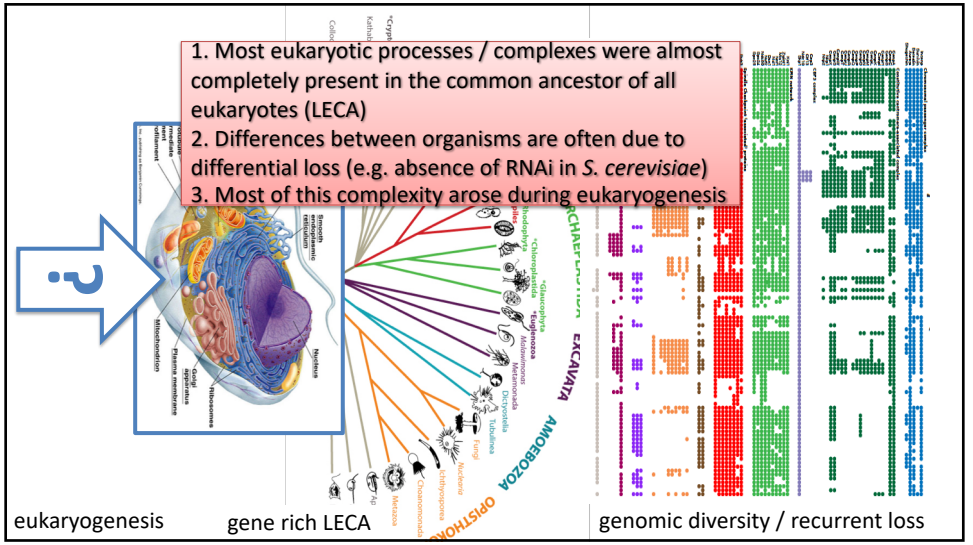
Homology (& domains)(& protein families)

- Introduction: on the importance of homology
- How to think about homology (what is homology, implications of homology, sequence evolution & selection)
- Methods for detecting diverged homologs
- What have we learned from (sensitive) homology searches?
- Homology & function
- How to think about and deal with function and evolution of non globular proteins
- Gene invention (i.e. absence of any homology)
- Summary and integration with automatic phylogeny methods

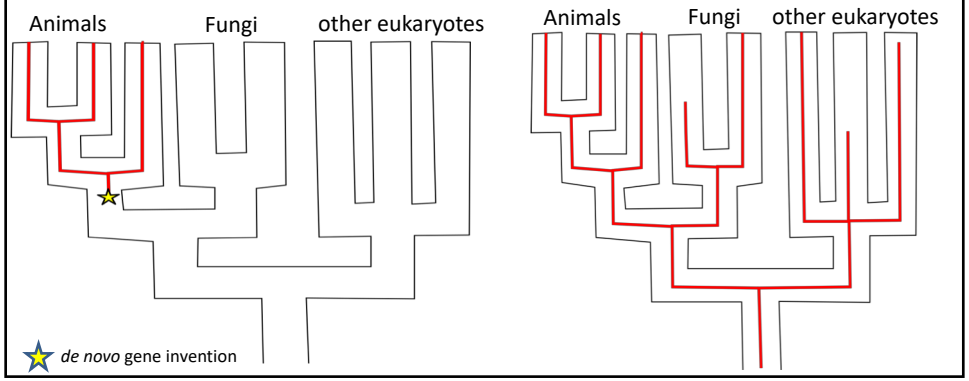


How the trend of a complex ancestor and independent loss was revealed

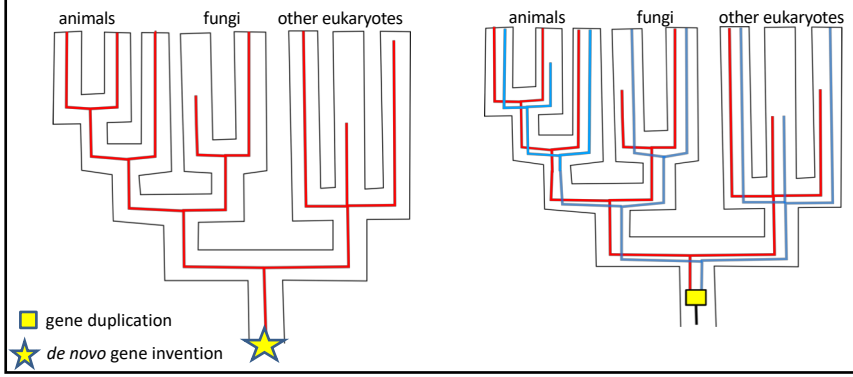
A combination of:

- New genomes at crucial positions
- **Improved sensitivity of sequence similarity searches (and homologs that are orthologs)**
- **Studying gene families with a lot of pre-LECA duplications**

Improved sensitivity of sequence similarity searches allows to distinguish between lineage specific genes and ancient genes with orthologs across eukaryotes



Improved sensitivity of sequence similarity searches allows to distinguish between genes invented in the ancestor or duplicated in the ancestor; i.e. *without* outparalogs or *with* outparalogs

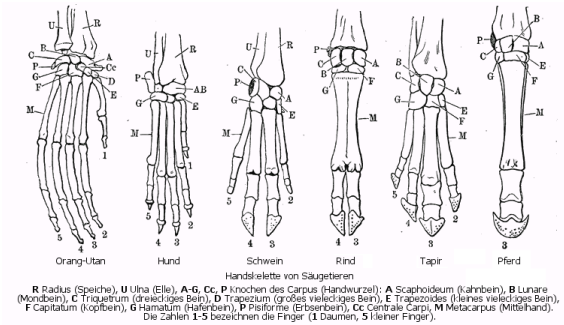


Homology is fundamental. Absolute basis of any comparative analysis

- in the examples on the previous slides, the tree of the lineage specific protein is correct for that part of the species tree, but it is wrong in the sense that is *incomplete*:
  - the tree does not describe the evolution of the entire family
  - we miss tons of orthologs
  - we think the protein originated in animals but in fact it is much older
  - But this not a problem of phylogenetic reconstruction or tree reconciliation it is a problem of homology detection!
- All the fancy tree reconciliation methods or fancy blast-graph methods fail to find orthologs *in the case* that homology goes unrecognized
- (Also, multiple sequence alignment is crucial for tree reconstruction and also here homology plays a key role)

