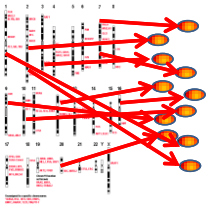
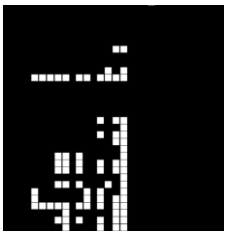
Homology (& domains)

Homology (& domains)

- Absolute basis of any comparative analysis, affects MSA and trees, detection still being improved,



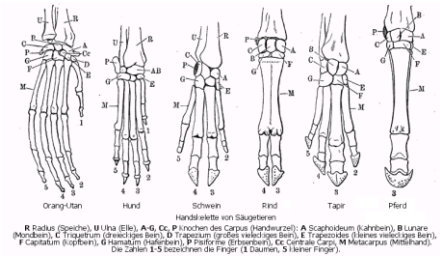


Homology and Domains "contents"

- What does homology mean (and how is it related to trees)
- How do we do it / why do the methods "work"
- Homology and function
- What does homolology mean: implications for transitivity and domains
- Domains and function
- Other mode of evolution: motifs, coiled coil, disordered regions
- How to automatically generate / obtain gene families
- Domains / families and genome evolution
  - Lineage specific familues
  - Domain / family abundance & promiscuity
- Multiple sequence alignments

Gene / protein sequence evolution: what is homology

- In evolutionary biology, **homology** refers to any similarity between characteristics of organisms that is due to their shared ancestry.



Handskelette von Säugetieren

R Radius (Speiche), **U** Ulna (Elle), **A**-**G**, **Cc**, **P** Knochen des Carpus (Handwurzel): **A** Scaphoideum (Kahnbein), **B** Lunare (Mondbein), **C** Triquetrum (dreieckiges Bein), **D** Trapezium (großes vieleckiges Bein), **E** Trapezoides (kleines vieleckiges Bein), **F** Capitatum (Kopfbein), **G** Hamatum (Hakenbein), **P** Pisiforme (Erbsenbein), **Cc** Centrale Carpi, **M** Metacarpus (Mittelhand). Die Zahlen **1**-**5** bezeichnen die Finger (**1** Daumen, **5** kleiner Finger).

Gene / protein sequence evolution: what is homology

- Definition homology (biology)
- structures are said to be homologous if they are alike because of shared ancestry.
- Classic: arms, ~ bird wings, ~ bat wings,
- Genes/proteins/stretches of dna: sequence and/or structural similarity because derived from the same ancestral sequence

Gene / protein sequence evolution: what is homology

- Homologous residues = alignment
- Parts of proteins can be homologous while others are not

- i.e. genes (or part thereof) share common ancestry: the nature of this ancestry could be speciation, duplication, horizontal gene transfer -> need trees to detect this

Trees vs blast, phylogeny vs homology

- Blast/hmm/psi-blast tell you
  - How likely it is that two (parts) of a sequence are homologous or not (and how high the similarity between a profile and a sequence of between two sequences is)
  - Which portions of the sequences are significantly similar, and thus helps to establish which section of which sequence is homologous to which section of which other sequence.
  - Homologous is a yes/no thing
- Trees/phylogeny tell you
  - How the sequences are related, i.e. In which order they diverged

Homology detection has to be done carefully: garbage in garbage out

- Non homologous sequences will be aligned by e.g. clustalx *and* any phylogeny program will make a tree
- Similarly unaligned sequences or very poorly sequences will nevertheless be turned into a tree by any phylogeny program

Homology and Domains "contents"

- What does homology mean (and how is it related to trees)
- How do we do it / why do the methods "work"
- Homology and function
- What does homolology mean: implications for transitivity and domains
- Domains and function
- Other mode of evolution: motifs, coiled coil, disordered regions
- How to automatically generate / obtain gene families
- Domains / families and genome evolution
  - Lineage specific familues
  - Domain / family abundance & promiscuity
- Multiple sequence alignments

How do we detect/define homology of proteins: classic

- By similarity of:
- 3D structure → most conserved aspect, yet not all structures are available. Structures are compared and classified by "eye" and software packages (Dali). (NB classical homology); criterion shared "idiosyncratic" features that are not strictly necessary for function + sequence features plus some degree of sequence similarity
- Sequence → less conserved, many sequences are however available. Homology determination is mainly based on models of sequence evolution and the likelihood that when you compare a sequence to a database you will find a sequence of at least that similarity by chance.
- NB Manually curated databases of 3D structure similarity are used as a benchmark for detection of homology by sequence similarity (SCOP)

?

