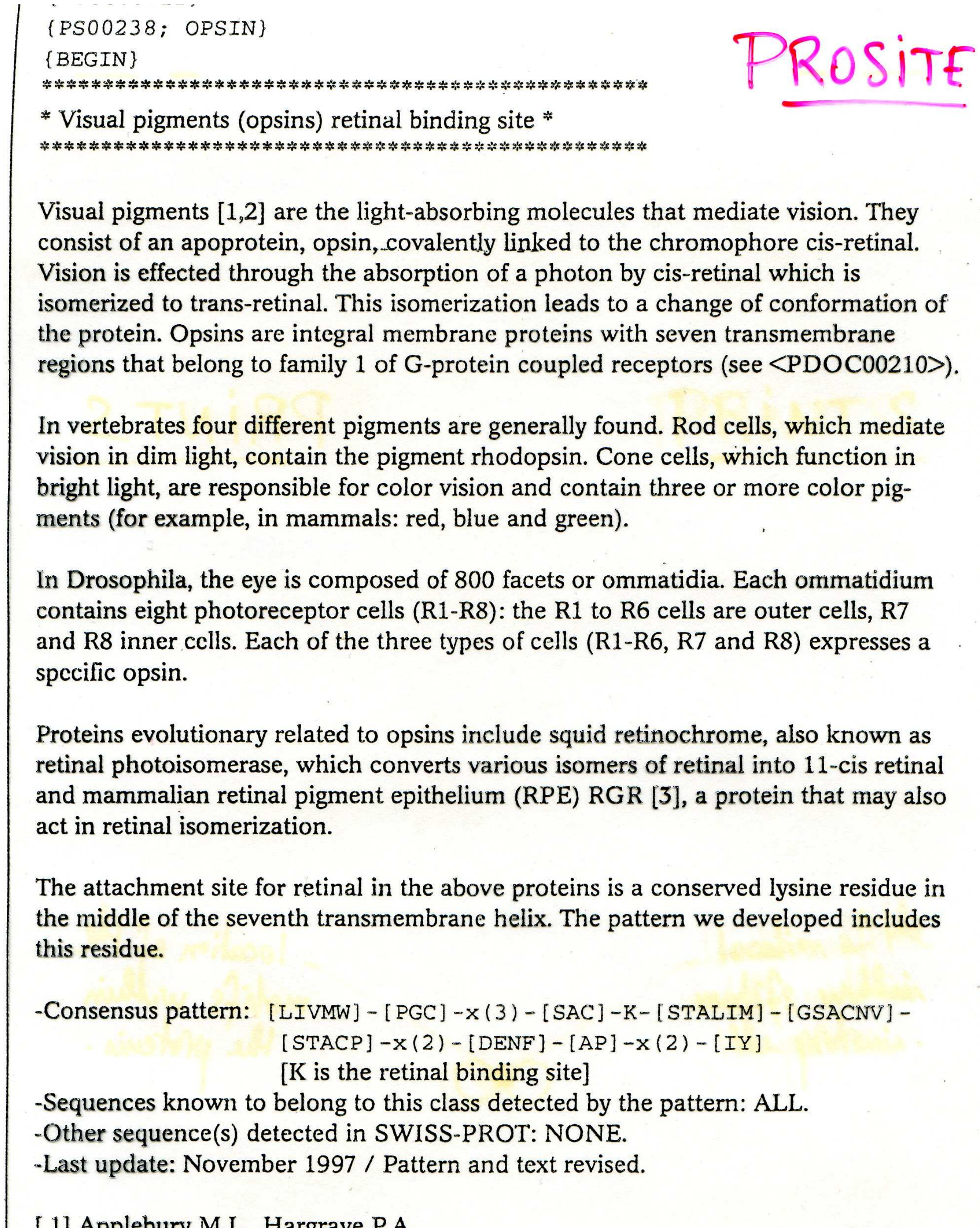
Bioinformatique M2: Lecture 1 – part B

P. Derreumaux

Function prediction from protein sequence

From a multiple sequence alignment we can derive:

- A motif
- A profile (PSSM)
- A Hidden Markov Model



{PS00238; OPSIN}
{BEGIN}

PROSITE

****

* Visual pigments (opsins) retinal binding site *

****

Visual pigments [1,2] are the light-absorbing molecules that mediate vision. They consist of an apoprotein, opsin, covalently linked to the chromophore cis-retinal. Vision is effected through the absorption of a photon by cis-retinal which is isomerized to trans-retinal. This isomerization leads to a change of conformation of the protein. Opsins are integral membrane proteins with seven transmembrane regions that belong to family 1 of G-protein coupled receptors (see <PDOC00210>).

In vertebrates four different pigments are generally found. Rod cells, which mediate vision in dim light, contain the pigment rhodopsin. Cone cells, which function in bright light, are responsible for color vision and contain three or more color pigments (for example, in mammals: red, blue and green).

In Drosophila, the eye is composed of 800 facets or ommatidia. Each ommatidium contains eight photoreceptor cells (R1-R8): the R1 to R6 cells are outer cells, R7 and R8 inner cells. Each of the three types of cells (R1-R6, R7 and R8) expresses a specific opsin.

Proteins evolutionary related to opsins include squid retinochrome, also known as retinal photoisomerase, which converts various isomers of retinal into 11-cis retinal and mammalian retinal pigment epithelium (RPE) RGR [3], a protein that may also act in retinal isomerization.

The attachment site for retinal in the above proteins is a conserved lysine residue in the middle of the seventh transmembrane helix. The pattern we developed includes this residue.

