*
Inference
Homology

- Commonly inferred by sequence similarity:
  - “All-against-all”
  - Profile-based search

- At low similarity (20-30% identity, “twilight zone”), protein structure tends to be better conserved, but

  - structure often unknown
  - only for conserved regions
Pairwise alignments statistics

Alignment score of 2 unrelated sequences are distributed according to Gumbel distribution (2 parameters, fat tail)

Params are estimated from seq lengths and scoring matrix (Karlin-Altschul theory)

Significance is assessed by its E-value (expected # spurious matches with score equal or higher), computed from probability density

http://www.math.ku.dk/~richard/courses/binf_project/Stinus-BLAST.pdf
Orthology

Bidirectional
Best-Hit

Gene/Species Tree
Reconciliation

Dufayard et al., Bioinformatics, 2005
Pairwise Methods
The basic idea

Between two species, orthologs are closer than paralogs.

• Closer genes usually have higher alignment scores → species-specific top scoring hit is likely to be an ortholog
• **Corresponding ortholog might be missing** → require symmetry

“Bidirectional best hit” (BBH)
Refinements

- Instead of score, use evolutionary distance “Reciprocal smallest distance” (RSD)
- Relax the top/smallest requirement to include more than one orthologs (e.g. 1:many orthology)
- Take into account statistical uncertainty of distance estimates
- Detect differential gene losses

Wall et al., Bioinformatics, 2003
Dessimoz et al., RECOMB CG Dublin, 2005
Fulton et al., BMC Bioinformatics, 2006
Dessimoz et al., Nucleic Acids Res, 2006
Roth et al., BMC Bioinformatics, 2008
Verification of stable pairs

Duplication

Speciation

Losses

\(X_1, X_2, Z_3, Z_4, Y_1, Y_2\)
Verification of stable pairs

- \((x_1, z_3)\) and \((y_2, z_4)\) are stable pairs
- \((x_1, z_3)\) signif. closer than \((x_1, z_4)\)
- \((y_1, z_4)\) signif. closer than \((y_2, z_3)\)
- \((x_1, z_4)\) and \((y_2, z_3)\) not signif. different

Dessimoz, Boeckmann, et al., Nucl Acid Res, 2006
How to group orthologs?

• If interested in particular gene x:
  - group all genes orthologous to x

• COGs database:
  - group “triangles” of orthologs
  - merge triangles with common face

• InParanoid (on pairs of genomes):
  - start with a pair of orthologs
  - add in-paralogs (w.r.t. the only speciation)

• OMA
  - all pairs in a given group are orthologs
All protein sequences from full genomes → All x All Comparison → Candidate Pairs → Formation of Stable Pairs → Stable Pairs → Verification of Stable Pairs → Verified Pairs → Orthologs → Group Pairs

<table>
<thead>
<tr>
<th>Pairs</th>
<th>Evolutionary Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Pairs (AP)</td>
<td>Any</td>
</tr>
<tr>
<td>Candidate Pairs (CP)</td>
<td>Homologs</td>
</tr>
<tr>
<td>Stable Pairs (SP)</td>
<td>Orthologs, Pseudo-Orthologs</td>
</tr>
<tr>
<td>Broken Pairs (BP)</td>
<td>Paralogs</td>
</tr>
<tr>
<td>Verified Pairs (VP)</td>
<td>Orthologs</td>
</tr>
<tr>
<td>Group Pairs (GP)</td>
<td>Close Orthologs</td>
</tr>
</tbody>
</table>

Roth et al., BMC Bioinformatics, 2008
http://omabrowser.org
Tree Reconciliation

Maximum Parsimony

Reconciliation
Map between G and S

Introduced implicitly: Goodman et al 1979
explicitly: Page 1994

Zmasek & Eddy 2001

For any $g \in G$, let $M(g) \in S$ be the smallest (lowest) node in $S$ satisfying $\gamma(g) \subseteq \sigma(M(g))$. That is, $M(g)$ points to the ancestral species in $S$ that (we infer) harbored ancestral gene $g$. 
• Mirkin, Muchnik, Smith (1996) conjecture that the map cost function coincide with number of gene duplication and losses.

• Zhang (1997) and Eulenstein (1997) independently prove it, and identify efficient algorithms to compute map: in $O(n)$ and $O(n^*\alpha(n))$ respectively.
Mapping
- Map leaves in $G$ to their species in $S$
- Map inner node $g_i$ in postfix order (from leaves to root):
  - Map $G_i$ to the lowest node $s_i$ such that the species below $g_i$ are all included below $s_i$

Duplication node assignment
- If $g_i$ maps in $S$ to the same node as one of its children, $g_i$ is a duplication node
Tree Rooting?