An alternative argument

- Amino acids are biochemically extremely versatile,
- The same *globular* structure/function could be made using many different "solutions" (e.g. why not simply reverse)
- So if proteins have the same globular structure and some significant degree of sequence similarity -> homologs

Gene / protein evolution: beyond blast, "distant homology"

- Not obvious by blast
- Substantial divergence, due to time **and/or speed**
- Use "profile"
- Profile works better because: is built from a multiple alignment of homologous sequences, contains more information about the sequence family than a single sequence. The profile allows one to distinguish between conserved positions that are important for defining members of the family and non-conserved positions that are variable among the members of the family. More than that, it describes exactly what variation in amino acids is possible at each position by recording the probability for the occurrence of each amino acid along the multiple alignment.

```
ECGHR ECGHR
ECNHR ECNHR
  C   R   G R
TCQQR SIGNR
```

(Also: e.g. is the F there because it is aromatic or because it is bulky hydrophobic)

"distant homology" in practice

- PSI-BLAST / jack-hmmer a multiple sequence alignment is generated on the fly to detect which residues/positions characterize the family.
- And/or use CDD, PFAM or SMART
  - Experts have collected representative and divergent members of a gene family and use HMMer or RPS-BLAST to see if your query sequence belongs to this gene family (i.e. is homologous to the members)
  - clearer/cleaner than psi-blast or blast. But limited to curated knowledge

Gene / protein evolution: Distant homology

- alignment-vs-alignment, Profile-vs-profile, HMM vs HMM comparison (whereas HHMer, PSI-BLAST compare a profile to a single sequence)
- "works" because

```
ACRNG ACRNG
ACGNR ACGNR
  C     C
TCQQL TCQQL

TFQQI TCILL
```

Used tools: HHsearch/hhpred, PRC or compass

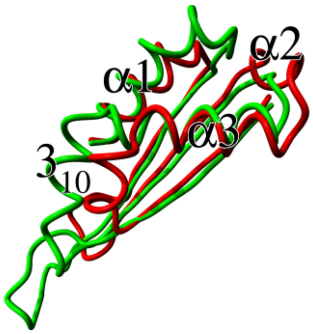
How do we know it works? Benchmark via manually curated database of superfamilies

- 3D structure comparison/alignment plus visual inspection of multiple sequence alignment by Alexey Murzin; emphasis on idiosyncratic similarities
- The results of this are stored in the SCOP database
- *Superfamily* same fold, shared ancestry VS *Fold* sharded ancestry not known / disproven
- (Blundel's bus)

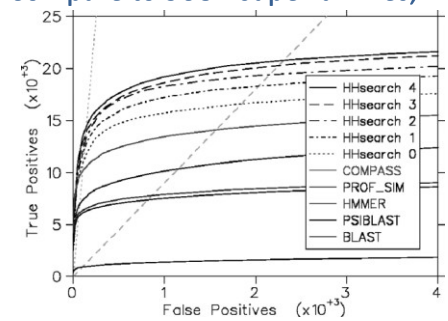
Structural alignment

Secondary structure elements

- Alpha-helices
- Beta strands (beta sheets)
- Loops



Compare to SCOP superfamilies, <20%

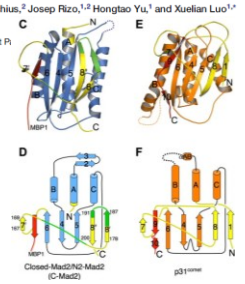


Bioinformatics. 2005 Apr 1;21(7):951-60. Epub 2004 Nov 5. **Protein homology detection by HMM-HMM comparison.** Söding J.

Cell

p31^{comet} Blocks Mad2 Activation through Structural Mimicry

Maojun Yang,[1] Bing Li,[1] Diana R. Tomchick,[2] Mischa Machius,[2] Josep Rizo,[1,2] Hongtao Yu,[1] and Xuelian Luo[1,*]
[1]Department of Pharmacology
[2]Department of Biochemistry
The University of Texas Southwestern Medical Center, 6001 Forest Pk
*Correspondence: xuelian.luo@utsouthwestern.edu
DOI 10.1016/j.cell.2007.08.048



Imply convergent evolution?
Same fold different origin?

Superfamily!

- Structural similarity unexpected, as p31 does not share obvious sequence similarity with Mad2 that is detectable by regular sequence-alignment algorithms.
- Structure-based sequence alignment: Mad2 and p31 do share limited sequence similarity,
- E.g. R35 and E98 are invariable residues in all Mad2 proteins. Form a buried salt bridge buried helping specify the Mad2 fold. R84 and E163 in p31 are equivalents. They also form an analogous (*????*) interior salt bridge conserved among p31 proteins
- The similarity between Mad2 and p31 sequences that specify their folds suggests that Mad2 and p31 have evolved from a common ancestor

Could this have been shown without structure guided alignment?