-Consensus pattern: [LIVMW]-[PGC]-x(3)-[SAC]-K-[STALIM]-[GSACNV]-
[STACP]-x(2)-[DENF]-[AP]-x(2)-[IY]
[K is the retinal binding site]
-Sequences known to belong to this class detected by the pattern: ALL.
-Other sequence(s) detected in SWISS-PROT: NONE.
-Last update: November 1997 / Pattern and text revised.

[1] Applebury M.L., Hargrave P.A.

False vs. true Positives

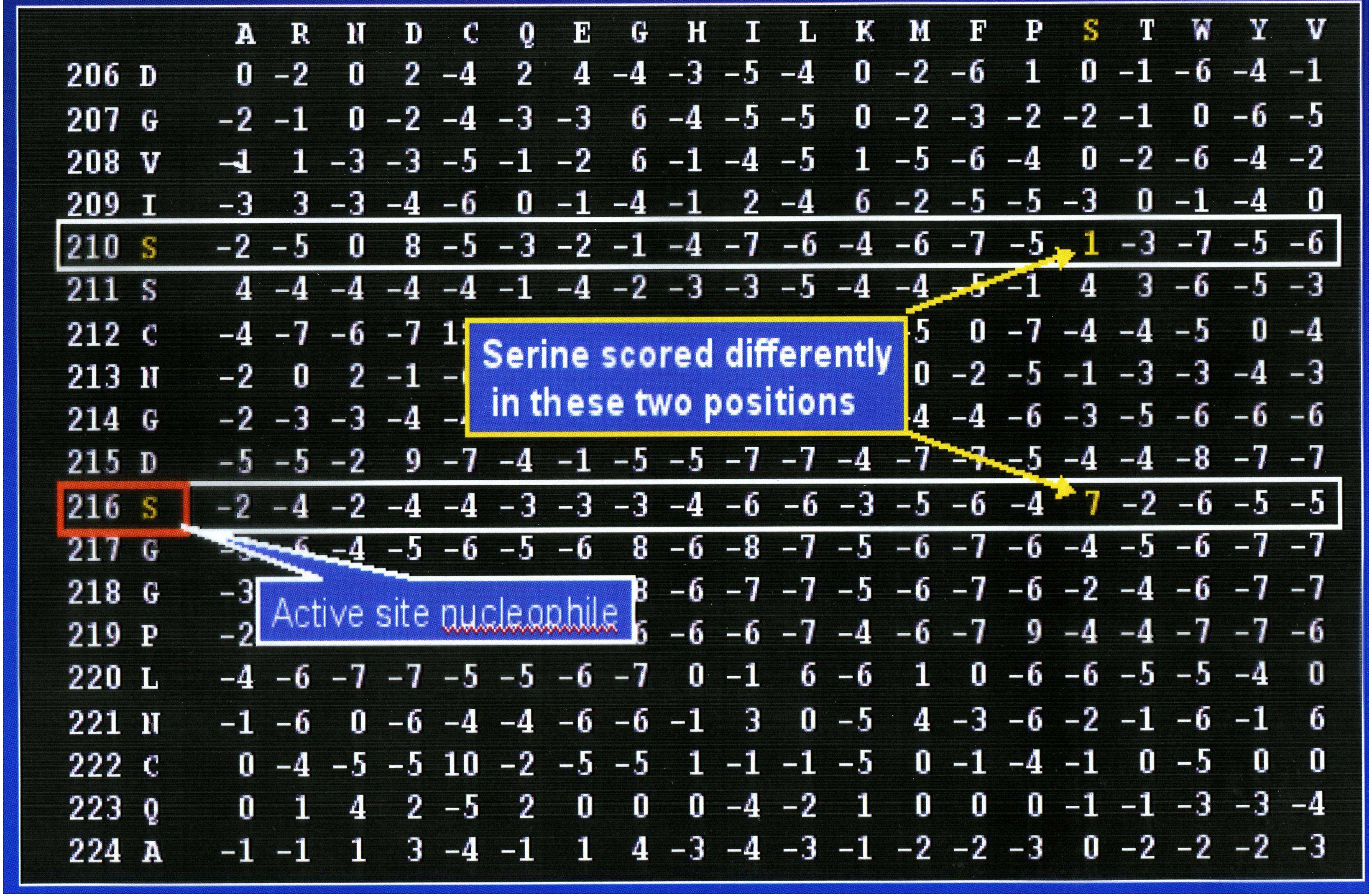
La syntaxe d'un pattern PROSITE (séquence proprement dite) suit les rêgles suivantes :

- **lettres A-Z** correspondant aux acides aminés (minuscules ou majuscules)
- **[] ambiguite inclusive** EX: [ILVM]
- **{} ambiguite exclusive** EX: {FWY}
- **X caractère positionnel indifférent**
- **(n) répétition n fixe** d'un sous-motif EX: [RD](2)
- **X(n,m) insertions min-max** (insertion variable) EX: X(2,4)
- **< au début du pattern** : le pattern est cadré à gauche de la séquence
- **> à la fin du pattern** : le pattern est cadré à droite de la séquence
- **le caractère '-'** sépare chaque position
- **le caractère '+'** indique que la suite du pattern continue à la ligne suivante

Exemples de motifs :

```
C-{CPWHF}-X(2,4)-C-H-{CFYW}
AXGHXXX[QST]{DR}
<P-x(2)-R-G-[STAIV](2)-x-N-[APK]-x-[DE]
[STANVF]-x(2)-F-x(4)-[DNS]-x(5,7)-[DENQTF]-Y-[HFY]-x(2)-[LIVMFY]-x(3)-+
[LIVM]-x(4)-[LIVM]-x(6,8)-Y-x(12,13)-[LIVM]-x(2)-N-[SACF]-x(2)-[FY]>
```

Position Specific Score Matrix (PSSM)



