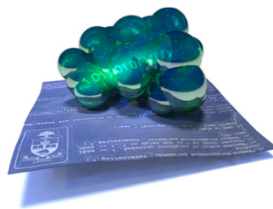


A
BIOINFORMATICS
COURSE

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

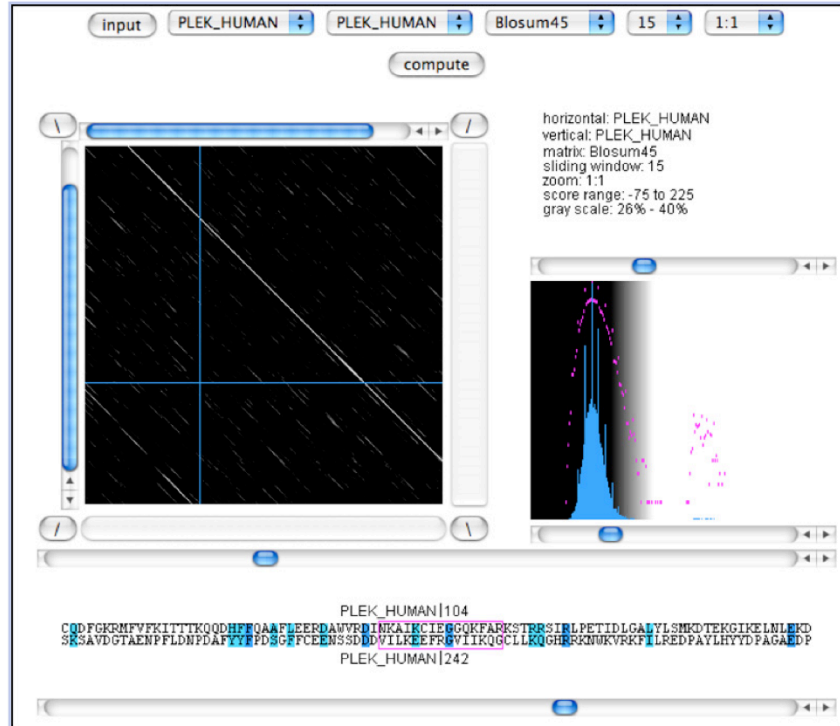
(SCOP, CATH classifications)

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 IREGYLVKKGSVFNTWKPMWVLLLEDG--IEFYKKKSDNSPKGMIPLKGS      53
   |::|.::|:|. . . . . | | . . . . :|. | | .   :::| . . . . . | . | . | :|.
245 IKQGCLLKQGHRRKNWKVRKFI LREDPAYLHYYDPAGAEDPLGAIHLRGC      294

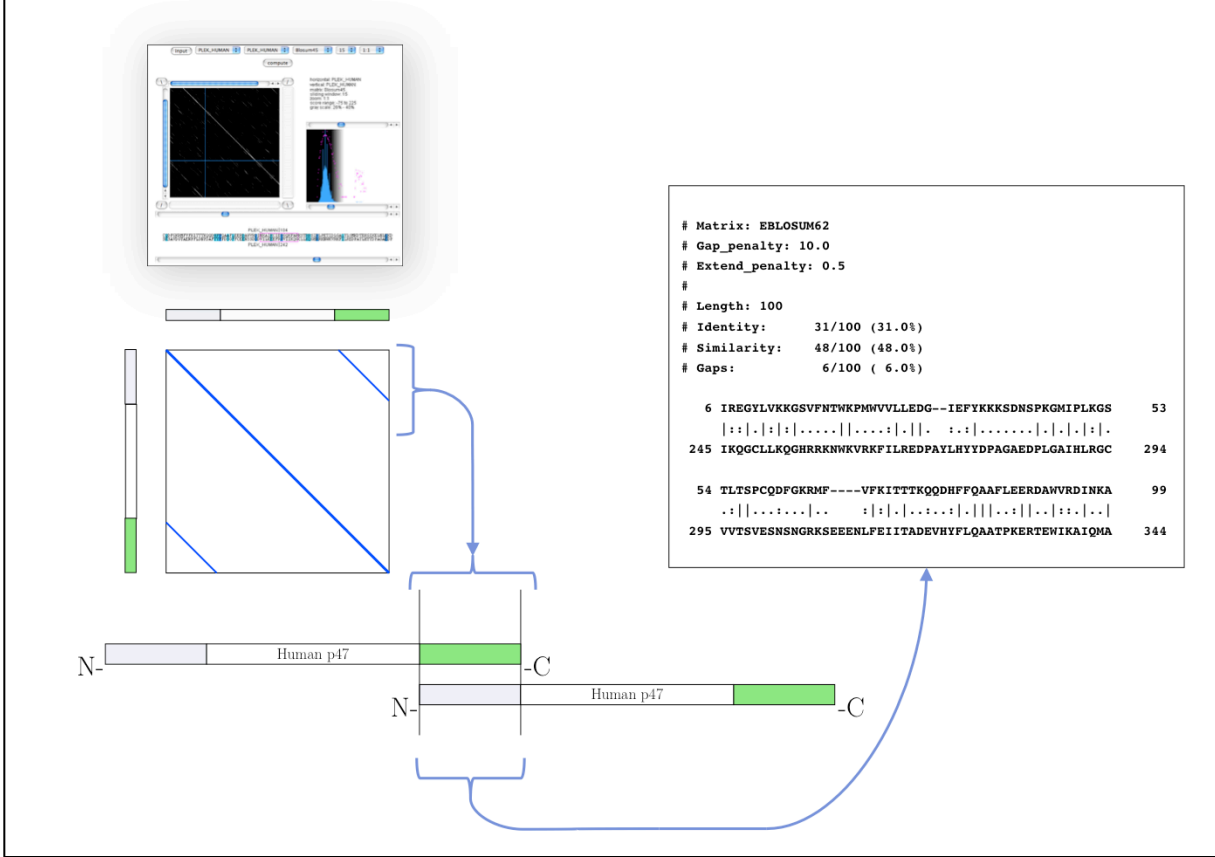
   54 TLTSPCQDFGKRMF-----VFKITTTKQDHFQAAFLEERDAWVRDINKA      99
   .:| | . . . : . . | . .   :|:|. | . . . . :|. | | | . . :| | . . : : . | . . |
295 VVTSVESNSNGRKSEEEENLFEIITADEVHYFLQAATPKERTEWIKAIQMA      344
```

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.

