

## Homology (& domains)(& protein families)

06/03/17 1

## Homology (& domains)

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

Gene originates in common ancestor... but evolves rapidly (coiled coil, disordered, very short globular domain)

Gene originates later... evolves normally (has decent length e.g. 200AA and globular fold). Few losses.

## 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.

Handeleite von Säugetieren  
R Radius (Speiche), U Ulna (Elle), A-G, CC, P Knochen des Carpus (Handwurzel): A Saphoidium (Hahnchen), B Lunare (Mondstein), C Trapezium (Gründelstein), D Trapezium (großes viereckiges Stein), E Trapezoides (Hahnens viereckiges Stein), F Capitulum (Kopfbirn), G Hamatum (Hahnstein), H Proximale (Erbsenbirn), CC Centrale Carpi, M Metacarpus (Metzhand). Die Zahlen 1-5 bezeichnen die Finger (1 Daumen, 2 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
- What is the history of my gene -> different parts can have different histories!

### 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

### 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
T	S
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
  - Used tools: HHsearch/hhpred, PRC, compass

A	C	R	N	G	A	C	R	N	G
A	C	G	N	R	A	C	G	N	R
C	C								
T	C	Q	Q	L	T	C	Q	Q	L
T	F	Q	I	T	C	I	L	L	

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

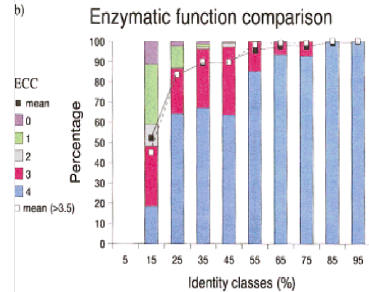
- 3D structure comparison/alignment plus visual inspection of multiple sequence alignment by Alexey Murzin; emphasis on idiosyncratic similarities
- The results of this are stored in the SCOP database
- *Superfamily* same fold, shared ancestry VS *Fold* shared ancestry not known / disproven
- (Blundel’s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



**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, RhoGAPs,
- Difficult with SH2, RING fingers,
- Even 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.**

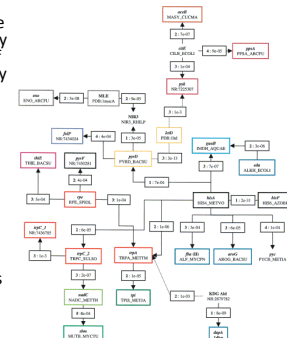
Cell. 1997 Jan 10;88(1):9-11.  
Y. Yuan, J. Schultz, M. Mlodzik, P. Bork

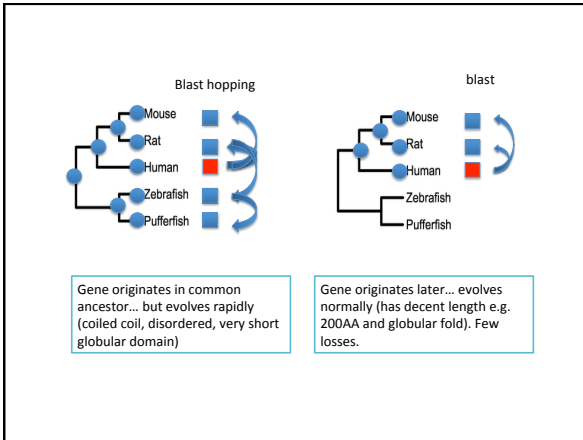
**Homology is transitive**

- 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 Rossmann folds (alpha/beta barrels) are likely homologous





### False positives, false negatives

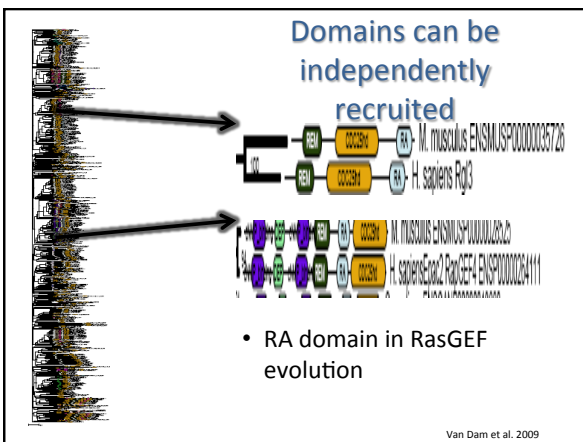
- The cut-off values for all sequence similarity searches are defined to eliminate FP's (and thus not by definition towards reducing FN's, despite HMMER vastly outperforming BLAST at sensitivity)
- Hence intuition the domain is simply there and FN for the PFAM
- However proper solution (still using the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