what is homology

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

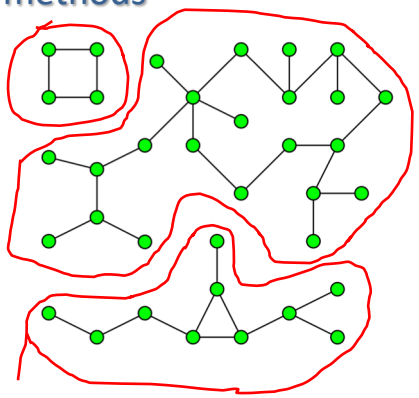


Gene / protein sequence evolution: what is homology

- 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

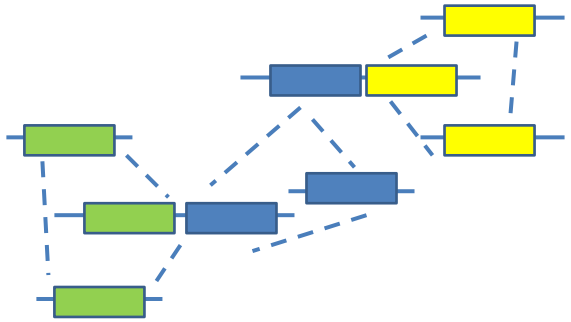
Homology is (in principle) transitive: rationale for network based methods

- i.e. if A is homologous to B and B is homologous to C, than A should be homologous C.
- when creating families for generating automatically trees or for phylogenetic profiles, you can just link them up by defining **connected components?**



https://en.wikipedia.org/wiki/Connected_component_(graph_theory)

In principle but fusion/fission



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 (bc of duplication and horizontal gene transfer need for "specification" of type of homology)
- What is the history of my gene -> different parts can have different histories!

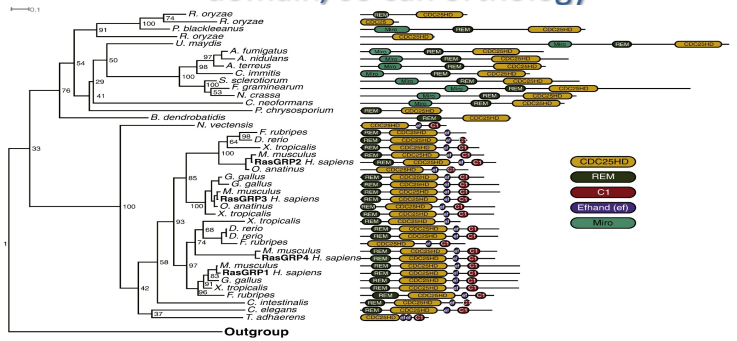
Trees vs blast, phylogeny vs homology

- 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 (e.g. orthology & paralogy)

Gene / protein sequence evolution: what is homology, implications for orthology

- Parts of proteins can be homologous while others are not
- Hence part of proteins can be orthologous while the rest is not

Orthologs can have different domain composition: (likely changed function); orthology is a specification of the homology relation and just like homology can span only a domain, so can orthology

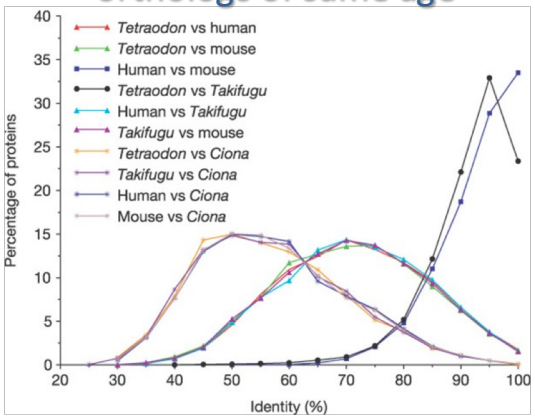


Methods for detecting distant homologs

A lot of (sequence) evolution is neutral

- Most accepted substitutions in sequence evolution are (nearly) neutral
- The percentage of conserved necessary to maintain the same fold and (biochemical) function differs enormously between proteins but it can be very low (e.g. 10% between orthologs) and just to maintain the fold it can be even lower

Big differences in sequence identities between orthologs of same age



https://slideplayer.com/slide/7221949/

Gene / protein evolution: beyond pairwise methods (e.g. blast), detecting "divergent homologs" by profile methods

- Not obvious by pairwise methods (BLAST, PHMMER, SMITH-WATERMAN)
- Substantial divergence, due to time **and/or speed of sequence evolution**
- Use "profile" (for example HMMER search or PSI-BLAST)
- 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.

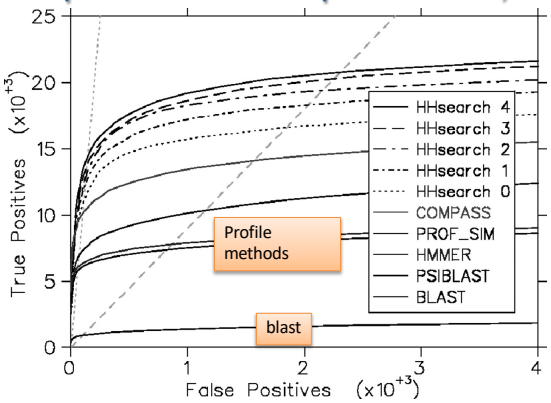
```
ECGHR ECGHR
ECNHR ECNHR
  C  R   G R
TCQQR SIGNR
```

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

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)

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.

“divergent homologs” in practice

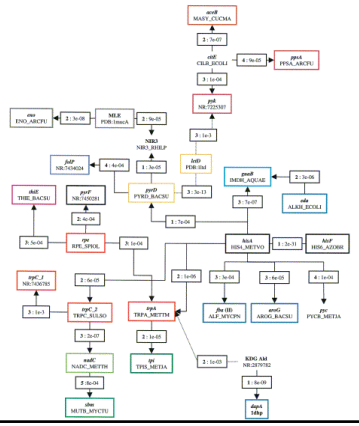
- Do it yourself:
  - PSI-BLAST (NCBI)/ jack-hmmer (EBI) a multiple sequence alignment is generated on the fly to detect which residues/positions characterize the family.
- Use what others have done. Conserved DomainDatabase Search (NCBI), PFAM (EBI) or SMART (EMBL)
  - Experts have collected representative 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 jackhmmer. But limited to curated knowledge

Homology is transitive

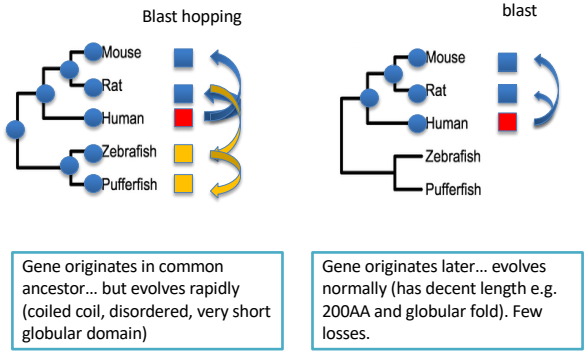
- i.e. if A is homologous to B and B is homologous to C, than A should be homologous C.

