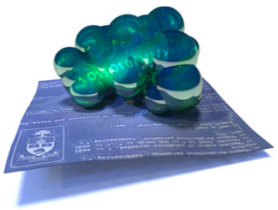
A
Bioinformatics
Course

Homology



Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto

homology

Almost all of bioinformatics is in some way derived from inference based on *homology*.

This was true when I wrote it 20 years ago, and it is still true today. There are interesting exceptions – such as inferences based on network relationships ... but for our everyday bioinformatics needs, these are truly exceptions.

Thus, since homology is important, we need to be really clear about what it means.

homology

Almost all of bioinformatics is in some way derived from inference based on *homology*.

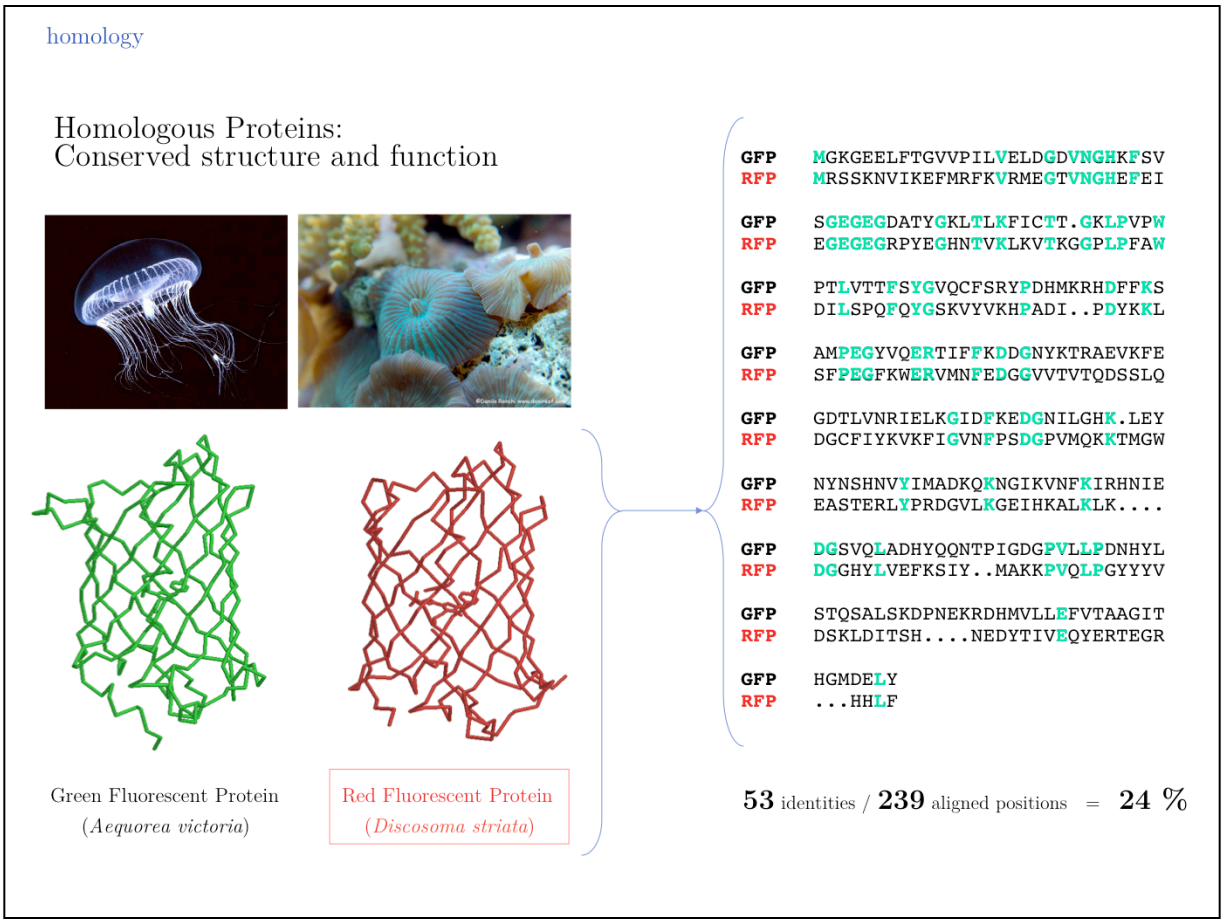
Two genes are **homologous** if they have diverged from a **common ancestor**.