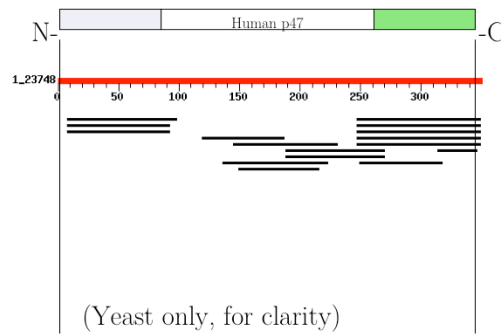
DOMAINS: UNITS OF INHERITANCE



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

SEQUENCE SEARCHES WITH DOMAINS CAN IMPROVE SPECIFICITY

BLAST search with *full-length* Pleckstrin sequence

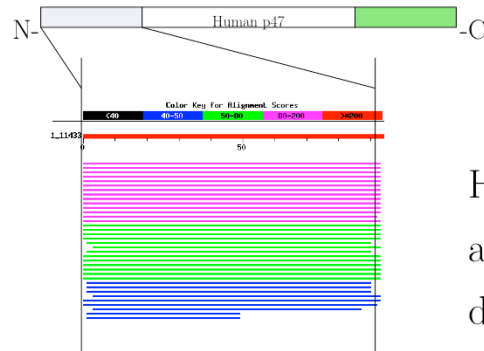


Hits extend over the entire domain. [PSI-BLAST](#) would be difficult ...

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: we are actually **adding information** by defining the domain boundaries, and the algorithm is not misled 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?

EMBL-EBI HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Protein: Q5KMQ9_CRYNJ9 (Q5KMQ9)

Summary

This is the summary of UniProt entry Q5KMQ9_CRYNJ9 (Q5KMQ9).

