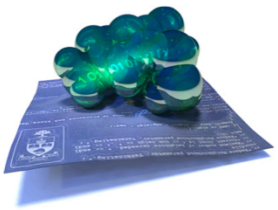
A
Bioinformatics
Course

Domain Annotation



Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto

Uses of Domain Information

Domains are units for ...

- separate folding
- distinct function
- modular inheritance

The *domain* is the natural unit of analysis of protein structure.

The discovery and catalogization of the universe of domains in protein sequences is a great achievement of profile– and model based sequence alignment experiments.

Uses of Domain Information

To ...

... identify regions of the polypeptide chain that fold independently, that are stable on their own
(folding units; initiation sites for folding)

... identify gene fusion or gene insertion events
from analysis of the 3D structure
(understand evolutionary history)

... understand protein mechanism as an additive/cooperative result of domain function
(CDART, SMART - domain architecture)

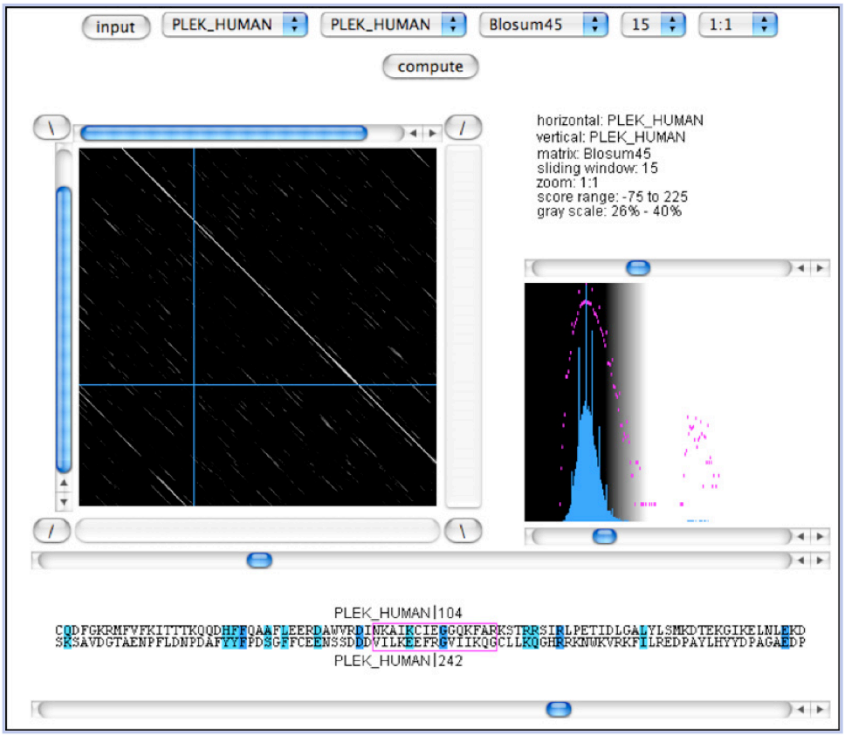
... allow for meaningful structural classification of proteins
(SCOP, CATH classifications)

Domains can be used for sequence analysis in many ways.

Domains: Units of Inheritance

Dotlet -

A dotplot of Pleckstrin (p47) reveals similarity between N-and C terminus !



Here is an example of how a domain was discovered from sequence alignments. A dotplot reveals significant similarity between the N- and the C- terminus of a protein.

