Comparative Genomics

- Multiple alignments
- Synteny
- Homologs
- Gene models

Evolution:

- Phylogeny
- Gene expression

Multiple sequence alignment

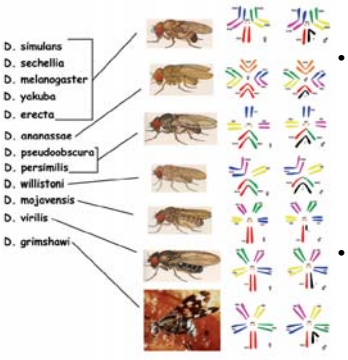
- Heuristic vs. global optimisation

- DP – v. v. slow
- Progressive alignment construction – e.g. Clustal family
- Iterative methods – e.g. MUSCLE
- Consensus methods

- HMMs e.g. HMMer
- Motif finding e.g. MEME – see Regulation lectures

- Not practical on large scale

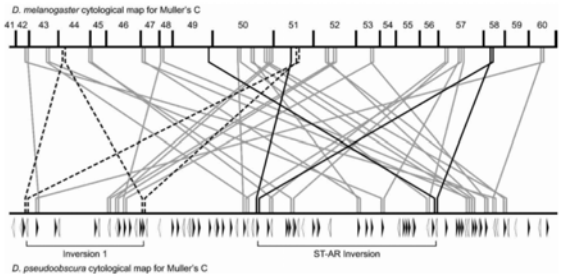
Synteny



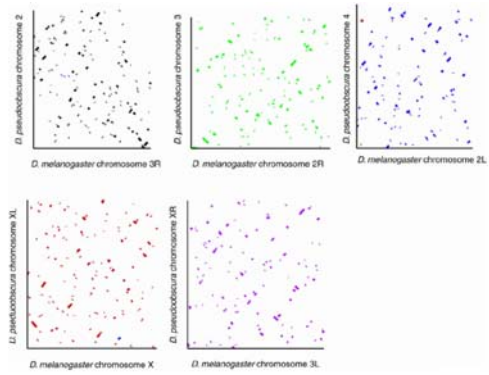
- organisms of relatively recent divergence show similar blocks of genes in the same relative positions in the genome

- different degrees of synteny

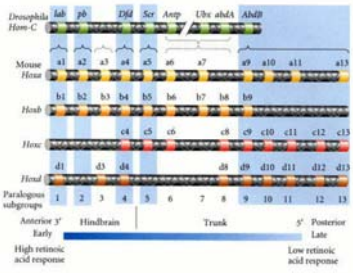
Synteny plot



Synteny – dot plots

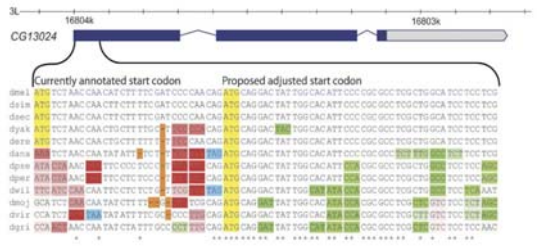


Synteny – Hox cluster



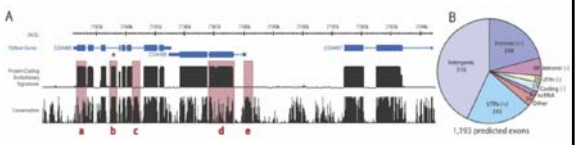
The most prominent syntenic group may be the Hox cluster which is even conserved between flies and mammals.

Making use of multiple genomes



Stark et al Nature. 2007 Nov 8;450(7167):219-232. PMID: 17994088
Lin et al Genome Research 2007 PMID: 17989253

Gene models

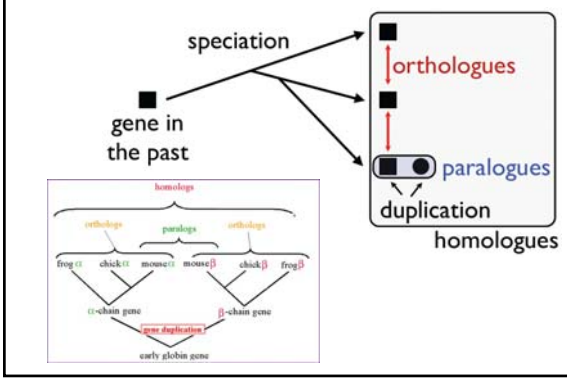


Comparative genomics can help to find additional exons even in the absence of cDNA evidence

