6.
Protein classifications

Common Motifs

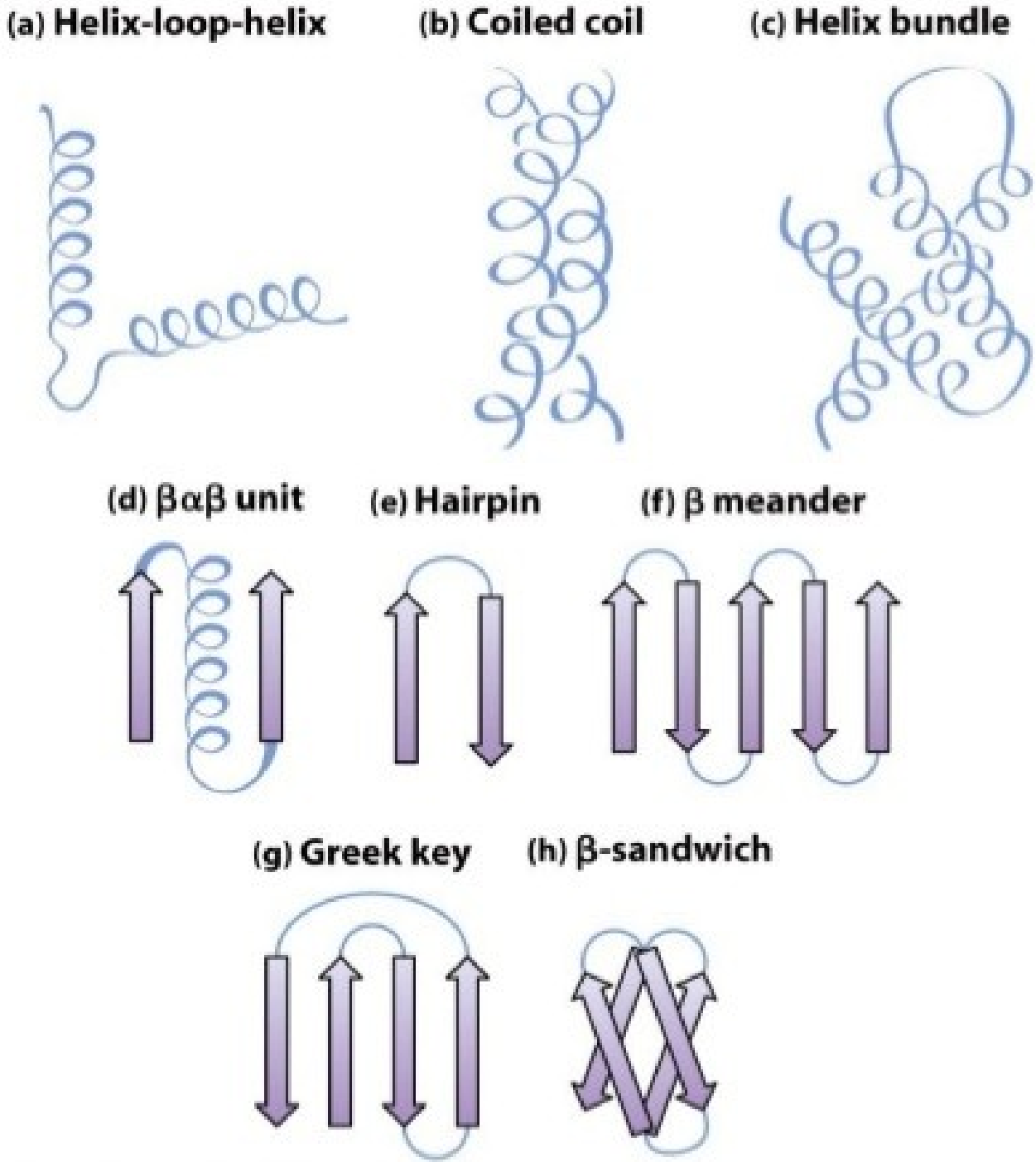
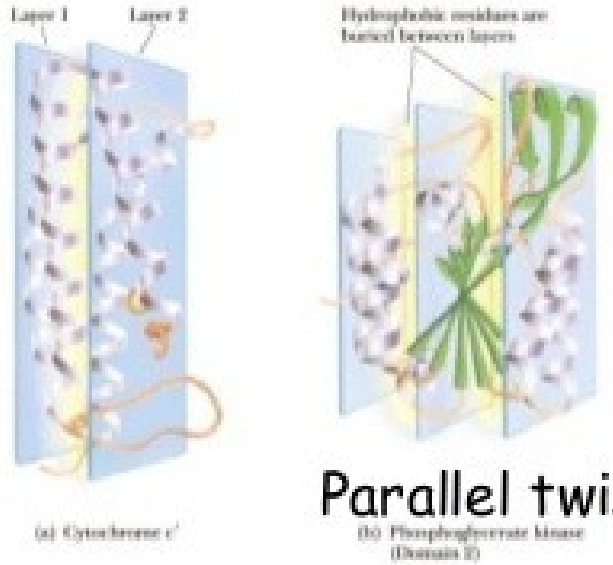


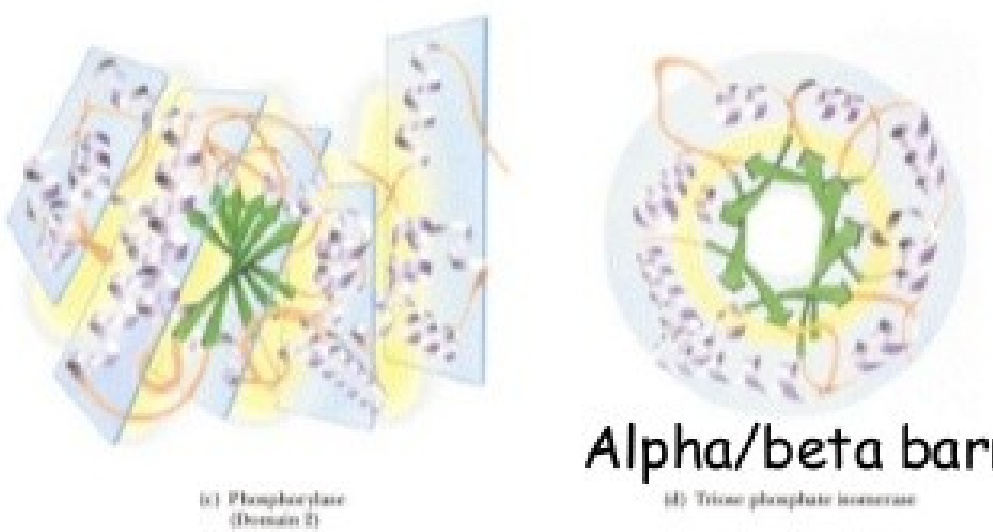
Figure 4-19 Principles of Biochemistry, 4/e

Motifs Combine to form Domains



Parallel twisted sheet

- Domains are independent folding units in a 3° structure of a protein
- Individual domains have specific function

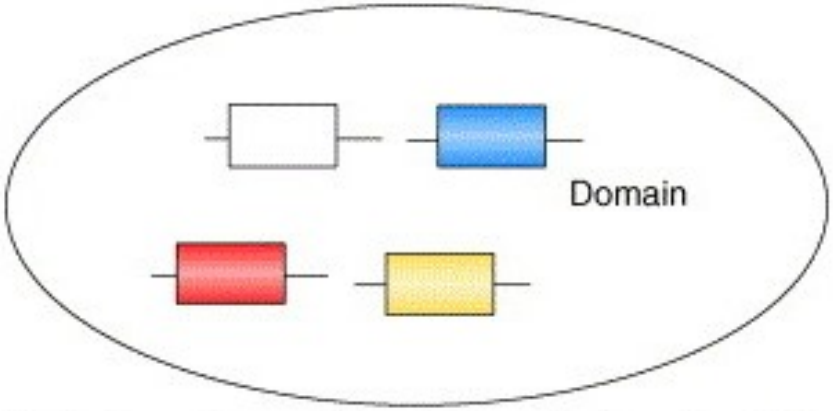


Alpha/beta barrel

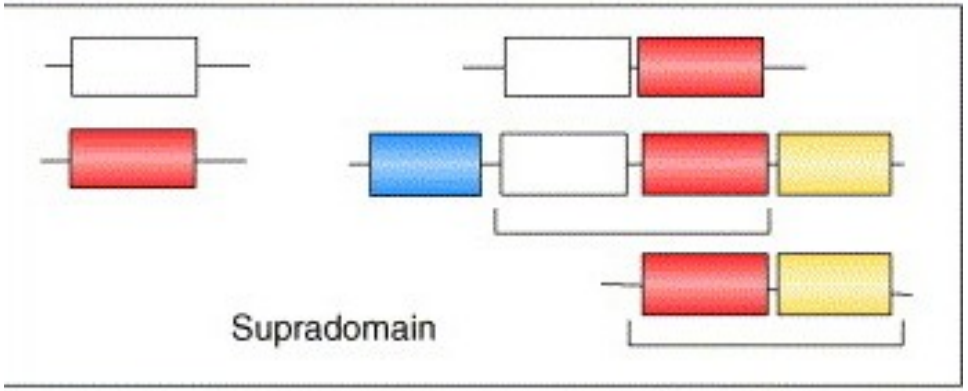
- Hydrophobic interactions are the major driving force in folding domains

Domains are reused and reinvented

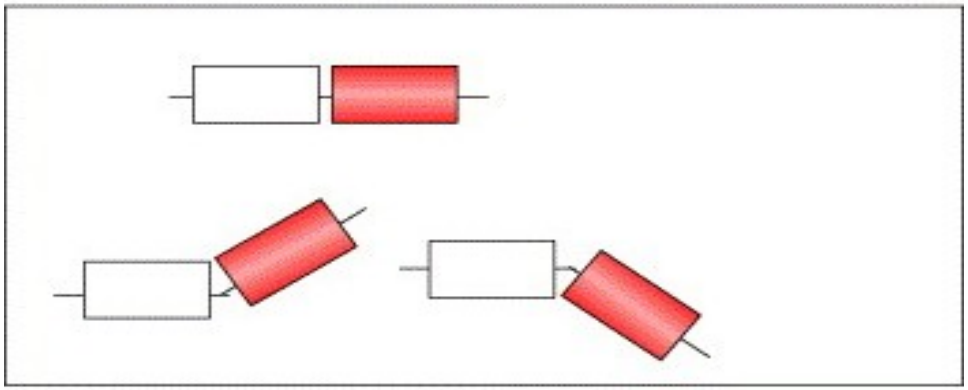
The repertoire of domain superfamilies...