Domains: Units of Inheritance

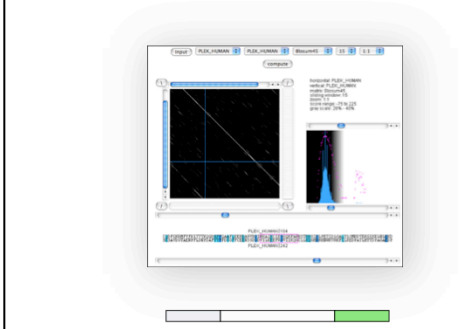
```
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 100
# Identity:      31/100 (31.0%)
# Similarity:    48/100 (48.0%)
# Gaps:           6/100 ( 6.0%)

   6 IREGYLVKKGSVFNTWKPMWVVLLEDG--IEFYKKKSDNSPKGMIPLKGS     53
     |::|.|:|:|.....||....:|.||.  :.:|.......|.|.|.|:|.
 245 IKQGCLLKQGHRRKNWKVRKFILREDPAYLHYYDPAGAEDPLGAIHLRGC    294

  54 TLTSPCQDFGKRMF----VFKITTTKQQDHFFQAAFLEERDAWVRDINKA     99
     .:||...:...|..    :|:|.|..:..:|.|||..:||..|::.|..|
 295 VVTSVESNSNGRKSEEENLFEIITADEVHYFLQAATPKERTEWIKAIQMA    344
```

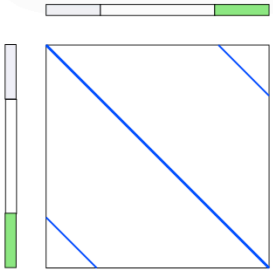
Optimal sequence alignment: 31% identity over ~100 amino acids.

(Alignment produced with the Needle program of the EMBOSS suite.)

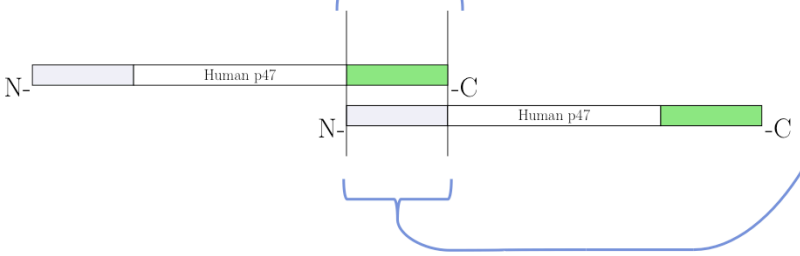
Sequence alignment shows high similarity between N- and C-terminus: at this level of similarity we are looking at two homologous domains, i.e. an "internal duplication" in the sequence.

Domains: Units of Inheritance





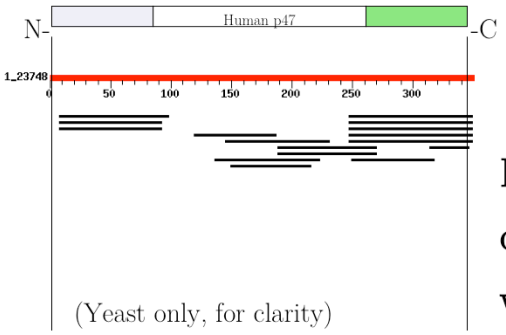
```
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 100
# Identity:      31/100 (31.0%)
# Similarity:    48/100 (48.0%)
# Gaps:           6/100 ( 6.0%)

   6 IREGYLVKKGSVFNTWKPMWVVLLEDG--IEFYKKKSDNSPKGMIPLKGS     53
     |::|.|:|:|.....||....:|.||.  :.:|.......|.|.|.|:|.
 245 IKQGCLLKQGHRRKNWKVRKFILREDPAYLHYYDPAGAEDPLGAIHLRGC    294

  54 TLTSPCQDFGKRMF----VFKITTTKQQDHFFQAAFLEERDAWVRDINKA     99
     .:||...:...|..    :|:|.|..:..:|.|||..:||..|::.|..|
 295 VVTSVESNSNGRKSEEENLFEIITADEVHYFLQAATPKERTEWIKAIQMA    344
```



Overlapping alignments define domain boundaries! We can search a database with this knowledge by retricting to our query to individual domains ...

