Protein Motifs, Domains and Families

Manuel J. Gómez

Centro de Astrobiologia

TWO CONCLUSSIONS CAN BE DERIVED FROM THE COMPARISON OF PROTEIN SEQUENCES

First

Proteins can be grouped into clusters, or families, on the basis of their sequence similarity.

Second

Sequence similarity may be detectable only in blocks of a multiple sequence alignment, what indicates that conservation is restricted to modules that are important from a functionally or structural point of view.

CONSERVED MOTIFS AND DOMAINS

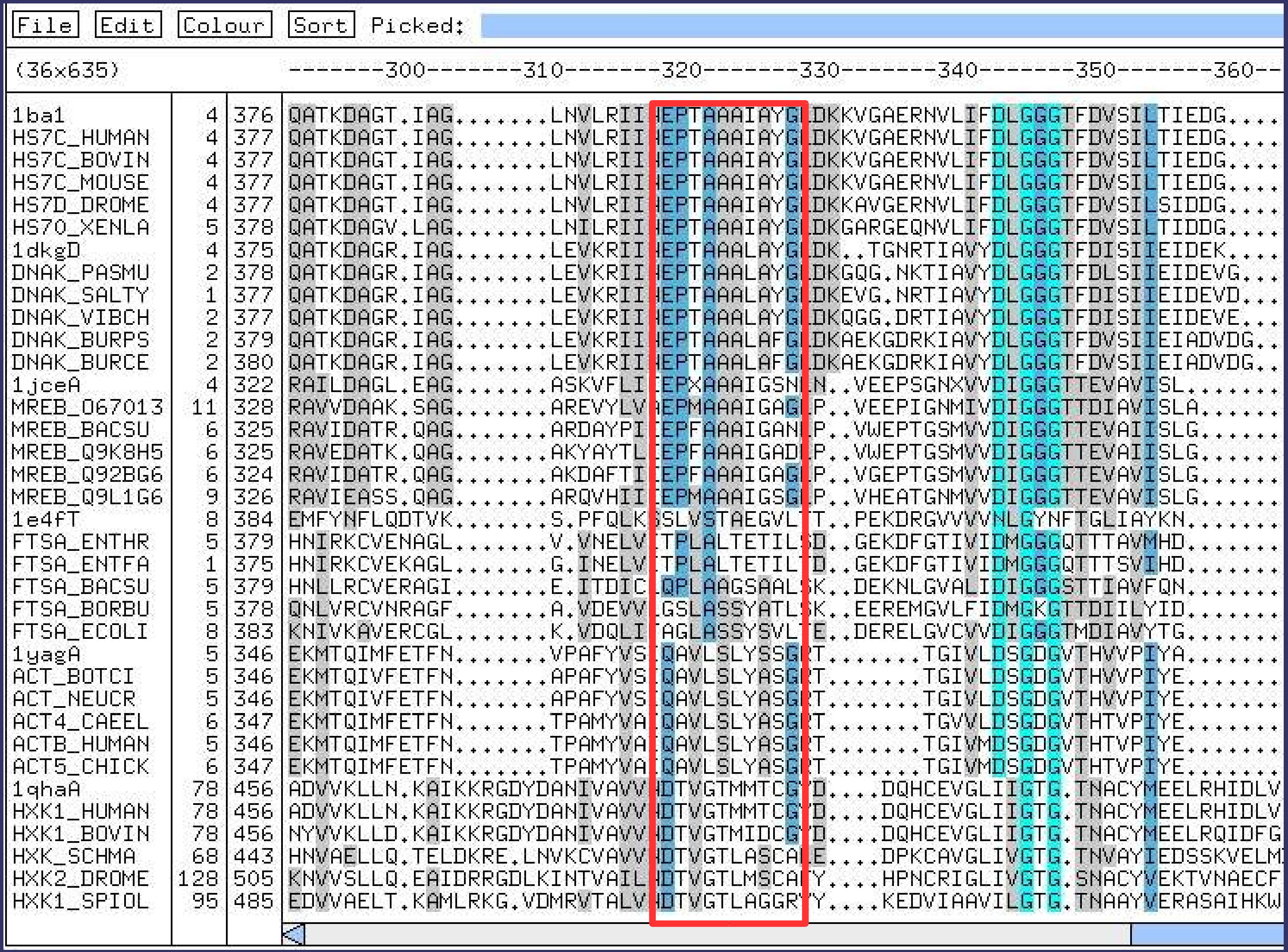
Conserved protein sub-sequences are often classified as:

- Motifs: short conserved sub-sequences that usually correspond to functional sites (active sites, binding sites, interaction sites). They may be part of bigger domains.

- Domains: stretches of secuence that appear as conserved modules in proteins that are not related, in global terms. They usually correspond to domains that can be defined using structural and/or functional definitions. Their average size is aprroaximately 100-150 aa. Domain shuffling is a mechanism of protein evolution, in some cases related with intron-exon architecture.

- Repeats: structurally or functionally interdependent modules.

ALIGNMENT OF ATP-BINDING PROTEINS

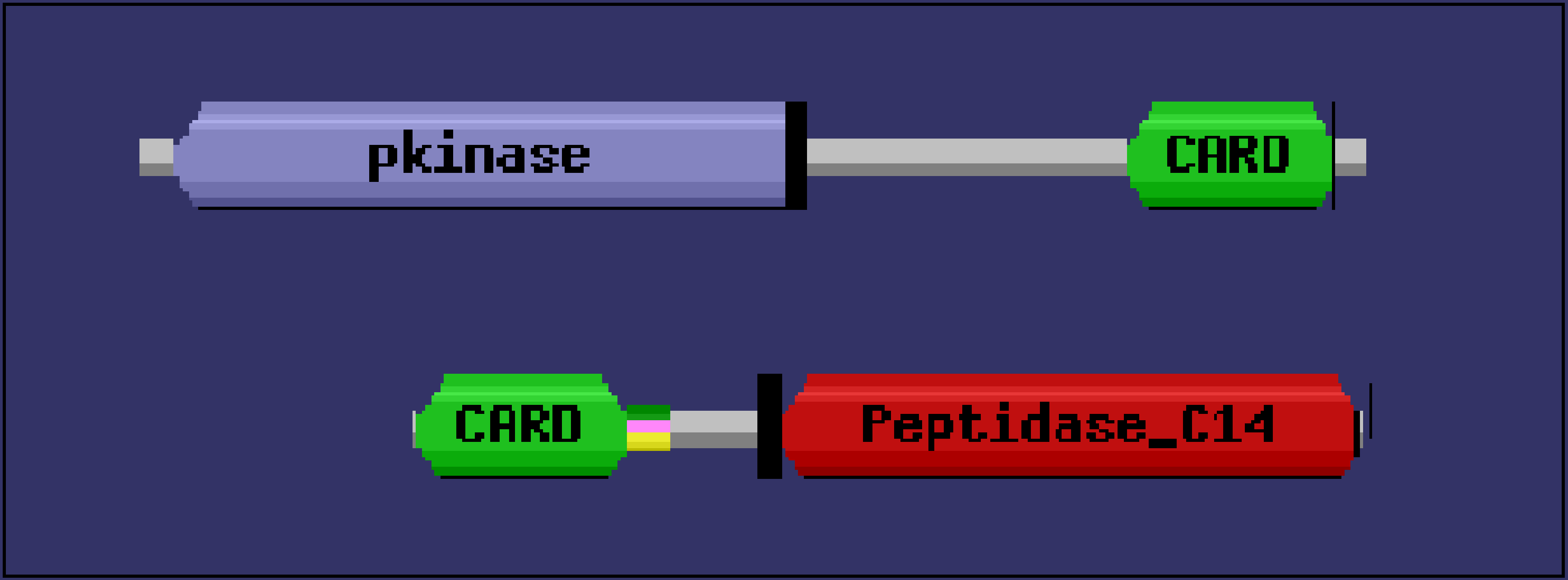


```
(36x635)                      -------300-------310-------320-------330-------340-------350-------360--
1ba1          4  376 QATKDAGT.IAG.......LNVLRIINEPTAAAIAYGLDKKVGAERNVLIFDLGGGTFDVSILTIEDG....
HS7C_HUMAN    4  377 QATKDAGT.IAG.......LNVLRIINEPTAAAIAYGLDKKVGAERNVLIFDLGGGTFDVSILTIEDG....
HS7C_BOVIN    4  377 QATKDAGT.IAG.......LNVLRIINEPTAAAIAYGLDKKVGAERNVLIFDLGGGTFDVSILTIEDG....
HS7C_MOUSE    4  377 QATKDAGT.IAG.......LNVLRIINEPTAAAIAYGLDKKVGAERNVLIFDLGGGTFDVSILTIEDG....
HS7D_DROME    4  377 QATKDAGT.IAG.......LNVLRIINEPTAAAIAYGLDKKAVGERNVLIFDLGGGTFDVSILSIDDG....
HS70_XENLA    5  378 QATKDAGV.LAG.......LNILRIINEPTAAAIAYGLDKGARGEQNVLIFDLGGGTFDVSILTIDDG....
1dkgD         4  375 QATKDAGR.IAG.......LEVKRIINEPTAAALAYGLDK..TGNRTIAVYDLGGGTFDISIIEIDEK....
DNAK_PASMU    2  378 QATKDAGR.IAG.......LEVKRIINEPTAAALAYGLDKGQG.NKTIAVYDLGGGTFDLSIIEIDEVG...
DNAK_SALTY    1  377 QATKDAGR.IAG.......LEVKRIINEPTAAALAYGLDKEVG.NRTIAVYDLGGGTFDISIIEIDEVD...
DNAK_VIBCH    2  377 QATKDAGR.IAG.......LEVKRIINEPTAAALAYGLDKQGG.DRTIAVYDLGGGTFDISIIEIDEVE...
DNAK_BURPS    2  379 QATKDAGR.IAG.......LEVKRIINEPTAAALAFGLDKAEKGDRKIAVYDLGGGTFDVSIIEIADVDG..
DNAK_BURCE    2  380 QATKDAGR.IAG.......LEVKRIINEPTAAALAFGLDKAEKGDRKIAVYDLGGGTFDVSIIEIADVDG..
1jceA         4  322 RAILDAGL.EAG.......ASKVFLIIEPXAAAIGSNLN..VEEPSGNXVVDIGGGTTEVAVISL.......
MREB_067013  11  328 RAVVDAAK.SAG.......AREVYLVAEPMAAAIGAGLP..VEEPIGNMIVDIGGGTTDIAVISLA......
MREB_BACSU    6  325 RAVIDATR.QAG.......ARDAYPIIEPFAAAIGANLP..VWEPTGSMVVDIGGGTTEVAIISLG......
MREB_Q9K8H5   6  325 RAVEDATK.QAG.......AKYAYTLIEPFAAAIGADLP..VWEPTGSMVVDIGGGTTEVAIISLG......
MREB_Q92BG6   6  324 RAVIDATR.QAG.......AKDAFTIIEPFAAAIGAGLP..VGEPTGSMVVDIGGGTTEVAVISLG......
MREB_Q9L1G6   9  326 RAVIEASS.QAG.......ARQVHIIIEPMAAAIGSGLP..VHEATGNMVVDIGGGTTEVAVISLG......
1e4fT         8  384 EMFYNFLQDTVK.......S.PFQLKGSLVSTAEGVLIT..PEKDRGVVVVNLGYNFTGLIAYKN.......
FTSA_ENTHR    5  379 HNIRKCVENAGL.......V.VNELVITPLALTETILSD..GEKDFGTIVIDMGGGQTTTAVMHD.......
FTSA_ENTFA    1  375 HNIRKCVEKAGL.......G.INELVITPLALTETILTD..GEKDFGTIVIDMGGGQTTTSVIHD.......
FTSA_BACSU    5  379 HNLLRCVERAGI.......E.ITDICLQPLAAGSAALSK..DEKNLGVALIDIGGGSTTIAVFQN.......
FTSA_BORBU    5  378 QNLVRCVNRAGF.......A.VDEVVLGSLASSYATLSK..EEREMGVLFIDMGKGTTDIILYID.......
FTSA_ECOLI    8  383 KNIVKAVERCGL.......K.VDQLILAGLASSYSVLTE..DERELGVCVVDIGGGTMDIAVYTG.......
1yagA         5  346 EKMTQIMFETFN.......VPAFYVSIQAVLSLYSSGRT.......TGIVLDSGDGVTHVVPIYA.......
ACT_BOTCI     5  346 EKMTQIVFETFN.......APAFYVSIQAVLSLYASGRT.......TGIVLDSGDGVTHVVPIYE.......
ACT_NEUCR     5  346 EKMTQIVFETFN.......APAFYVSIQAVLSLYASGRT.......TGIVLDSGDGVTHVVPIYE.......
ACT4_CAEEL    6  347 EKMTQIMFETFN.......TPAMYVAIQAVLSLYASGRT.......TGVVLDSGDGVTHTVPIYE.......
ACTB_HUMAN    5  346 EKMTQIMFETFN.......TPAMYVAIQAVLSLYASGRT.......TGIVMDSGDGVTHTVPIYE.......
ACT5_CHICK    6  347 EKMTQIMFETFN.......TPAMYVAIQAVLSLYASGRT.......TGIVMDSGDGVTHTVPIYE.......
1qhaA        78  456 ADVVKLLN.KAIKKRGDYDANIVAVVNDTVGTMMTCGYD....DQHCEVGLIIGTG.TNACYMEELRHIDLV
HXK1_HUMAN   78  456 ADVVKLLN.KAIKKRGDYDANIVAVVNDTVGTMMTCGYD....DQHCEVGLIIGTG.TNACYMEELRHIDLV
HXK1_BOVIN   78  456 NYVVKLLD.KAIKKRGDYDANIVAVVNDTVGTMIDCGYD....DQHCEVGLIIGTG.TNACYMEELRQIDFG
HXK_SCHMA    68  443 HNVAELLQ.TELDKRE.LNVKCVAVVNDTVGTLASCALE....DPKCAVGLIVGTG.TNVAYIEDSSKVELM
HXK2_DROME  128  505 KNVVSLLQ.EAIDRRGDLKINTVAILNDTVGTLMSCAYY....HPNCRIGLIVGTG.SNACYVEKTVNAECF
HXK1_SPIOL   95  485 EDVVAELT.KAMLRKG.VDMRVTALVNDTVGTLAGGRYY....KEDVIAAVILGTG.TNAAYVERASAIHKW
```

Family, or Subfamily specific motifs

RICK PROTEIN KINASE AND CASPASE-9

Domain organization of proteins that contain a CARD domain, from Pfam.



MOTIF AND DOMAIN DESCRIPTION

- In the case of Motifs, their small size and the lack of perfect consrvation make not possible the use of BLAST, for example, to identify proteins that contain a given motif, in a database.

- Conserved domains can be identified with sequence alignment tools, such as BLAST. However, only with the development of more sensitive algorithms based on sequence profiles, it has been possible to capture the widespread distribution of some domains in unrelated protein families, as result of domain shuffling.

MOTIF AND DOMAIN DESCRIPTION

FOUR strategies are usually considered to describe, or represent, conserved motifs or domains :

- Consensus sequences
- Patterns (Regular expressions)
- PSSM (PSWM) Profiles
- HMM profiles

CONSENSUS SEQUENCES

Consensus sequences

```
      ALRDFATHDDF
      SMTAEATHDSI
      ECDQAATHEAS

80%   XXXXXATHXXX

50%   XXXXXATHDXX
```

REGULAR EXPRESSIONS

Regular expression

ALRDF**ATHD**DF
SMTAE**ATHD**SI
ECDQA**ATHE**AS

NNNNN**ATH[DE]**NN

•Any aminoacid: x	
•Ambiguity: [A,B…]	A, or B...
or {A,B..}	anything except A, B…
•Repetition: A(2,4)	A-A o A-A-A o A-A-A-A
•N terminal: <	
•C-terminal: >	

Regular expression are the basis of Prosite

[AC]-x-V-x(4)-{E,D}

[Ala or Cys]-any-Val-any-any-any-any-{any but Glu or Asp}

PSSM (PSWM) PROFILES
(Position Specific Scoring / Weight Matrices)

Much more sensitive than regular expressions.

```
F   K   L   L   S   H   C   L   L   V

F   K   A   F   G   Q   T   M   F   Q
Y   P   I   V   G   Q   E   L   L   G
F   P   V   V   K   E   A   I   L   K
F   K   V   L   A   A   V   I   A   D
L   E   F   I   S   E   C   I   I   Q
F   K   L   L   G   N   V   L   V   C
```

A	-18	-10	**-1**	-8	8	-3	3	-10	-2	-8
C	-22	-33	-18	-18	-22	-26	22	-24	-19	-7
D	-35	0	-32	-33	-7	6	-17	-34	-31	0
E	-27	15	-25	-26	-9	23	-9	-24	-23	-1
F	60	-30	12	14	-26	-29	-15	4	12	-29
G	-30	-20	-28	-32	28	-14	-23	-33	-27	-5
H	-13	-12	-25	-25	-16	14	-22	-22	-23	-10
I	3	-27	21	25	-29	-23	-8	33	19	-23
K	-26	25	-25	-27	-6	4	-15	-27	-26	0
L	14	-28	19	27	-27	-20	-9	33	26	-21
M	3	-15	10	14	-17	-10	-9	25	12	-11
N	-22	-6	-24	-27	1	8	-15	-24	-24	-4
P	-30	24	-26	-28	-14	-10	-22	-24	-26	-18
Q	-32	5	-25	-26	-9	24	-16	-17	-23	7
R	-18	9	-22	-22	-10	0	-18	-23	-22	-4
S	-22	-8	-16	-21	11	2	-1	-24	-19	-4
T	-10	-10	-6	-7	-5	-8	2	-10	-7	-11
V	0	-25	22	25	-19	-26	6	19	16	-16
W	9	-25	-18	-19	-25	-27	-34	-20	-17	-28
Y	34	-18	-1	1	-23	-12	-19	0	0	-18

The weights reflect not only the frequencies of each amino acid at that position in the alignment, but also the expected frecuencies of each aminoacid and substitution probabilities.

A shows a lower preference than M because, although is not at that position in the alignment, it is a more likely replacement for L,I,V and F.

Prosite uses profiles, in addition to Regular Expressions.

PSI-BLAST

PSI-BLAST (Position Specific Iterated BLAST) is one of the programs related with BLAST that is accesible at the NCBI web server (it is also part of the BLAST package).

- Like BLAST, PSI-BLAST takes a query sequence as input to perform a similarity search agains a chosen database.

- Then, a multiple sequence alignment and a profile are constructed from significant local alignments .

- The profile is then used to search the database again, and any new significant hits are incorporated to the profile.

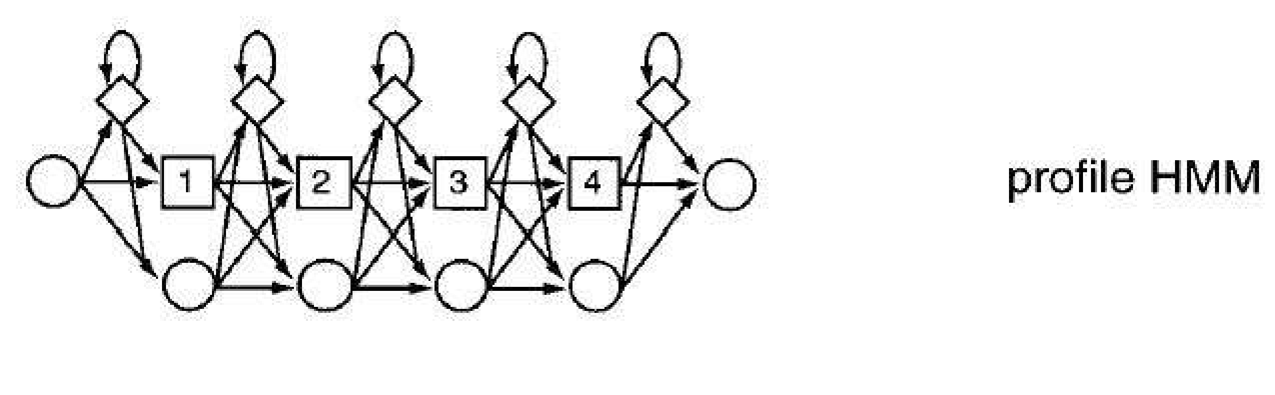
- The process iterates an arbitrary number of times or until convergence (no new sequences can be found in the database that match the profile).

HMM PROFILES

Hidden Markov model (HMM) profiles are statistical models of the primary structure of aminoacid sequences.

They are similar to PSWMs, but the weights are calculated according to a probabilistic model that explicitly take into account insertions and deletions, and that also may take into account previous positions.

They are the basis of Pfam and SMART, among other domain family databases.



DEVELOPMENT OF DOMAINS AND MOTIFS DATABASES

Known motifs

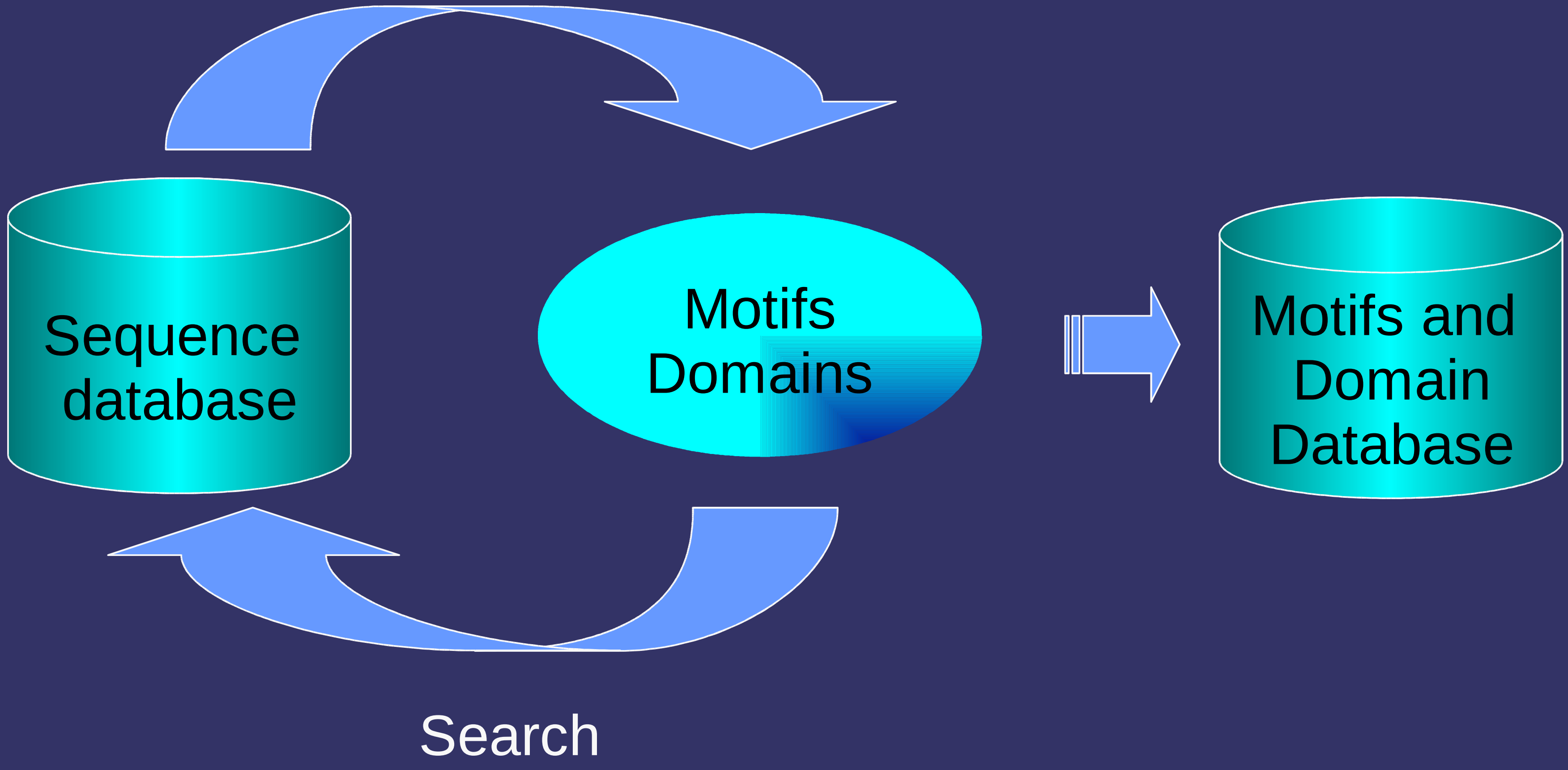
Multiple sequence alignments

Representation

Regular expression

Profile

HMM



Sequence database

Motifs Domains

Motifs and Domain Database

Search

Protein family databases

- Proteins are seen as evolutionary units
- Examples: COG, ProtoMap

vs

Domain family databases

- Domains are the evolutionary units
- Examples: Pfam, SMARt

SOME DOMAINS AND MOTIFS DATABASES

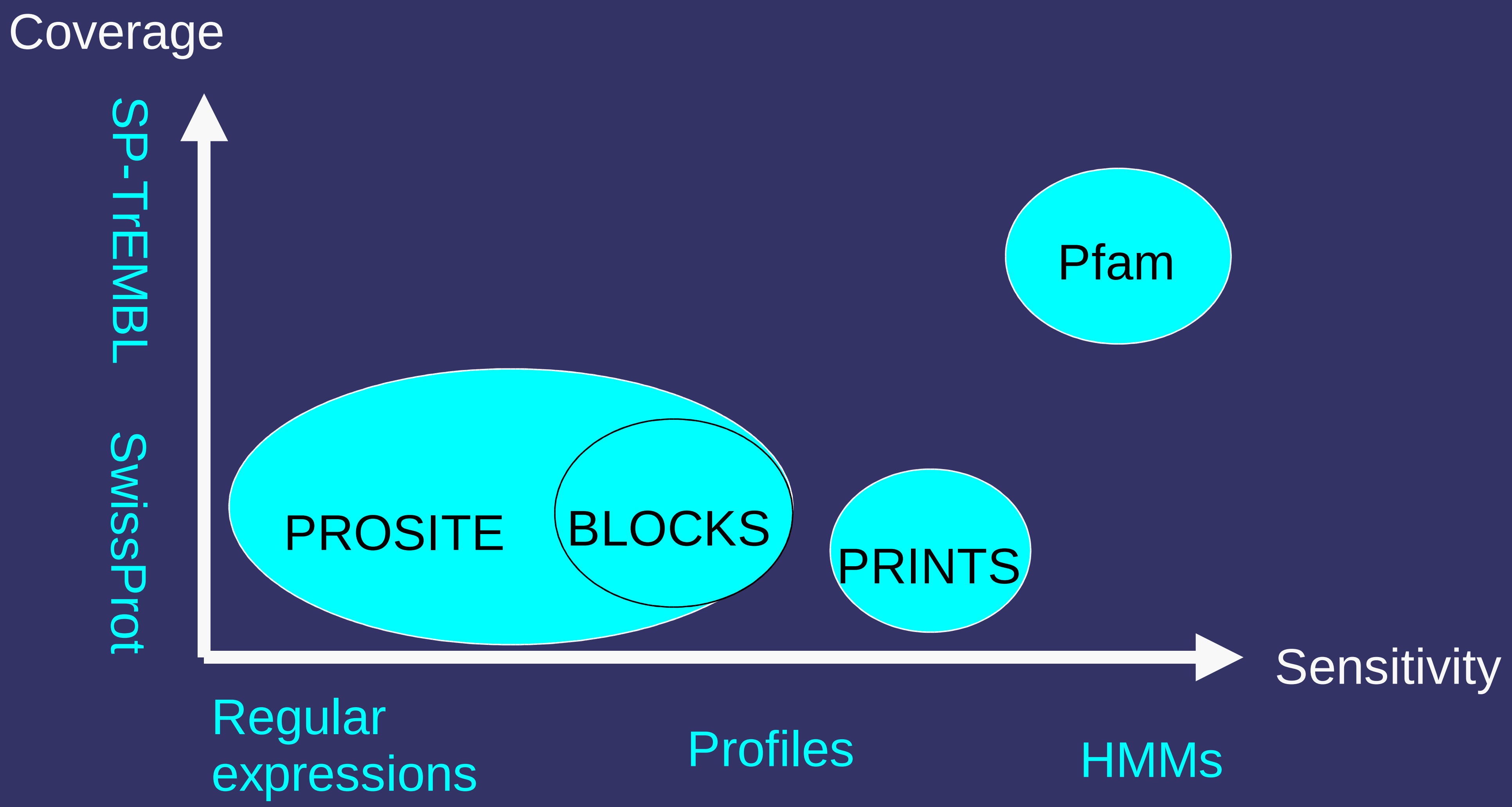
**** PROSITE**. Database of protein families and domains, defined by patterns and profiles, at ExPASy.

**** Pfam,** at Sanger. Multiple sequence alignments and HMMs of protein domains and families, at Sanger Institute.
**** SMART**. Simple Modular Architecture Research Tool, at EMBL.

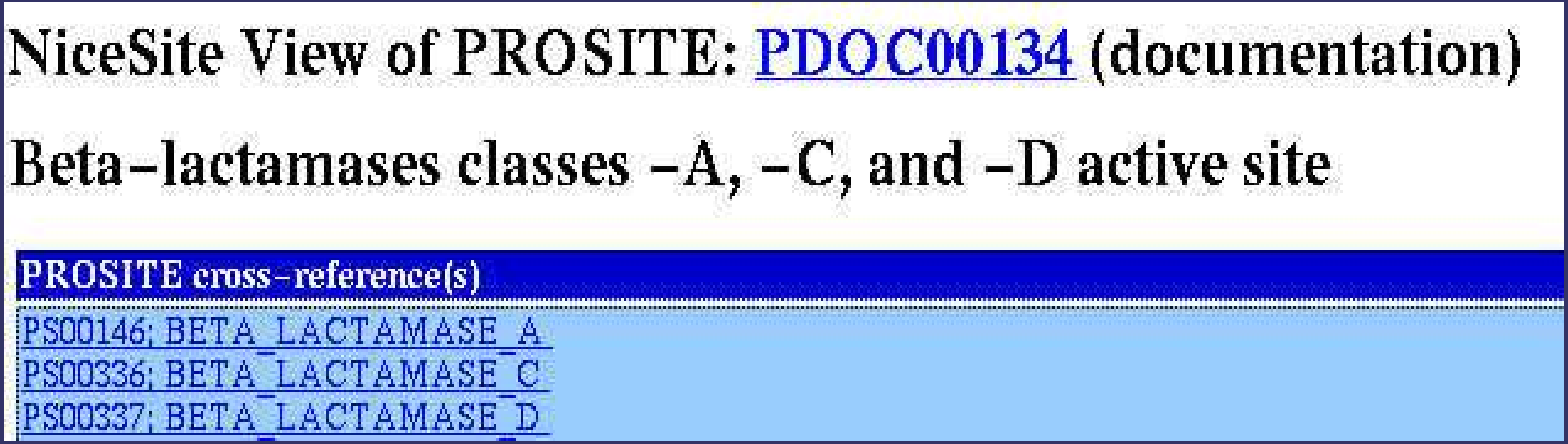
MetaFam. Comprehensive database of protein family information.
Blocks. Multiple alignments of the most highly conserved regions of groups of proteins documented in InterPro.
PRINTS. Database of groups of conserved motifs, or protein fingerprints.
ProDom. Protein domain families automatically generated from SWISS-PROT and TrEMBL.

**** InterPro**. Integrated view of commonly used motif databases.

COMPARISON OF DIFFERENT STRATEGIES AND DATABASES THAT COLLECT INFORMATION ABOUT PROTEIN MOTIFS AND DOMAINS



beta-Lactamase PATTERNS, from Prosite

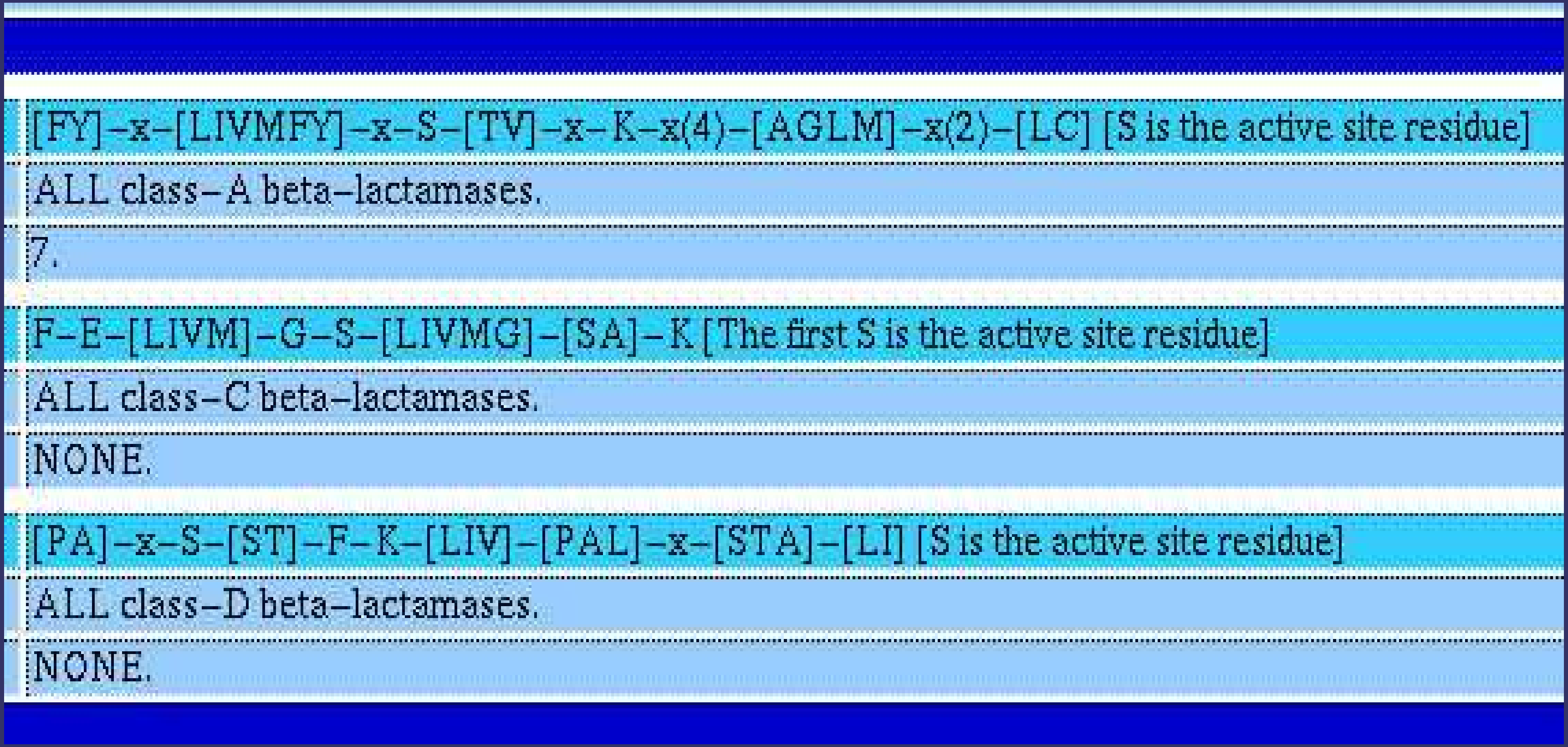


NiceSite View of PROSITE: PDOC00134 (documentation)

Beta–lactamases classes –A, –C, and –D active site

PROSITE cross-reference(s)

PS00146; BETA_LACTAMASE_A
PS00336; BETA_LACTAMASE_C
PS00337; BETA_LACTAMASE_D



[FY]–x–[LIVMFY]–x–S–[TV]–x–K–x(4)–[AGLM]–x(2)–[LC] [S is the active site residue]

ALL class–A beta–lactamases.

7.

F–E–[LIVM]–G–S–[LIVMG]–[SA]–K [The first S is the active site residue]

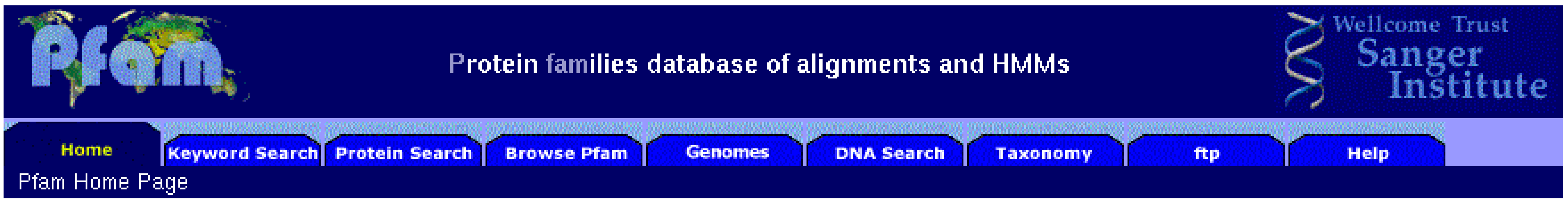
ALL class–C beta–lactamases.

NONE.

[PA]–x–S–[ST]–F–K–[LIV]–[PAL]–x–[STA]–[LI] [S is the active site residue]

ALL class–D beta–lactamases.

NONE.



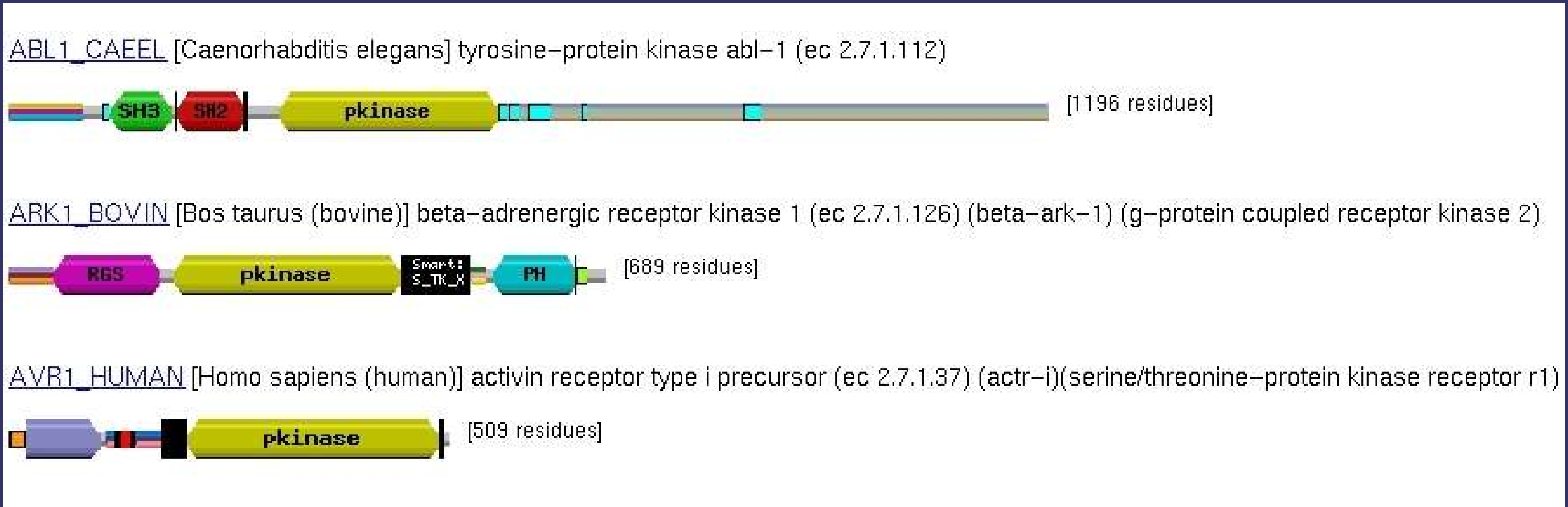
Pfam

Protein families database of alignments and HMMs

Wellcome Trust Sanger Institute

Pfam Home Page

ORGANIZATION OF SEVERAL PROTEINS THAT CONTAIN A PROTEIN KINASE DOMAIN, from Pfam

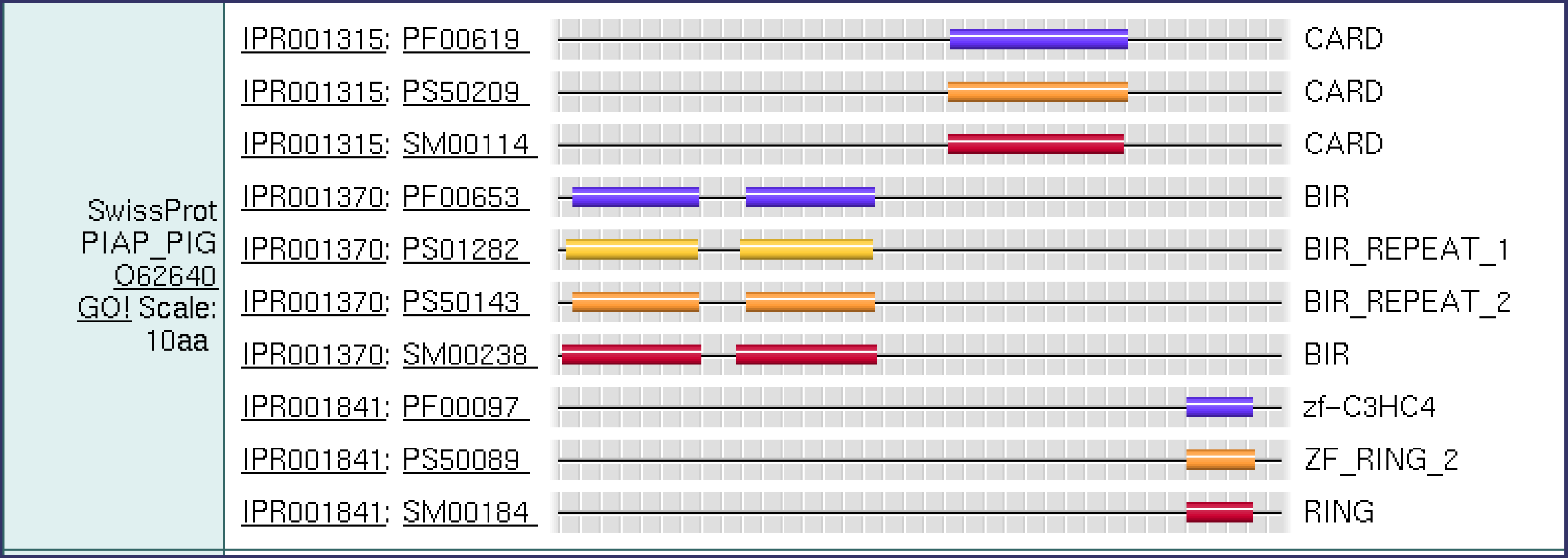


ABL1_CAEEL [Caenorhabditis elegans] tyrosine-protein kinase abl-1 (ec 2.7.1.112)

ARK1_BOVIN [Bos taurus (bovine)] beta-adrenergic receptor kinase 1 (ec 2.7.1.126) (beta-ark-1) (g-protein coupled receptor kinase 2)

AVR1_HUMAN [Homo sapiens (human)] activin receptor type i precursor (ec 2.7.1.37) (actr-i)(serine/threonine-protein kinase receptor r1)

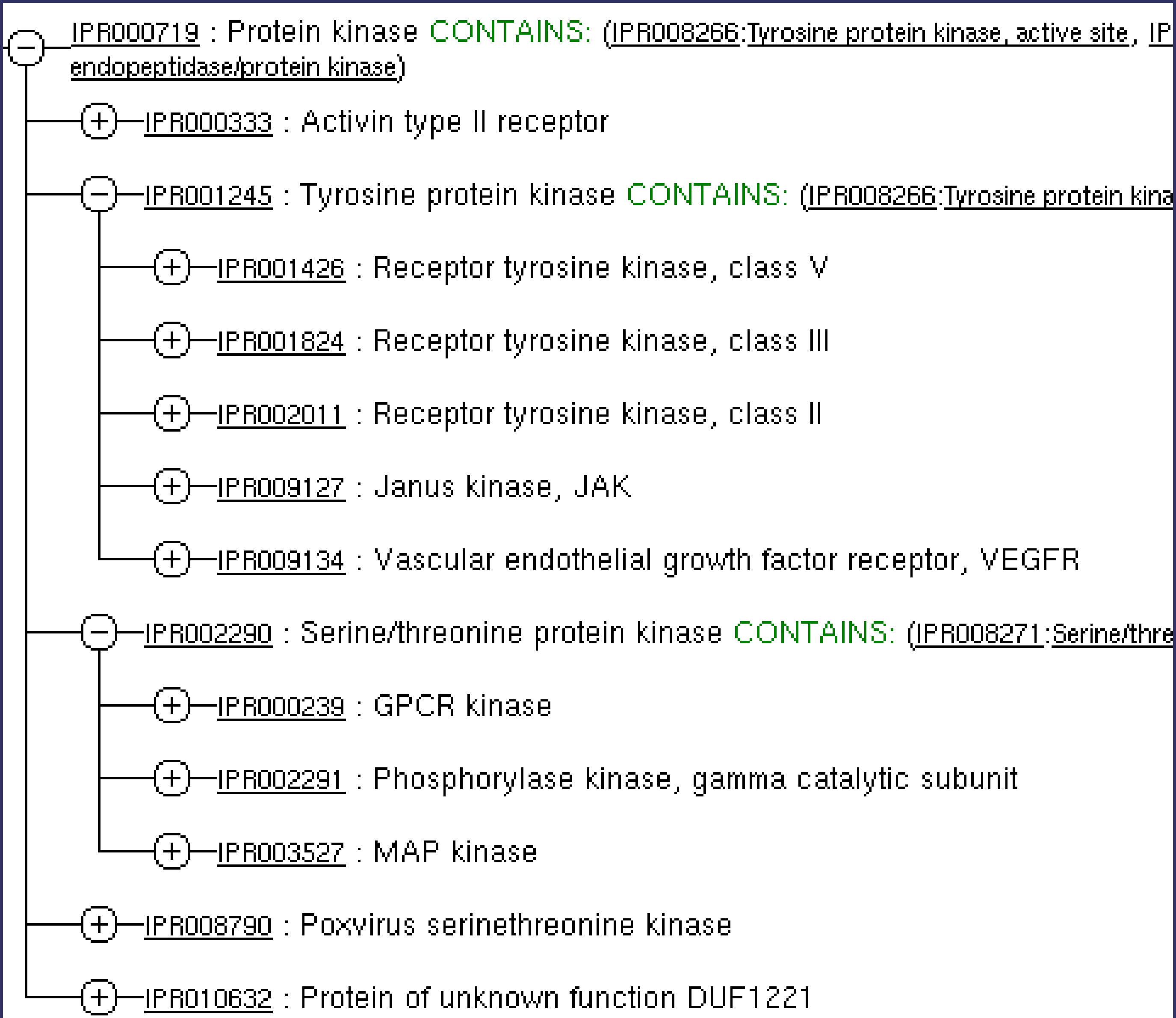
INTERPRO
DOMAIN ORGANIZATION FOR PIAP-PIG



SwissProt
PIAP_PIG
O62640
GO! Scale:
10aa

IPR001315: PF00619	CARD
IPR001315: PS50209	CARD
IPR001315: SM00114	CARD
IPR001370: PF00653	BIR
IPR001370: PS01282	BIR_REPEAT_1
IPR001370: PS50143	BIR_REPEAT_2
IPR001370: SM00238	BIR
IPR001841: PF00097	zf-C3HC4
IPR001841: PS50089	ZF_RING_2
IPR001841: SM00184	RING

Parent-Child Tree for InterPro Entry IPR000719



- IPR000719 : Protein kinase CONTAINS: (IPR008266:Tyrosine protein kinase, active site, IP endopeptidase/protein kinase)
  - IPR000333 : Activin type II receptor
  - IPR001245 : Tyrosine protein kinase CONTAINS: (IPR008266:Tyrosine protein kina
    - IPR001426 : Receptor tyrosine kinase, class V
    - IPR001824 : Receptor tyrosine kinase, class III
    - IPR002011 : Receptor tyrosine kinase, class II
    - IPR009127 : Janus kinase, JAK
    - IPR009134 : Vascular endothelial growth factor receptor, VEGFR
  - IPR002290 : Serine/threonine protein kinase CONTAINS: (IPR008271:Serine/thre
    - IPR000239 : GPCR kinase
    - IPR002291 : Phosphorylase kinase, gamma catalytic subunit
    - IPR003527 : MAP kinase
  - IPR008790 : Poxvirus serinethreonine kinase
  - IPR010632 : Protein of unknown function DUF1221

SOME TOOLS FOR THE ANALYSIS OF DOMAINS AND MOTIFS

SCANPROSITE. Scans a sequence to find matches to PROSITE or SWISSPROT and TrEMBL with a user provided pattern.
PRATT. Generates conserved patterns from a series of unaligned proteins.

PSI-BLAST. Position-Specific Iterated BLAST.
BIOACCELERATOR. Generation of PSSMS and database search.
MOTIF. Scans a sequence against several databases of patterns and profiles, but also, scans databases with user provided profiles, and also generates profiles from sequences provided by the user.

PFAM. Scans a sequence against the Pfam database of protein domains (defined as HMM profiles).
SMART. Scans a sequence against the SMART database (and other, like Pfam) of protein domains (defined as HMM profiles).

INTERPROSCAN. Scans a sequence against the InterPro database of patterns and profiles (which integrates information from several other databases)

APPLICATIONS OF DATABASES AND ALGORITHMS BASED ON MOTIFS AND DOMAINS

Identification of remote homologs: by means of identifying sequences that share a particular motif.

Sequence clustering: the presence of conserved motifs or domains allow the definition of protein families.

Function prediction: by means of the identification of characterized motifs in the sequence of interest

This presentation contains material from:

Federico Abascal, CNB
Oswaldo Trelles, UMA
Joaquín Dopazo, CNIO
Paulino Gómez Puertas, CAB
Manuel J Gómez, CNB