		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
206	D	0	-2	0	2	-4	2	4	-4	-3	-5	-4	0	-2	-6	1	0	-1	-6	-4	-1
207	G	-2	-1	0	-2	-4	-3	-3	6	-4	-5	-5	0	-2	-3	-2	-2	-1	0	-6	-5
208	V	-1	1	-3	-3	-5	-1	-2	6	-1	-4	-5	1	-5	-6	-4	0	-2	-6	-4	-2
209	I	-3	3	-3	-4	-6	0	-1	-4	-1	2	-4	6	-2	-5	-5	-3	0	-1	-4	0
210	S	-2	-5	0	8	-5	-3	-2	-1	-4	-7	-6	-4	-6	-7	-5	1	-3	-7	-5	-6
211	S	4	-4	-4	-4	-4	-1	-4	-2	-3	-3	-5	-4	-4	-5	-1	4	3	-6	-5	-3
212	C	-4	-7	-6	-7	1[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]5	0	-7	-4	-4	-5	0	-4
213	N	-2	0	2	-1	-[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	-2	-5	-1	-3	-3	-4	-3
214	G	-2	-3	-3	-4	-[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]4	-4	-6	-3	-5	-6	-6	-6
215	D	-5	-5	-2	9	-7	-4	-1	-5	-5	-7	-7	-4	-7	-7	-5	-4	-4	-8	-7	-7
216	S	-2	-4	-2	-4	-4	-3	-3	-3	-4	-6	-6	-3	-5	-6	-4	7	-2	-6	-5	-5
217	G	[illegible]	[illegible]	-4	-5	-6	-5	-6	8	-6	-8	-7	-5	-6	-7	-6	-4	-5	-6	-7	-7
218	G	-3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]8	-6	-7	-7	-5	-6	-7	-6	-2	-4	-6	-7	-7
219	P	-2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]6	-6	-6	-7	-4	-6	-7	9	-4	-4	-7	-7	-6
220	L	-4	-6	-7	-7	-5	-5	-6	-7	0	-1	6	-6	1	0	-6	-6	-5	-5	-4	0
221	N	-1	-6	0	-6	-4	-4	-6	-6	-1	3	0	-5	4	-3	-6	-2	-1	-6	-1	6
222	C	0	-4	-5	-5	10	-2	-5	-5	1	-1	-1	-5	0	-1	-4	-1	0	-5	0	0
223	Q	0	1	4	2	-5	2	0	0	0	-4	-2	1	0	0	0	-1	-1	-3	-3	-4
224	A	-1	-1	1	3	-4	-1	1	4	-3	-4	-3	-1	-2	-2	-3	0	-2	-2	-2	-3

Serine scored differently in these two positions

Active site nucleophile

Profile Hidden Markov Model (profile HMM)

- An MSA can be described by a HMM
- HMM is a probabilistic model of the MSA consisting of a number of interconnected states
- The different states are match (M), delete (D) or insert (I). They are associated with symbol emission probability distributions and connected by transition state probabilities
- Each position is modeled independently
- The concatenation of the probabilistic models of the positions is the protein model.
- The Viterbi (DP) algorithm identifies the optimal path, the forward algorithm is used for scoring sequences.

The Protein Domain Database

ProDom

WARNING: new procedure for ProDom construction

July 1998 (ProDom 35)

The ProDom protein domain database consists of an automatic compilation of homologous domains. Current versions of ProDom are built using a novel procedure based on recursive PSI-BLAST searches (Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W & Lipman DJ, 1997, *Nucleic Acids Res.*, **25**:3389-3402; Gouzy J., Corpet F. & Kahn D., 1999, *Computers and Chemistry* **23**:333-340.) Large families are much better processed with this new procedure than with the former DOMAINER program (Sonnhammer, E.L.L. & Kahn, D., 1994, *Protein Sci.*, **3**:482-492).

March 2001 (ProDom 2001.1)