The terms **homology** and **similarity** are often confused and used incorrectly.

Homology is a *quality*. Two genes can either be homologous, or not. There is no such thing is **highly homologous* or 50% homologous. People who speak like that do not fully understand what *homologous* means. Being homologous is like being "pregnant" in that sense: you can't be 50% pregnant. Looking forward in time this is completely intuitive: homology describes two examples of the very same gene (common ancestor) that evolved independently and diverged. Looking backward, this is less intuitive. Now we have two different genes and need to consider their evolutionary trajectory. But either there is a common ancestor, or there is not. It doesn't make sense to speak of "common ancestry" over only part of the evolutionary history.

Similarity on the other hand is a *quantity*. It can be measured, quantified, graded, and compared. Often, **homologous genes have similar sequences**. This implies that it is possible to discover homologous genes by measuring sequence similarity.

Also consider the term **analogous**. This describes similarty of function or structure or some other property, but not through homology – i.e. descendance from a common ancestor – but by convergent evolution. It is perhaps remarkable that there is no sequence similarity between analogous genes, except for those residues that may be directly involved in a function. (Cf. the analogous subtilisin and trypsin hydrolases that both have a catalytic triad.)



But why do we care? Why is it so important to know that two proteins are homologues? That's because:

Common ancestry implies similar structure and function.

Many obviously homologous genes have very low similarity. In this example, the aligned sequences of green- and red- fluorescent protein share only 57 of 239 residues, i.e. their pairwise sequence identity is 23.8%. more then three quarters of the amino acids in the two sequences are not identical!

Howvere, the two organisms share evolutionary ancestry and it is a reasonable hypothesis that the two fluorescent proteins have evolved from the same ancestral sequence. Strikingly, despite 78% amino acid differences in the sequence, the structures of the two proteins are virtually identical and their functions (autocatalytic cyclization and oxidation of a conjugated ssytem of double bonds from a polypeptide precursor) are very similar.

evolution

Mechanisms of variation and change

- Conjugation
- Sexual replication
- Point mutations (on a genetic code optimized for evolution)
- Gene loss (pseudogenes)
- Segmental inversion
- Segmental duplication
- Gene fusion
- Polysomy
- Polyploïdy
- Transposable elements (more than 30% of human genome)
- Retroviruses
- Horizontal gene transfer
- Gene transfer from acquired endosymbionts
- ...

But why is that the case? Don't protein sequences diverge precisely because this allows them to acquire new functions?

Yes and no.

In the ultimate outcome this is true. But considering the mechanism, we have a trajectory of gradual, stepwise change, oiften under continuous selective pressure that prevents the protein becoming completely non-functional.

Of course, disruptive changes happen too – but they are likely to result in pseudogenes – sequence fossils, that rapidly acquire further nonsense and mis-sense mutations.

Overall this results in structure being conserved (no major refolding), function being conserved (unlikely to acquire entirely new activities), and other functional features too being conserved.

definitions

Homologous := **Diverged from a common ancestor**

Orthologous := Homologous, diverging after speciation

Paralogous := Homologous, diverging after duplication

Analogous := *Similar, but not homologous*

Similar := **Sharing a property**

Important conjecture:
Function is conserved between *orthologues* and unconserved between *paralogues*!

There are two subcategories of homologous genes: those that arise from speciation, and those that arise from duplication events.

Orthologues:

Genes that have diverged through *speciation*.

Changes on the evolutionary trajectory occur under selective pressure.

Function ususally is conserved.

Orthologues are the closest analogue of each other in different species.

Although, even identical sequences can't really be said to have the **same** function in different species, after all, they operate in a different context.

Paralogues:

Genes that have diverged through *duplication*.

Changes on the evolutionary trajectory occur under reduced or absent selective pressure.

Consequences:

Function ususally is not conserved:

- *Neo*functionalization
- *Sub*functionalization

*Neo*functionalization: acquisition of a **new** function.

*Sub*functionalization: expression of the **original** function as a response to different signals, during different times, and/or in different tissues.

Homology is not a quantity but a quality.

Homology is commutative.

$$A \otimes B \Longrightarrow B \otimes A$$

Homology is transitive.