The fact that Homology is transitive has also helped to detect diverged homologs and thereby 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



Transitivity allows blast hopping



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.

Most sensitive: detecting diverged homologs by profile-vs-profile searches

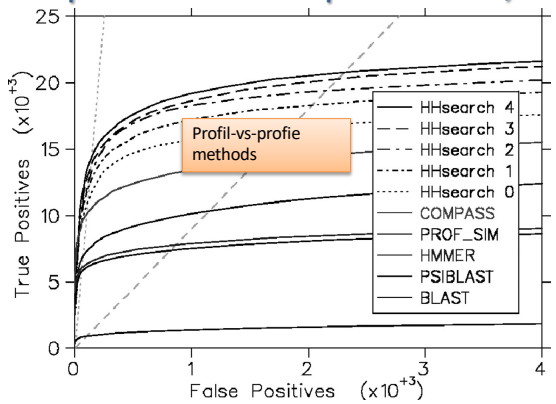
- alignment-vs-alignment, Profile-vs-profile, HMM vs HMM comparison (whereas HHMer, PSI-BLAST compare a profile to a single sequence)
- "works" because

```
ACRNG ACRNG
ACGNR ACGNR
  C     C
TCQQL TCQQL

TFQQI TCILL
```

Used tools: HHsearch/hhpred, PRC, compass

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.

What have we learned from (sensitive) homology searches?

- Histories:
  - Found undetected orthologs (CAMSAP, COX14)
  - Found inter-"domain of life" homologies:
    - homologs of eukaryotes proteins in prokaryotess: (ftsZ-tubulin)
    - Origin of viral capsid proteins
  - Found undetected ancient paralogs:(i.e. duplications from feca-2-leca)
    - p31 and mad2
    - RWD proteins
- "Genome evolution"
  - powerlaw
- NB Detecting previously undetected homologies will, make proteins older, find more duplicates, more orthologs, more losses, and less inventions

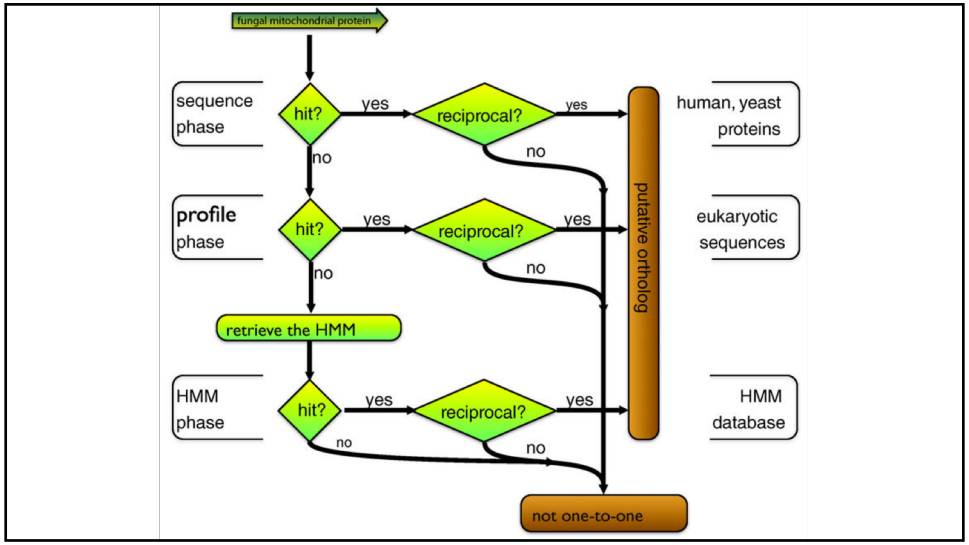
Szklarczyk *et al. Genome Biology* 2012, **13**:R12
http://genomebiology.com/content/13/2/R12

Genome Biology

RESEARCH Open Access

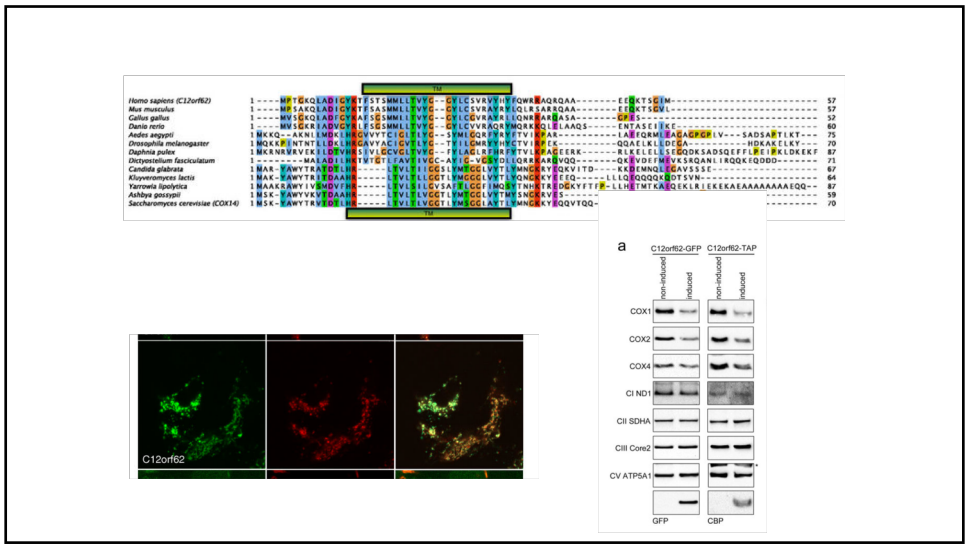
Iterative orthology prediction uncovers new mitochondrial proteins and identifies C12orf62 as the human ortholog of COX14, a protein involved in the assembly of cytochrome *c* oxidase

Radek Szklarczyk[1*†], Bas FJ Wanschers[1,2†], Thomas D Cuypers[1,3], John J Esseling[4], Moniek Riemersma[2], Mariël AM van den Brand[2], Jolein Gloerich[5], Edwin Lasonder[1], Lambert P van den Heuvel[2], Leo G Nijtmans[2] and Martijn A Huynen[1*]