- PRC searches of p31 profile versus a database of PFAM profiles and Mad2 profiles and reciprocal searches of Mad2 profile versus a database of PFAM profiles and p31 profile.
- Best hit of p31 is Mad2 at e=0.019, best hit of the Mad2 is p31 at 0.038.
- Although these are borderline hits they are significant, the alignments are nearly full-length and they are each others reciprocal best hits.
- Retrieve "salt-bridge"
- p31comet is an ancient duplication of Mad2 from before the last eukaryotic common ancestor.
- (NB I expect normally duplications from before LECA do not require PRC/hhpred, e.g. kinases, small-GTPases)

HHpred alignment

```
Q Thu_Jan_27_11:   65 SQEGCCQFTCEL----LKHIMYQRQQLPLPYEQLKHFYRKPSPQAEEMLKKKPRATTEVSSRKCQQALAELESVLSHLED  14
Q Consensus        65 t-e-C-rfv-EL----LK-LLYqR-QIPIpYd-Lk--v-K------d----k--------q-rk----l--le-ll--L--  14
                      |.+++..+|.++    +-.|||+|.#.|--+#+-+.-+.F.....+.                    ++.+=|+|+.++..+..
T Consensus         1 t---S---v---l--ai--Ily-RqiyP---F--------l-v-------------------------l---i-------v-d   6
T pfam02301         1 TLKQSLELVKEPLEVAINSILYLRGIYPEESFEDRKKYNLPVLVSEDP-----------------QLIDYLEKVLSGVFD   6


Q Thu_Jan_27_11:  141 FFARTLVPRVLILLGGNA----LSPKEFYELDLSLLAPYSVDQSL-----STAACLRRLFRAIFMADAF-SELQAPPLMG  21
Q Consensus       141 -F--s-V--V1iLfGsT-------sPKE-Y-I-lp-------e--l-----st---lRkL-R-L-t-d-l-s-l-s-plt-  21
                      +.....++++.|.+-...    -.+.|.|.++|.-...+.....-     .++.-+|.++|+|+.+-+ ..+..-....
T Consensus        64 aL-k--L--l-l-I-----------lE-y-F-------------------------lr-l------L--LP------  14
T pfam02301        64 ALEKGYLKKLVLVIYEDDPEKENEVLERYQFDFSYFPSGGNSSDSEKTEDETRQEIRALLRQLIALVTFLPFLPEDRTCT  14

Q Thu_Jan_27_11:  211 TVVMAQGHRNCGEDWFRP  228 (274)
Q Consensus       211 t-Vl-q--r-c---wF-P  228 (274)
                      .-|+|...|.|+..+.|.+
T Consensus       144 --l---tp-dy-pp-f--  161 (189)
T pfam02301       144 FKLLYYTPFDYEPPGFKW  161 (189)
```

Homology and Domains "contents"

- What does homology mean (and how is it related to trees)
- How do we do it / why do the methods "work"
- Homology and function
- What does homology mean: implications for transitivity and domains
- Domains and function
- Other mode of evolution: motifs, coiled coil, disordered regions
- How to automatically generate / obtain gene families
- Domains / families and genome evolution
  - Lineage specific familues
  - Domain / family abundance & promiscuity
- Multiple sequence alignments

Homology and fold ok; what about function?

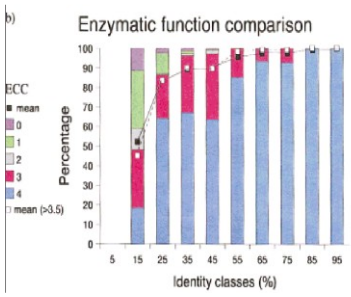
- To what extent do homologs/"proteins in a protein family", have the same function?
- Structure determines function? Fold != exact structure
- Relevant for function prediction
- Relevant for evolution of function

E(nzyme) C(ode) number: a hierarchical system to describe enzymatic function

- **EC 1 Oxidoreductases**
- **EC 2 Transferases**
- **EC 3 Hydrolases**
- **EC 4 Lyases**
- **EC 5 Isomerases**
- **EC 6 Ligases**

- **EC 2.7 Transferring phosphorus-containing groups**
- **EC 2.7.7 Nucleotidyltransferases**
- EC 2.7.7.6 DNA-directed RNA polymerase

Homology ~ molecular function



Homology ~ molecular function

- Protein kinases, SH2, RING fingers,
- More difficult with WD40, TPR

Using distant homology for function prediction: example from (just) before PSI-BLAST & HMMer

Secreted Fringe-like Signaling Molecules May Be Glycosyltransferases.