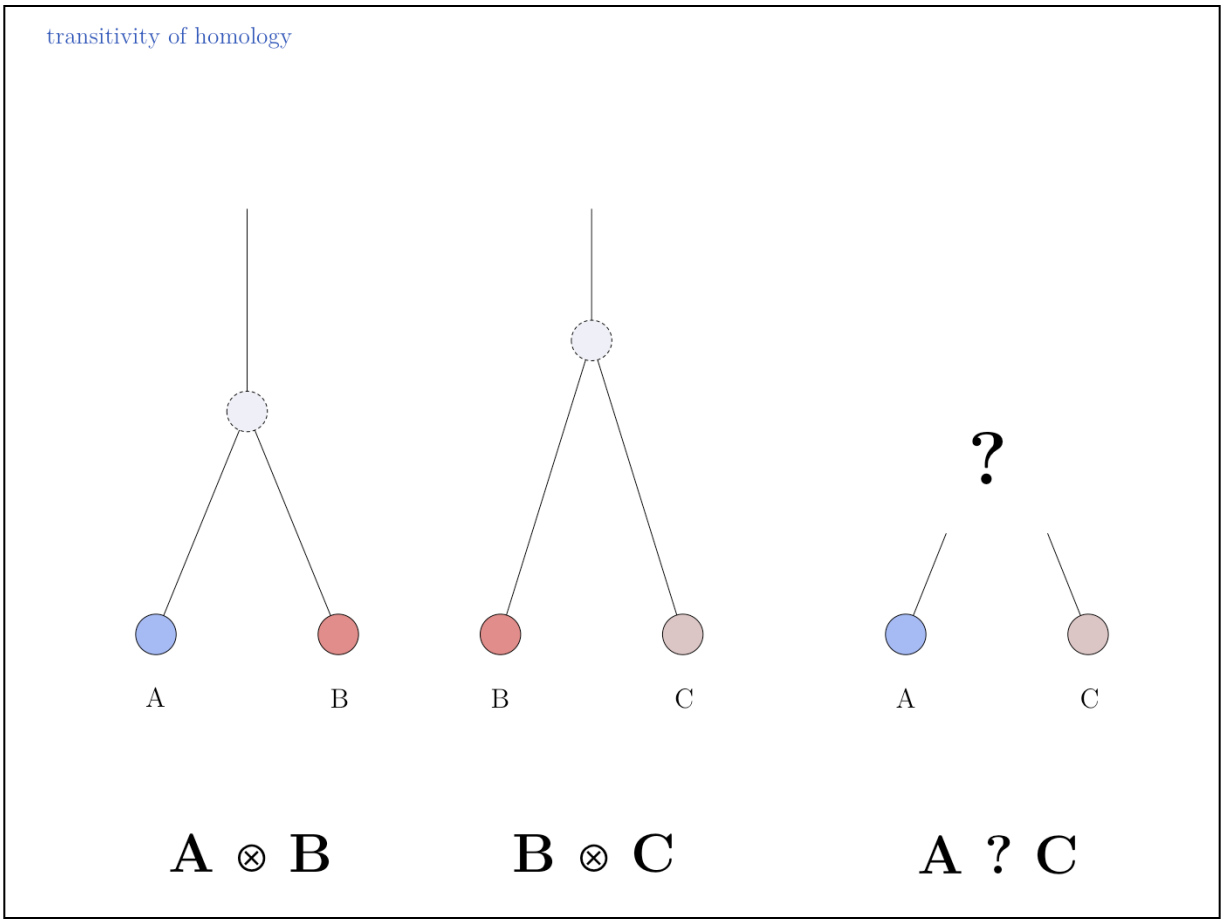
$$A \otimes B, B \otimes C \therefore A \otimes C$$

Three important properties of homology.