Sequence Searches With Domains can ImpROVE Specificity

BLAST search with *full-length* Pleckstrin sequence

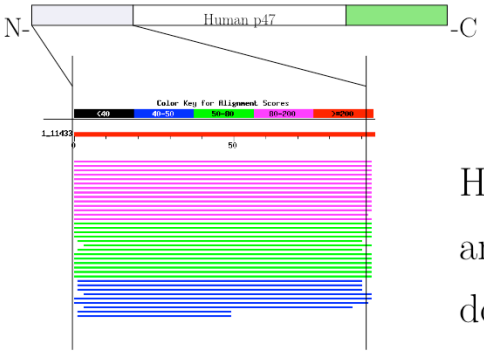


Hits extend over the entire domain. PSI -BLAST would be difficult ...

While full length searches are rather non-specific ...

Sequence Searches With Domains can ImpROVE Specificity

BLAST search with Pleckstrin Domain *only* ...



Hits are smoothly bounded and extend over the entire domain. PSI-BLAST is straightforward.

486 hits ... etc.

... searches with the domain sequences are highly specific and discover a large family of related sequences.

This is intriguing – why do we get better results if we search with **less** information?

The answer is this is not less information at all: e are actually **adding information** by defining the domain boundaries, and the algorithm is not mislead into trying to incorporate spurious similarities from irrelevant sequences at the termini into the aligned region, which degrades the results.

Domain discovery:

– HMMER profiles

– Pfam and SMART databases

Alternative: NCBI's RBS BLAST at CDD.

Domain discovery on a large scale has been made possible through Hidden Markov Model alignments, implemented in Sean Eddy's HMMER program. This has been used to compile large databases like **Pfam** that curate domain profiles. These profiles can be scanned against an unknown sequence, thus allowing the annotation of the sequence with the domains it contains. In many cases this allows to assign at least a coarse description of function and mechanism.

Sequence Based Domain Annotation:

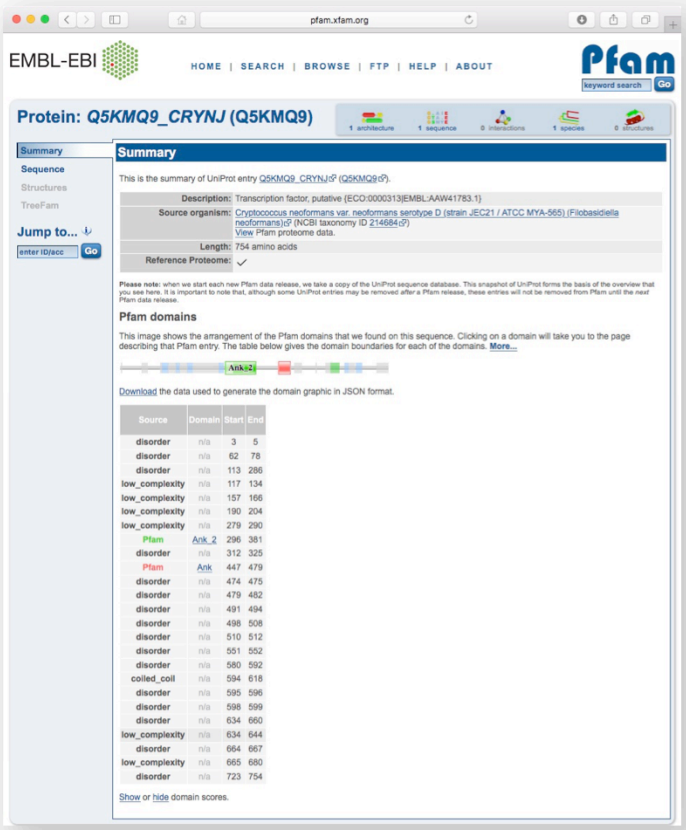
Pfam

http://pfam.xfam.org/

Pfam annotations for UniProt ID `Q5KMQ9`, the CRYNE Mbp1 ortholog.

Pfam makes the annotation data available in JSON format.

But where is the APSES domain?



Pfam is a large collection of protein domain families, based on automated multiple sequence alignments and Hidden Markov Models that represent the information in an alignment.