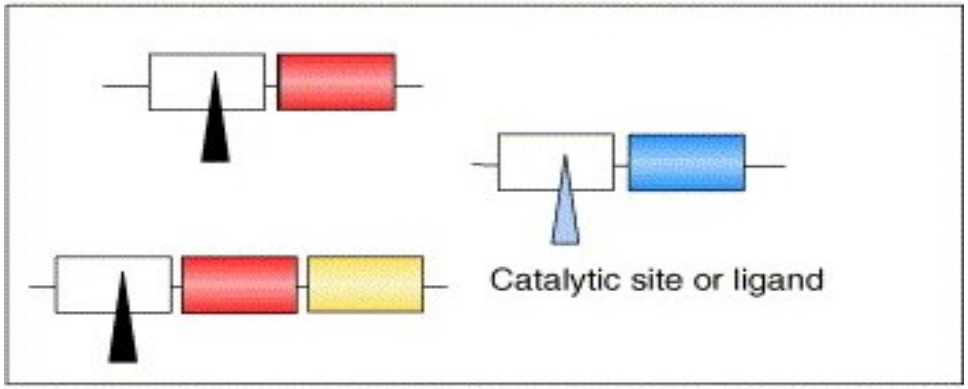
...duplicates and recombines to form single and multi-domain proteins.



The same combination can adopt different geometries...



...and/or different functions.