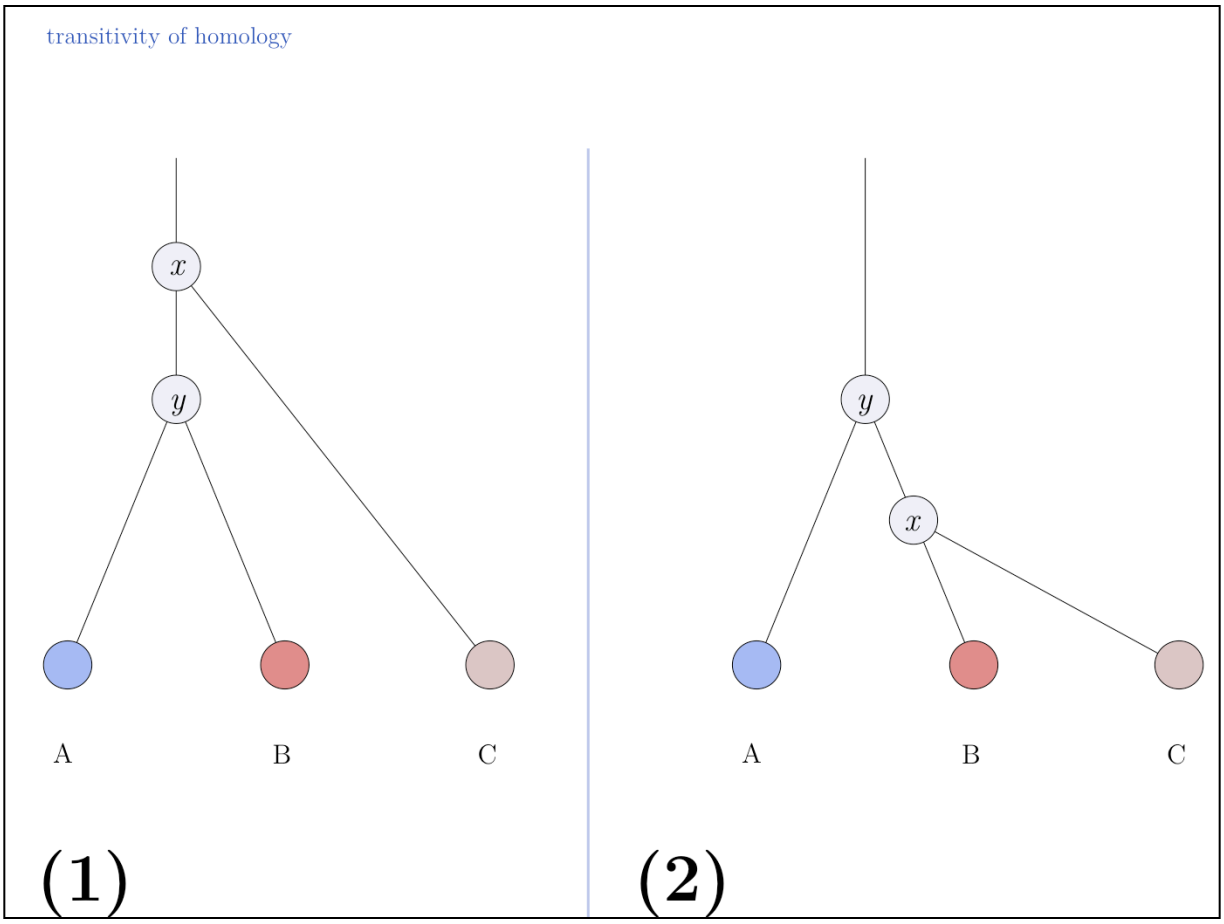
We have already mentioned that homology is a **quality**, that describes common ancestry.

The second property – **commutatitvity** – should be obvious since it is an immediate consequence of the definition: if A is homologous to B, then B is homologous to A.

Whether homology also must be **transitive** requires more a bit more consideration.



Is it necessarily the case that two proteins are homologous if both of them are (perhaps distantly) related to the same third protein?

