**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?

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!

In principle but fusion/fission

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.

(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%
“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 jackhammer. 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-Blat “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) Few losses.

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

<table>
<thead>
<tr>
<th>A</th>
<th>G</th>
<th>R</th>
<th>N</th>
<th>G</th>
<th>A</th>
<th>C</th>
<th>R</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>C</td>
<td>G</td>
<td>N</td>
<td>R</td>
<td>A</td>
<td>C</td>
<td>G</td>
<td>N</td>
</tr>
<tr>
<td>C</td>
<td>T</td>
<td>C</td>
<td>Q</td>
<td>Q</td>
<td>I</td>
<td>T</td>
<td>C</td>
<td>I</td>
</tr>
</tbody>
</table>

Used tools: HHsearch/hhpred, PRC, compass

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 prokaryotes: (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

Compare to SCOP superfamilies, <20%
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/sap4 Family of Animal Proteins

Anthony J. Raine,‡ Paola A. Bigiòne,‡ Mikayla D.A. King,* Aliyn M. Maggs,‡ Pauline M. Bennetts, Jennifer C. Finder,‡ and Gareth W. Phillips*‡

‡Departments of Biosciences, University of Kent, Canterbury, Kent, United Kingdom; *Center for Biomedical Informatics, University of Kent, Canterbury, Kent, United Kingdom; and [Randal Division of Cell and Molecular Biophysics, King’s College London, New Kent’s House, London, United Kingdom.

We describe a structural domain common to proteins related to human calcium-regulated spectrin-associated proteins (CAMSAPs). Analysis of the sequence of CAMSAP1 identified a domain near the C-terminus common to CAMSAP1 and two other mammalian proteins (KIAA0755 and KIAA1053), which we term a CKK domain. This domain was also present in invertebrate CAMSAP homologues and was found in all available cnidarian genomes (including cnidarians), but not in the phylum Placozoa, nor in any annelid or nematode. Analysis of conserved alignments by the online likelihood ratio method gave evidence for strong purifying selection on all codons in the mammalian CKK domain, potentially indicating conserved function. Interestingly, the Drosophila homologue of the CAMSAP family is modified by a tandem repeat of the CKK domain, suggesting that this domain may have evolved as a result of localised gene duplication in the distributed actin cytoskeleton. In vitro, human CAMSAP1 enhanced green fluorescent protein (EGFP) and fragments including the CKK domain were expressed in HeLa cells. Both wild-type CAMSAP1 and the CKK domain showed localization coincident with microtubules, whereas fragments lacking the CKK domain were cytoplasmic. In vivo, both wild-type CAMSAP1 and CKK-GST bound to microtubules, reminiscent of the in vitro results. We conclude that the CKK domain defines a domain that evolved with the metazoa.
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.

What is their scenario? Imly convergent evolution? Same fold different origin?
mimicry


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 conservation 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 other's 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)
**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**

- Protein kinases, RhoGAPs, (enzymatic activity)

  - Difficult with SH2 (bind to tyr-P), \textbf{Cys}\textsubscript{2}\textbf{His}\textsubscript{2} ZINC fingers, (DNA & RNA binding)

  - Even more difficult with WD40, TPR (scaffolding / structural roles)
Using distant homology for function prediction: example from (just) before PSI-BLAST & HMMer

Secreted Fringe-like Signaling Molecules May Be Glycosyltransferases.
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

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 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