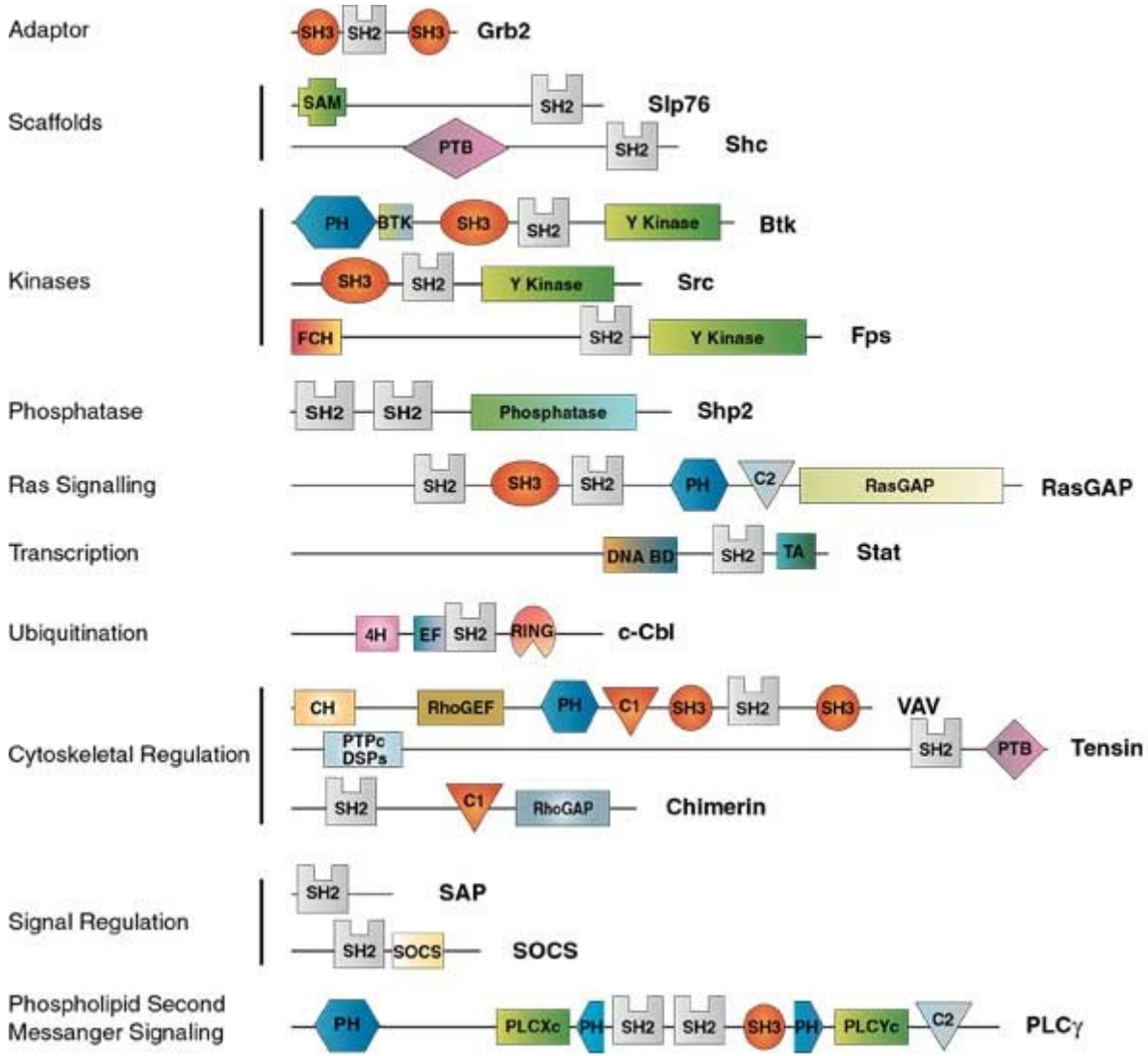
SH2 domain

Involved in PPIs

Involved in signal transduction

Binds phosphorylated Tyr residues

SH2 domain

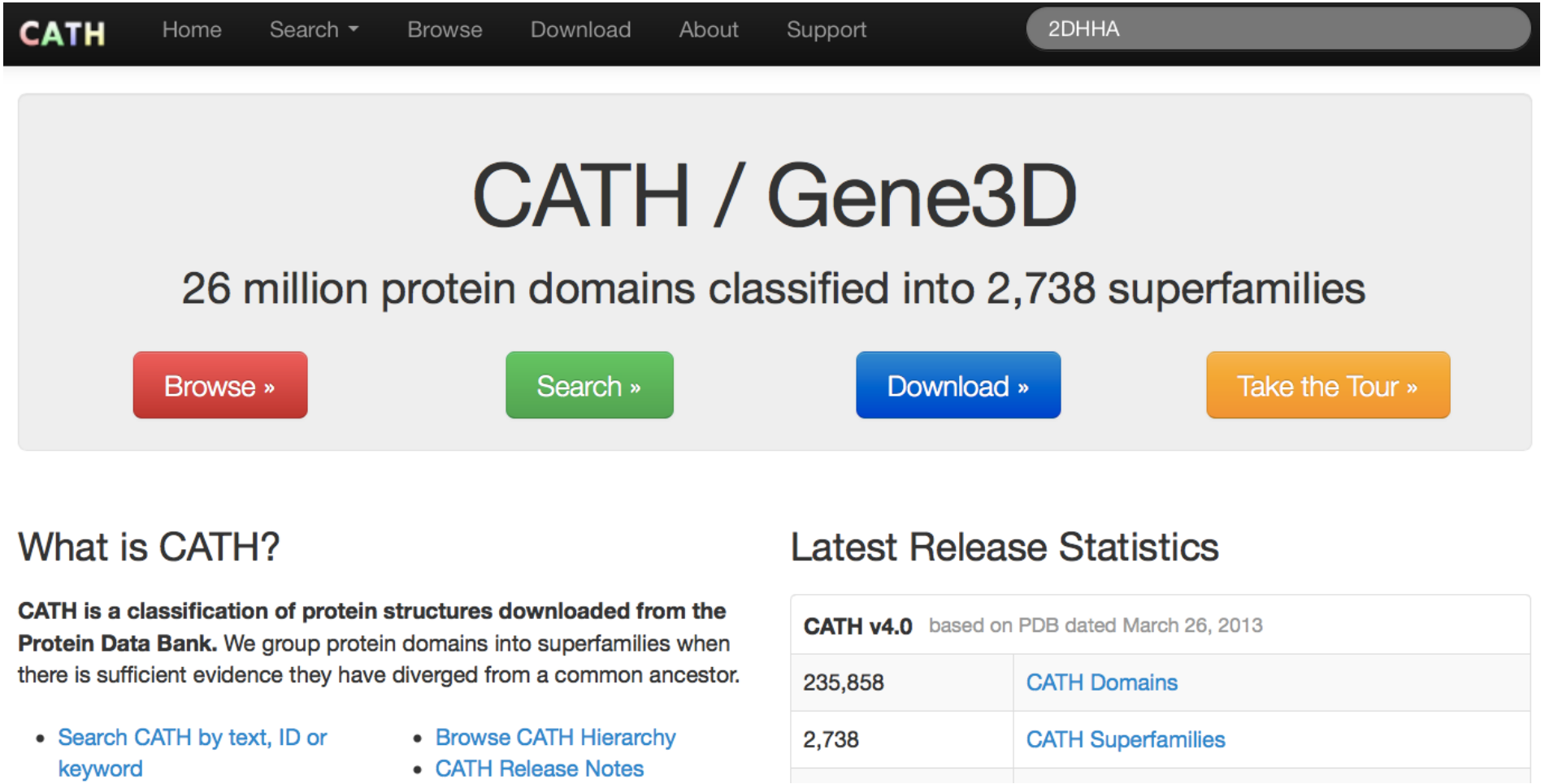

Adaptor
SH3
SH2
SH3
Grb2
Scaffolds
SAM
SH2
Slp76
PTB
SH2
Shc
Kinases
PH
BTK
SH3
SH2
Y Kinase
Btk
SH3
SH2
Y Kinase
Src
FCH
SH2
Y Kinase
Fps
Phosphatase
SH2
SH2
Phosphatase
Shp2
Ras Signalling
SH2
SH3
SH2
PH
C2
RasGAP
RasGAP
Transcription
DNA BD
SH2
TA
Stat
Ubiquitination
4H
EF
SH2
RING
c-Cbl
Cytoskeletal Regulation
CH
RhoGEF
PH
C1
SH3
SH2
SH3
VAV
PTPc
DSPs
SH2
PTB
Tensin
SH2
C1
RhoGAP
Chimerin
Signal Regulation
SH2
SAP
SH2
SOCS
SOCS
Phospholipid Second
Messanger Signaling
PH
PLCXc
PH
SH2
SH2
SH3
PH
PLCYc
C2
PLCγ

Protein classification

- Structural domains
- Protein families
- Sequence signatures

CATH

manually-curated hierarchical classification of protein domain structures.



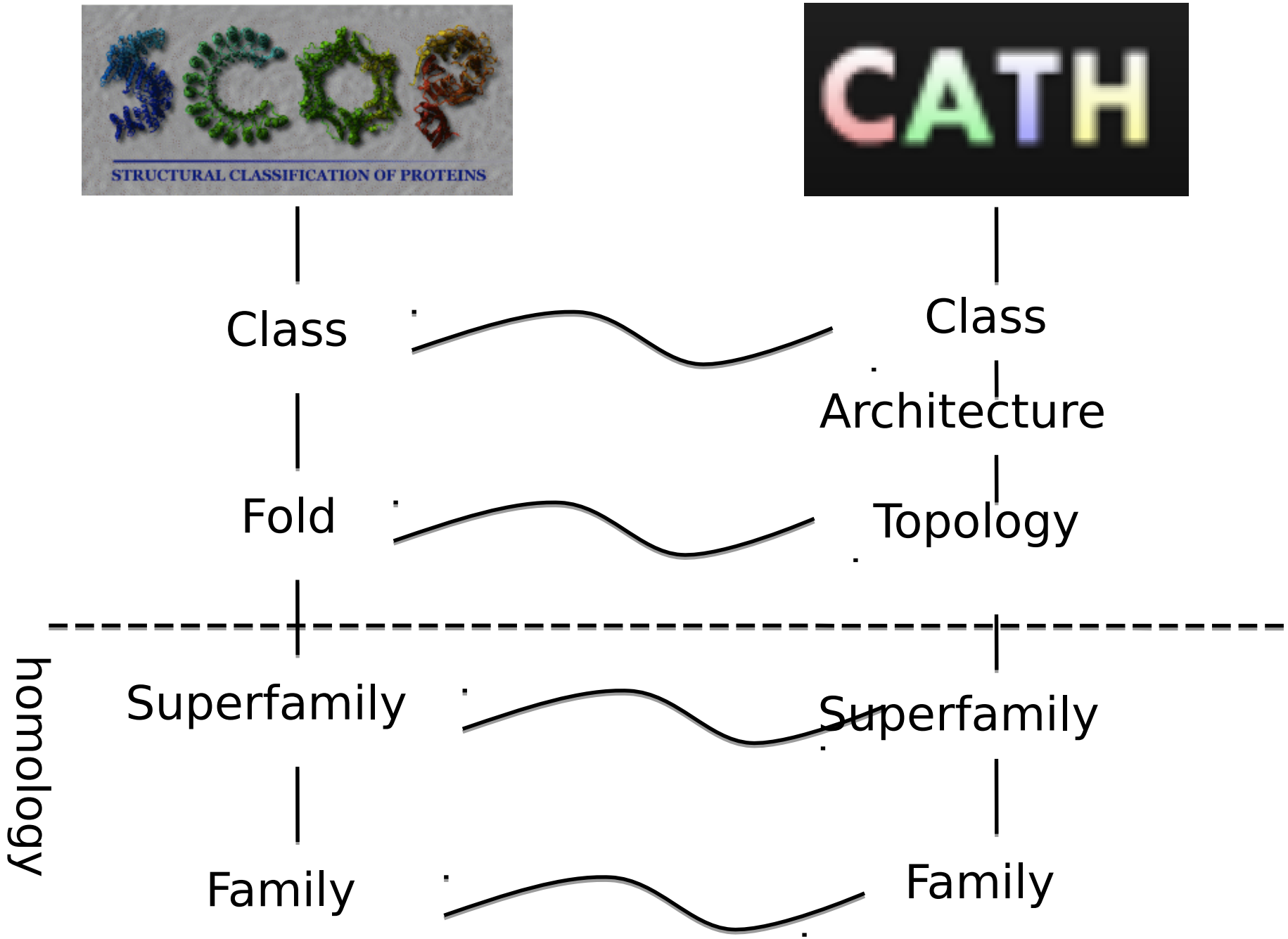
http://www.cathdb.info

Class: secondary structure content (e.g. mainly-alpha, mainly-beta, mixed alpha/beta or 'few secondary structures');

Architecture: general arrangement of the secondary structures

Topology (fold) takes into account the connectivity of secondary structures in the chain;

Homologous Superfamily: domains that are believed to be related by a common ancestor.


SCOP
STRUCTURAL CLASSIFICATION OF PROTEINS
CATH
Class
Class
Architecture
Fold
Topology
homology
Superfamily
Superfamily
Family
Family

Protein sequence families

Protein families

- Defined based on MSA
- Identification of functional amino acids
- Diversity -> detecting remote homologues
- Identification of parts of the sequence space which have no functional annotation

Pfam database

Pfam: Family: Avidin (PF01382) - Mozilla Firefox

Pfam: Family: Avidin (P...

pfam.xfam.org/family/PF01382

Most Visited | Google | pubmed | szotarak | Disorder | ANCHOR | UCSC Genome | PDBsum home page | Printer

EMBL-EBI

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam

keyword search Go

Family: *Avidin* (PF01382)

18 architectures | 142 sequences | 1 interaction | 81 species | 454 structures

- Summary
- Domain organisation
- Clan
- Alignments
- HMM logo
- Trees
- Curation & model
- Species
- Interactions
- Structures

Jump to...

enter ID/acc Go

Summary: Avidin family

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

Wikipedia: Avidin | Pfam | InterPro

This tab holds the annotation information that is stored in the Pfam database. As we move to using Wikipedia as our main source of annotation, the contents of this tab will be gradually replaced by the Wikipedia tab.

Avidin family

Provide feedback

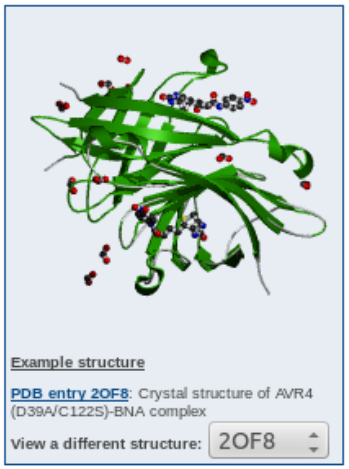
No Pfam abstract.

Internal database links

SCOOP:	DUF1143

External database links

HOMSTRAD:	avi
PANDIT:	PF01382
PROSITE:	PDOC00499
Pseudofam:	PF01382
SCOP:	1slf
SYSTERS:	Avidin



Example structure

PDB entry 2OF8: Crystal structure of AVR4 (D39A/C122S)-BNA complex

View a different structure: 2OF8

Comments or questions on the site? Send a mail to pfam-help@ebi.ac.uk.