Table 1. Categorization of existing gene annotations according to comparative evidence

	Total	Confirmed	Unclear	Rejected*
Named genes	4711	4566 (96.9%)	105 (2.2%)	40 (0.8%)
Well-studied genes	891	882 (98.8%)	8 (0.9%)	1 (0.1%)
Other named genes	3,818	3684 (96.5%)	97 (2.5%)	37 (1.0%)
CGid-only genes	9022	7879 (87.3%)	729 (8.1%)	414 (4.6%)
GO-annotated	4373	3897 (89.1%)	278 (6.4%)	198 (4.5%)
Uncharacterized	4649	3982 (85.7%)	451 (9.7%)	216 (4.6%)
All genes	13,733	12,445 (50.6%)	834 (6.1%)	454 (3.4%)
Noncoding regions	15,564	1 (0.0%)	131 (0.8%)	13,430 (99.1%)

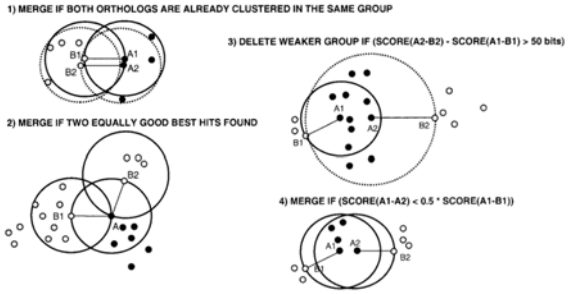
Homologs



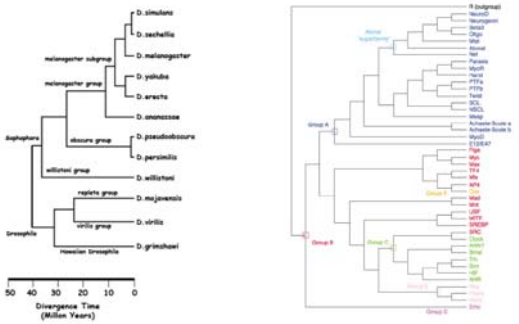
Finding homology

- BLAST bi-directional best hit of protein sequences and minimal sequence identity of 30%
- Protein family clusters (looser criterion, includes paralogy within the family), implemented in
  - ENSEMBL Compara
  - Inparanoid
  - OrthoMCL / MCL

Inparanoid



Phylogeny

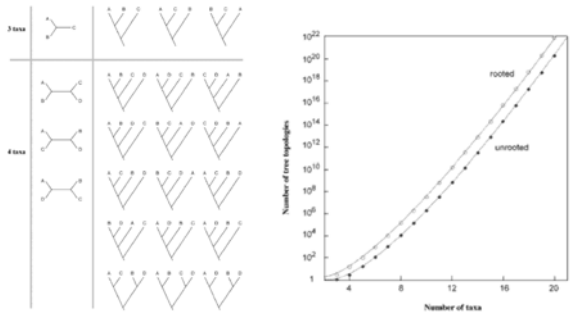


Of species

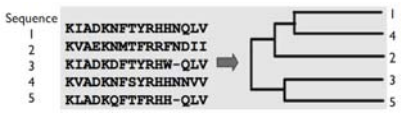
Of proteins

- Phylogenetic trees are usually based on DNA or protein sequences.
- Comparisons are possible even between animals with no physical resemblance.
- Ideally, a phylogenetic reconstruction is unambiguous…

Graph theory shows how difficult tree growing is



General principle



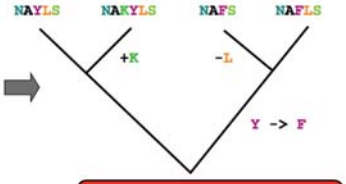
1. Make a multiple sequence alignment.
2. Determine the distance between the sequences.
3. Use these differences to infer the phylogenetic relationship.

- closely related species: use DNA.
- in most other cases: use protein

Which hypothesis is true?