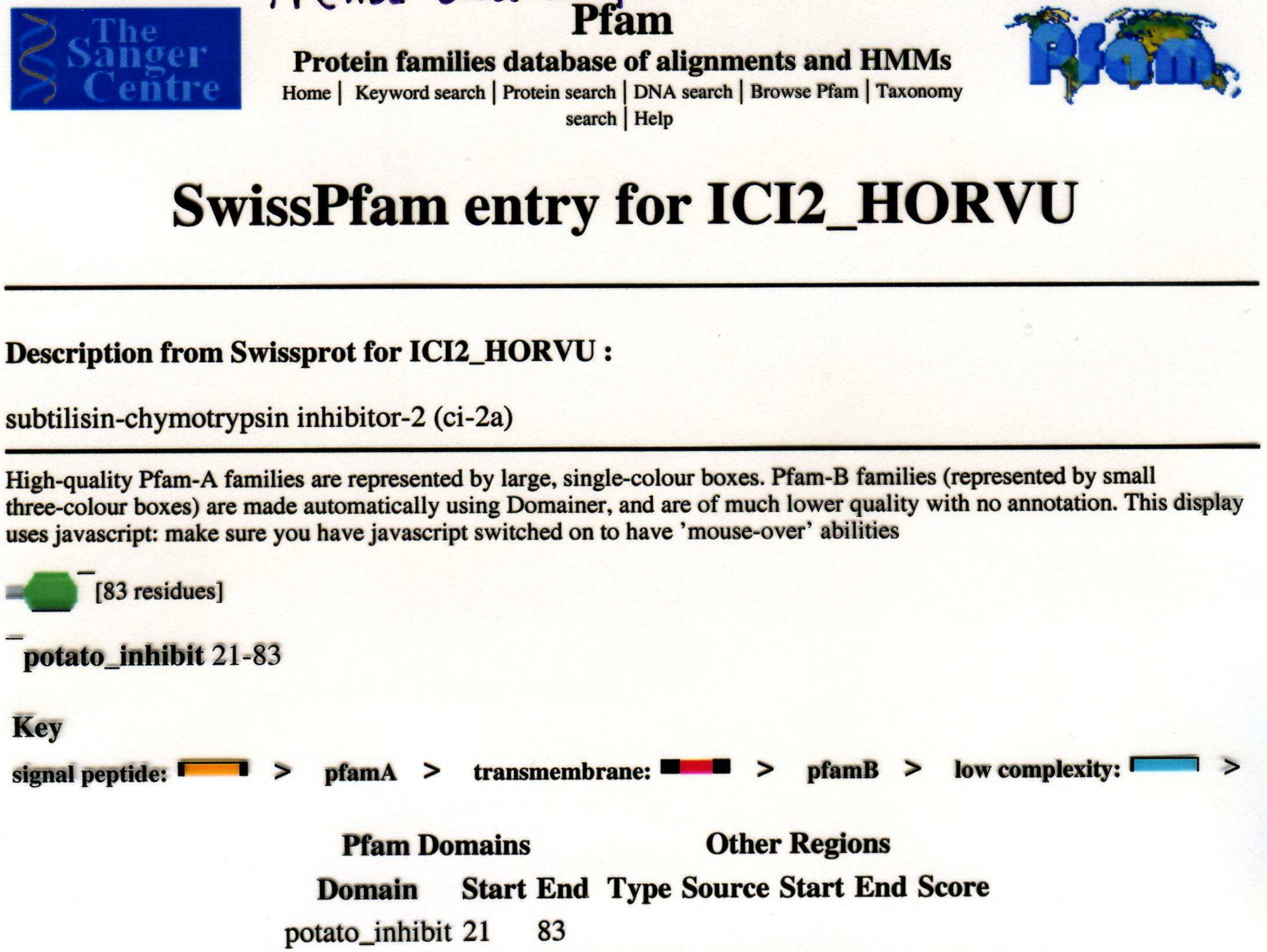
390 ProDom families were generated automatically using PSI-BLAST with a profile built from the seed aligments of Pfam-A 4.3 families. (internal repeat detection

Last ProDom update:	March 30th, 2001
Current ProDom release:	ProDom 2001.1 / (Statistics)
built from	non fragmentary sequences from SWISS-PROT 39 + TREMBL + TREMBL updates - December 8, 2000

Both the ProDom database and this server have been designed as a tool to help analyze domain arrangements of proteins and protein families *(Corpet F, Servant F, Gouzy J, Kahn D (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. Nucleic Acids Res. 28:267-269)*. Strong emphasis has been put on the graphical user interface which allows for interactive analysis of protein homology relationships. Here is a brief outline of what the ProDom server can do for you:

Pfam is a database of multiple alignments of protein domains or conserved protein region.

- Pfam-A are accurate human crafted multiple alignments
- Pfam-B is an automatic clustering of the rest of SwissProt and TrEMBL derived from the Prodom Database.



The Sanger Centre

Pfam

Protein families database of alignments and HMMs

Home | Keyword search | Protein search | DNA search | Browse Pfam | Taxonomy search | Help

SwissPfam entry for ICI2_HORVU

Description from Swissprot for ICI2_HORVU :

subtilisin-chymotrypsin inhibitor-2 (ci-2a)

High-quality Pfam-A families are represented by large, single-colour boxes. Pfam-B families (represented by small three-colour boxes) are made automatically using Domainer, and are of much lower quality with no annotation. This display uses javascript: make sure you have javascript switched on to have 'mouse-over' abilities

[83 residues]