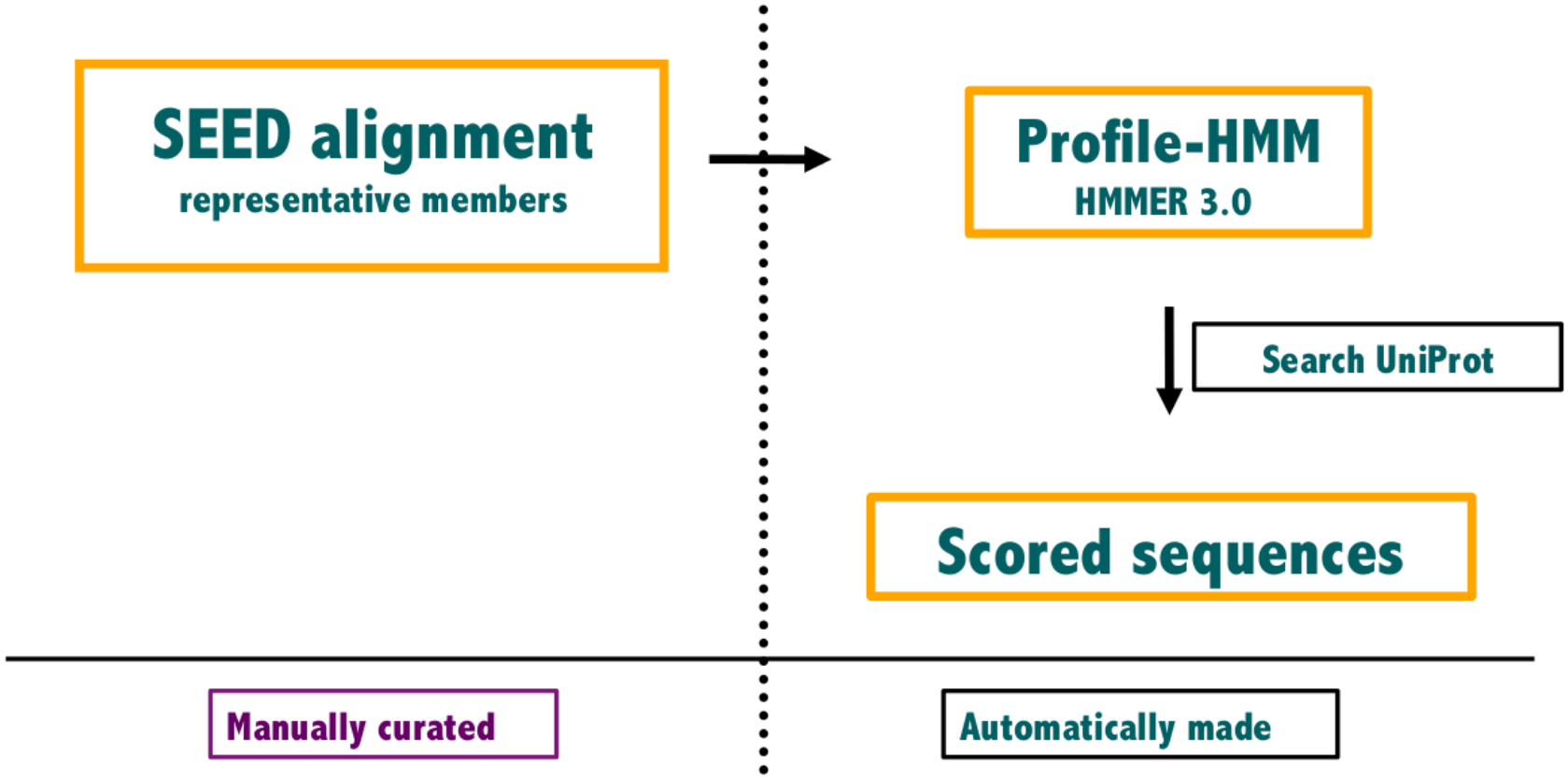
European Molecular Biology Laboratory

Pfam sequence families

family a group of evolutionary related proteins and/or protein regions

```
A6VZD9_MARMS/3-149      NY...WLMK......SEPDA.........FSIDDL...KRLKH..SPWDGVRNYQARN..FMK.EMNEGDLVFFYHSS.
THYN1_HUMAN/55-219      SH...WLMK......SEPESRLEKGVDVKFSIEDLK.AQPKQT..TCWDGVRNYQARN..FLR.AMKLGEEAFFYHSN.
A3ZLP5_9PLAN/16-165     RY...WLLK......TEPES.........YSIDDLA.NEKKQT..TFWSGVRNYQARN..FMRDDMKVGDEVFFYHSN.
Q312I8_DESDG/2-150      RY...WLMK......SEPEC.........FSLEDLV.NAPEQT..TPWDGVRNYQARN..FMRDEMRPGDKVLFYHSG.
Q1MR11_LAWIP/2-150      QY...WLFK......SDTDC.........YSINDLQ.SAPNQT..TSWDGVRNYQARN..FMRDEMRIGDLGFFYHSG.
Q729D5_DESVH/2-149      NY...WLFK......SETDC.........FSVDDLA.ASPDAT..SSWDGVRNYQARN..FMR.TMRKGDLGFFYHSG.
Q3A4Z6_PELCD/6-155      RY...WLMK......SEPGC.........FSIDDLK.DCPDGI..SPWDGVRNYQARN..LLRDEIKAGDGVLFYHSN.
Q74AS6_GEOSL/6-154      RY...WLFK......SEPSC.........FSFDDLG.SRPNGT..EHWDGVRNFQARN..LLRDEIKPGDGVLFYHSN.
B3E7D4_GEOLS/2-150      RY...WLFK......SEPGC.........FSFQDLQ.ARPNAT..EQWDGVRNFQARN..FLRDEIKPGDRVLFYHSS.
A1ATY5_PELPD/2-150      NY...WLFK......TEPGC.........FSFDNLK.NRPNMT..EPWDGVRNFQARN..YLRDTVKVGDLVLFYHSN.
Q054Y1_LEPBL/2-151      KH...WLFK......TEPDV.........FSIDDLY.KAPSRI..APWEGVRNYQARN..FLRDSIQKGDLILFYHSR.
Q5KKF6_CRYNE/7-162      EF...WQTA......NG......................KFCRV..SPWDGVRNHEAKK..IMREKMKLGDKVLFYHSN.
Q7UMR5_RHOBA/2-156      KY...WLMK......TEPNT.........FSIDDLA.EQPEQI..TCWEGVRNYQARN..LLRDEIEEGDQVLFYHSA.
Q60BW0_METCA/2-151      RY...WLMK......TEPGE.........FGIDDLA.ARPAQT..EPWDGVRNYQARN..MMRDEMKVGDGVLFYHSN.
A4BS86_9GAMM/2-151      SY...WVMK......SEPSV.........YGIDDLA.AQPSQT..DHWEGVRNYQARN..MLRDQMRPGDLALLYHSN.
A1WW93_HALHL/2-151      NR...WVMK......SEPDV.........FGIDDLA.AAPQGT..DRWDGVRNYQVRN..MIRDHMRPGDAAFFYHSN.
Q1K2S0_DESAC/2-151      NY...WLMK......SEPEA.........FGIDDLQ.QMPEQT..EHWDGVRNYQARN..MMRDDMKIGDLAFFYHSN.
Q0A5L0_ALHEH/2-151      NY...WLMK......SEPDE.........FGIEDLK.QRPDQI..EPWDGVRNYQARN..MMRDQMKVGDLAFFYHSN.
Q83BN8_COXBU/2-151      NY...WLLK......SEPTS.........YSIDDLF.REKNKI..TRWDGVRNYQARN..FMRDGMKKGDLAFFYHSN.
Q7S515_NEUCR/116-280    QY...WLLK......AEPLPRLENGYDVHFSIDDL..AARTSP..EPWDGIRNYSARN..NLR.SMRVGDLAFFYHSN.
A6SM74_BOTFB/112-267    QY...WLMK......AEPESRIEKGHDIKFSIDDL..AAKTEP..EPWD......ARN..NLR.AMKKGDLAFFYHSS.
Q0CLU5_ASPTN/109-262    SY...WLMK......AEPESRIEKGVDVKFSIDDL..RERTKP..EPWD......ARN..NMR.EMKKGDYAFFYHSN.
Q0U623_PHANO/118-277    VF...WLLK......AEPLPRYENGVNVAFSIDDL..AACTVP..EPWGGVRNPQARN..NMQ.AMRKGDLGFFYHSN.
```