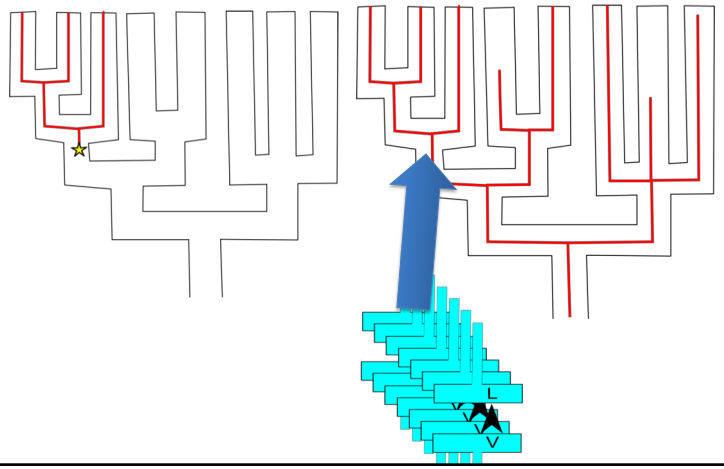


COX14 and C12orf62

- regulates cytochrome c oxidase assembly
- Found in the HHM-vs-HMM phase



These uncovered diverged homologs that turn out to be orthologs and have the same function, shows that sequence evolution can be highly neutral



The CKK Domain (DUF1781) Binds Microtubules and Defines the CAMSAP/*ssp4* Family of Animal Proteins

Anthony J. Baines,† Paola A. Bignone,*[1] Mikayala D.A. King,* Alison M. Maggs,‡ Pauline M. Bennett,‡ Jennifer C. Pinder,‡[2] and Gareth W. Phillips*‡[3]*

*Department of Biosciences, University of Kent, Canterbury, Kent, United Kingdom; †Centre for Biomedical Informatics, University of Kent, Canterbury, Kent, United Kingdom; and ‡Randall Division of Cell and Molecular Biophysics, King's College London, New Hunt's House, London, United Kingdom

We describe a structural domain common to proteins related to human calmodulin-regulated spectrin-associated protein1 (CAMSAP1). Analysis of the sequence of CAMSAP1 identified a domain near the C-terminus common to CAMSAP1 and two other mammalian proteins KIAA1078 and KIAA1543, which we term a CKK domain. This domain was also present in invertebrate CAMSAP1 homologues and was found in all available eumetazoan genomes (including cnidaria), but not in the placozoan *Trichoplax adherens*, nor in any nonmetazoan organism. Analysis of codon alignments by the sitewise likelihood ratio method gave evidence for strong purifying selection on all codons of mammalian CKK domains, potentially indicating conserved function. Interestingly, the *Drosophila* homologue of the CAMSAP family is encoded by the *ssp4* gene, which is required for normal formation of mitotic spindles. To investigate function of the CKK domain, human CAMSAP1-enhanced green fluorescent protein (EGFP) and fragments including the CKK domain were expressed in HeLa cells. Both whole CAMSAP1 and the CKK domain showed localization coincident with microtubules. In vitro, both whole CAMSAP1-glutathione-s-transferase (GST) and CKK-GST bound to microtubules. Immunofluorescence using anti-CAMSAP1 antibodies on cerebellar granule neurons revealed a microtubule pattern. Overexpression of the C[illegible] conc[illegible]

We conclude that the CKK domain binds microtubules and represents a domain that evolved with the metazoa.

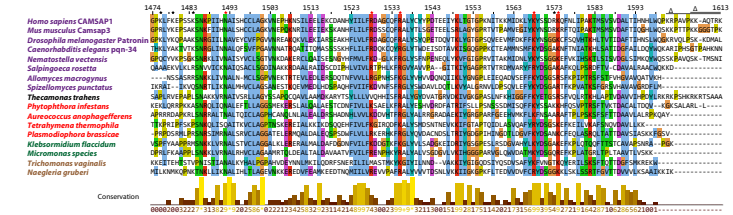
Mol. Biol. Evol. 26(9):2005–2014. 2009
doi:10.1093/molbev/msp115
Advance Access publication June 9, 2009

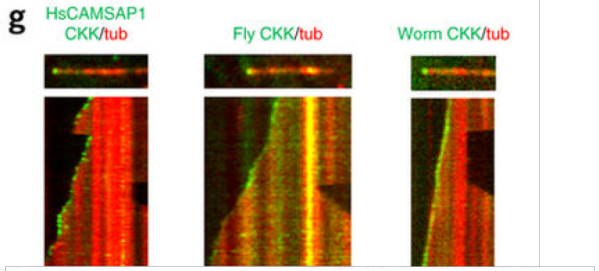
But:

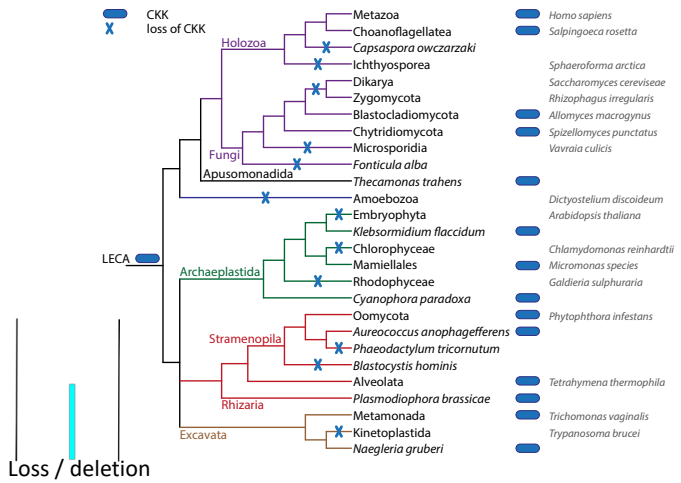
ARTICLES
nature structural & molecular biology

A structural model for microtubule minus-end recognition and protection by CAMSAP proteins

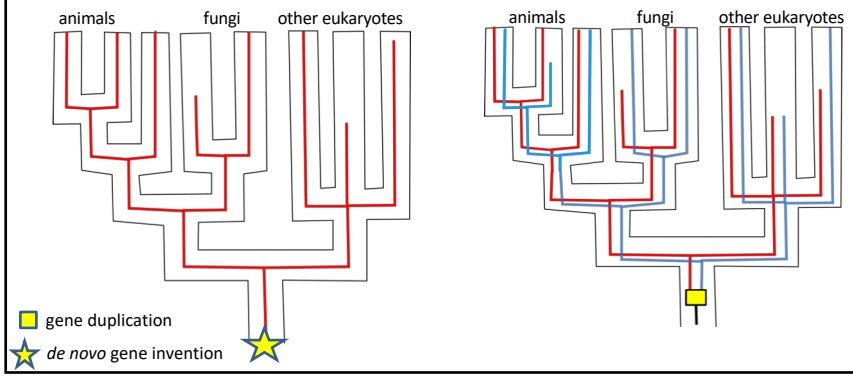
Joseph Atherton[1,11], Kai Jiang[2,11], Marcel M Stangier[3], Yanzhang Luo[4], Shasha Hua[2], Klaartje Houben[4], Jolien J E van Hooff[5–7], Agnel-Praveen Joseph[1], Guido Scarabelli[8], Barry J Grant[8], Anthony J Roberts[1], Maya Topf[1], Michel O Steinmetz[9,10], Marc Baldus[4], Carolyn A Moores[1] & Anna Akhmanova[2]