Cell. 1997 Jan 10;88(1):9-11.

Y. Yuan, J. Schultz, M. Mlodzik, P. Bork

Homology and Domains "contents"

- What does homology mean (and how is it related to trees)
- How do we do it / why do the methods "work"
- Homology and function
- What does homolology mean: implications for transitivity and domains
- Other mode of evolution: motifs, coiled coil, disordered regions
- How to automatically generate / obtain gene families
- Domains / families and genome evolution
  - Lineage specific familues
  - Domain / family abundance & promiscuity
- Multiple sequence alignments

Homology and Domains "contents"

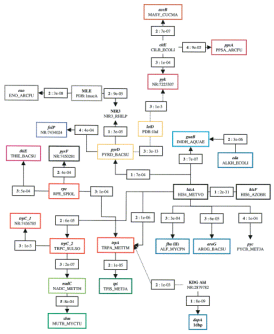
- What does homology mean (and how is it related to trees)
- How do we do it / why do the methods "work"
- Homology and function
- What does homolology mean: implications for transitivity and domains
- Domains and function
- Other mode of evolution: motifs, coiled coil, disordered regions
- Domains / families and genome evolution
- Multiple sequence alignments

Homology is transitive

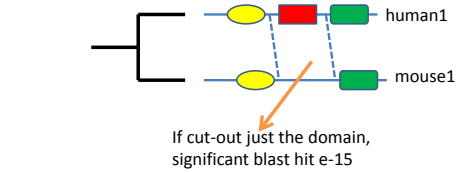
- i.e. if A is homologous to B and B is homologous to C, than A should be homologous C.

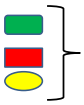
Homology is transitive helps to define superfamilies

- When two protein families are homologous but the homology is not obvious they are part of the same so called superfamily
- How to detect:
- In depth PSI-BLAST
- Reciprocal
- Use of right seed
- Psi-Blast "hopping"
- Used to show that all Rosmann folds (alpha/beta barrels) are likely homologous



Homology is transitive /"schnipsel" approach?



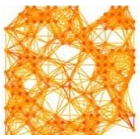


False positives, false negatives

- The cut-off values for all sequence similarity searches are defined to eliminate FP's (and thus not by definition towards reducing FN's, despite HMMER vastly outperforming BLAST at sensitivity)
- Hence intuition the domain is simply there and FN for the PFAM
- However proper solution (still using the transitivity line of reasoning but less dirty), include close relative in the profile, i.e. improve PFAM model

Homology is transitive

- So when creating families for generating automatically trees or for phylogenetic profiles, you can just link them up or not?



- No: domains / fusion
- No: coiled coil etc.
- No: minor fraction of FP's -> huge connected component

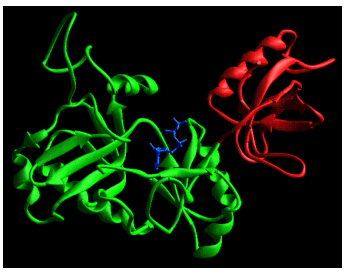
b

Homology and Domains "contents"

- What does homology mean (and how is it related to trees)
- How do we do it / why do the methods "work"
- Homology and function
- What does homolology mean: implications for transitivity and domains
- Domains
- Other mode of evolution: motifs, coiled coil, disordered regions
- How to automatically generate / obtain gene families
- Domains / families and genome evolution
  - Lineage specific familues
  - Domain / family abundance & promiscuity
- Multiple sequence alignments

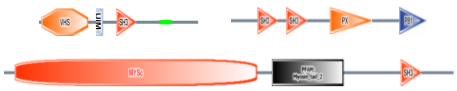
Protein domains: structural definition: separate in structure

a structural domain ("domain") is an element of overall structure that is self-stabilizing and often folds independently of the rest of the protein chain

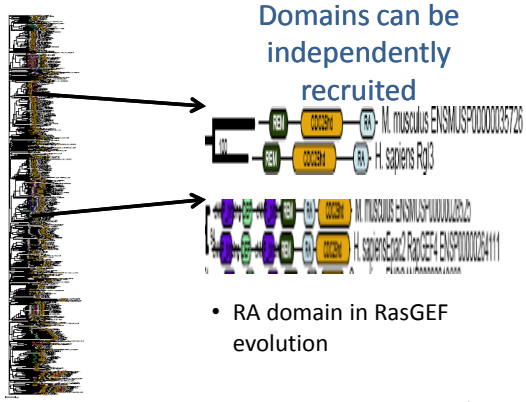


Protein domains: sequence/evolutionary definition: Separate in "evolution"

- Homologous parts of proteins that occur with different "partners"
- Mobile
- Modules
- Almost always same as structural definition



Domains can be independently recruited



- RA domain in RasGEF evolution

Van Dam et al. 2009

Implications of domains for homology:

- The shared ancestry is not a property of the whole gene but only of part of the gene.
- When studying the evolution of gene families, consider fusions / domain combinations (also when making trees etc.)