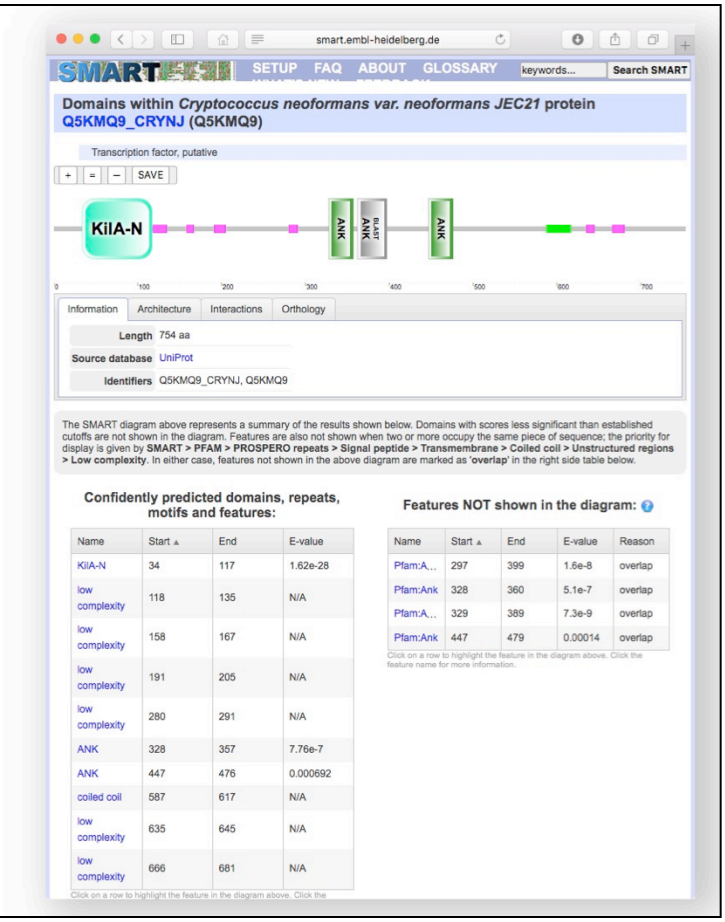
Sequence Based Domain Annotation:

SMART

http://smart.embl-heidelberg.de/

SMART annotations for UniProt ID **Q5KMQ9**, the CRYNE Mbp1 ortholog.

SMART does not support downloading the annotations.



Note that this SMART database result page has several tabs that lead to additional tools which provide context to the protein's function.