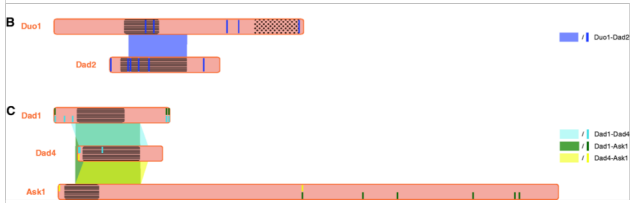




Improved sensitivity of sequence similarity searches allows to distinguish between genes invented in the ancestor or duplicated in the ancestor; i.e. *without* outparalogs or *with* outparalogs



Intra-complex homologies predicted from profile-profile searches suggests pre-LECA duplication



These homologies (paralogies) were confirmed by cryoEM and in addition even more homologies were detected.

RESEARCH

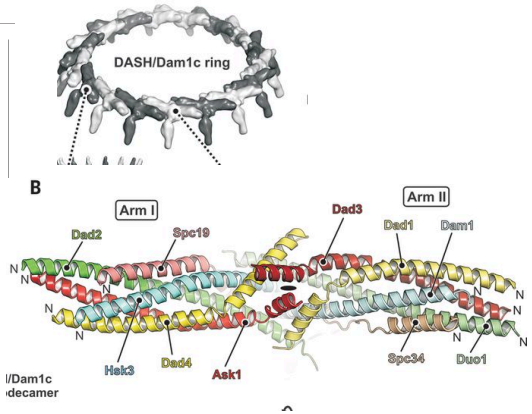
STRUCTURAL BIOLOGY

Structure of the DASH/Dam1 complex shows its role at the yeast kinetochore-microtubule interface

Simon Jenni[1] and Stephen C. Harrison[1,2]*

Kinetochores connect mitotic-spindle microtubules with chromosomes, allowing microtubule depolymerization to pull chromosomes apart during anaphase while resisting detachment as the microtubule shortens. The heterodecameric DASH/Dam1 complex (DASH/Dam1c), an essential component of yeast kinetochores, assembles into a microtubule-encircling ring. The ring associates with rodlike Ndc80 complexes to organize the kinetochore-microtubule interface. We report the cryo–electron microscopy

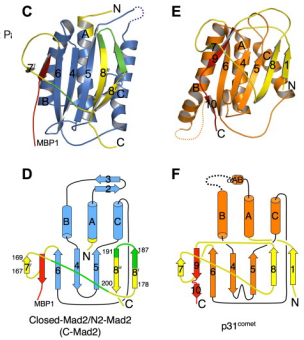
Jenni *et al.*, *Science* **360**, 552–558 (2018) 4 May 2018



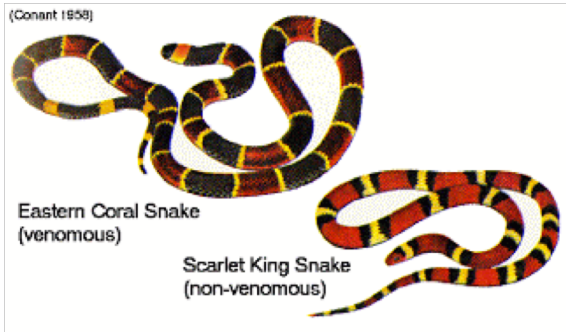
p31^{comet} Blocks Mad2 Activation through Structural Mimicry

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 Pi
*Correspondence: xuelian.luo@utsouthwestern.edu
DOI 10.1016/j.cell.2007.08.048

What is their scenario?
Imply convergent evolution?
Same fold different origin?



mimicry



http://falkenblog.blogspot.nl/2010/07/batesian-mimicry-explanation-of.html

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  140 (274)
Q Consensus         65 t~e~C~rfv~EL----LK~LLYqR~QIPfpYd~Lk~~v~K~~~~~d~~~~k~~~~~~q~rk~~~~l~~le~ll~~L~~  140 (274)
                       |.+++..+|.++    +-.|||+|.=.|--+=+-+..+.=......+.                  ++.+=|+.++..+..
T Consensus          1 t~~~S~~~v~~~l~~ai~~Ily~RgiyP~~~F~~~~~~~l~v~~~~~~~~~~~~~~~~~~~~~~~~l~~~i~~~~~~v~d   63 (189)
T pfam02301          1 TLKQSLELVKEFLEVAINSILYLRGIYPEESFEDRKKYNLPVLVSEDP----------------QLIDYLEKVLSGVFD   63 (189)

Q Thu_Jan_27_11:   141 FFARTLVPRVLILLGGNA----LSPKEFYELDLSLLAPYSVDQSL-----STAACLRRLFRAIFMADAF-SELQAPPLMG  210 (274)
Q Consensus        141 ~F~~s~V~~VliLfGsT~~~~~sPKE~Y~I~lp~~~~~~~e~~l-----st~~~lRkL~R~L~t~d~l-s~l~s~plt~  210 (274)
                       +.....++++.|.+-...    -.+.|.|.++|.-...+.....-     .++.-+|.++|+|+.+-.+ ..+..-....
T Consensus         64 aL~k~~L~~l~l~I~~~~~~~~~lE~y~F~~~~~~~~~~~~~~~~~~~~~~lr~l~~~~~L~~LP~~~~~  143 (189)
T pfam02301         64 ALEKGYLKKLVLVIYEDDPEKENEVLERYQFDFSYFPSGGNSSDSEKTEDETRQEIRALLRQLIALVTFLPPLPEDRTCT  143 (189)