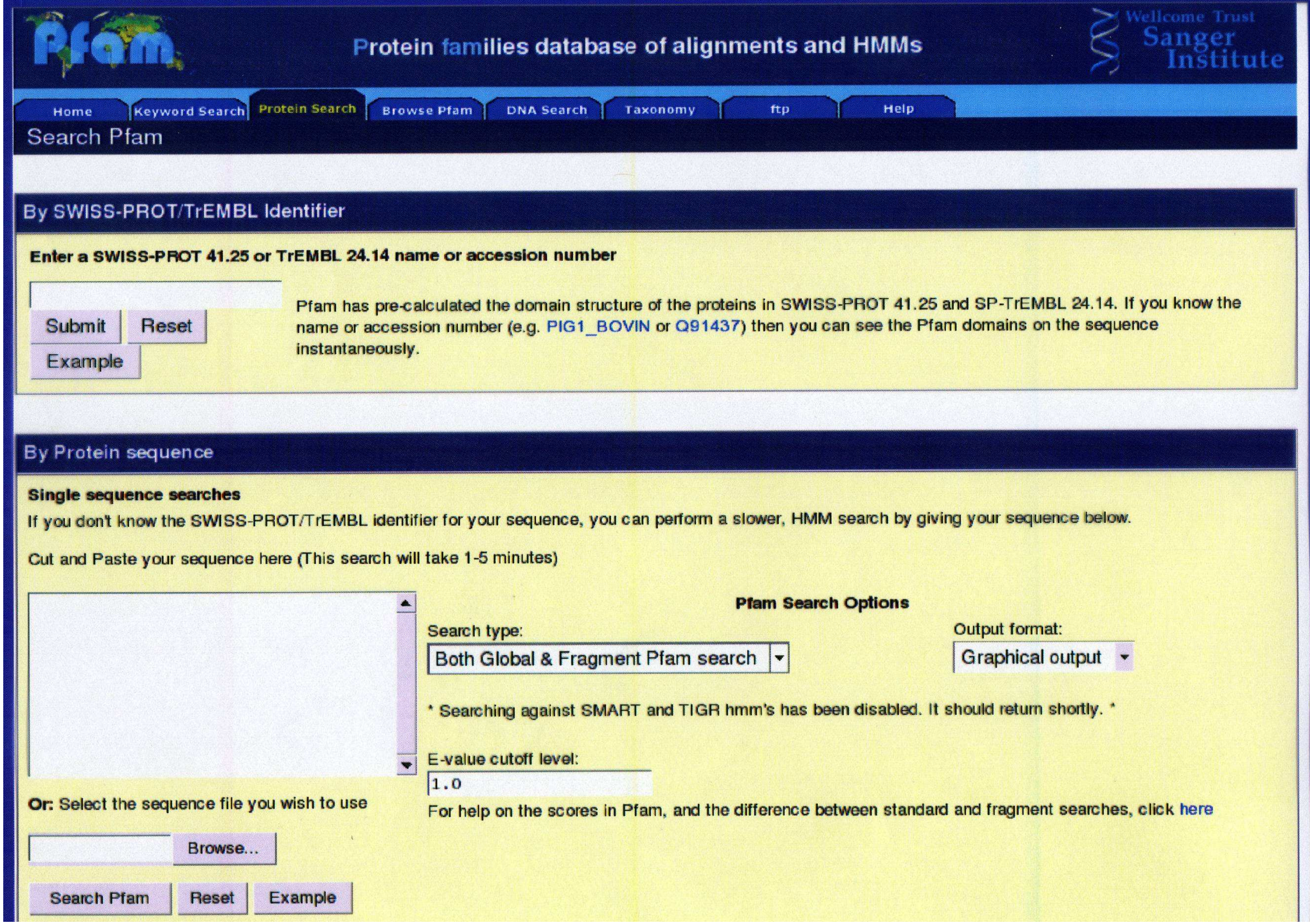
potato_inhibit 21-83

Key

signal peptide: 🟧 > **pfamA** > **transmembrane:** 🟥 > **pfamB** > **low complexity:** 🟦 >

Pfam Domains			Other Regions			
Domain	**Start**	**End**	**Type**	**Source**	**Start**	**End**	**Score**
potato_inhibit	21	83					

Cribler une protéine vs domaines Pfam


Pfam
Protein families database of alignments and HMMs
Wellcome Trust Sanger Institute
Home
Keyword Search
Protein Search
Browse Pfam
DNA Search
Taxonomy
ftp
Help

Search Pfam

By SWISS-PROT/TrEMBL Identifier

Enter a SWISS-PROT 41.25 or TrEMBL 24.14 name or accession number

Pfam has pre-calculated the domain structure of the proteins in SWISS-PROT 41.25 and SP-TrEMBL 24.14. If you know the name or accession number (e.g. PIG1_BOVIN or Q91437) then you can see the Pfam domains on the sequence instantaneously.

By Protein sequence

Single sequence searches

If you don't know the SWISS-PROT/TrEMBL identifier for your sequence, you can perform a slower, HMM search by giving your sequence below.

Cut and Paste your sequence here (This search will take 1-5 minutes)

Pfam Search Options

Search type: Both Global & Fragment Pfam search

Output format: Graphical output

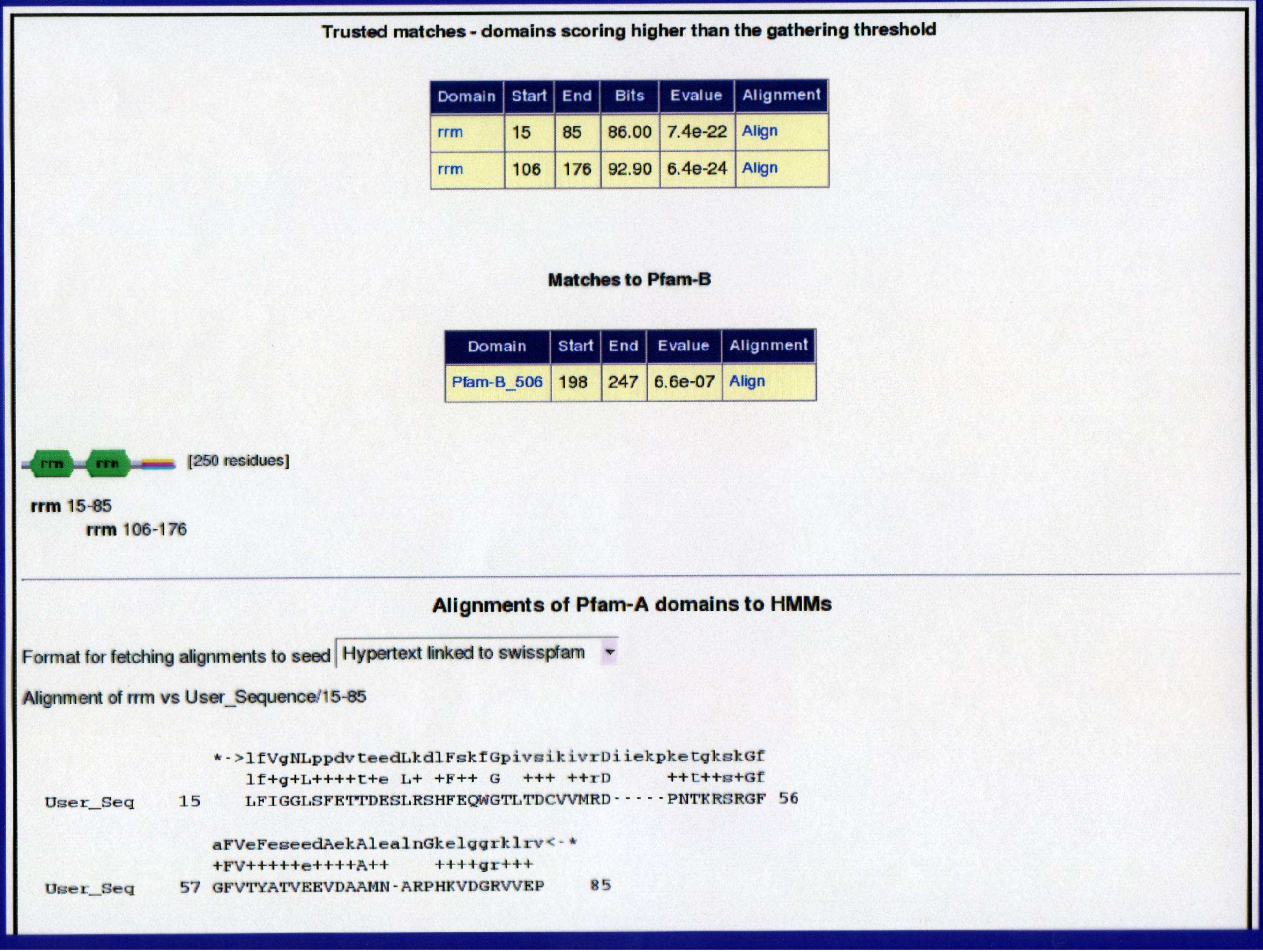
* Searching against SMART and TIGR hmm's has been disabled. It should return shortly. *