<table>
<thead>
<tr>
<th>Kingdom organism</th>
<th>Number of sequences</th>
<th>Disorder frequency</th>
<th>Length &gt;30</th>
<th>Length &gt;50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archaea</td>
<td>1841</td>
<td>4.7</td>
<td>2.1</td>
<td>0.5</td>
</tr>
<tr>
<td>Archaea</td>
<td>2809</td>
<td>2.8</td>
<td>0.9</td>
<td>0.2</td>
</tr>
<tr>
<td>Archaea</td>
<td>2862</td>
<td>6.2</td>
<td>4.0</td>
<td>0.3</td>
</tr>
<tr>
<td>Archaea</td>
<td>1994</td>
<td>2.8</td>
<td>1.0</td>
<td>0.3</td>
</tr>
<tr>
<td>Archaea</td>
<td>1769</td>
<td>3.0</td>
<td>4.4</td>
<td>0.7</td>
</tr>
<tr>
<td>Archaea</td>
<td>1997</td>
<td>3.2</td>
<td>1.0</td>
<td>0.3</td>
</tr>
<tr>
<td>Bacteria</td>
<td>5286</td>
<td>6.4</td>
<td>5.7</td>
<td>2.0</td>
</tr>
<tr>
<td>Bacteria</td>
<td>1957</td>
<td>3.3</td>
<td>1.9</td>
<td>0.4</td>
</tr>
<tr>
<td>Bacteria</td>
<td>1111</td>
<td>6.2</td>
<td>4.4</td>
<td>2.3</td>
</tr>
<tr>
<td>Bacteria</td>
<td>2248</td>
<td>5.1</td>
<td>3.3</td>
<td>0.5</td>
</tr>
<tr>
<td>Bacteria</td>
<td>1075</td>
<td>6.1</td>
<td>4.4</td>
<td>2.6</td>
</tr>
<tr>
<td>Bacteria</td>
<td>21,482</td>
<td>16.8</td>
<td>33.8</td>
<td>19.0</td>
</tr>
<tr>
<td>Bacteria</td>
<td>28,506</td>
<td>35.9</td>
<td>27.5</td>
<td>15.4</td>
</tr>
<tr>
<td>Bacteria</td>
<td>13,913</td>
<td>21.6</td>
<td>36.6</td>
<td>22.1</td>
</tr>
<tr>
<td>Bacteria</td>
<td>26,385</td>
<td>21.6</td>
<td>35.2</td>
<td>21.9</td>
</tr>
<tr>
<td>Bacteria</td>
<td>42,915</td>
<td>17.0</td>
<td>31.2</td>
<td>19.3</td>
</tr>
<tr>
<td>Archaea</td>
<td>11,742</td>
<td>3.6</td>
<td>2.0</td>
<td>0.7</td>
</tr>
<tr>
<td>Bacteria</td>
<td>35,389</td>
<td>5.7</td>
<td>4.2</td>
<td>1.6</td>
</tr>
<tr>
<td>Bacteria</td>
<td>18,535</td>
<td>10.9</td>
<td>33.0</td>
<td>19.6</td>
</tr>
<tr>
<td>PDB (non-redundant at 95% sequence identity)</td>
<td>7169</td>
<td>3.2</td>
<td>0.5</td>
<td>0.1</td>
</tr>
</tbody>
</table>

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.
Diverged orthologous IDRs recapitulate S. cerevisiae IDR functions compared with the 5A mutant.
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 many CC 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?

- APC7
- APC16
- SPINDLY
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!

So they conclude …

• High correlation between amino acid substitutions and novelty, (stronger than other factors correlating 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. (???)
“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

Irrespectively 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.
Distant homology / iterative or clustered homology 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

Gene fusion

Domain cassettes

Domains

Single structural elements?

Gene fusion

Domain cassettes

Domains

Single structural elements?

Gene fusion

Domain cassettes

Domains

Single structural elements?

Gene fusion

Domain cassettes

Domains

Single structural elements?

Gene fusion

Domain cassettes

Domains

Single structural elements?

Gene fusion

Domain cassettes

Domains

Single structural elements?

Gene fusion

Domain cassettes

Domains

Single structural elements?

Gene fusion

Domain cassettes

Domains

Single structural elements?

Gene fusion

Domain cassettes

Domains

Single structural elements?

Gene fusion

Domain cassettes

Domains

Single structural elements?

Gene fusion

Domain cassettes

Domains

Single structural elements?

Gene fusion

Domain cassettes

Domains

Single structural elements?

Gene fusion

Domain cassettes

Domains

Single structural elements?

Gene fusion

Domain cassettes

Domains

Single structural elements?

Gene fusion

Domain cassettes

Domains

Single structural elements?

Gene fusion

Domain cassettes

Domains

Single structural elements?

Gene fusion

Domain cassettes

Domains

Single structural elements?

Gene fusion

Domain cassettes

Domains

Single structural elements?