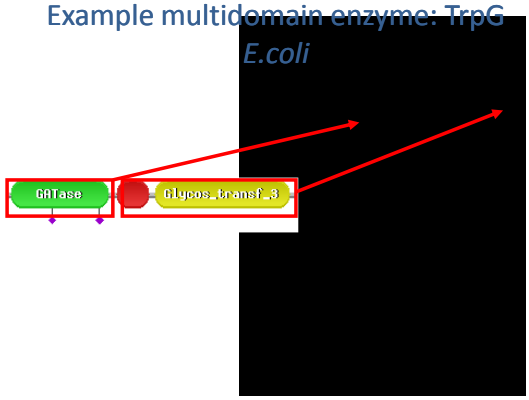
Implications of domains for doing homology searches when doing blast do psi-blast, cdd / pfam instead /also.

- Rather than discover the domain structure by blast yourself, use e.g. SMART / PFAM / CDD to do it for you
- NB CDD

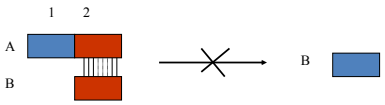
Ramifications for function prediction & understanding of cellular processes: "one domain one (molecular) function" (in contrast to one gene one function)

- This bit does this and that bit does that
- E.g.
  - multidomain enzymes
  - Signalling proteins

Example multidomain enzyme: TrpG *E.coli*



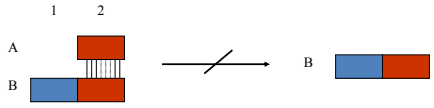
Ramifications for function prediction when doing blast: mind the domains



Protein B is wrongly annotated as having the function of domain 1, based on homology with the multidomain protein A, but not with domain 1

(multi-domain architecture problem for annotating proteins via blast)

Ramifications for function prediction when doing blast: mind the domains



Protein B is incompletely annotated as having the function of domain 2, based on homology with the single domain protein A, the second domain is missed in the annotation

Homology and Domains "contents"

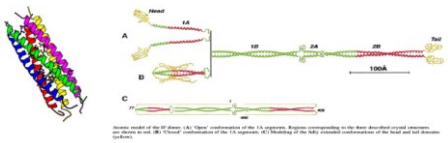
- What does homology mean (and how is it related to trees)
- How do we do it / why do the methods "work"
- Homology and function
- What does homolology mean: implications for transitivity and domains
- Domains
- Other mode of evolution: motifs, coiled coil, disordered regions
- How to automatically generate / obtain gene families
- Domains / families and genome evolution
  - Lineage specific familues
  - Domain / family abundance & promiscuity
- Multiple sequence alignments

Disclaimer: non-globular regions

- Low complexity
- Unstructured, Elongated (as opposed to globular)
- Many polar/charged residues; few hydrophobic residues
- parts of proteins that do not posses a clear 3D structure
- Convergence
- Do not obey PAM or BLOSUM

Disclaimer: Coiled coil

- All alpha: thought to arise independently (convergence)
- Hypothesis: reservoir for "new" folds: all alpha folds (Koonin EV)
- E.g. ras / rho / rab / ran / -GAPs

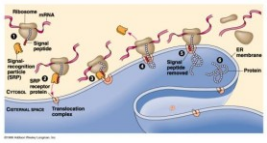


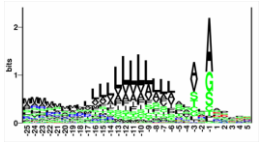
How to deal with coiled-coil proteins in homology / orthology searches?

- No one really knows / no accepted method / but needed for evolutionary cell biology
- Coiled coil is especially a problem for iterative methods (psi-blast / jack-hmmer) i.e. if you see e.g. myosin / dynein / spectrin; ABORT
- Only use globular & non-coiled coil part of the protein.
- Use blast hopping?

Disclaimer: Other protein motifs

- Signal peptides
- Lipid anchoring
- Convergence yet still important to predict
- Trans-membrane?