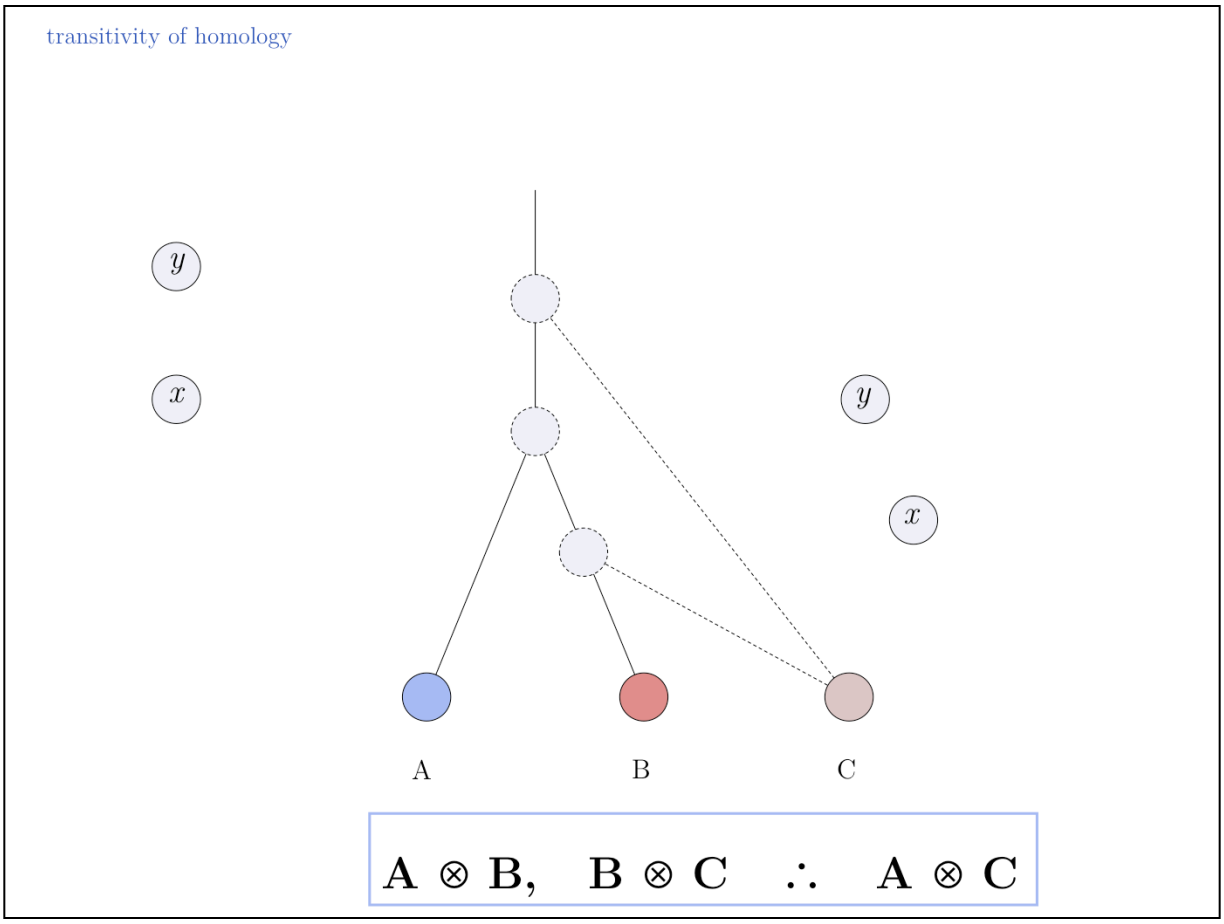


Yes, absolutely.

There are two possible evolutionary trees that describe the situation.

Both have a common ancestor: either x or y in our sketch above.

If we draw the evolutionary tree, all three genes are related to the same ancestor. The ordering of their descent (the topology of their evolutionary tree) may be different: but this only relates to **where** nodes x and y insert into the tree relative to each other, not whether there exists such a node in the first place..



This is how we mine sequence space for information.

Transitivity of homology justifies inferences across very distant evolutionary relationships, as long as connections via recognizably homologous genes can be made. This is the basis of advanced alignment algorithms that compare sequences against profiles or probabilistic models: a group of genes are **all** homologues if there is a path of homology relationships between **any** pair – however long hat path may be.

But note that this holds **only for domains**, not necessarily for entire genes with their patchwork of (possibly) independently inherited domains.

inferences

Homologous proteins always have similar structure.

Homologous proteins usually have similar function†.

Homology **can't be proven** since we can't observe ancestral sequences. However: ...

... **sequence similarity** can be measured.

Homologous proteins frequently have similar sequence.

† ... including similar localization, modification, processing, expression patterns, interactions etc.

Homologous proteins always have similar structure.