### 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

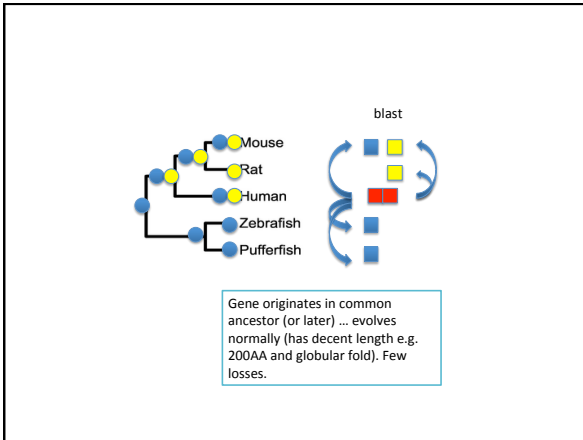
### Protein domains: sequence/evolutionary definition: Separate in "evolution"

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



### 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.

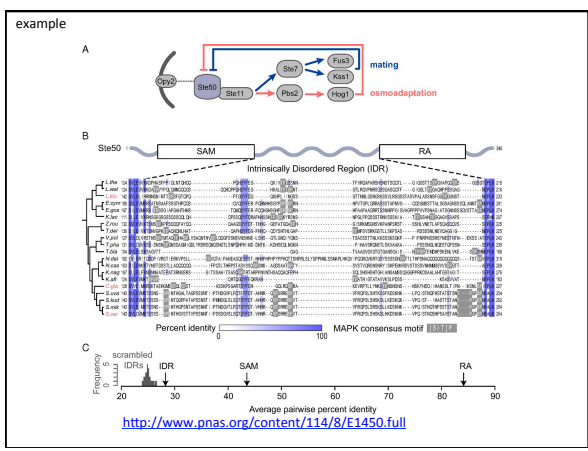
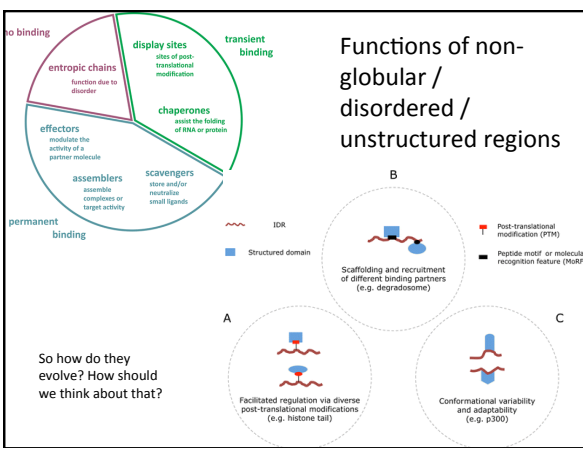
- Rather than discover the domain structure by blast yourself, use e.g. SMART / PFAM / CDD to do it for you
- NB CDD

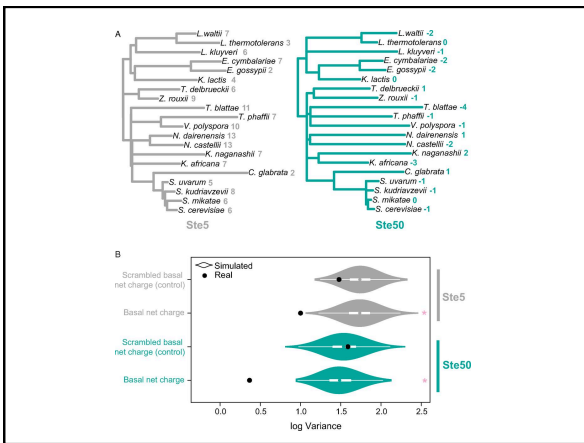
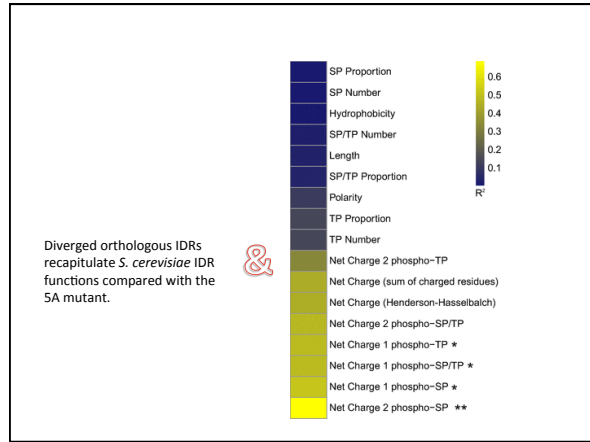
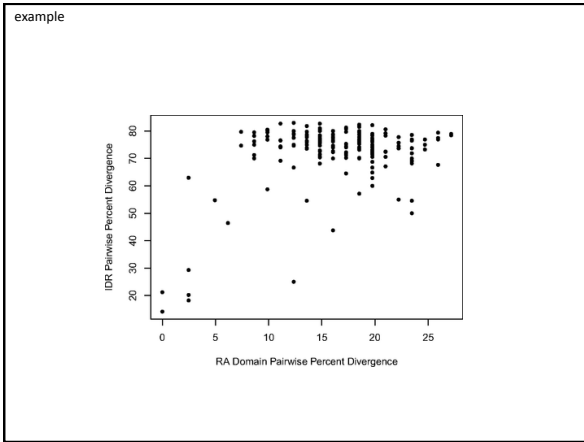
### Ramifications for function prediction & understanding of cellular processes: "one domain one (molecular) function" (in contrast to one gene one function)