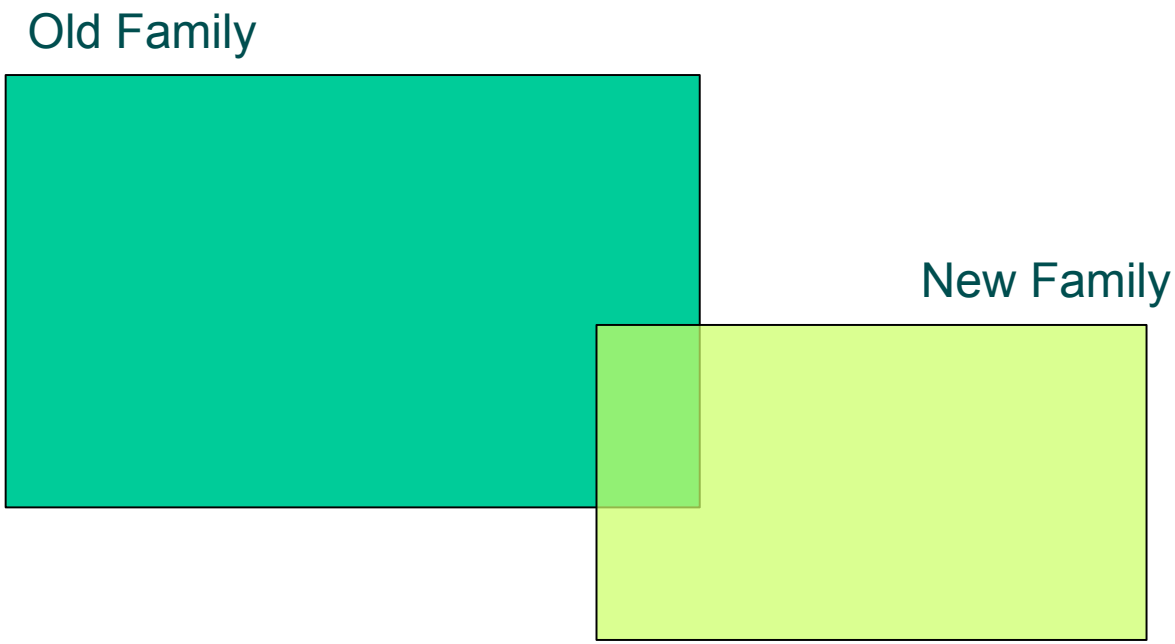
PFAM generálása



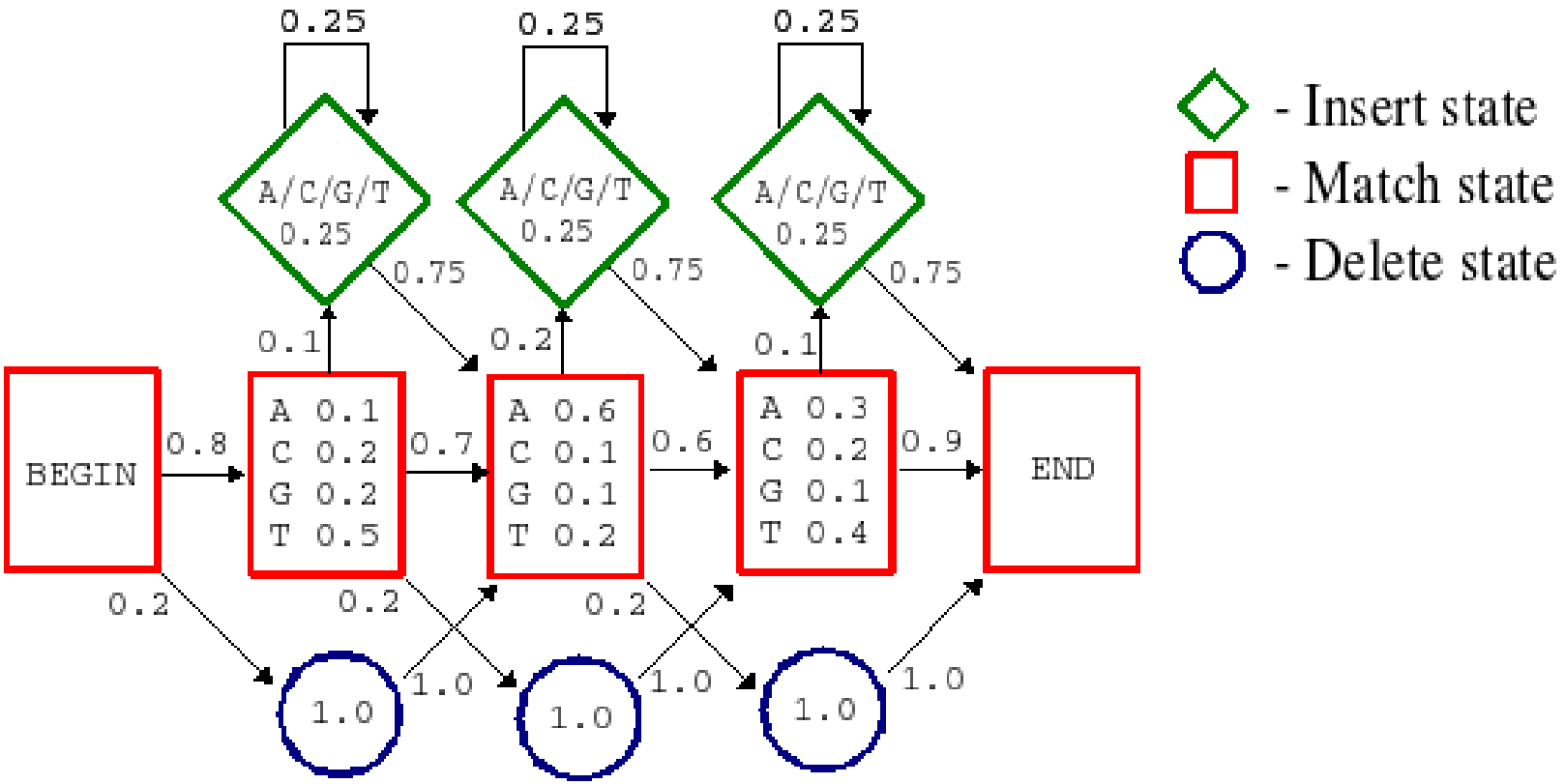
A description of the family, includes thresholds you to create the full alignment

<u>Rules</u> – No false positives. A family is not allowed to overlap with any other family