Sequence Based Domain Annotation: CDD Database

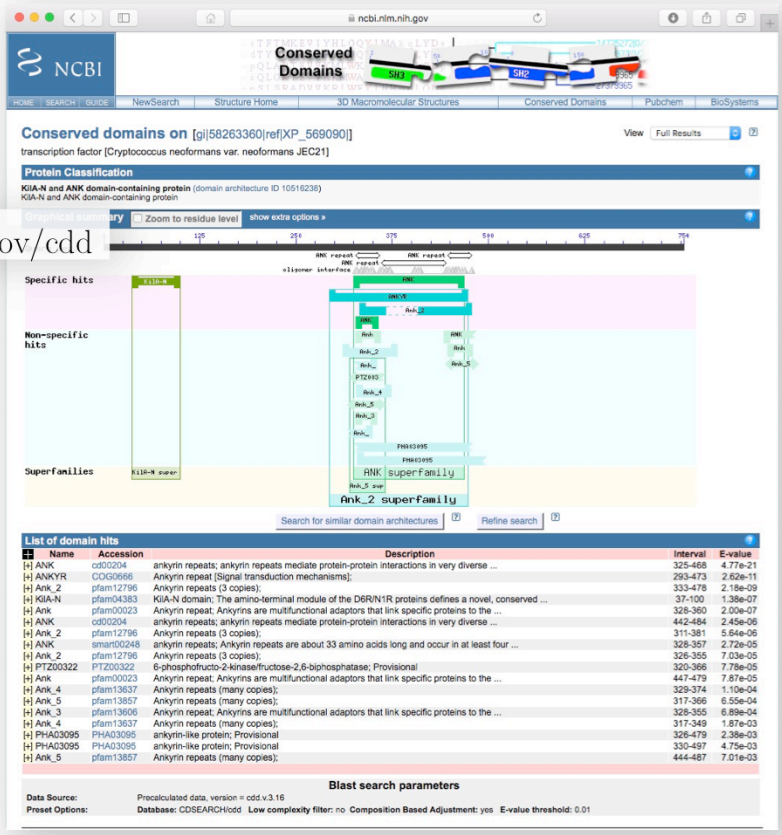
CDD

https://www.ncbi.nlm.nih.gov/cdd

CDD "Full Results" for RefSeqID **XP_569090**, the CRYNE Mbp1 ortholog.

This window is accessed via a side-bar link in the NCBI Protein database.

Note the detailed ankyrin domain annotations in "Full Results" view.



The NCBI CDD database identifies domain hits via a reverse BLAST algorithm (RPS-BLAST) in which it searches database profiles against a quaery sequence.

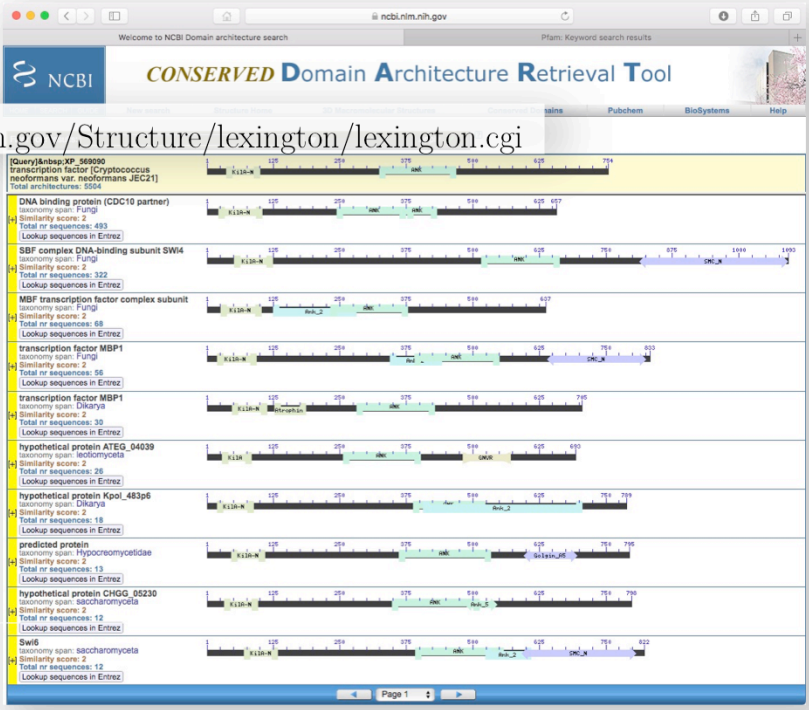
We see a very well resolved, fine grained annotation of the overlapping ankyrin repeats. However, there is no simple way to download nnotation data from here.

There is a related utility, the Batch serach facility (https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi) which supports download of annotation details and ranges in various formats, however the data is a bit awkward to use since it does not contain definitions, just NCBI internal IDs (GI), and it does not produce valid JSON but requires editing of the download file (includes more than one top-level object).

Sequence Based Domain Annotation: CDART Database

CDART

https://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi



CDART annotations for RefSeqID `XP_569090`, the CRYNE Mbp1 ortholog.

Annotations can be downloaded as csv files.

Geer L et al. (2002), "CDART: protein homology by domain architecture.", Genome Res.12(10)1619-23

The shared domain architectures discovered by CDART are invaluable to place the function of a protein into a larger context. Note that all hits share the KilA-N and the Ankyrin repeat domains, but there are several other domains in related families that have ancillary known domain annotations. These domains might also be pressent – albeit more highly diverged and thus below the cutoff of the algorithm – in our target 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

ExPASy/ProSite

http://www....

Ref ...

ExPASy...

RPS-BLAST

http://www....

Ref ...

RPS-BLAST...