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 FKLLYYTPPDYEPPGFKW  161 (189)
```

Homology and fold ok; what about function?

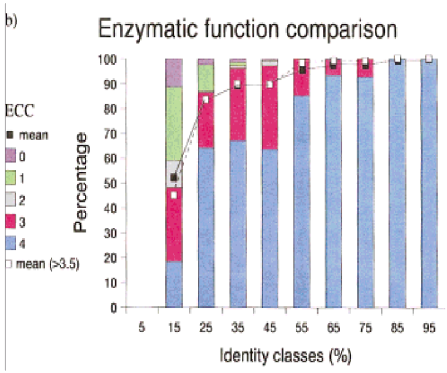
- To what extent do homologs/"proteins in a protein family", have the same "function"?
- Structure determines function? Fold != exact structure
- If distant homologs are orthologs likely "the same" function (i.e. CAMSAP/CKK, COX14)
- Relevant for function prediction
- Relevant for evolution of function

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, RhoGAPs, (enzymatic activity)
- Difficult with SH2 (bind to tyr-P), **Cys_2His_2** ZINC fingers, (DNA & RNA binding)
- Even more difficult with WD40, TPR (scaffoliding / sturcutral roles)

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.

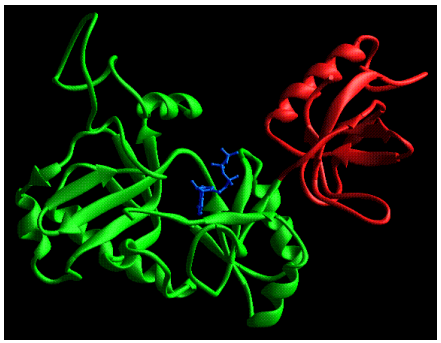
Y. Yuan, J. Schultz, M. Mlodzik, P. Bork

When detecting diverged homologies many homologies turn out to be restricted to small parts of the protein: domains

- Domains emphasize the fact that bits of protein can duplicate and recombine into “novel” proteins
- Gene families emphasize that duplications expands the number of homologs within a genome

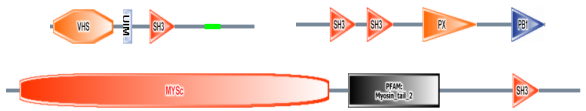
Protein domains: structural definition: separate in structure

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

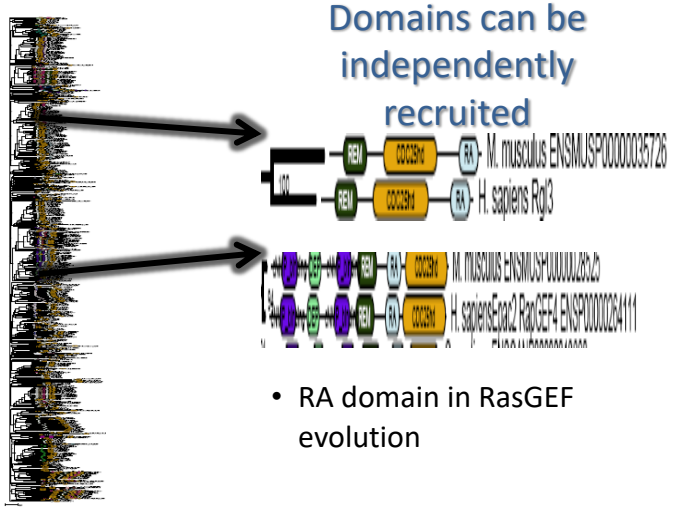


Protein domains: sequence/evolutionary definition: Separate in “evolution”

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



Domains can be independently recruited

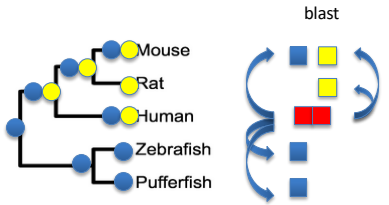


- RA domain in RasGEF evolution

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.)



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

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

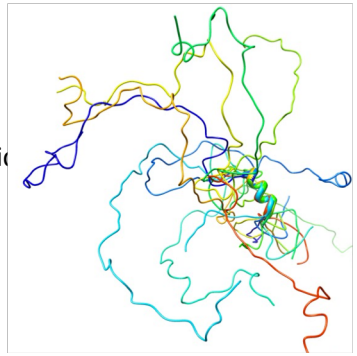
- Rather than discover the domain structure by blast yourself, use e.g. SMART / PFAM / CDD to do it for you
- NB Conserved Domain Database

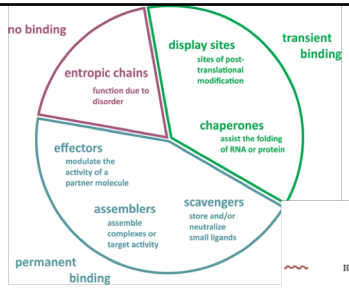
Beyond globular domains

- The preceding (and 99% of protein / structural bioinformatics) deals with “globular domains”
- However sometimes you also want to study the evolution of non-globular protein sequences

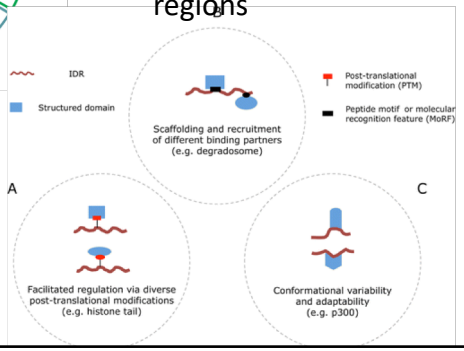
Disclaimer 1: intrinsically disordered proteins

- 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





Functions of non-globular / disordered / unstructured regions



So how do they evolve? How should we think about that?

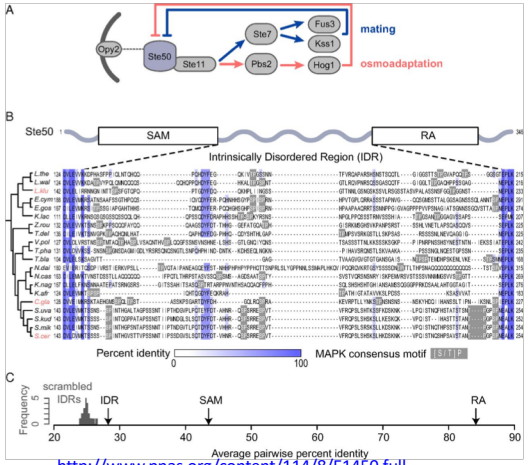
Table 2. Estimated disorder frequencies

Kingdom organism	Number of sequences	Disorder frequency	Length >30	Length >50
Archaea *Aeropyrum pernix*	1841	4.7	2.1	0.5
Archaea *Archaeoglobus fulgidis*	2409	2.8	0.9	0.2
Archaea *Halobacterium* sp.	2442	6.2	5.0	1.9
Archaea *Methanococcus jannaschi*	1784	2.8	1.0	0.3
Archaea *Pyrococcus abyssi*	1769	3.0	1.4	0.7
Archaea *Thermoplasma volcanium*	1497	3.2	1.0	0.3
Bacteria *Agrobacterium tumefaciens* C58	5288	6.4	5.7	2.0
Bacteria *Aquifex aeolicus* VF5	1557	3.3	1.9	0.4
Bacteria *Chlamydophila pneumoniae* AR39	1111	6.2	4.8	2.3
Bacteria *Chlorobium tepidum* TLS	2248	5.1	3.3	0.5
Bacteria *Treponema pallidum*	1035	6.1	6.4	2.6
Eukaryota *Arabidopsis thaliana*	21,482	16.8	33.8	19.0