Family overlaps

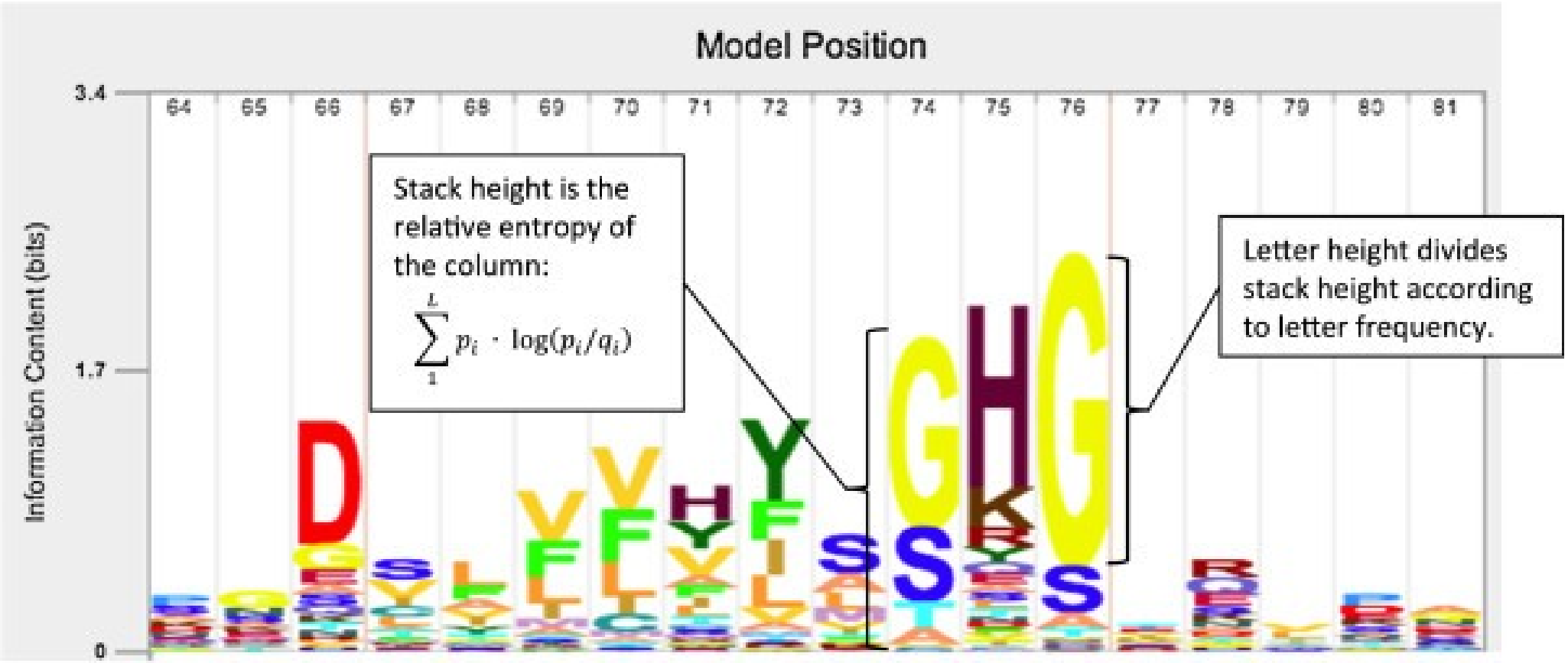


Profile Hidden Markov Models - Encapsulate diversity



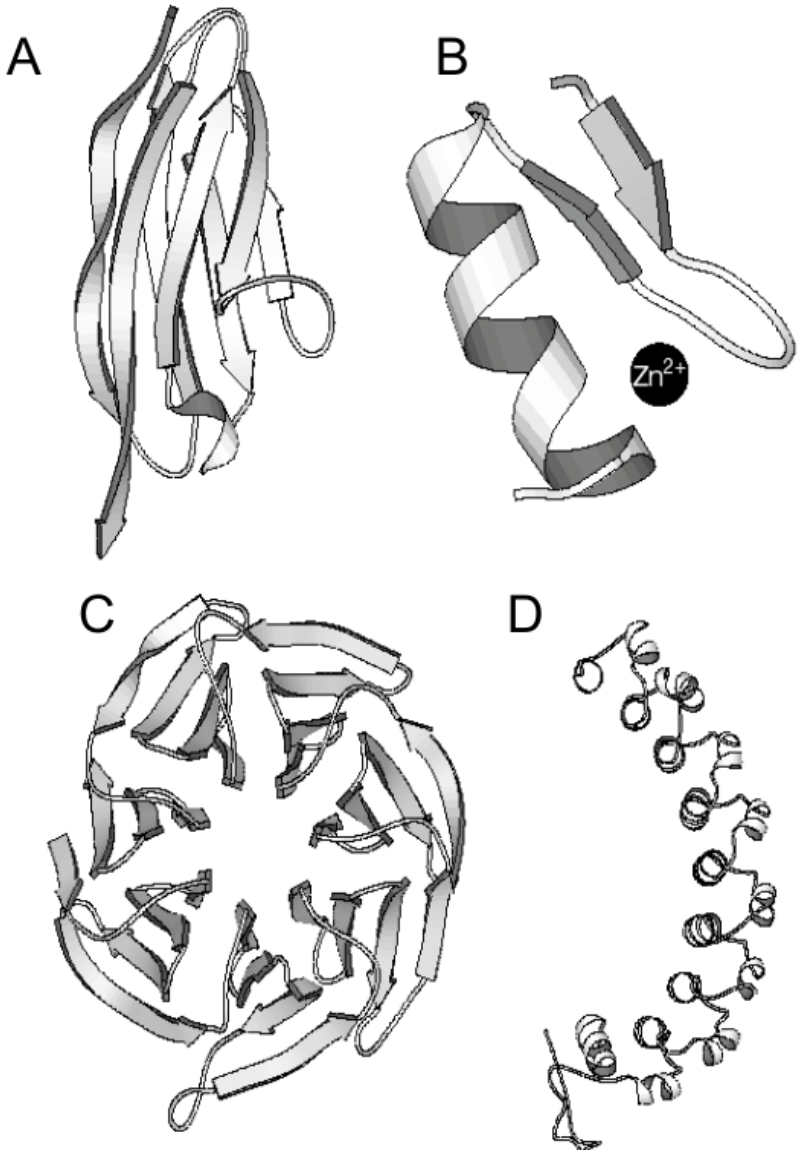
Hmmer

Logos

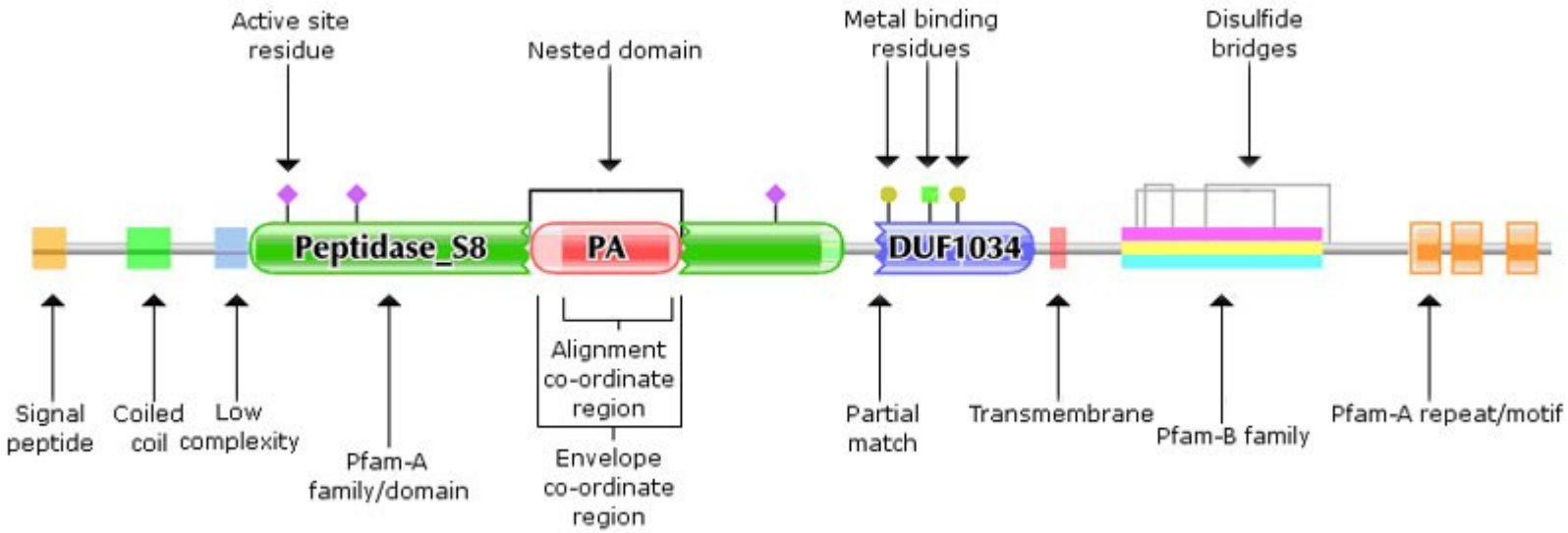

Model Position
Information Content (bits)
3.4
1.7
0
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
Stack height is the relative entropy of the column:
$\sum_{1}^{L} p_i \cdot \log(p_i/q_i)$
Letter height divides stack height according to letter frequency.

Types of PFAM families

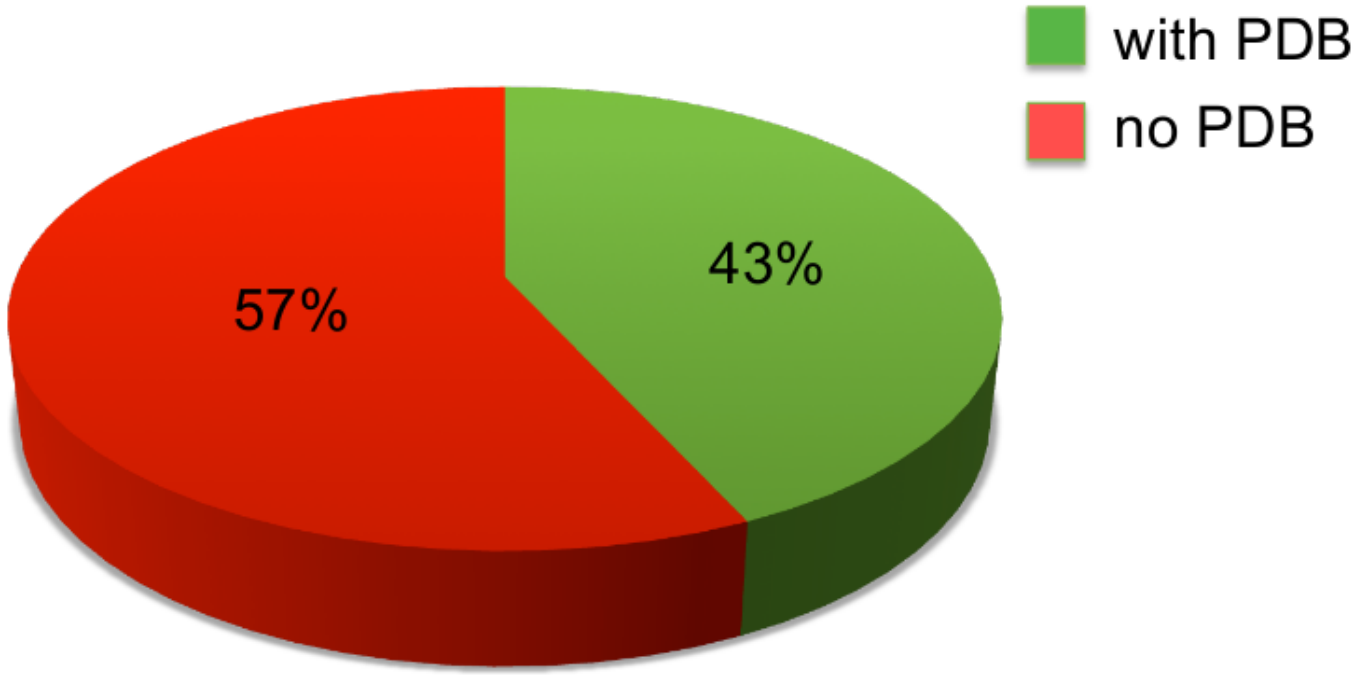
- **PFAM**
  - Domains –
    structural information
  - Families
  - Repeats
  - Motifs