E-value cutoff level: 1.0

For help on the scores in Pfam, and the difference between standard and fragment searches, click here

Or: Select the sequence file you wish to use

Résultats Pfam & E-value



Trusted matches - domains scoring higher than the gathering threshold

Domain	Start	End	Bits	Evalue	Alignment
rrm	15	85	86.00	7.4e-22	Align
rrm	106	176	92.90	6.4e-24	Align

Matches to Pfam-B

Domain	Start	End	Evalue	Alignment
Pfam-B_506	198	247	6.6e-07	Align

[250 residues]

rrm 15-85

rrm 106-176

Alignments of Pfam-A domains to HMMs

Format for fetching alignments to seed: Hypertext linked to swisspfam

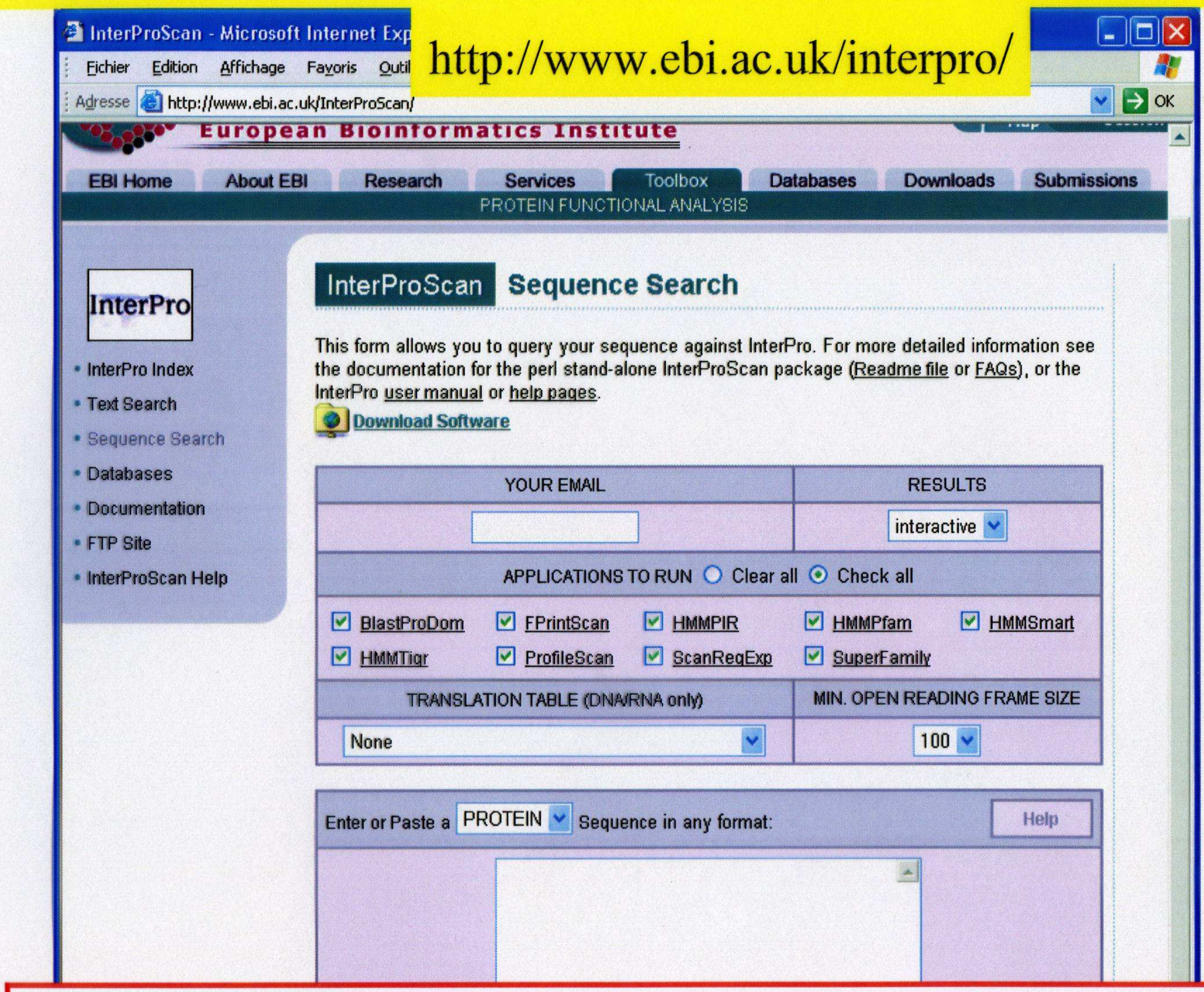
Alignment of rrm vs User_Sequence/15-85

```
                  *->lfVgNLppdvteedLkdlFskfGpivsikivrDiiekpketgkskGf
                     lf+g+L++++t+e L+ +F++ G  +++ ++rD     ++t++s+Gf
  User_Seq     15    LFIGGLSFETTDESLRSHFEQWGTLTDCVVMRD-----PNTKRSRGF 56

                  aFVeFeseedAekAlealnGkelggrklrv<-*
                  +FV+++++e++++A++     ++++gr+++
  User_Seq     57 GFVTYATVEEVDAAMN-ARPHKVDGRVVEP    85
```