Homology and Domains "contents"

- What does homology mean (and how is it related to trees)
- How do we do it / why do the methods "work"
- Homology and function
- What does homolology mean: implications for transitivity and domains
- Domains
- Other mode of evolution: motifs, coiled coil, disordered regions
- How to automatically generate / obtain gene families
- Domains / families and genome evolution
  - Lineage specific families
  - Domain / family abundance & promiscuity
- Multiple sequence alignments

Automatic methods to obtain homologous protein / gene families

- Should be easy, homology = transitivity etc.?
- Hence "single linkage" (in the network sense), but outcome "connected component"
- Problem 1) : false positives FP's statistics/e-value true but "multiple testing"/bad luck & disorder & coiled coil
  - solutions: very conservative e-values, filter low complexity / take low complexity into p-value into account (modern blast), filter coiled / coil (infrequent), filter disorder (never seen done). work at restricted taxon sets (e.g. ensembl COMPARA)
- Problem: 2) fusion & fission, violates transitivity
  - "Disallow fusion proteins to bring in stuff"/ work at restricted taxon sets (e.g. ensembl COMPARA)

Automatic methods to obtain use curated homologous protein / gene families

- Just use PFAM? Works fairly well, but …
  - Novel gene families (e.g. !!! Hyelanoperesonora)
  - False negatives (e.g. schnipsel)

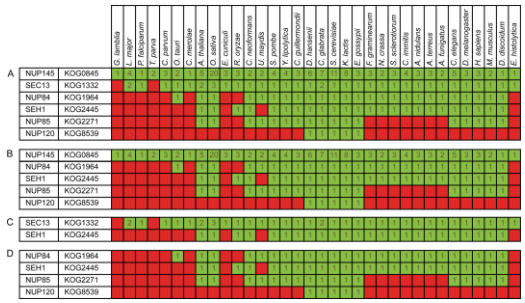
So what to do

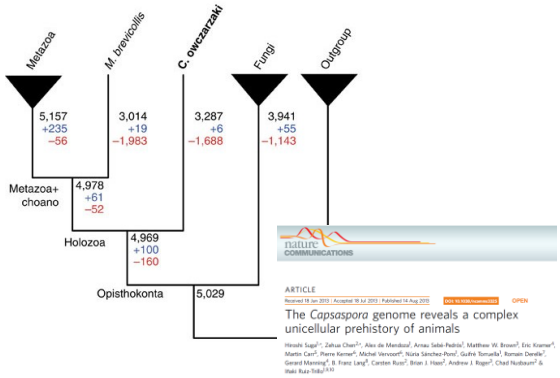
- Do – all and integrate
- Wait (e.g. hmm3) / proactively improve pfam
- Do not mind the problems if they are not too big and consider it noise in your analysis
- Exclude too big families
- Restrict yourself to a taxon,
- Do only case-by-case basis
- Any combination of the above … / depends on your question …

Homology and Domains "contents"

- What does homology mean (and how is it related to trees)
- How do we do it / why do the methods "work"
- Homology and function
- What does homolology mean: implications for transitivity and domains
- Domains
- Other mode of evolution: motifs, coiled coil, disordered regions
- How to automatically generate / obtain gene families
- Domains / families and genome evolution
  - Lineage specific families
  - Domain / family abundance & promiscuity
- Multiple sequence alignments

Apparent lineage specific (LS) genes?

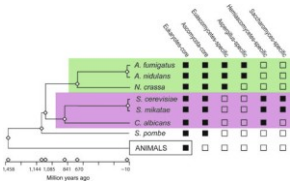




What about apparent lineage specific genes? (LS)

Four possibilities are generally proposed

1. Loss in all but one lineage: unlikely and where did the gene come from in the first place.
2. LS genes formed by the recombination/duplication of exons/ORFS from other genes i.e. ~ duplication but I would not call them LS and we would still see homology unless option 4
3. from random ORFs. Should show similarity to non coding DNA in other species, semantics (still homolog)! is unlikely that such a protein would be functional. But has been shown to happen at least for extensions i.e. 3' shift of stop codon, 5' shift of start codon + now many other cases
4. Some genes evolve at a rapid rate and so can no longer be recognized as orthologues of the genes they diverged from after a certain time span. OR after duplication!

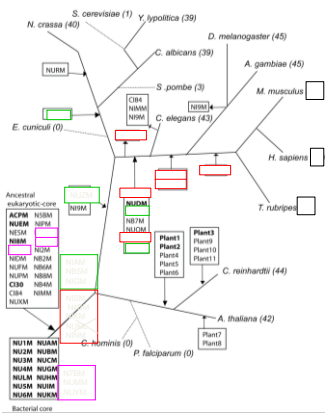


J Mol Evol. 2006 Jul;63(1):1-11. Epub 2006 Jun 3.
Accelerated evolutionary rate may be responsible for the emergence of lineage-specific genes in ascomycota.
Cai JJ, Woo PC, Lau SK, Smith DK, Yuen KY.