Pfam pictograms

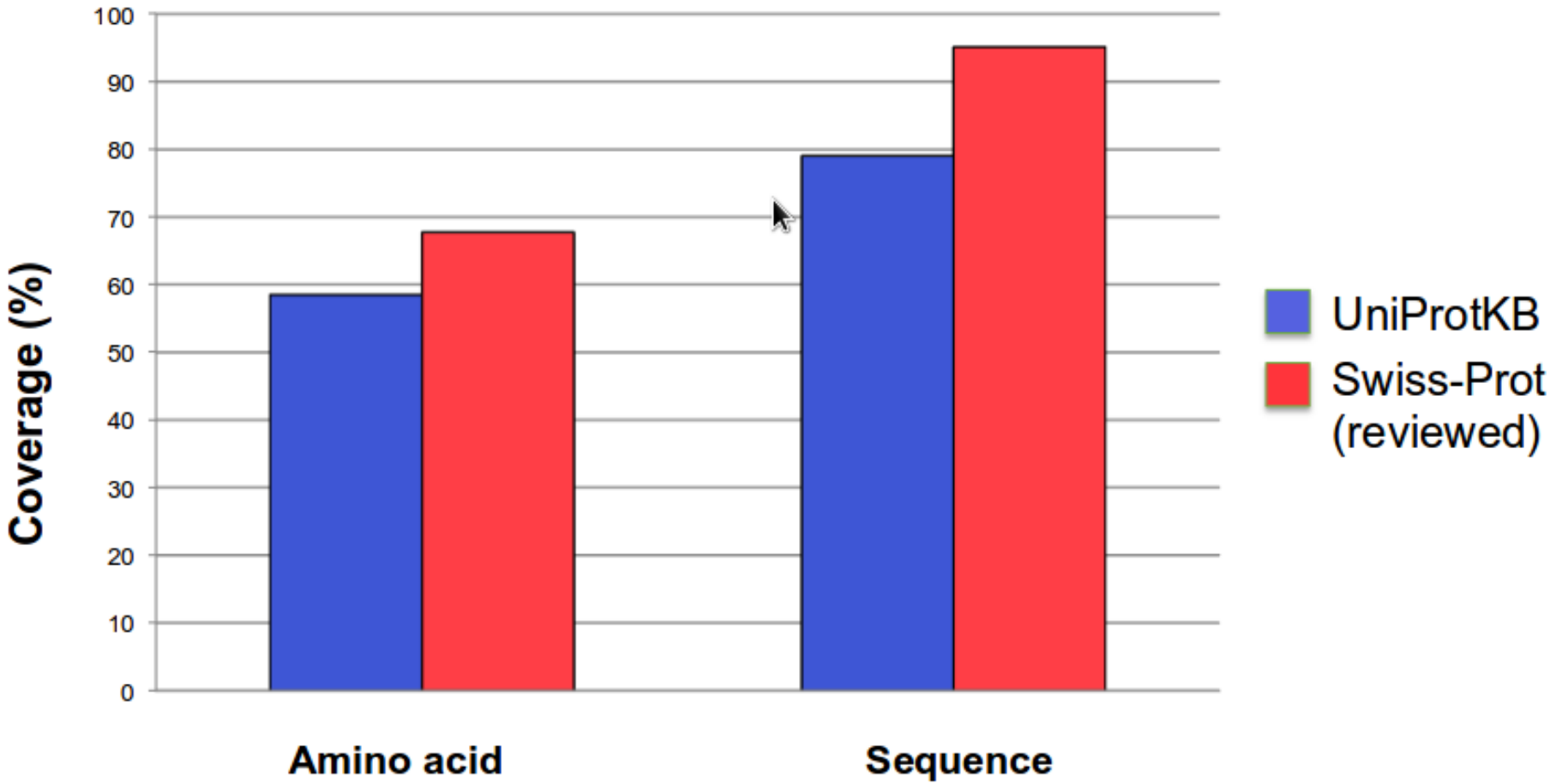


Pfam families with PDB structure

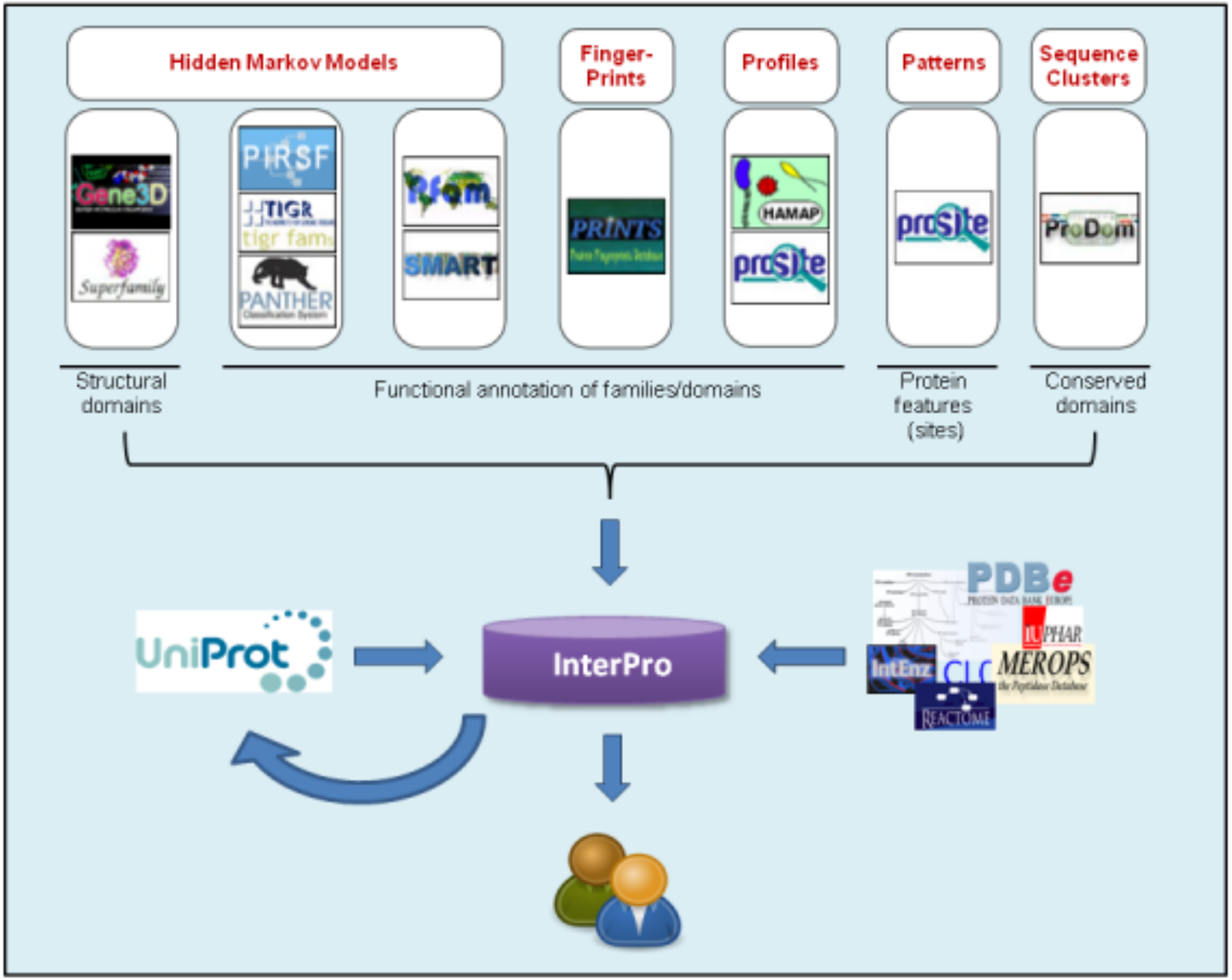


100%=all Pfam families

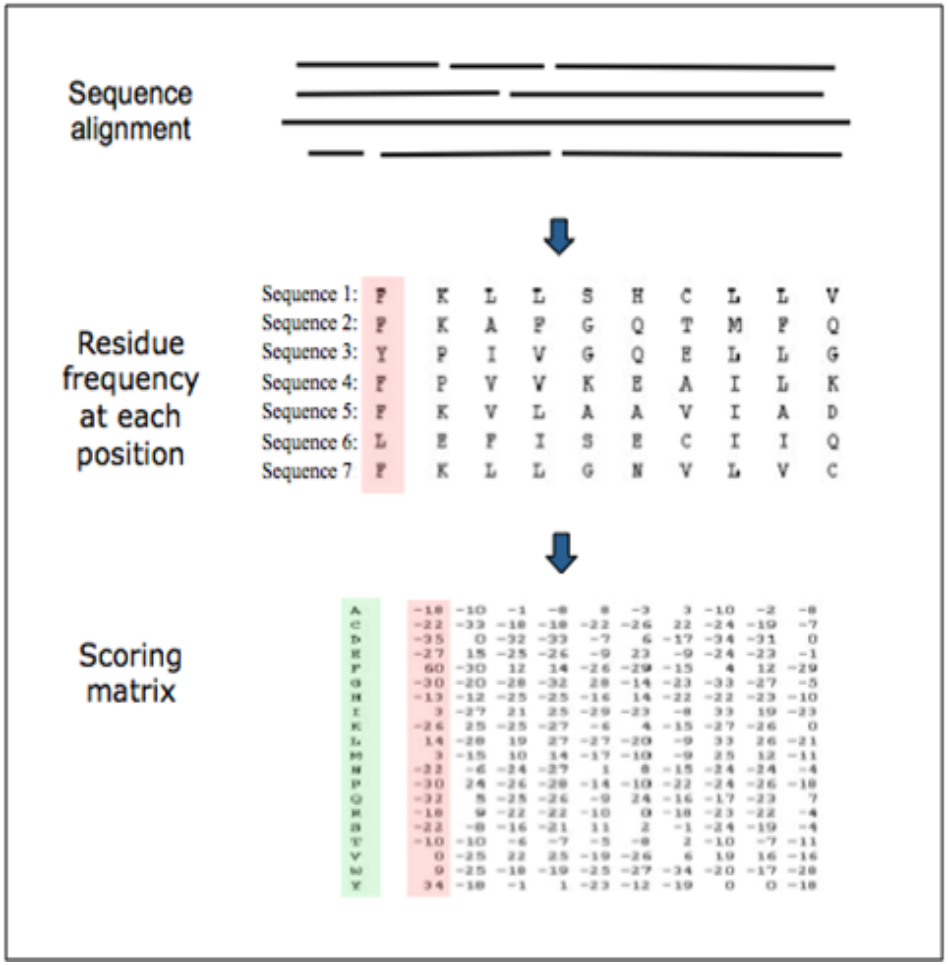
Pfam covarege



Interpro


Hidden Markov Models
Finger-Prints
Profiles
Patterns
Sequence Clusters
Gene3D
Superfamily
PIRSF
TIGR
tigr fams
PANTHER
Pfam
SMART
PRINTS
HAMAP
proSite
proSite
ProDom
Structural domains
Functional annotation of families/domains
Protein features (sites)
Conserved domains
UniProt
InterPro
PDBe
IUPHAR
IntEnz
CLC
MEROPS
REACTOME