Eukaryota *Caenorhabditis elegans*	20,506	15.9	27.5	15.6
Eukaryota *Drosophila melanogaster*	13,913	21.6	36.6	22.1
Eukaryota *Homo sapiens*	26,385	21.6	35.2	21.9
Eukaryota *S. cerevisiae*	6245	17.0	31.2	19.3
Archaea	11,742	3.8	2.0	0.7
Bacteria	35,389	5.7	4.2	1.6
Eukaryota	88,531	18.9	33.0	19.6
PDB (non-redundant at 95% sequence identity)	7169	3.2	0.5	0.1

The columns show the number of sequences, the percentage of residues predicted as being disordered and the percentage of chains with contiguous disordered segments of length greater than 30 and 50 residues, respectively.

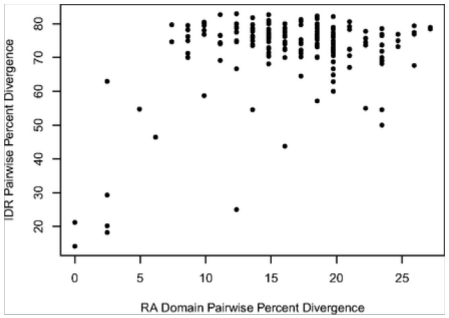
https://www.sciencedirect.com/science/article/pii/S0022283604001482

example



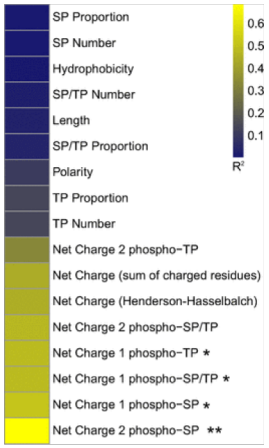
http://www.pnas.org/content/114/8/E1450.full

example

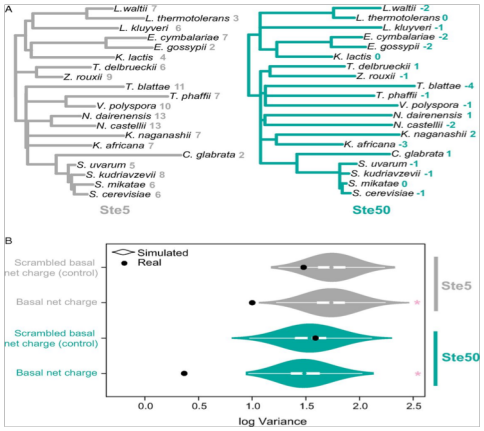


Diverged orthologous IDRs recapitulate *S. cerevisiae* IDR functions compared with the 5A mutant.

&

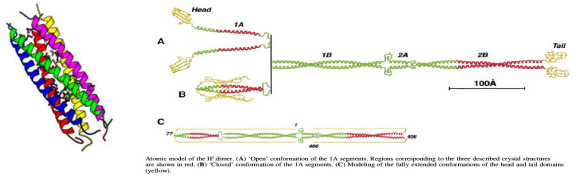


Conservative selection in net charge, but not on sequence



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

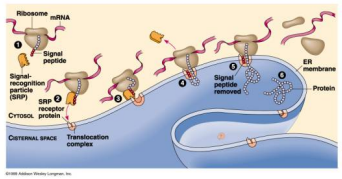


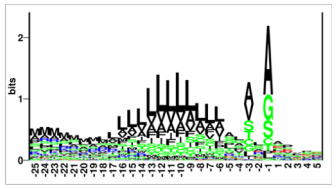
How to deal with coiled-coil (CC) proteins in homology / orthology searches?

- No one really knows / no accepted method / but needed for evolutionary cell biology
- Coiled coil is A VERY BIG problem for iterative methods (psi-blast / jack-hmmer) i.e. if you see e.g. myosin / dynein / spectrin; ABORT in profile-vs-profile searches many CC proteins are significantly similar to manyCC proteins
- Only use globular & non-coiled coil part of the protein.
- Use blast hopping?

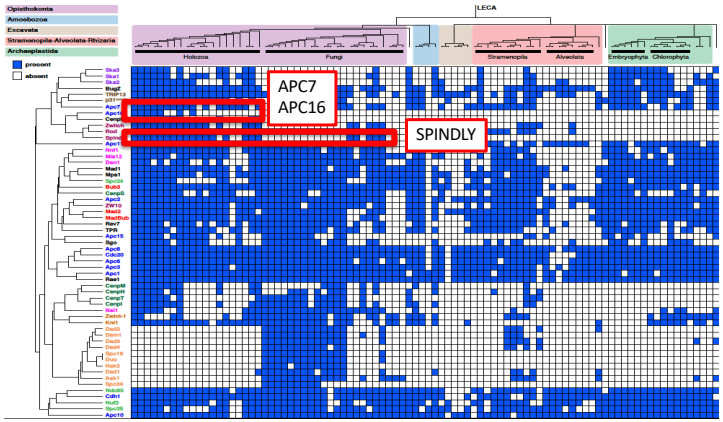
Disclaimer 3: protein motifs

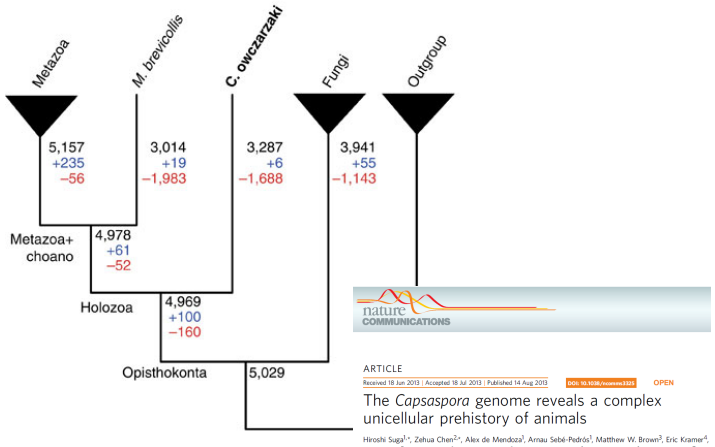
- Signal peptides
- Lipid anchoring
- Trans-membrane
- Kinase consensus motifs
- Can convergently evolve yet still important to predict





Apparent lineage specific (LS) genes?





nature COMMUNICATIONS

ARTICLE

Received 18 Jun 2013 | Accepted 18 Jul 2013 | Published 14 Aug 2013 DOI: 10.1038/ncomms3353 OPEN

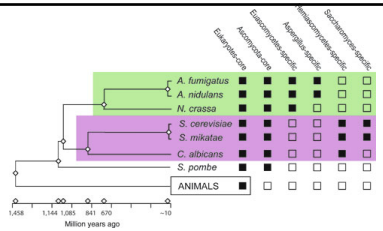
The *Capsaspora* genome reveals a complex unicellular prehistory of animals

Hiroshi Suga[1,*], Zehua Chen[2,*], Alex de Mendoza[3], Arnau Sebé-Pedrós[3], Matthew W. Brown[3], Eric Kramer[4], Martin Carr[5], Pierre Kerner[6], Michel Vervoort[6], Núria Sánchez-Pons[1], Guifré Torruella[1], Romain Derelle[7], Gerard Manning[4], B. Franz Lang[8], Carsten Russ[2], Brian J. Haas[2], Andrew J. Roger[3], Chad Nusbaum[2] & Iñaki Ruiz-Trillo[1,9,10]

What about apparent lineage specific genes? (LS)

Four possibilities are implicitly or explicitly 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 randomly emerging ORFs in non coding DNA. 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 ("Proto-genes and de novo gene birth", https://www.nature.com/articles/nature11184). (Also non globular!)
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

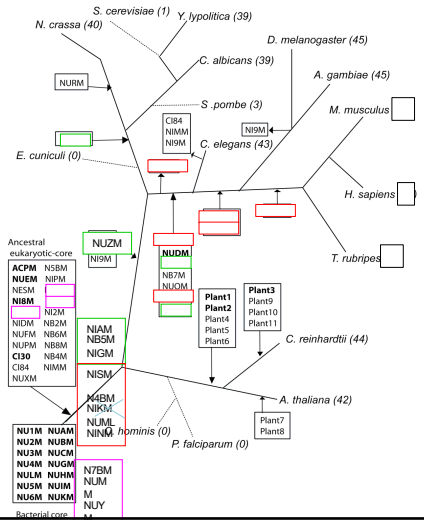
Table 2. Average nonsynonymous substitution rate (K_a), synonymous substitution rate (K_s), and K_a/K_s ratio among LS classes

LS class	No. of gene pairs	Mean (SD)		
		K_a^a	K_s^b	K_a/K_s^a
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; $p > 0.01$.