- Center of gravity
  - Storm & Sonnhammer 2002

- Min # of duplications
  - Hallett & Lagergren 2000
  - Zmasek & Eddy 2002
    (min height to break ties)

- Outgroup (tricky)
  - Huerta-Cepas et al. 2007
Tree Inference Errors?

- Bootstrapping
  - Storm & Sonnhammer 2002
  - Zmasek & Eddy 2002
- Multifurcation (unresolved branches)
  - Dufayard et al. 2005
  - Berglund-Sonnhammer et al. 2006
  - Durand et al. 2006
High Duplication/Loss Rates?

- Maximum parsimony criterion may be inappropriate
Case study: Ensembl Compara

Evaluating Orthology Predictions

Previous work

Benchmarking ortholog identification methods using functional genomics data
Tim Hulsen*, Martijn A Huynen*, Jacob de Vlieg†† and Peter MA Groenen†
Genome Biology 2006, 7:R31

Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes
Feng Chen1,3, Aaron J. Mackey2,3*, Jeroen K. Vermunt4, David S. Roos2,3*
April 2007 | Issue 4 | e383
Comparison Approach

• For each project, map sequences to OMA
  7.16 million sequences in total
  329.2 million orthology relations

• Intersection over all projects: $\emptyset$!
  ➔ “pairwise” tests with OMA
  ➔ “intersection” tests with subset
## Assessment of Orthologs

<table>
<thead>
<tr>
<th>Phylogeny</th>
<th>Conserved function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species Tree Discordance</td>
<td>Gene Ontology</td>
</tr>
<tr>
<td>Phylogenetic Analyses from Literature</td>
<td>Enzyme Classification</td>
</tr>
<tr>
<td></td>
<td>Gene Expression</td>
</tr>
<tr>
<td></td>
<td>Genomic Context</td>
</tr>
</tbody>
</table>
Species Tree Discordance

Homo sapiens
Other Primates
Other Mammals
Other Vertebrates
Protosomes (Worms & co)
Fungi
Species Tree Discordance

5-6 eukaryotes, ML trees

Avg. fraction of correct splits

EggNOG, OrthoMCL, Inparanoid, Homologene, Ensembl Compara, RoundUp, OMA Group, RSD, BBH, OMA Pairwise

lundi, 7 juin 2010
Species Tree Discordance

6-7 bacteria, distance trees

fraction of correct splits
Panther Ontology Conservation

- Relative difference in mean Panther similarity compared to OMA
- Relative difference in number of ortholog predictions compared to OMA

Better values are closer to 0.00.
Limitations and Future Directions
Limits of Pairwise Relations

• Useful if focused on a particular gene

• But several shortcomings:
  - Evolutionary distance?
  - Function conservation?
  - Grouping strategy?

Review
Large-scale assignment of orthology: back to phylogenetics?
Toni Gabaldón

Bioinformatics and Genomics Program, Center for Genomic Regulation, Doctor Aiguader, 88, 08003 Barcelona, Spain. Email: tngabaldon@erg.es

Published: 30 October 2008
Limits of Model

- Lateral gene transfer?
- Gene fusion/fission?
- Domain shuffling?
- Heterogeneous population?
- Hybridization?

Dessimoz et al., RECOMB 2008
Limits of Orthology for function inference

How confident can we be that orthologs are similar, but paralogs differ?

Romain A. Studer and Marc Robinson-Rechavi

Department of Ecology and Evolution, Biophore, Lausanne University, CH-1015 Lausanne, Switzerland and Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland

Homologous genes are classified into orthologs and paralogs, depending on whether they arose by speciation or duplication. It is widely assumed that orthologs share similar functions, whereas paralogs are expected to diverge more from each other. But does this assumption hold up on further examination? We present evidence that orthologs and paralogs are not so different in either their evolutionary rates or their mechanisms of divergence. We emphasize the importance of appropriately designed studies to test models of gene evolution between orthologs and between paralogs. Thus, functional change between orthologs might be as common as between paralogs, and future studies should be designed to test the impact of duplication against this alternative model.

But the assumption that changes in function are commonly associated with duplication has rarely been explicitly tested. Although there have been many studies of comparative genomics focused on the role of duplication (for a review, see Ref. [1]), few have compared the evolution of paralogs with the evolution of orthologs. However, these studies repeatedly find little, if any, specific impact of duplication. This pattern is surprising if the standard model is correct.

This ‘standard model’ makes two predictions. First, paralogs are expected to diverge more per unit of time than orthologs. Second, paralogs are expected to diverge frequently in ways that are rarely observed between orthologs; for example, different substrate specificities. Divergence can concern different aspects of gene function [3],
Limits of Computational Inference

Growth of GO Annotations

![Bar graph showing the growth of GO annotations from 2006 to 2010. The graph compares experimental and automated, uncurated annotations. Experimental annotations show a steady increase, while automated, uncurated annotations show a significant increase in 2010.]
Questions?