Profiles

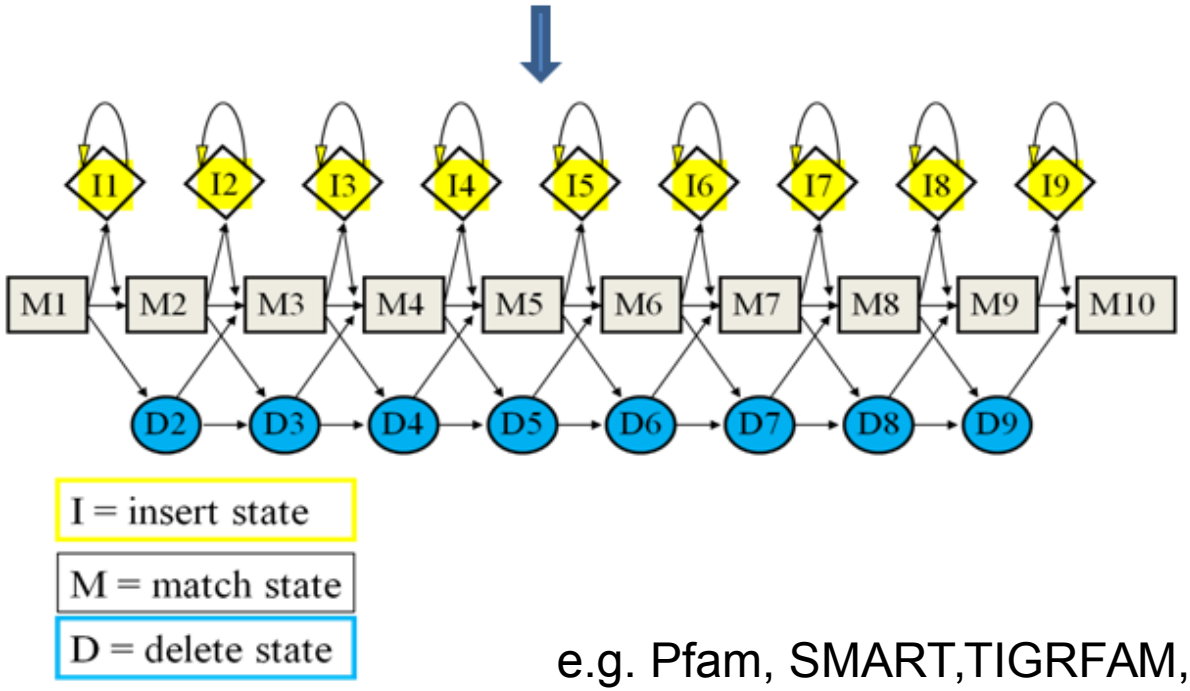


e.g. HAMAP, PROSITE,PRODOM

HMM

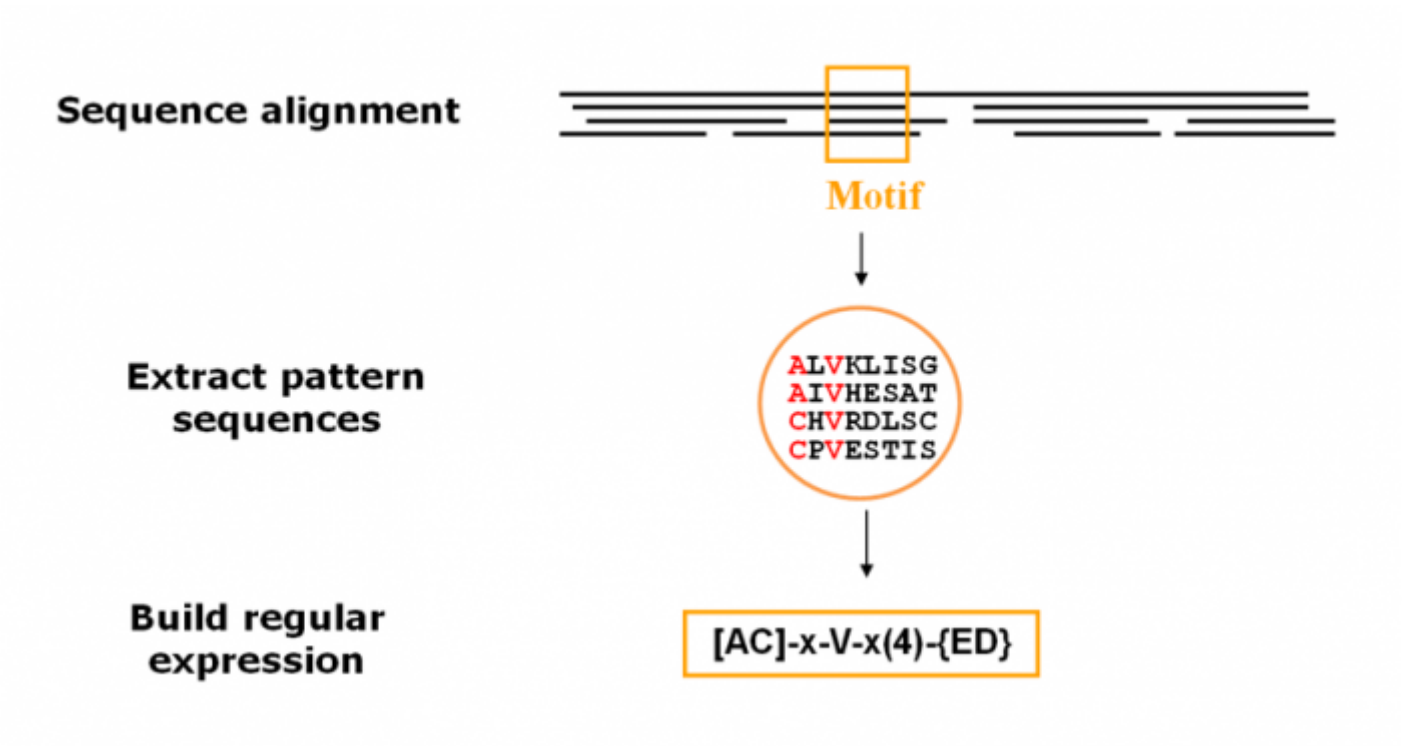
Multiple sequence alignment

Sequence 1:	F	K	L	L	S	H	C	L	L	V
Sequence 2:	F	K	A	F	G	Q	T	M	F	Q
Sequence 3:	Y	P	I	V	G	Q	E	L	L	G
Sequence 4:	F	P	V	V	K	E	A	I	L	K
Sequence 5:	F	K	V	L	A	A	V	I	A	D
Sequence 6:	L	E	F	I	S	E	C	I	I	Q
Sequence 7:	F	K	L	L	G	N	V	L	V	C



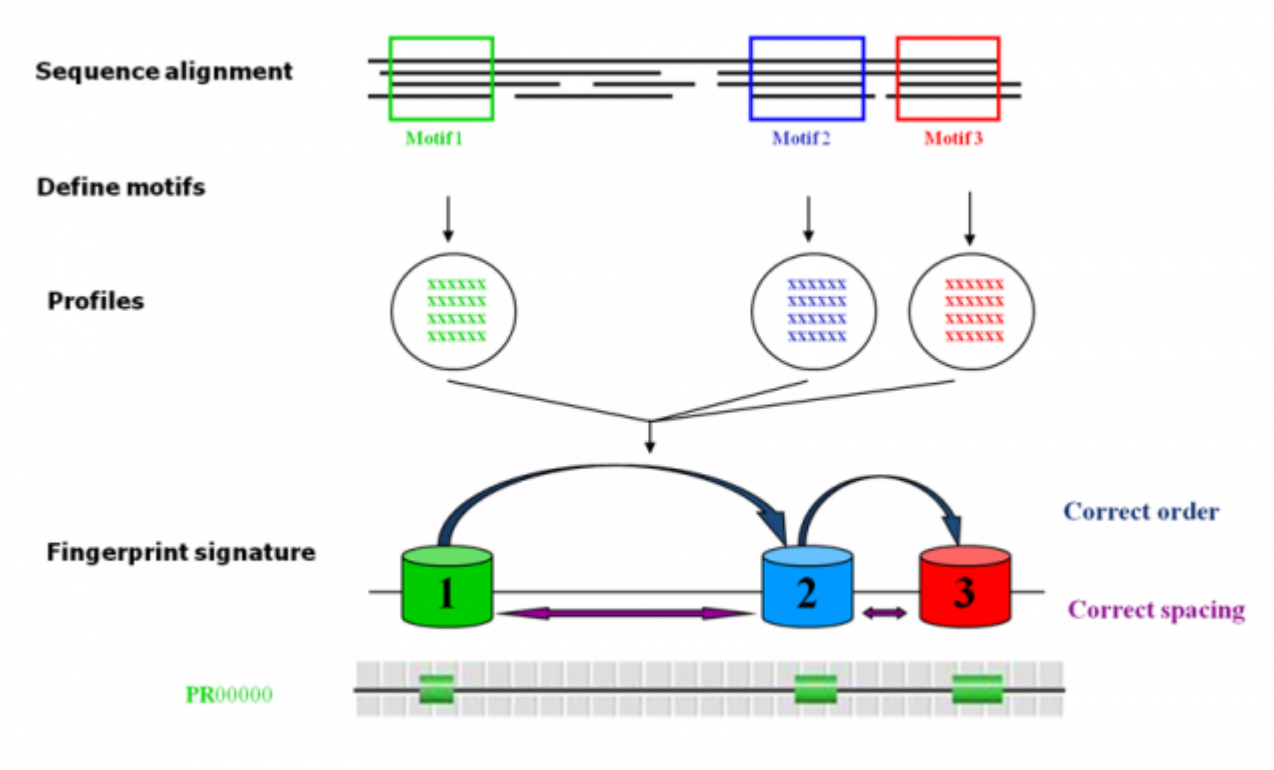
e.g. Pfam, SMART,TIGRFAM,PANTHER..

Patterns



e.g. PROSITE

Fingerprint



e.g. PRINTS

Domain architectures