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

### Disclaimer 1: non-globular regions

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





### Disclaimer 2: 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

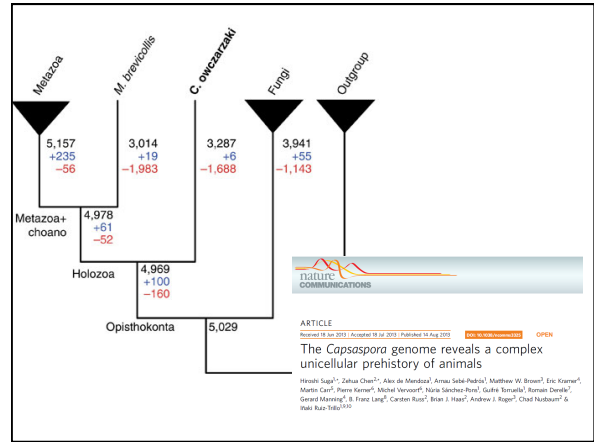
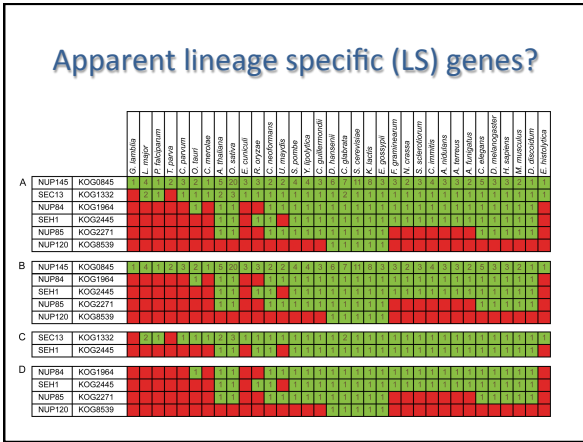
### How to deal with coiled-coil proteins in homology / orthology searches?

- 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-hmmmer) 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 3: protein motifs

- Signal peptides
- Lipid anchoring
- Trans-membrane
- Kinase consensus motifs
- Can convergently evolve yet still important to predict

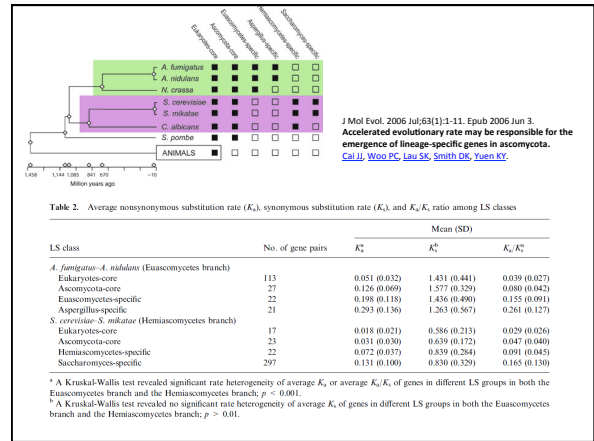




### What about apparent lineage specific genes? (LS)

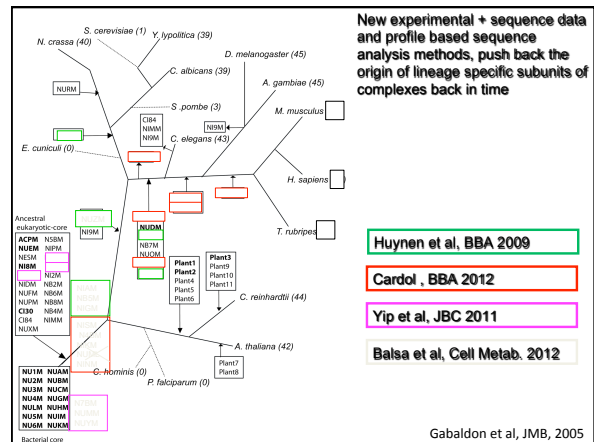
Four possibilities are generally proposed

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 for extensions i.e. 3' shift of stop codon, 5' shift of start codon. & recently for small ORFs
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!



### 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

- Some lineage specific genes/families are the result of
  - coding becoming non-coding,
- And others from
  - 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, Paredes 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.

- Distant homology / iterative or clustered homology searches lead to
  - “Protein families”
  - “Protein domains”
  - They are the same thing but emphasize different aspects
- (blackboard)