Example 1:

```
SeqA  N A - F L S
SeqB  N A - F - S
SeqC  N A K Y L S
SeqD  N A - Y L S
```



Example 2:

```
seqA  TCAGACGATTG (11)
seqB  TCGGAGCTG (9)

I.    TCAG-ACG-ATTG
      TC-GGA-GC-T-G

II.   TCAGACGATTG
      TCGGAGCTG--

III.  TCAG-ACGATTG
      TC-GGA--GCTG
```

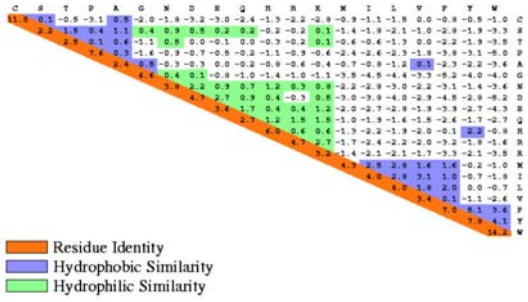
In both cases, you're making an assumption about how evolution has taken place!

perfectly matching alignment

no internal gaps

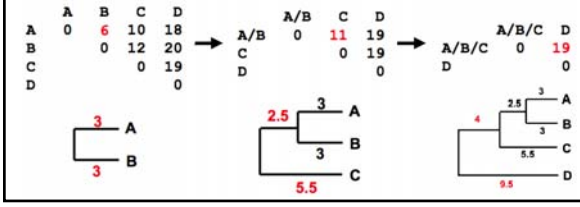
a mixture of both

Gonnet Pam250 Matrix



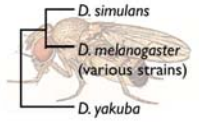
Tree building

- Huge range of different methods around, here only the most simple one: UPGMA.
- Unweighted pair group method using arithmetic means => this is similar to average distance clustering of gene expression data.

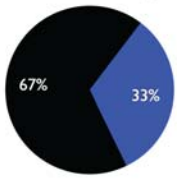


Gene expression

Half of the genes show developmental dynamics



Based on microarray data:

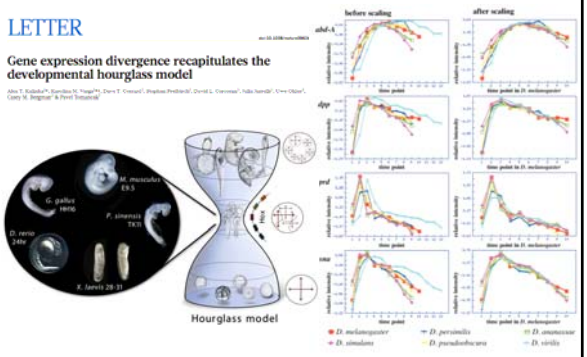


gene expression between species

- conserved
- changes

Adapted from:
Rifkin et al., Nature Genetics 33: 138-144 (2003)

Hourglass model



References

- Multiple sequence alignment - http://www.ebi.ac.uk/Tools/msa/
  Durbin *et al.* - Biological Sequence Analysis -Cambridge University Press
- Inparanoid
  - Remm *et al.* - Automatic clustering of orthologs and in-paralogs from pairwise species comparisons - J. Mol. Biol. 2001
    http://inparanoid.sbc.su.se/cgi-bin/index.cgi
- OrthoMCL / MCL
  - Chen *et al.* - OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups - Nucl Acids Res. 2006
    http://www.orthomcl.org/orthomcl/
  - Enright *et al.* - An efficient algorithm for large-scale detection of protein families - Nucl Ac Res. 2002
    http://micans.org/mcl/
- Kalinka *et al.* - Gene expression divergence recapitulates the developmental hourglass model – Nature 2010

Networks References

- Buchanan et al (eds.) - Networks in Cell Biology - Cambridge University Press
- Barabási and Oltvai - Network biology: understanding the cell's functional organization - Nature Reviews Genetics 2004
- Boone *et al.* - Exploring genetic interactions and networks with yeast. - Nat Rev Genet. 2007
- Kelley and Ideker - Systematic interpretation of genetic interactions using protein networks. - Nat Biotechnol. 2005
- Markowetz and Spang - Inferring cellular networks – a review - BMC Bioinformatics 2007
- Schuster *et al.* - A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks - Nat Biotechnol. 2000