Description: Transcription factor, putative (EC0:000313|EMBL:AAW41783.1)

Source organism: [Cryotococcus neoformans var. neoformans serotype D \(strain JEC21 / ATCC MYA-565\) \(Fibrobacterales: neoformans\) \(NCBI taxonomy ID 214684-9\)](#)
[View Pfam proteome data.](#)

Length: 754 amino acids

Reference Proteome:

Please note: when we start each new Pfam data release, we take a copy of the UniProt sequence database. This snapshot of UniProt forms the basis of the overview that you see here. It is important to note that, although some UniProt entries may be removed after a Pfam release, these entries will not be removed from Pfam until the next Pfam data release.

Pfam domains

This image shows the arrangement of the Pfam domains that we found on this sequence. Clicking on a domain will take you to the page describing that Pfam entry. The table below gives the domain boundaries for each of the domains. [More...](#)

Download the data used to generate the domain graphic in JSON format.

Source	Domain	Start	End
disorder		3	5
disorder		62	75
disorder		113	286
low_complexity		117	134
low_complexity		157	166
low_complexity		190	204
low_complexity		279	290
Pfam	Ark_2	296	381
disorder		312	325
Pfam	Ark	447	479
disorder		474	475
disorder		479	482
disorder		491	494
disorder		498	508
disorder		510	512
disorder		551	552
disorder		580	592
collod_coll		594	618
disorder		595	596
disorder		598	599
disorder		634	660
low_complexity		634	644
disorder		664	667
low_complexity		665	680
disorder		723	754

Show or [hide](#) domain scores.

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.

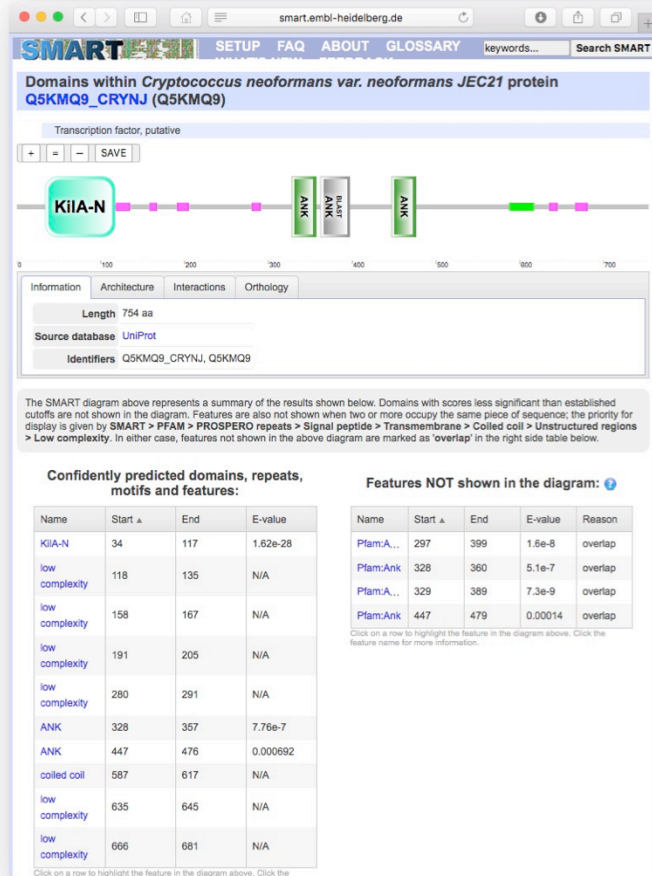
SEQUENCE BASED
DOMAIN ANNOTATION:

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

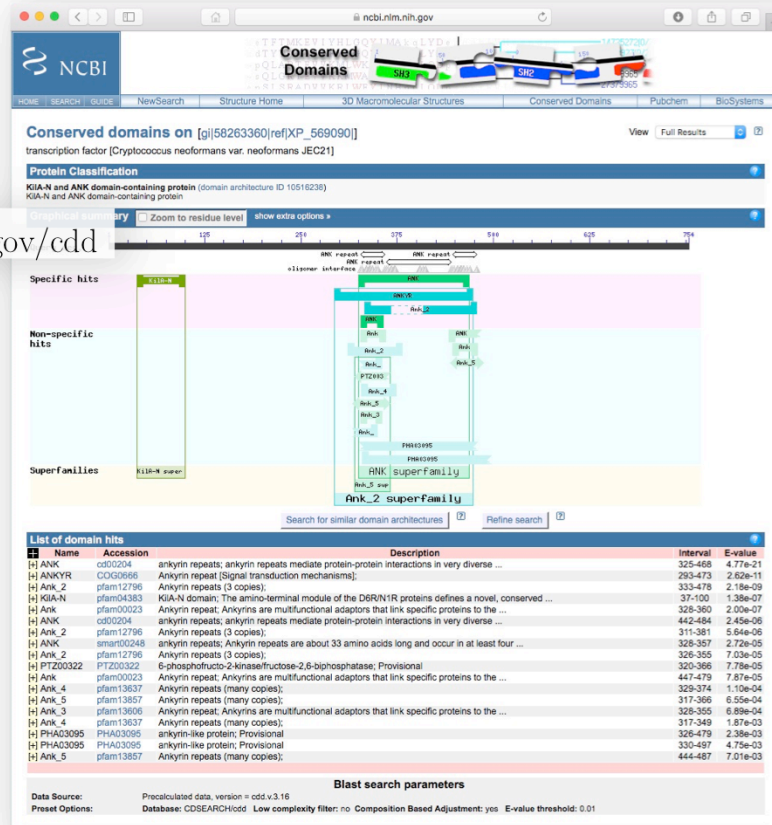
CDD

<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 query sequence.

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

There is a related utility, the Batch search 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).

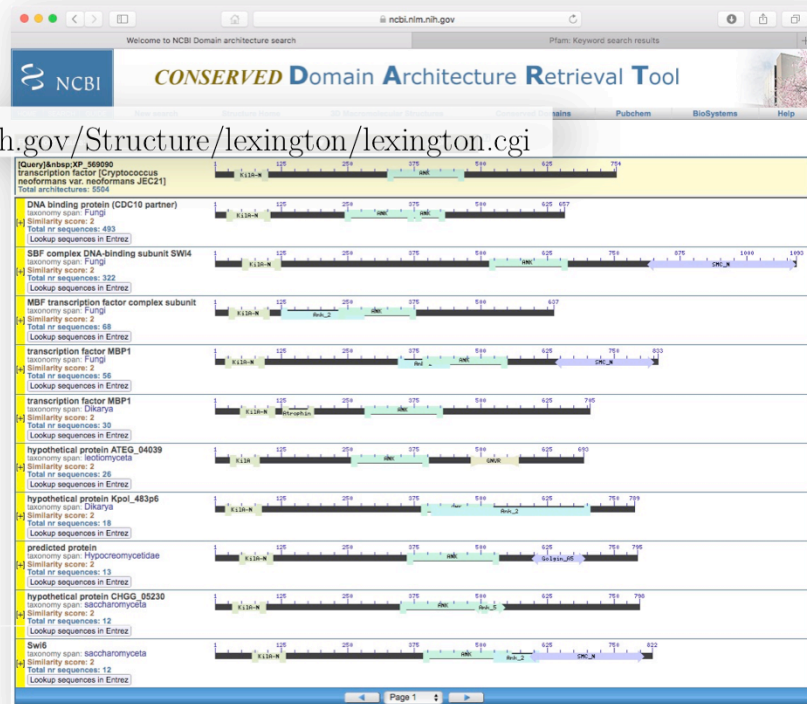
SEQUENCE BASED DOMAIN ANNOTATION: CDART DATABASE

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 present – 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

SEQUENCE BASED DOMAIN ANNOTATION: EXPASY DATABASE

ExPASy/ProSite

[http://www....](http://www...)

Ref ...

ExPASy...

SEQUENCE BASED DOMAIN ANNOTATION: RPS-BLAST

RPS-BLAST

[http://www....](http://www...)

Ref ...

RPS-BLAST...