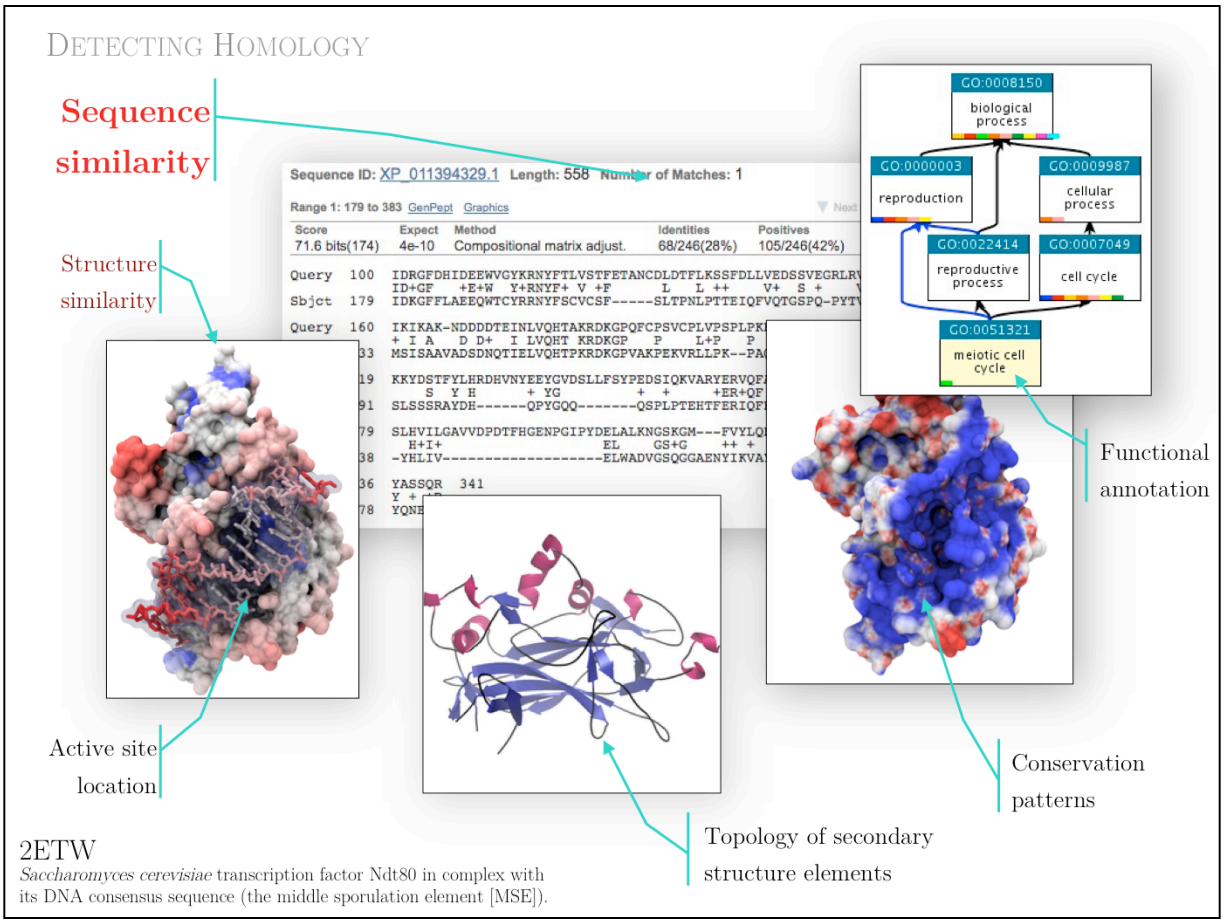
Homologous proteins usually have similar function†.

Homology can't be proven since we can't observe ancestral sequences.

However: sequence similarity can be measured. Homologous proteins frequently have similar sequence.

That said, how do we find sequences that are homologous, or, how do we measure similarity?

† ... including similar localization, modification, processing, expression patterns, interactions etc.



Any aspect of a protein that can be compared quantitatively or qualitatively can be considered:

Sequence similarity is the most important one: the number of possible sequences is so large, that there is virtually no other explanation for sequences that are more than about 25% identical, than homology.