Table 2. Average nonsynonymous substitution rate (K_a), synonymous substitution rate (K_s), and K_a/K_s ratio among LS classes

LS class	No. of gene pairs	Mean (SD) K_a	Mean (SD) K_s	Mean (SD) K_a/K_s
A. fumigatus–*A. nidulans* (Euascomycetes branch)				
Eukaryotes-core	113	0.051 (0.032)	1.431 (0.441)	0.039 (0.027)
Ascomycota-core	27	0.126 (0.069)	1.577 (0.329)	0.080 (0.042)
Euascomycetes-specific	22	0.198 (0.118)	1.436 (0.490)	0.155 (0.091)
Aspergillus-specific	21	0.293 (0.136)	1.263 (0.567)	0.261 (0.127)
S. cerevisiae–*S. mikatae* (Hemiascomycetes branch)				
Eukaryotes-core	17	0.018 (0.021)	0.586 (0.213)	0.029 (0.026)
Ascomycota-core	23	0.031 (0.030)	0.639 (0.172)	0.047 (0.040)
Hemiascomycetes-specific	22	0.072 (0.037)	0.839 (0.284)	0.091 (0.045)
Saccharomyces-specific	297	0.131 (0.100)	0.830 (0.329)	0.165 (0.130)

[a] A Kruskal-Wallis test revealed significant rate heterogeneity of average K_a or average K_a/K_s of genes in different LS groups in both the Euascomycetes branch and the Hemiascomycetes branch; $p < 0.001$.
[b] A Kruskal-Wallis test revealed no significant rate heterogeneity of average K_s of genes in different LS groups in both the Euascomycetes branch and the Hemiascomycetes branch; $p > 0.01$.



New experimental + sequence data and profile based sequence analysis methods, push back the origin of lineage specific subunits of complexes back in time

Huynen et al, BBA 2009

Cardol , BBA 2012

Yip et al, JBC 2011

Balsa et al, Cell Metab. 2012

Gabaldon et al, JMB, 2005

But …

- New genes have low expression (Carvunis *et al.* 2012 Nature)
- Low expression leads to fast sequence evolution (Drummond and Wilke 2008 Cell)
- So chicken and egg …

"Anything goes" in (genome) evolution

- Lineage specific genes/families are the result of
  - coding becoming non-coding,
- Or
  - extreme sequence (and structure?) divergence after duplication or speciation

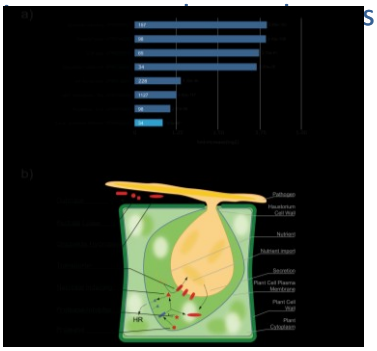
Irrespective of important source of innovation in genome evolution is novel gene families, which NB reveal that novel gene families play pivotal role in eukaryogenesis

The genome of Naegleria gruberi illuminates early eukaryotic versatility.
Fritz-Laylin LK, Prochnik SE, Ginger ML, Dacks JB, Carpenter ML, Field MC, Kuo A, Paredez A, Chapman J, Pham J, Shu S, Neupane R, Cipriano M, Mancuso J, Tu H, Salamov A, Lindquist E, Shapiro H, Lucas S, Grigoriev IV, Cande WZ, Fulton C, Rokhsar DS, Dawson SC.
Cell. 2010 Mar 5;140(5):631-42.

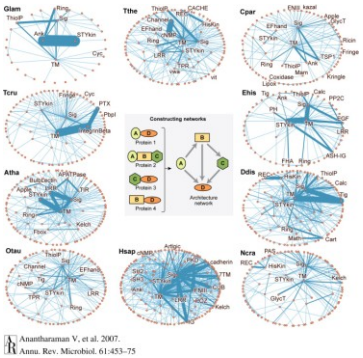
Homology and Domains "contents"

- What does homology mean (and how is it related to trees)
- How do we do it / why do the methods "work"
- Homology and function
- What does homolology mean: implications for transitivity and domains
- Domains
- Other mode of evolution: motifs, coiled coil, disordered regions
- How to automatically generate / obtain gene families
- Domains / families and genome evolution
  - Lineage specific families
  - Domain / family abundance & promiscuity
- Multiple sequence alignments

Domains and study of genome evolution: Domain overrepresentation



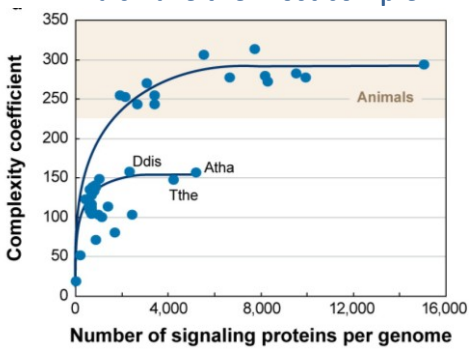
Animals have the most complex proteins?



Anantharaman V, et al. 2007.
Annu. Rev. Microbiol. 61:453–75

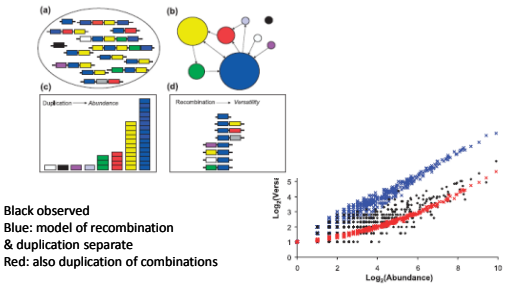
Anantharaman *et al.* 2007

Animals have the most complex



Anantharaman *et al.* 2007

Interesting result on protein evolution regarding domains and duplications: neutral?



Black observed
Blue: model of recombination & duplication separate
Red: also duplication of combinations

b

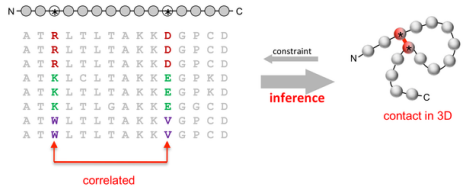
Homology and Domains "contents"

- What does homology mean (and how is it related to trees)
- How do we do it / why do the methods "work"
- Homology and function
- What does homology mean: implications for transitivity and domains
- Domains and function
- Other mode of evolution: motifs, coiled coil, disordered regions
- How to automatically generate / obtain gene families
- Domains / families and genome evolution
  - Lineage specific familues
  - Domain / family abundance & promiscuity
- Multiple sequence alignments

Multiple sequence alignments

- Needed for phylogenies
- Homologous residues = alignment
- Basis for profile(vs-profile) methods and db's like PFAM
- Functionally important residues
- Secondary structure prediction
- New: Tertiary structure prediction EVfold

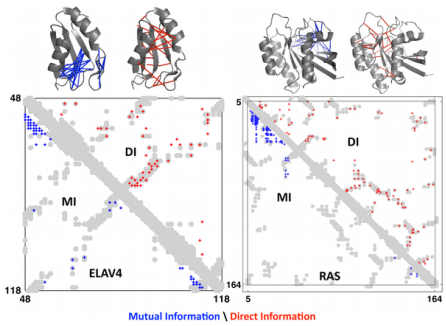
EVfold: basic idea



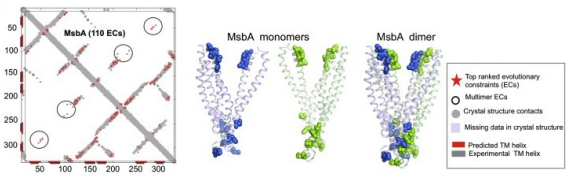
Compared to before: (more seqs,) solves "chaining", "partial correlation" for mutual information = "direct information" / "maximum entropy condition", global frequency counts instead of local frequency counts …

Protein 3D structure computed from evolutionary sequence variation.
Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C.
PLoS One. 2011;6(12):e28766. doi: 10.1371/journal.pone.0028766. Epub 2011 Dec 7.
PMID: 22163331

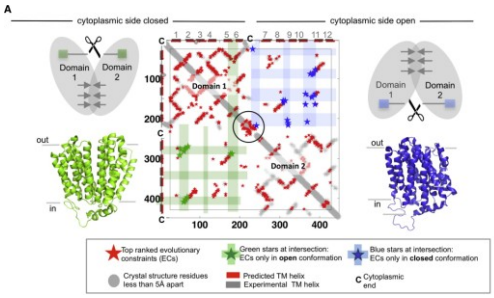
Direct information predicts structure contacts



Evolutionary coupled residues that do not fit the structure point to dimers



Evolutionary coupled residues that do not fit the structure point to alternative confirmations



EVfold and deep sequence similarity detection?

- Profile(vs-profile) searches assume independence between sites
- EVfold maps the dependencies between sites → will / should be used for remote sequence similarity detection