Recherche de motifs et de domaines avec InterProScan

http://www.ebi.ac.uk/interpro/



InterProScan Sequence Search

This form allows you to query your sequence against InterPro. For more detailed information see the documentation for the perl stand-alone InterProScan package (Readme file or FAQs), or the InterPro user manual or help pages.

YOUR EMAIL:

RESULTS: interactive

APPLICATIONS TO RUN ○ Clear all ● Check all

☑ BlastProDom ☑ FPrintScan ☑ HMMPIR ☑ HMMPfam ☑ HMMSmart
☑ HMMTigr ☑ ProfileScan ☑ ScanRegExp ☑ SuperFamily

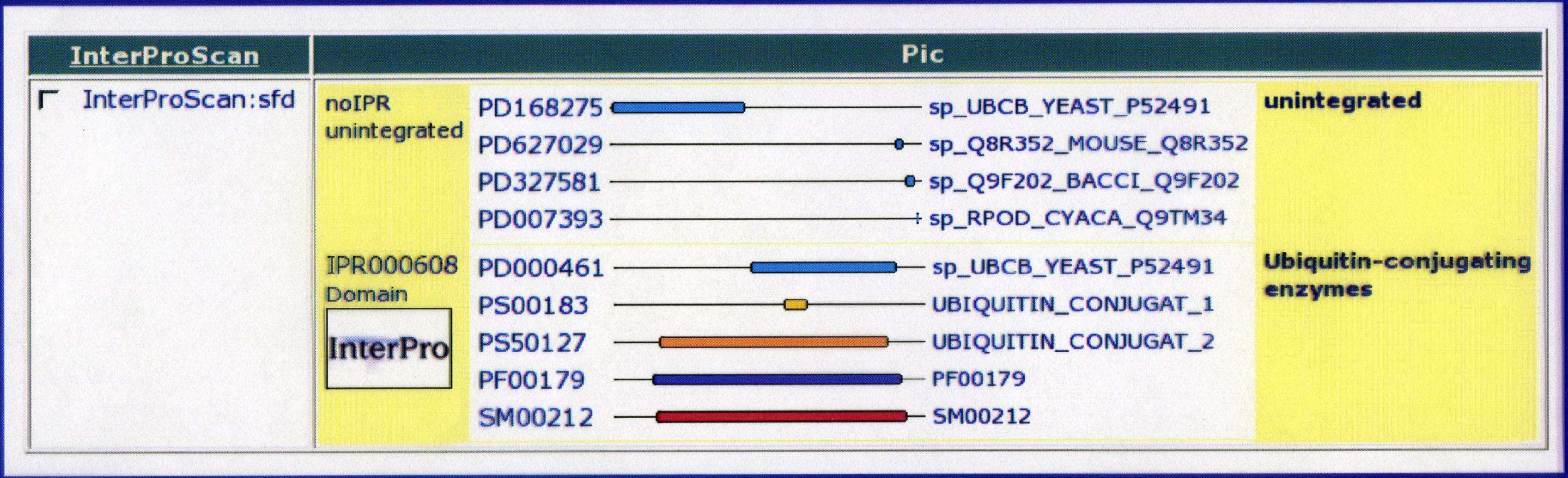
TRANSLATION TABLE (DNA/RNA only): None

MIN. OPEN READING FRAME SIZE: 100

Enter or Paste a PROTEIN Sequence in any format:

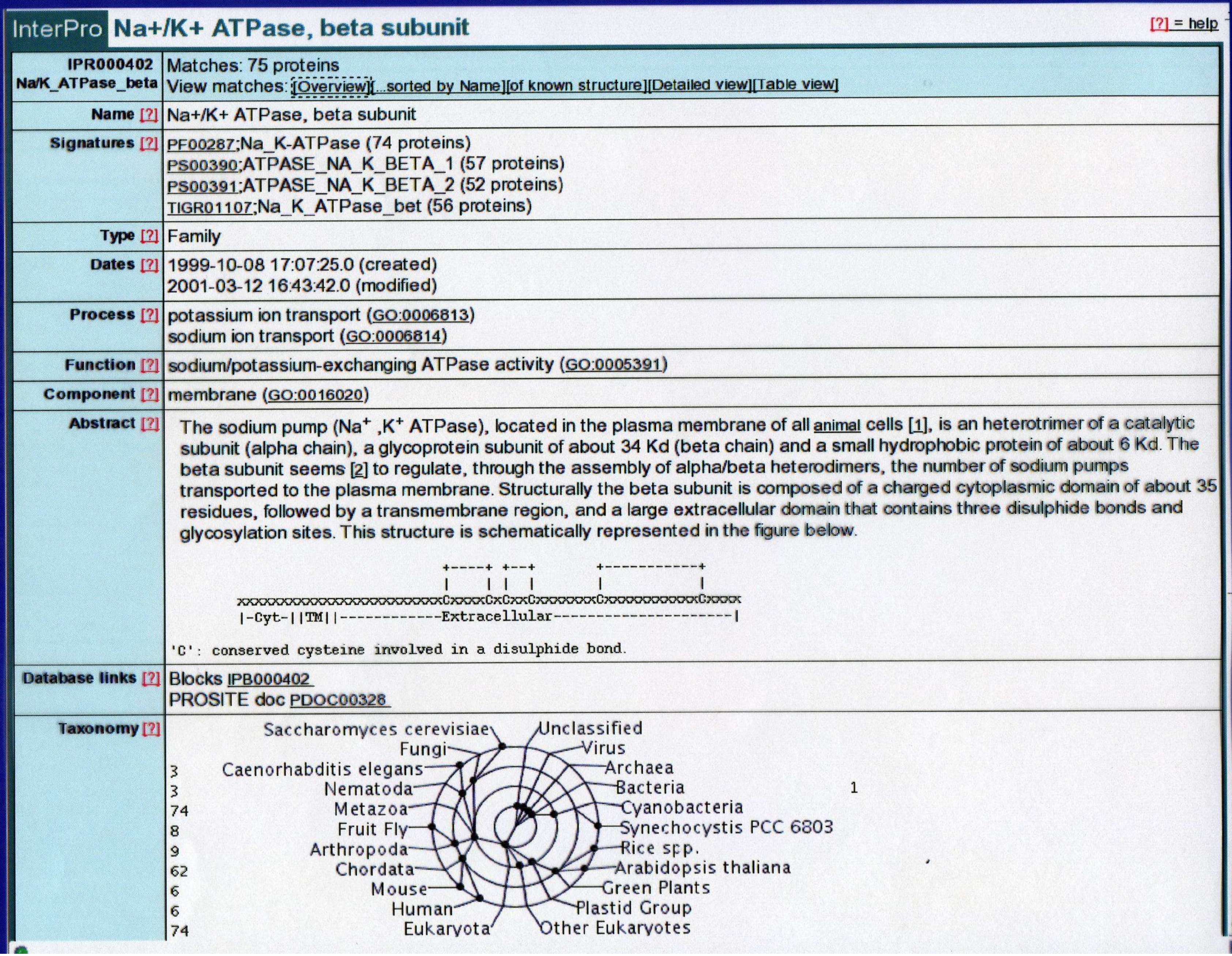
>1A1F:A DSNR ZINC FINGER
MERPYACPVESCDRRFSDSSNLTRHIRIHTGQKPFQCRICMRNFSRSDHLTTHIRTHTGEKPFA
CDICGRKFARSDERKRHTKIHLRQKD

Résultat criblage InterPRO



InterProScan	Pic			
⌜ InterProScan:sfd	noIPR unintegrated	PD168275 PD627029 PD327581 PD007393	sp_UBCB_YEAST_P52491 sp_Q8R352_MOUSE_Q8R352 sp_Q9F202_BACCI_Q9F202 sp_RPOD_CYACA_Q9TM34	unintegrated
	IPR000608 Domain InterPro	PD000461 PS00183 PS50127 PF00179 SM00212	sp_UBCB_YEAST_P52491 UBIQUITIN_CONJUGAT_1 UBIQUITIN_CONJUGAT_2 PF00179 SM00212	Ubiquitin-conjugating enzymes

Autre vue fiche InterPro



InterPro **Na+/K+ ATPase, beta subunit** [?] = help

IPR000402
Na/K_ATPase_beta
Matches: 75 proteins
View matches: [Overview][...sorted by Name][of known structure][Detailed view][Table view]

Name [?] Na+/K+ ATPase, beta subunit

Signatures [?]
PF00287;Na_K-ATPase (74 proteins)
PS00390;ATPASE_NA_K_BETA_1 (57 proteins)
PS00391;ATPASE_NA_K_BETA_2 (52 proteins)
TIGR01107;Na_K_ATPase_bet (56 proteins)

Type [?] Family

Dates [?]
1999-10-08 17:07:25.0 (created)
2001-03-12 16:43:42.0 (modified)

Process [?]
potassium ion transport (GO:0006813)
sodium ion transport (GO:0006814)

Function [?] sodium/potassium-exchanging ATPase activity (GO:0005391)

Component [?] membrane (GO:0016020)

Abstract [?]

The sodium pump (Na^+ ,K^+ ATPase), located in the plasma membrane of all animal cells [1], is an heterotrimer of a catalytic subunit (alpha chain), a glycoprotein subunit of about 34 Kd (beta chain) and a small hydrophobic protein of about 6 Kd. The beta subunit seems [2] to regulate, through the assembly of alpha/beta heterodimers, the number of sodium pumps transported to the plasma membrane. Structurally the beta subunit is composed of a charged cytoplasmic domain of about 35 residues, followed by a transmembrane region, and a large extracellular domain that contains three disulphide bonds and glycosylation sites. This structure is schematically represented in the figure below.

```
                        +----+ +--+       +-----------+
                        |    | |  |       |           |
xxxxxxxxxxxxxxxxxxxxxxxxCxxxxCxCxxCxxxxxxxCxxxxxxxxxxxCxxxx
|-Cyt-||TM||------------Extracellular---------------------|

'C': conserved cysteine involved in a disulphide bond.
```

Database links [?]
Blocks IPB000402
PROSITE doc PDOC00328

Taxonomy [?]