So they conclude …

- High correlation between amino acid substitutions and novely, (stronger than other factors tcorrelating with rate such as expression, essentiality, dispensability, or number of protein-protein interactions.
- The accelerated evolutionary rates of genes with higher LS may reflect the influence of selection and adaptive divergence during the emergence of orphan genes. These analyses suggest that accelerated rates of gene evolution may be responsible for the emergence of apparently orphan genes. (???)



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

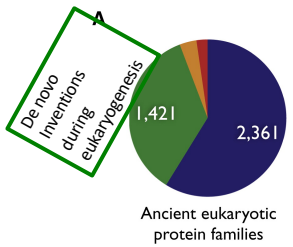
Balsa et al, Cell Metab. 2012

Gabaldon et al, JMB, 2005

"Anything goes" in (genome) evolution

- Some lineage specific genes/families are the result of coding becoming non-coding
  - evolutionary/ transcriptional noise which is marginally functional, but provides substrate for evolution that infrequently becomes "really" functional
- And others from extreme sequence (and structure?) divergence after duplication **or** speciation
  - technically maybe not novel but until structure solved and subsequent analysis suggest superfamily relation they would classify as "de novo origin".
- Note: The better we are able to detect homology, the less de novo we think we see

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.

Summary & connection to automatic phylogeny methods

- Distant homology / iterative or clustered homoloy searches lead to
  - "Protein families"
  - "Protein domains"
  - They are the same thing but emphasize different aspects
  - Families emphasize duplication (and HGT, secondary endosymbiosis, WGD)
  - Domains emphasize gene family fusion/recombination after duplication)

  - (blackboard)

When to do what

- Sometimes sequence similarity is the bottle neck for finding orthologs e.g. med11, apc15???, spindly
  - Fulfill separated by speciation and bi-directional best hit criterion
  - are occasionally found via experiments rather than sequence
- Sometimes gene duplications are the problem
  - Make "informative" trees
- Sometimes domain recombinations or motifs are "the problem"

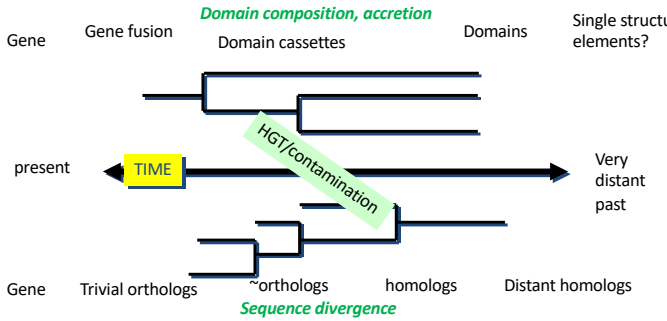
Automatic methods to obtain use curated homologous protein / gene families

- Just use PFAM? Works fairly well, but …
  - Misses novel gene families (e.g. taxon specific families in e.g. oomycetes)
  - False negatives (e.g. schnipsel)
  - Certain families are "too much like a domain" to go into an e.g. tree pipeline / are not what people would consider a domain.
    - Too promiscious
    - Families too big
    - Sequences too short

Implication of coupling between duplication & domain accretion for evolution (ortholog) and function prediction

- for some genes life is easy 1:1:1 orthologs, no fusion / domains, couple of losses. For a minority of families **but a large** proportion of proteins it is a formidable challenge. Domain permutations, duplications and unrecognized homology make "life complicated"

The too ambitious comparative genomics dilemma: duplication/speciation vs domains



i.e. genome comparison between close species:
no domain considerations, sub-sub-ortholog. Between distant
Homologs, loads of domain considerations
(gene trees, problematic)