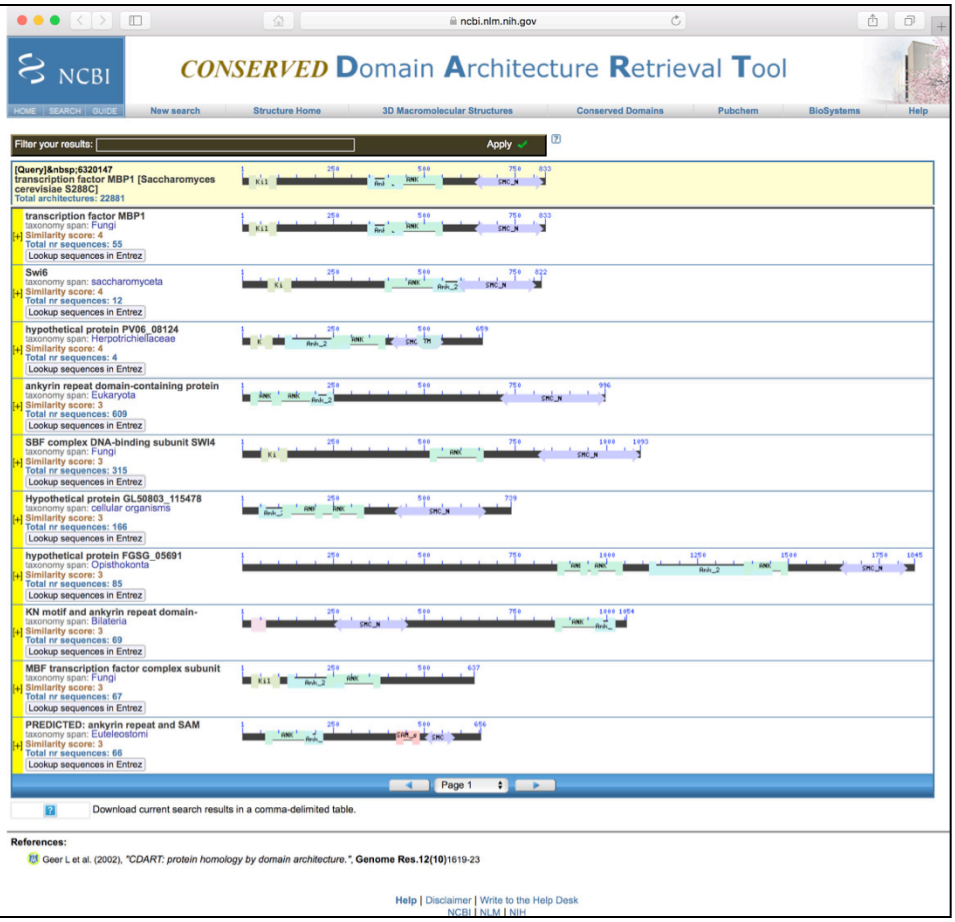
Proteins that share less sequence identity may (somtimes surprisingly) have **similar structure**, which still supports homology;

... especially if additional features can be shown to be equivalent like

- location of active sites;
- the global fold and arrangement of secondary structure elements;
- patterns of functionally conserved residues;
- functional annotations.

CDART

Access via **Domain Relatives** in the right-hand-menu list of **Related Information** of a protein's NCBI RefSeq database page.



While sequence similarity is the most important consequence of homology for practical purposes, matters are not as simple in actual proteins. Homology can manifest at the level of domains, and domains can be (freely) mixed in a combinatorial fashion. Thus proteins need not be homologous over their entire length!

Each domain may have its own, partially independent evolutionary history. And we frequently don't know exactly where the domain boundaries are. This tremendously complicates analysis and inference, because a given protein may be homologous in different parts to other proteins that are themselves not related at all. Our transitivity relation of homology only holds for domains, not necessarily for multi-domain proteins.

Databases such as CDART at the NCBI make this information available and explicit. But knowledege of domain distributions can tell us even more: combinations of functional domains (identified by homology to a domain family) can give mechanistic insight into a protein's function. For example, a DNA binding domain combined with a protein-protein interaction domain makes a good candidate for a transcription factor. This is good because similar arrangement of homologous domains is itself a hallmark of homology.

Yet, partial homology is a problem because it can lead to inappropriate annotations. A functional annotation at the domain level might be incorrect for the full-length protein.

http://steipe.biochemistry.utoronto.ca/abc

B O R I S . S T E I P E @ U T O R O N T O . C A

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA