H3ABioNet

Pan African Bioinformatics Network for H3Africa

Introduction to Bioinformatics Online Course:IBT

# Introduction to Databases and Resources

# Protein Classification and Resources

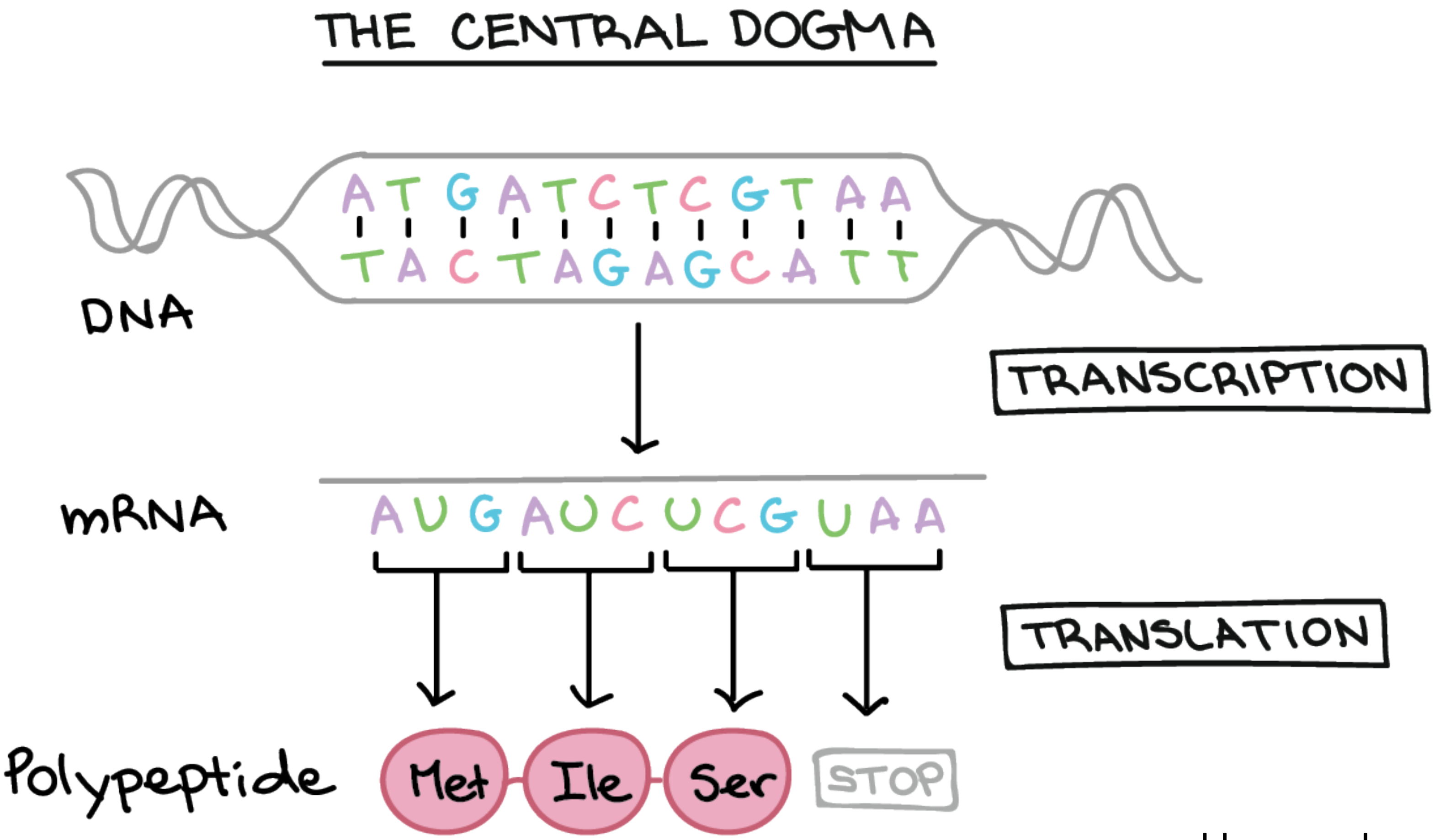
# Learning Objectives

- Understand how protein sequences are annotated
- Understand the different levels of protein classification
- Identify the key resources used for classifying protein sequences

# Learning Outcomes

- Differentiate between the different protein classification methods
- Use the appropriate tools to annotate a protein sequence of interest
- Access and retrieve information of interest from protein resources

# Central Dogma



www.khanacademy.org

# Protein Resources

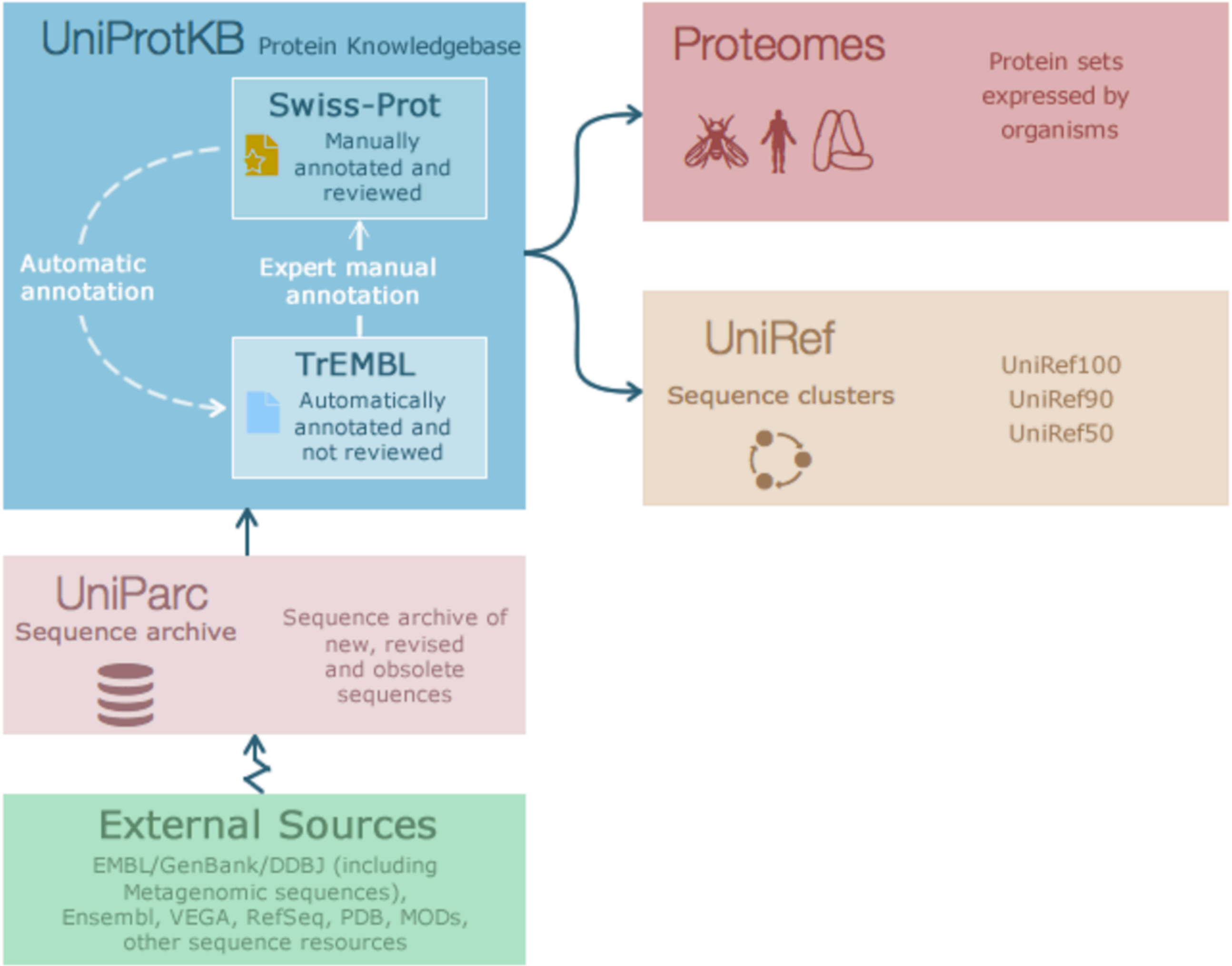
- A variety of protein resources online
- Several websites/ resources dedicated to providing a single interface to multiple resources
- Important to differentiate between databases and resources

PDBe
Protein Data Bank in Europe
UniProt
SIB
swissprot
Swiss Institute of Bioinformatics
PIR
Pfam
InterPro
RCSB PDB
PROTEIN DATA BANK

# Protein Databases

- Sequence and information databases
  - ✓NCBI Protein Database – contains protein sequences from GenBank, RefSeq , as well as records from SwissProt, PIR, PRF, and PDB
  - ✓EBI - UniProtKB – the "Protein knowledgebase", a comprehensive set of protein sequences. Functional information on proteins, with accurate, consistent, and rich annotation, the amino acid sequence, protein name or description, taxonomic data and citation information. Divided into two parts: Swiss-Prot and TrEMBL

# Protein Databases



UniProtKB Protein Knowledgebase
Swiss-Prot
Manually annotated and reviewed
Automatic annotation
Expert manual annotation
TrEMBL
Automatically annotated and not reviewed
Proteomes
Protein sets expressed by organisms
UniRef
Sequence clusters
UniRef100
UniRef90
UniRef50
UniParc
Sequence archive
Sequence archive of new, revised and obsolete sequences
External Sources
EMBL/GenBank/DDBJ (including Metagenomic sequences), Ensembl, VEGA, RefSeq, PDB, MODs, other sequence resources

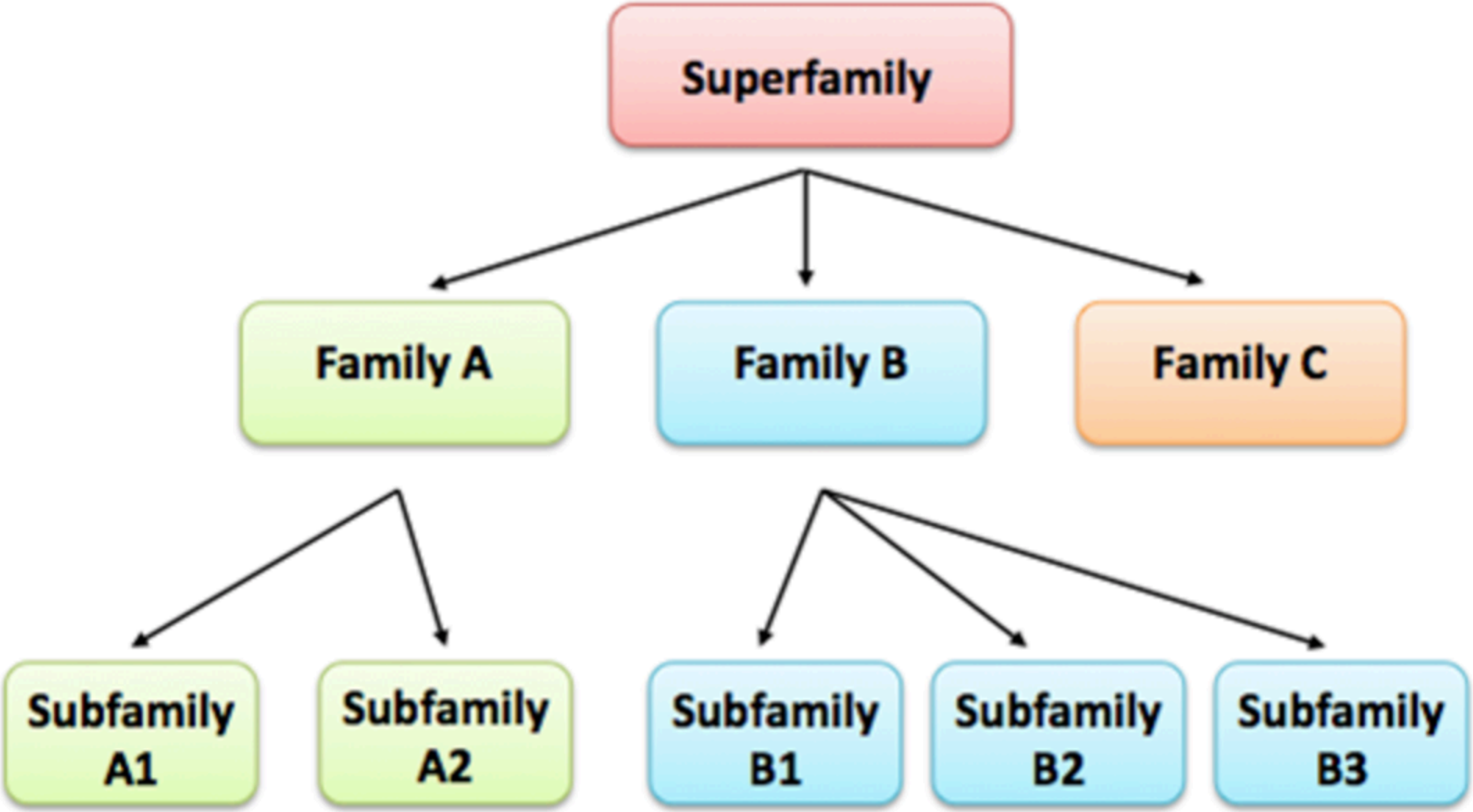
# Protein Classification Concepts

- Classification methods group proteins based on:
  - ✓Sequence similarity
  - ✓Structural similarity
- Most groups already contain a set of well characterised proteins whose function is known

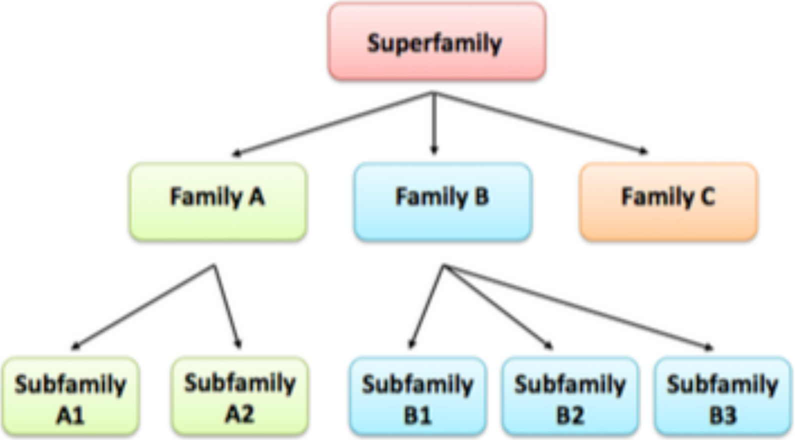
# Protein Classification Concepts

- Proteins can be classified into different groups based on:
  - ✓The families to which they belong
  - ✓The domains they contain
  - ✓The sequence features they possess
- Protein families share a common evolutionary origin, based on their related functions an similarities in sequence or structure

# Protein Classification

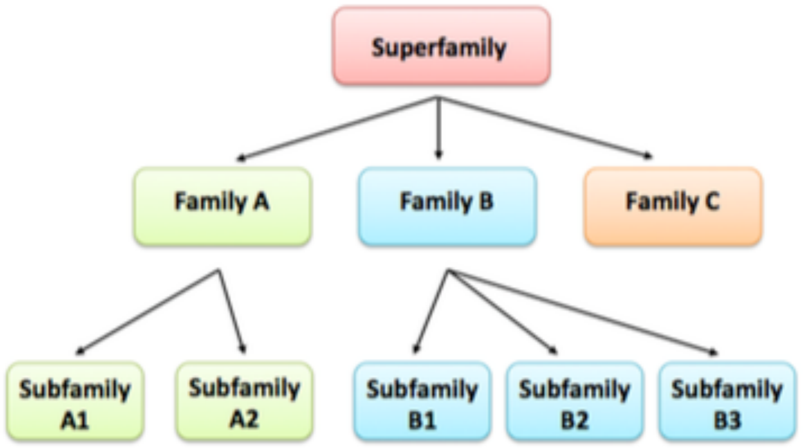


# Protein Classification



- Superfamily
  - ✓ A large group of distantly related proteins

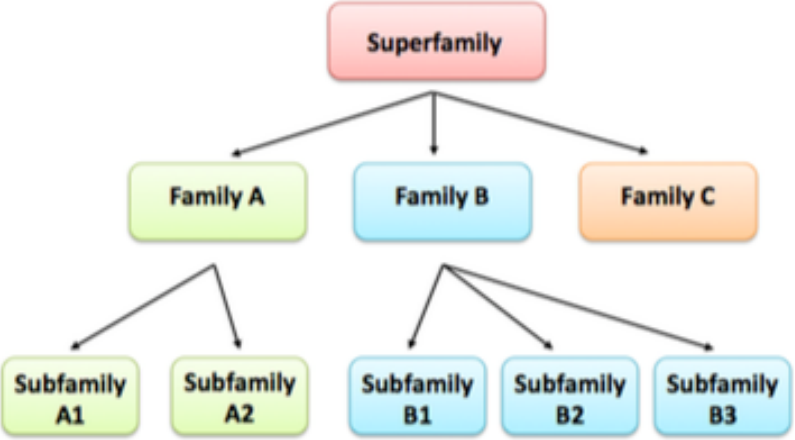
# Protein Classification



- Family
  - ✓ Group of evolutionarily related proteins that share one or more domains/repeats

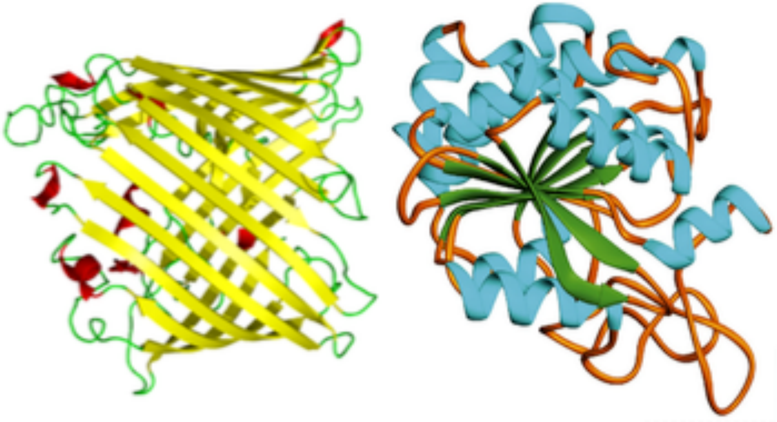
# Protein Classification

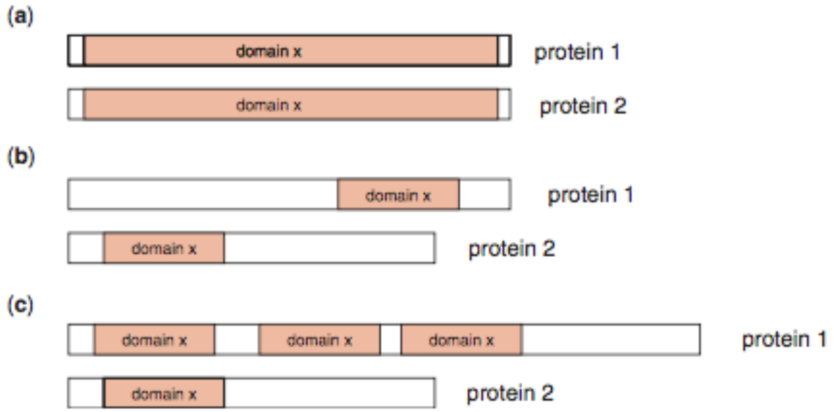
- Subfamily
  - ✓ A small group of closely related proteins



# Protein Domains

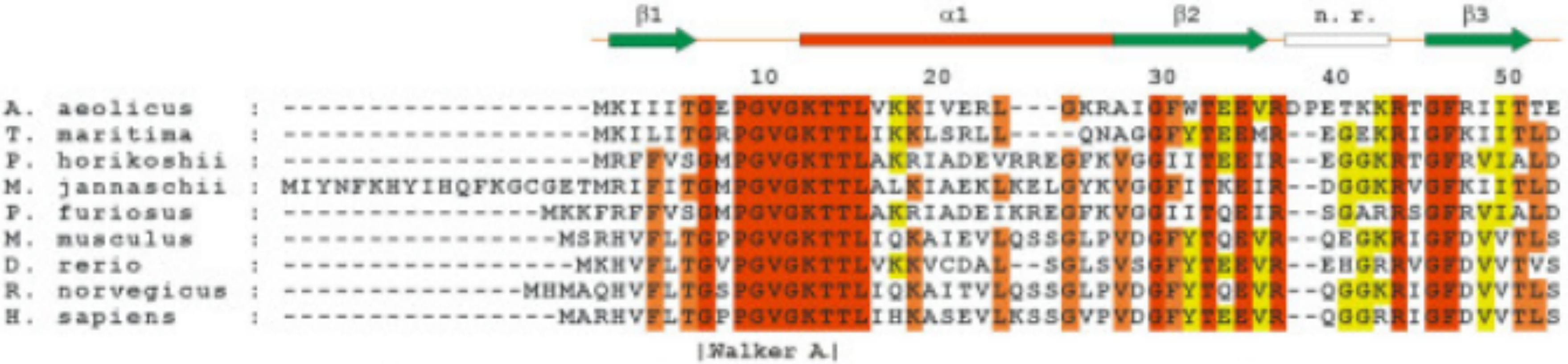
- Domain
  - ✓ Discrete structural unit that is assumed to fold independently of the rest of the protein and to have its own function.
  - ✓ It can be composed of 20 – 100s of amino acid residues.
  - ✓ Similar domains can be found in proteins with different functions





# Protein Sequence Features

- Motifs
  - ✓Short conserved regions and frequently are the most conserved regions of a domain. Motifs are critical for the domain to function – in enzymes, for example, the contain the active sites

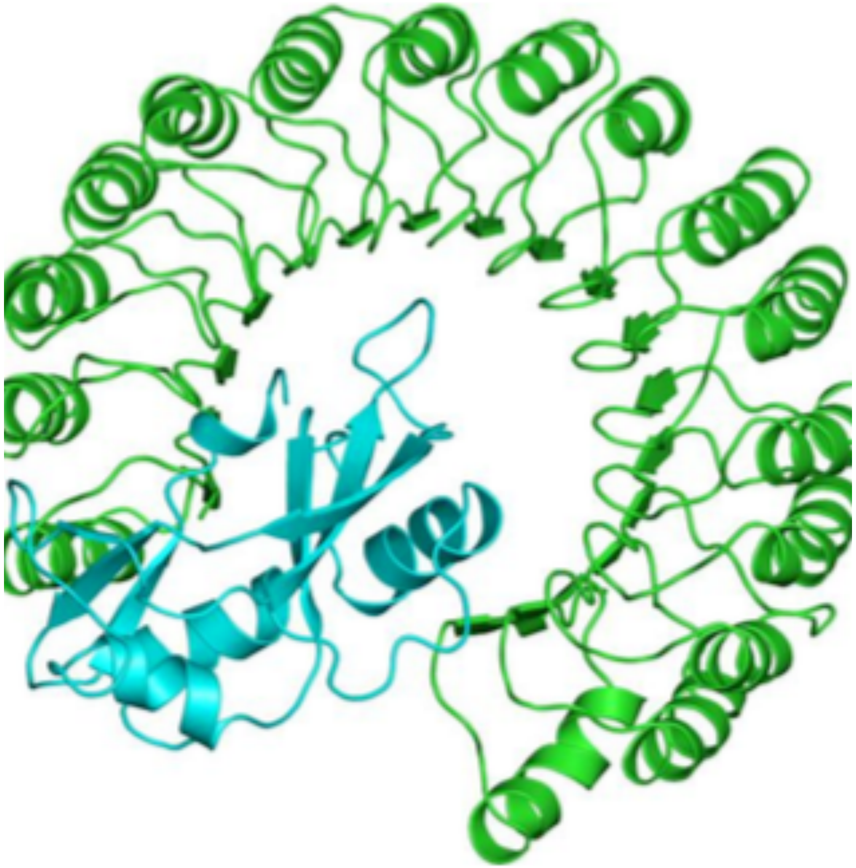


```
A. aeolicus    : ------------------MKIIITGEPGVGKTTLVKKIVERL---GKRAIGFWTEEVRDPETKKRTGFRIIITTE
T. maritima    : ------------------MKILITGRPGVGKTTLIKKLSRLL----QNAGGFYTEEMR--EGEKRIGFKIITLD
P. horikoshii  : ------------------MRFFVSGMPGVGKTTLAKRIADEVRREGFKVGGIITEEIR--EGGKRTGFRVIALD
M. jannaschii  : MIYNFKHYIHQFKGCGETMRIFITGMPGVGKTTLALKIAEKLKELGYKVGGFITKEIR--DGGKRVGFKIITLD
P. furiosus    : ---------------MKKFRFFVSGMPGVGKTTLAKRIADEIKREGFKVGGIITQEIR--SGARRSGFRVIALD
M. musculus    : ----------------MSRHVFLTGPPGVGKTTLIQKAIEVLQSSGLPVDGFYTQEVR--QEGKRIGFDVVTLS
D. rerio       : -----------------MKHVFLTGVPGVGKTTLVKKVCDAL--SGLSVSGFYTEEVR--EHGRRVGFDVVTVS
R. norvegicus  : --------------MHMAQHVFLTGSPGVGKTTLIQKAITVLQSSGLPVDGFYTQEVR--QGGKRIGFDVVTLS
H. sapiens     : ----------------MARHVFLTGPPGVGKTTLIHKASEVLKSSGVPVDGFYTEEVR--QGGRRIGFDVVTLS
```

H3ABioNet
Pan African Bioinformatics Network for H3Africa

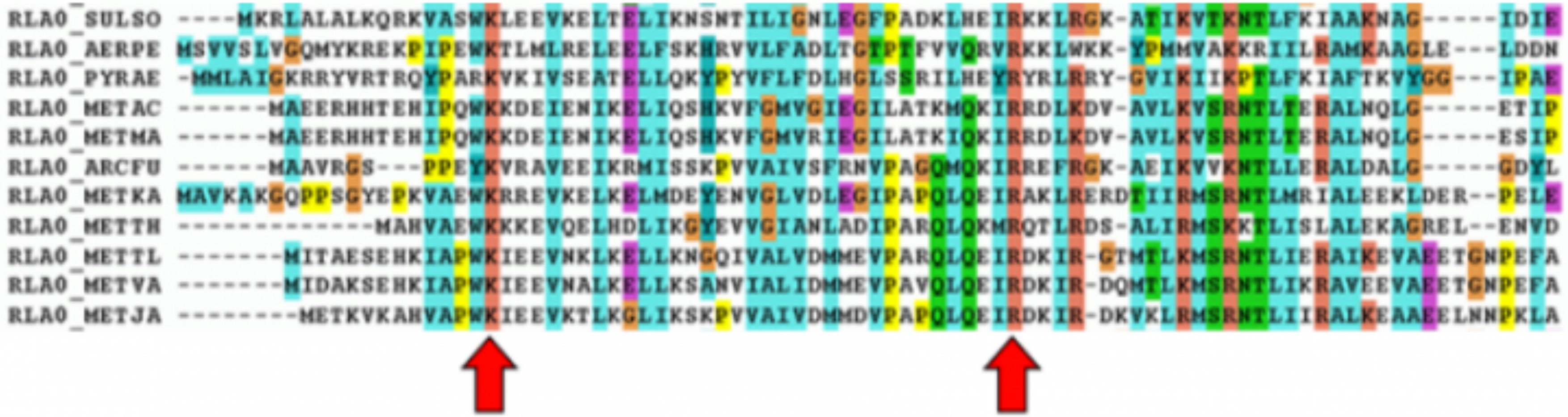
# Protein Sequence Features

- Repeat
  - ✓ Stretch of amino acid sequence that gets repeated a number of times along the length of the sequence. Many domains are constituted from repeats
  - ✓ Repeats may contain binding sites and contribute to structural properties of the protein



# Protein Sequence Features

- Consensus site/post-translation modification site (PTM)
  - ✓ A conserved position(s) among homologous sequences. Position can be theoretically modified, for example, by phosphorylation or glycosylation. An asparagine followed by any amino acid followed by serine or threonine, for example, is a consensus site for N-linked glycosylation

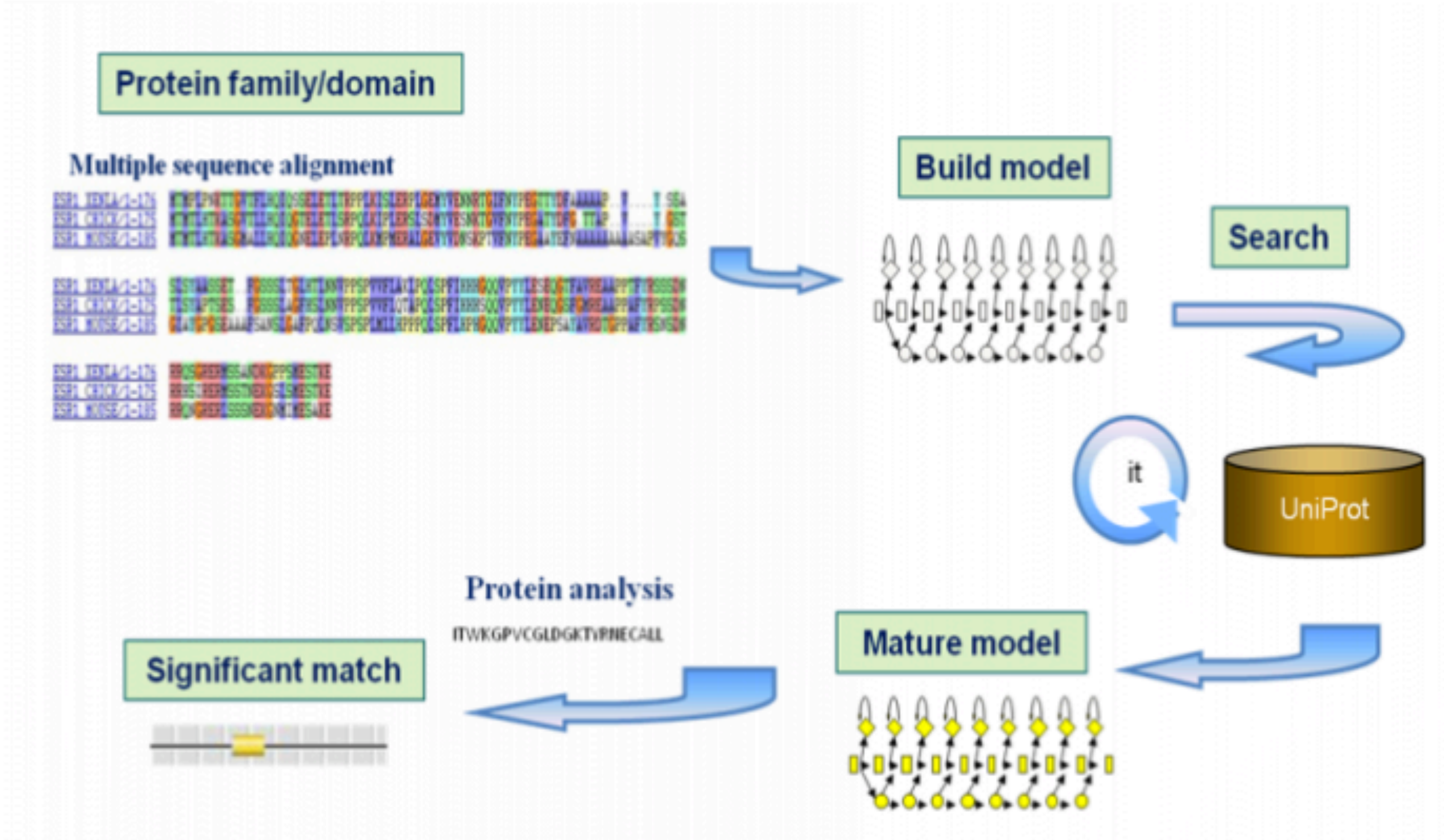


```
RLA0_SULSO  ----MKRLALALKQRKVASWKLEEVKELTELIKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFKIAAKNAG-----IDIE
RLA0_AERPE  MSVVSLVGQMYKREKPIPEWKTLMLRELEELFSKHRVVLFADLTGTPTFVVQRVRKKLWKK-YPMMVAKKRIILRAMKAAGLE---LDDN
RLA0_PYRAE  -MMLAIGKRRYVRTRQYPARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEYRYRLRRY-GVIKIIKPTLFKIAFTKVYGG---IPAE
RLA0_METAC  ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVGIEGILATKMQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ETIP
RLA0_METMA  ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVRIEGILATKIQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ESIP
RLA0_ARCFU  ------MAAVRGS---PPEYKVRAVEEIKRMISSKPVVAIVSFRNVPAGQMQKIRREFRGK-AEIKVVKNTLLERALDALG-----GDYL
RLA0_METKA  MAVKAKGQPPSGYEPKVAEWKRREVKELKELMDEYENVGLVDLEGIPAPQLQEIRAKLRERDTIIRMSRNTLMRIALEEKLDER--PELE
RLA0_METTH  -------------MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPARQLQKMRQTLRDS-ALIRMSKKTLISLALEKAGREL--ENVD
RLA0_METTL  -------MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPARQLQEIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNPEFA
RLA0_METVA  -------MIDAKSEHKIAPWKIEEVNALKELLKSANVIALIDMMEVPAVQLQEIRDKIR-DQMTLKMSRNTLIKRAVEEVAEETGNPEFA
RLA0_METJA  --------METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLQEIRDKIR-DKVKLRMSRNTLIIRALKEAAEELNNPKLA
```

# Protein Signatures

- Protein signature are computational models used to classify protein properties:
  - ✓ Protein families
  - ✓ Domains
  - ✓ Conserved sites
  - ✓ Protein sequence features
- Built from multiple sequence alignments (MSA) of proteins
  - ✓ Proteins belonging to the same family or sharing a domain
  - ✓ Predictive model built
  - ✓ Trained on new data
  - ✓ Used for protein sequence analysis

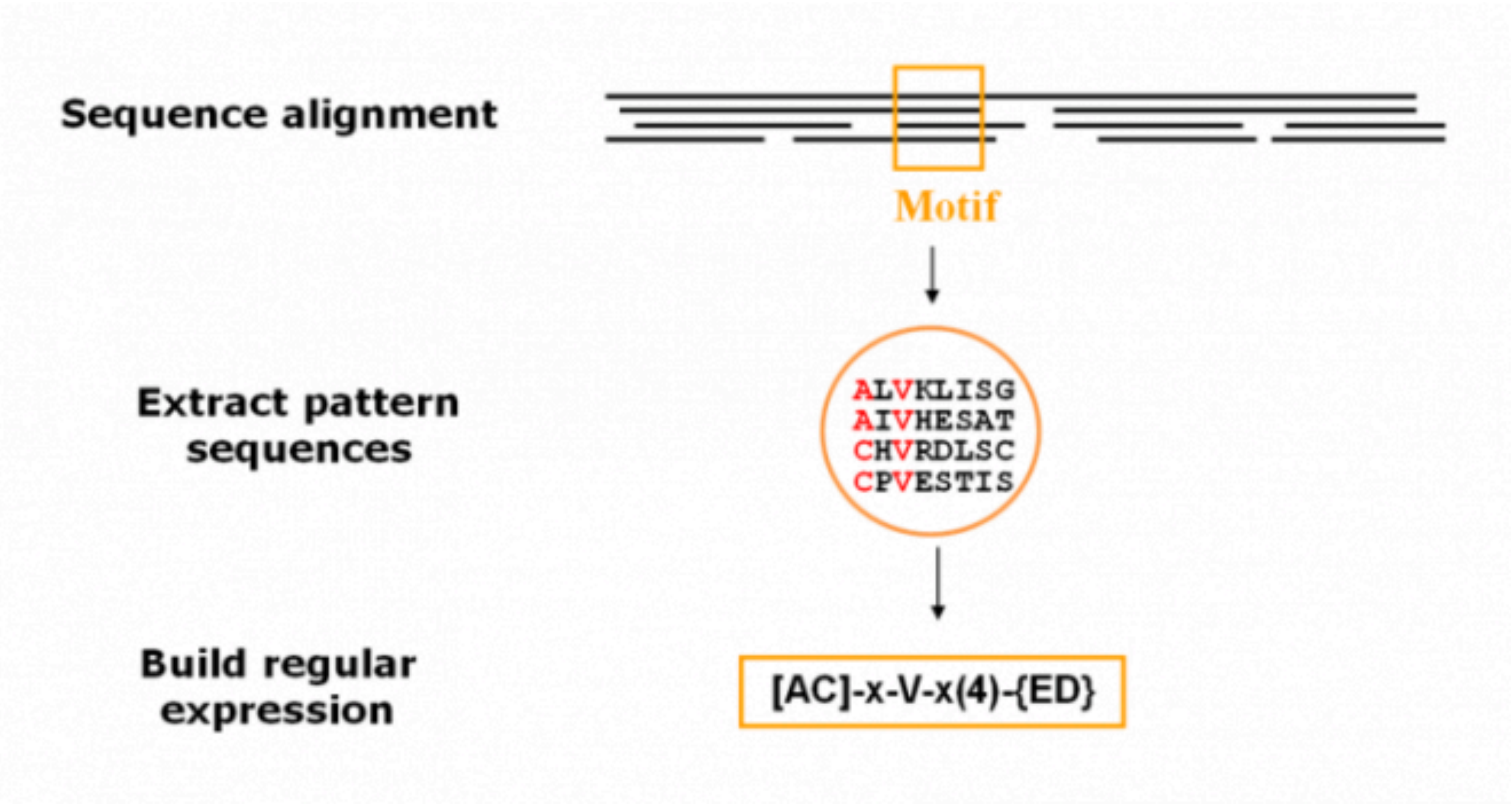
# Protein Signature Models



# Types of Protein Signatures

- Pattern
  - ✓ Functional sites such as binding/active sites usually consist of a few conserved amino acids
  - ✓ These conserved patterns are identified from MSAs
  - ✓ Modeled as a short, contiguous stretch of protein using regular expressions. E.g D[DE]X is a pattern composed of amino acid D, followed by either D or E, followed by any amino acid

# Types of Protein Signatures

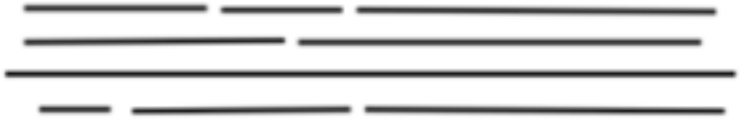


# Types of Protein Signatures

- Profile
  - ✓Used to model protein families and domains
  - ✓A profile is built from MSAs and is a matrix or table that describes the probability of finding a particular amino acid at at certain position.
  - ✓The matrix is generated based on the frequency at which an amino acid occurs at each position.
  - ✓Hidden Markov Models (HMMs) can be used to create a more powerful statistical profile from MSAs

# Types of Protein Signatures

Sequence alignment



Residue frequency at each position

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 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 |

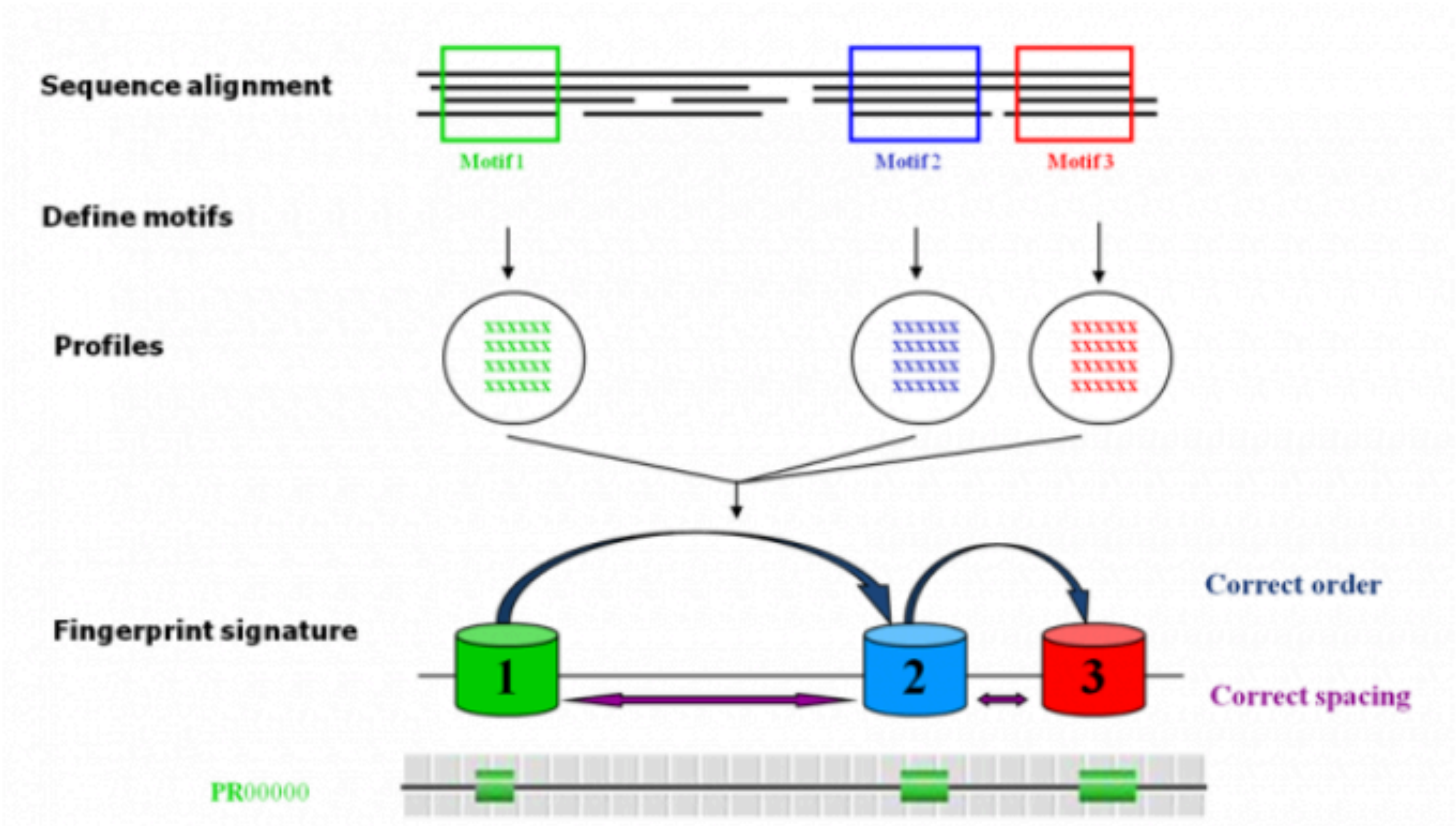
Scoring matrix

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 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 |

# Types of Protein Signatures

- Fingerprints
  - ✓Used to identify several conserved motifs
  - ✓Multiple short conserved motifs, are drawn from sequence alignments.
  - ✓Each motif is converted into an individual profile to create a fingerprint signature.
  - ✓Useful for identifying small differences between closely related proteins.

# Types of Protein Signatures



# PROTEIN RESOURCES

# Pfam

- Collection of protein families and domains
- Represented by
  - ✓Multiple sequence alignments
  - ✓Hidden Markov Models (HMMs)

# Pfam

- Two components to Pfam:
  - Pfam-A entries: High quality, manually curated families
  - Pfam-B entries: Automatically generated
- Generation of higher-level groupings of related families, known as clans (collection of Pfam-A entries which are related by similarity of sequence, structure or profile-HMM
- http://pfam.xfam.org

# SMART

- Simple Modular Architecture Research Tool
  - ✓ Identification and annotation of protein domains
  - ✓ Analysis of protein domain architectures
  - ✓ Manually curated models for the prediction of protein domains
  - ✓ http://smart.embl-heidelberg.de

# PRINTS

- Collection of protein family "fingerprints" (group of conserved motifs used to characterise a protein family)
- Prediction of functional families in uncharacterised protein sequences
- http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/index.php

# ExPASY (https://www.expasy.org/)

- Expasy (Swiss Institute of Bioinformatics)
  - ✓ UniProt, PROSITE, homology modelling, docking, many many other tools doing protein sequences and identication, mass spectrometry and 2-DE data,  protein characterisation and function families,  patterns and profiles, post-translational modication, protein structure, protein-protein interaction, similarity search/alignment, drug design, molecular modelling

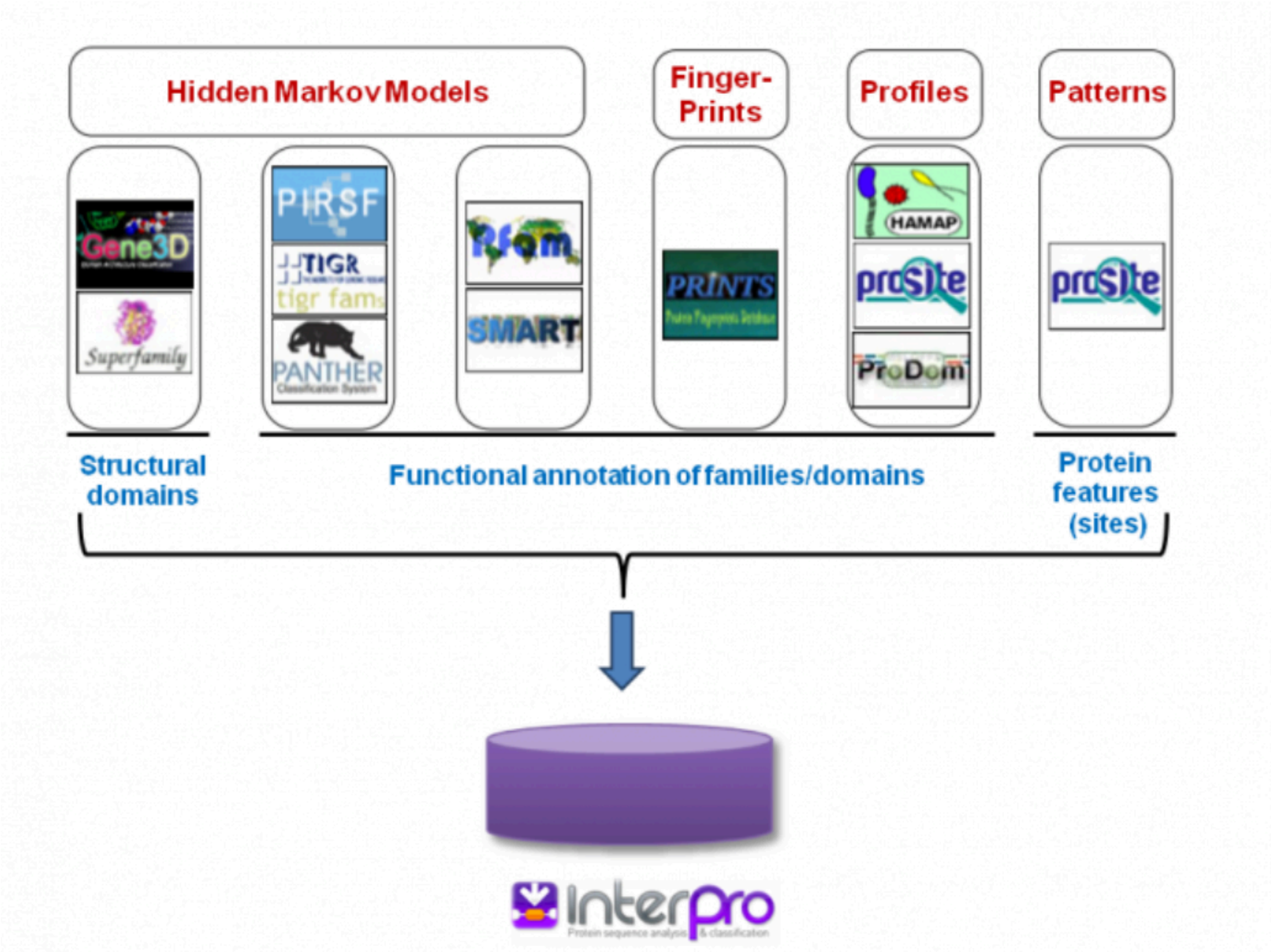
# Protein Information Resource

- PIR
  - ✓Protein ontology
  - ✓ProClass: Reports for UniProtKB
  - ✓ProLink: Literature, Text Mining
  - ✓http://pir.georgetown.edu/

# InterPro

- Designed to integrate signature databases
  - ✓Protein families, domain and functional sites
  - ✓http://www.ebi.ac.uk/interpro/

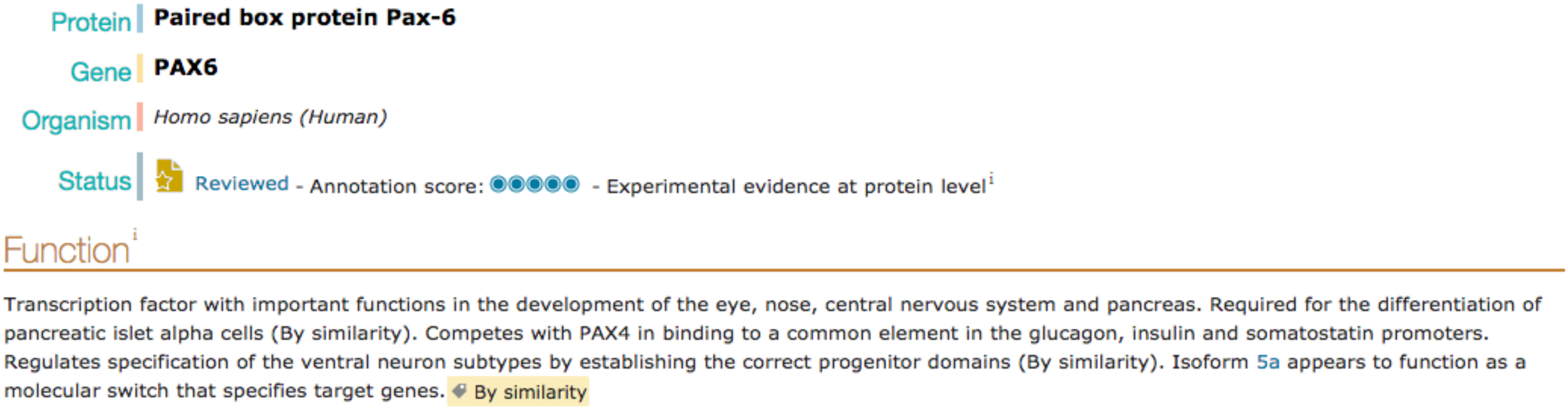
# InterPro


Hidden Markov Models
Finger-Prints
Profiles
Patterns
Gene3D
Superfamily
PIRSF
TIGR
tigr fams
PANTHER
Pfam
SMART
PRINTS
HAMAP
prosite
ProDom
prosite
Structural domains
Functional annotation of families/domains
Protein features (sites)
InterPro
Protein sequence analysis & classification

# InterPro

- Signatures describing the same protein family, domain or functional site grouped into a single InterPro identifier
- InterProScan tool
  - ✓ Integrate signature recognition methods into a single application
  - ✓ Find signatures that match a protein sequence of interest
  - ✓ Web-based version of InterProScan
  - ✓ http://www.ebi.ac.uk/interpro/

# Uniprot – Example Pax-6 protein



Protein | **Paired box protein Pax-6**

Gene | **PAX6**

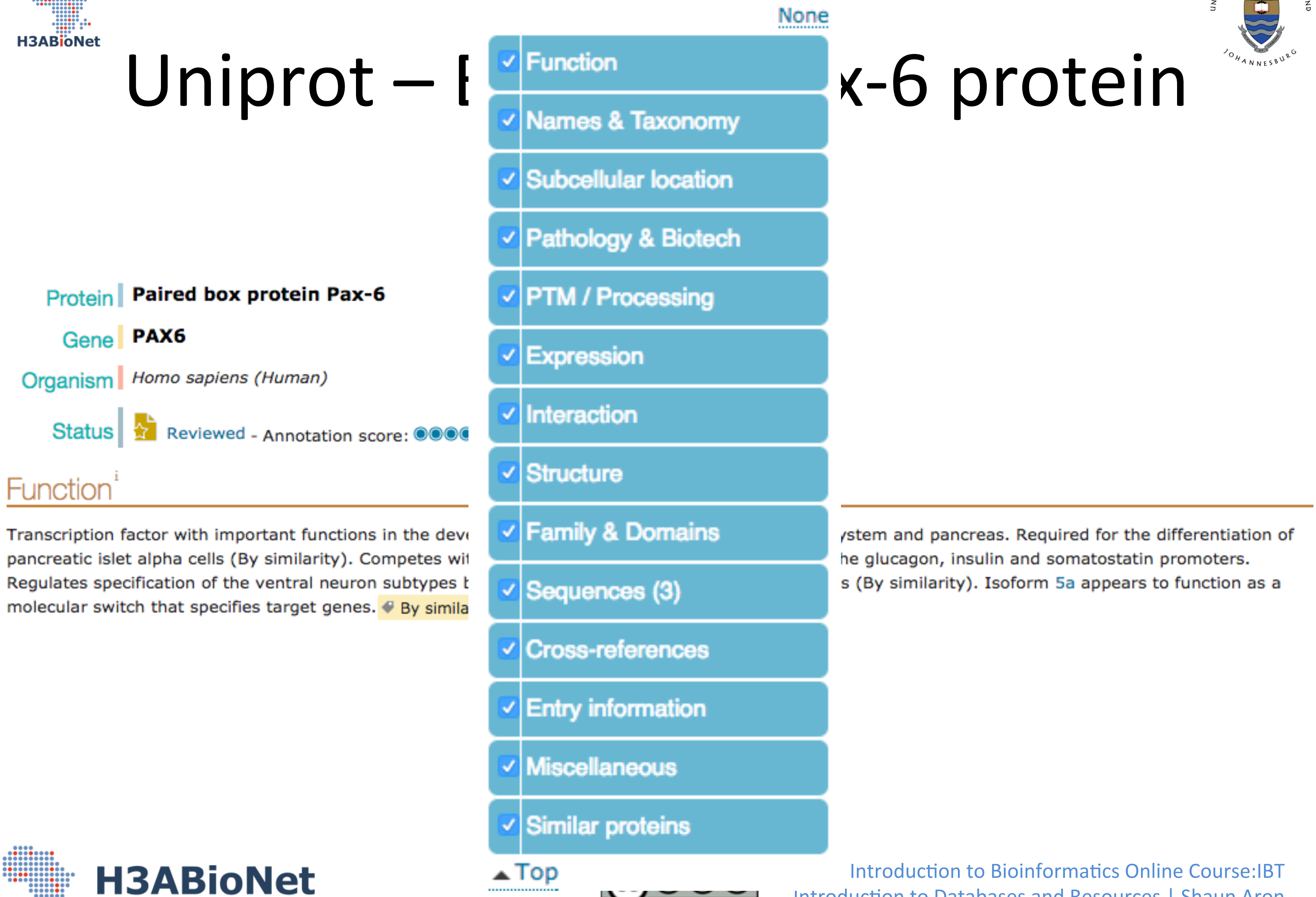
Organism | *Homo sapiens (Human)*

Status | Reviewed - Annotation score: ●●●●● - Experimental evidence at protein level[i]

## Function[i]

Transcription factor with important functions in the development of the eye, nose, central nervous system and pancreas. Required for the differentiation of pancreatic islet alpha cells (By similarity). Competes with PAX4 in binding to a common element in the glucagon, insulin and somatostatin promoters. Regulates specification of the ventral neuron subtypes by establishing the correct progenitor domains (By similarity). Isoform 5a appears to function as a molecular switch that specifies target genes. By similarity

# Uniprot – x-6 protein



Protein | **Paired box protein Pax-6**

Gene | **PAX6**

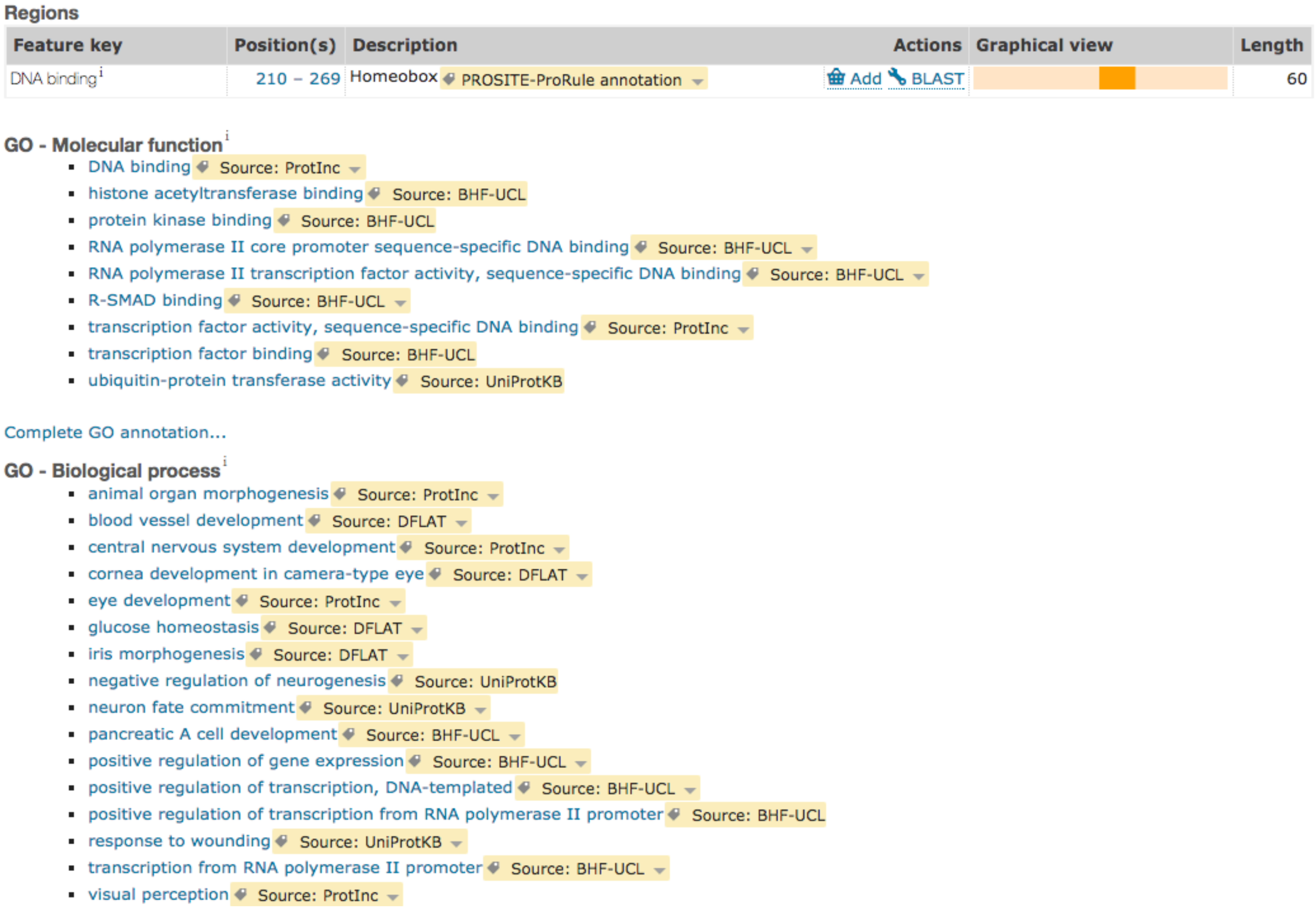
Organism | *Homo sapiens (Human)*

Status | Reviewed - Annotation score: ●●●●

## Function[i]

Transcription factor with important functions in the dev ystem and pancreas. Required for the differentiation of
pancreatic islet alpha cells (By similarity). Competes wi he glucagon, insulin and somatostatin promoters.
Regulates specification of the ventral neuron subtypes t s (By similarity). Isoform 5a appears to function as a
molecular switch that specifies target genes. By simila

# Uniprot – Example Pax-6 protein



### Regions

| Feature key | Position(s) | Description | Actions | Graphical view | Length |
|---|---|---|---|---|---|
| DNA binding i | 210 – 269 | Homeobox PROSITE-ProRule annotation | Add BLAST | | 60 |

### GO - Molecular function i

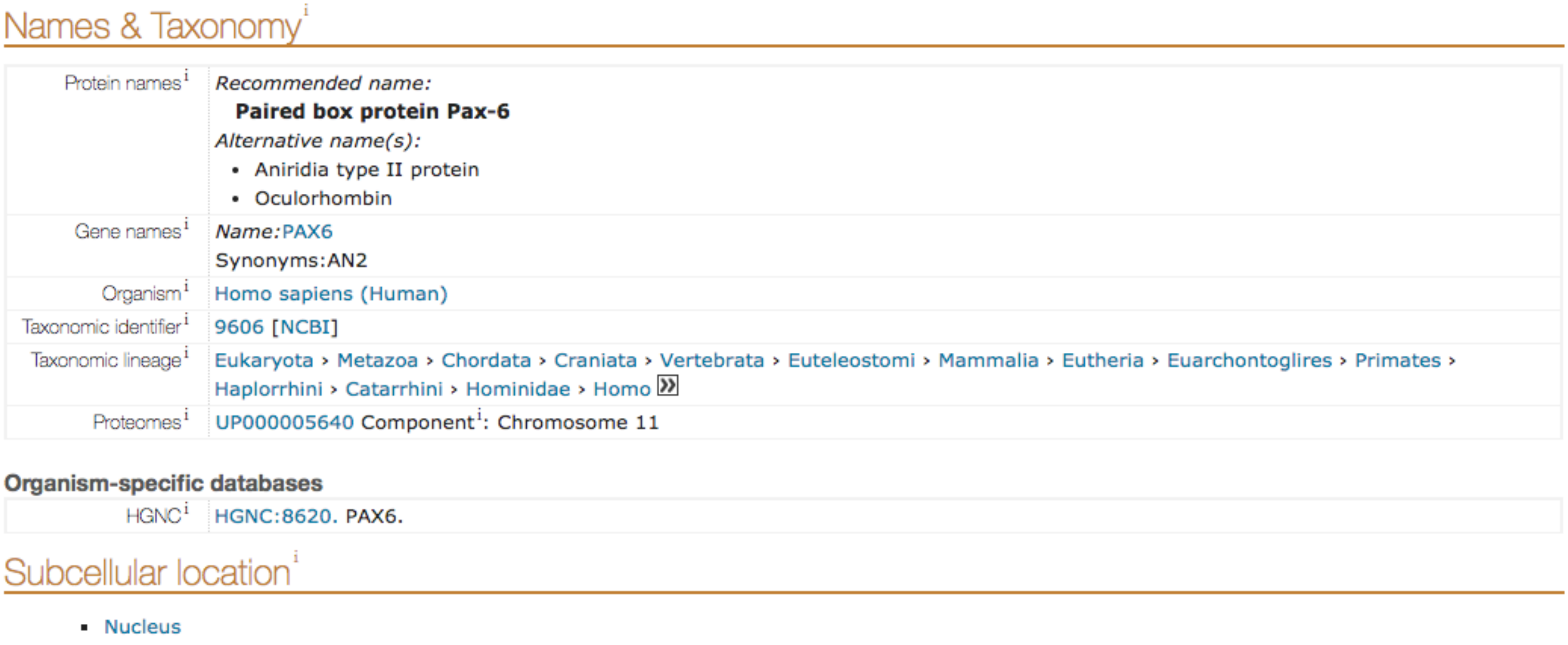
- DNA binding Source: ProtInc
- histone acetyltransferase binding Source: BHF-UCL
- protein kinase binding Source: BHF-UCL
- RNA polymerase II core promoter sequence-specific DNA binding Source: BHF-UCL
- RNA polymerase II transcription factor activity, sequence-specific DNA binding Source: BHF-UCL
- R-SMAD binding Source: BHF-UCL
- transcription factor activity, sequence-specific DNA binding Source: ProtInc
- transcription factor binding Source: BHF-UCL
- ubiquitin-protein transferase activity Source: UniProtKB

Complete GO annotation...

### GO - Biological process i

- animal organ morphogenesis Source: ProtInc
- blood vessel development Source: DFLAT
- central nervous system development Source: ProtInc
- cornea development in camera-type eye Source: DFLAT
- eye development Source: ProtInc
- glucose homeostasis Source: DFLAT
- iris morphogenesis Source: DFLAT
- negative regulation of neurogenesis Source: UniProtKB
- neuron fate commitment Source: UniProtKB
- pancreatic A cell development Source: BHF-UCL
- positive regulation of gene expression Source: BHF-UCL
- positive regulation of transcription, DNA-templated Source: BHF-UCL
- positive regulation of transcription from RNA polymerase II promoter Source: BHF-UCL
- response to wounding Source: UniProtKB
- transcription from RNA polymerase II promoter Source: BHF-UCL
- visual perception Source: ProtInc

# Uniprot – Example Pax-6 protein



## Names & Taxonomy

| | |
|---|---|
| Protein names | *Recommended name:*<br>**Paired box protein Pax-6**<br>*Alternative name(s):*<br>• Aniridia type II protein<br>• Oculorhombin |
| Gene names | *Name:*PAX6<br>Synonyms:AN2 |
| Organism | Homo sapiens (Human) |
| Taxonomic identifier | 9606 [NCBI] |
| Taxonomic lineage | Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Mammalia › Eutheria › Euarchontoglires › Primates › Haplorrhini › Catarrhini › Hominidae › Homo |
| Proteomes | UP000005640 Component: Chromosome 11 |

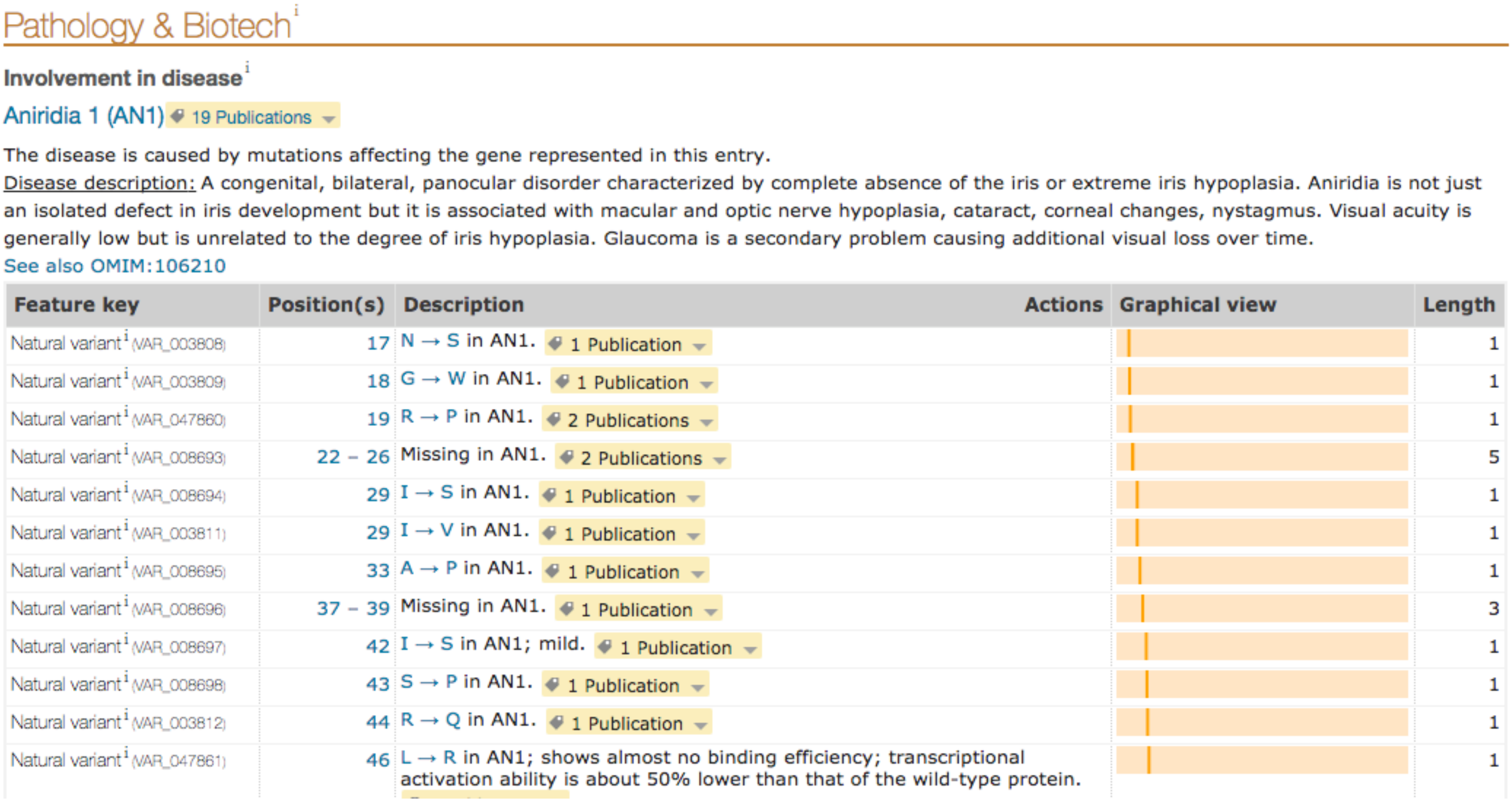
**Organism-specific databases**

| | |
|---|---|
| HGNC | HGNC:8620. PAX6. |

## Subcellular location

- Nucleus

# Uniprot – Example Pax-6 protein



## Pathology & Biotech

### Involvement in disease

Aniridia 1 (AN1) 19 Publications

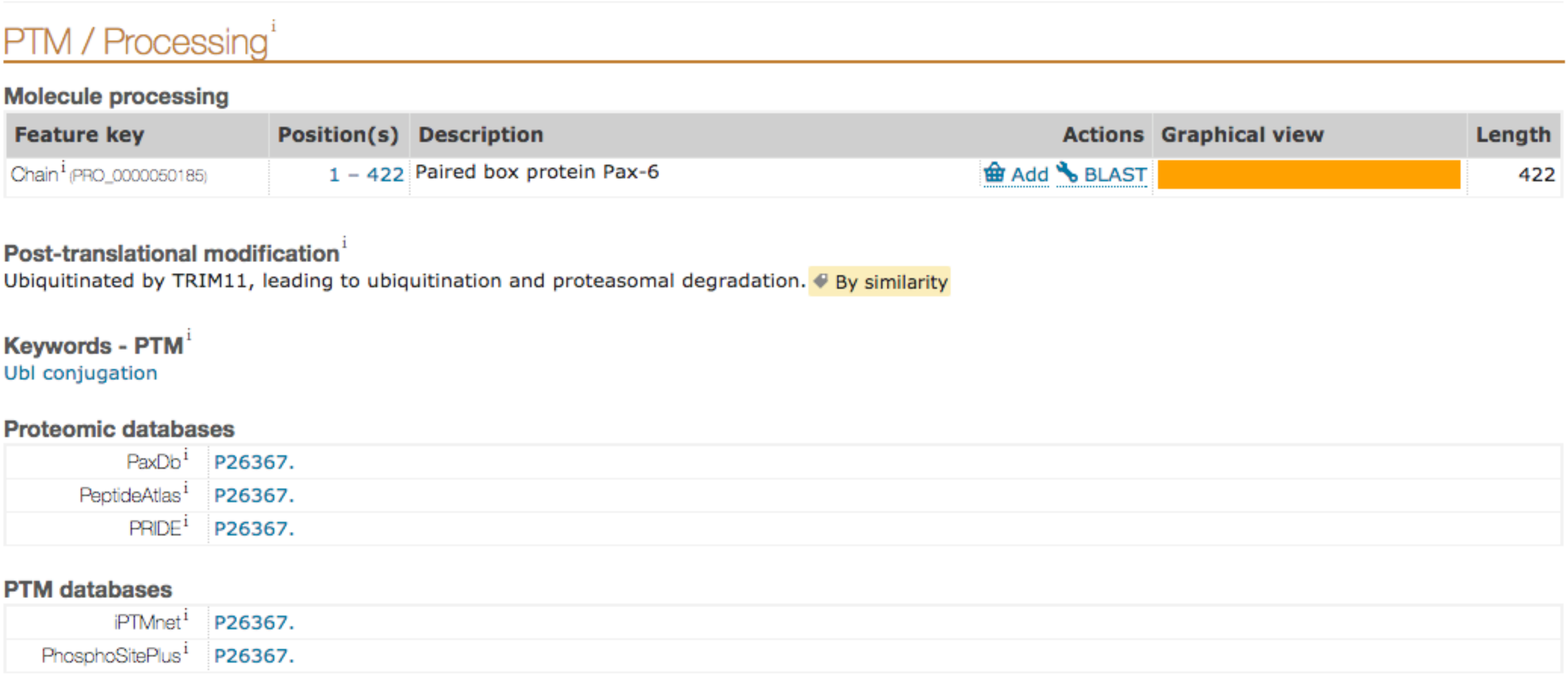
The disease is caused by mutations affecting the gene represented in this entry.

Disease description: A congenital, bilateral, panocular disorder characterized by complete absence of the iris or extreme iris hypoplasia. Aniridia is not just an isolated defect in iris development but it is associated with macular and optic nerve hypoplasia, cataract, corneal changes, nystagmus. Visual acuity is generally low but is unrelated to the degree of iris hypoplasia. Glaucoma is a secondary problem causing additional visual loss over time.

See also OMIM:106210

| Feature key | Position(s) | Description | Actions | Graphical view | Length |
|---|---|---|---|---|---|
| Natural variant (VAR_003808) | 17 | N → S in AN1. 1 Publication | | | 1 |
| Natural variant (VAR_003809) | 18 | G → W in AN1. 1 Publication | | | 1 |
| Natural variant (VAR_047860) | 19 | R → P in AN1. 2 Publications | | | 1 |
| Natural variant (VAR_008693) | 22 – 26 | Missing in AN1. 2 Publications | | | 5 |
| Natural variant (VAR_008694) | 29 | I → S in AN1. 1 Publication | | | 1 |
| Natural variant (VAR_003811) | 29 | I → V in AN1. 1 Publication | | | 1 |
| Natural variant (VAR_008695) | 33 | A → P in AN1. 1 Publication | | | 1 |
| Natural variant (VAR_008696) | 37 – 39 | Missing in AN1. 1 Publication | | | 3 |
| Natural variant (VAR_008697) | 42 | I → S in AN1; mild. 1 Publication | | | 1 |
| Natural variant (VAR_008698) | 43 | S → P in AN1. 1 Publication | | | 1 |
| Natural variant (VAR_003812) | 44 | R → Q in AN1. 1 Publication | | | 1 |
| Natural variant (VAR_047861) | 46 | L → R in AN1; shows almost no binding efficiency; transcriptional activation ability is about 50% lower than that of the wild-type protein. | | | 1 |

# Uniprot – Example Pax-6 protein



## PTM / Processing

### Molecule processing

| Feature key | Position(s) | Description | Actions | Graphical view | Length |
|---|---|---|---|---|---|
| Chain (PRO_0000050185) | 1 – 422 | Paired box protein Pax-6 | Add BLAST | 🟧 | 422 |

### Post-translational modification

Ubiquitinated by TRIM11, leading to ubiquitination and proteasomal degradation. By similarity

### Keywords - PTM

Ubl conjugation

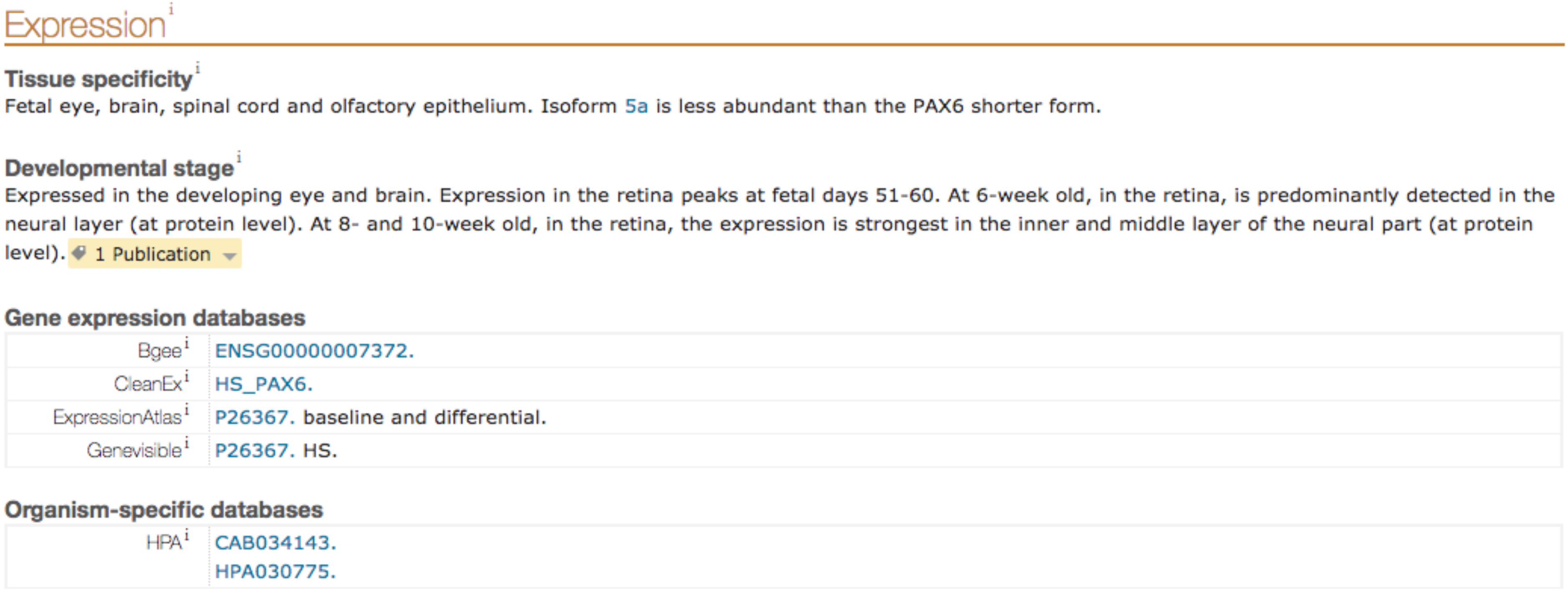
### Proteomic databases

| | |
|---|---|
| PaxDb | P26367. |
| PeptideAtlas | P26367. |
| PRIDE | P26367. |

### PTM databases

| | |
|---|---|
| iPTMnet | P26367. |
| PhosphoSitePlus | P26367. |

# Uniprot – Example Pax-6 protein



## Expression

### Tissue specificity

Fetal eye, brain, spinal cord and olfactory epithelium. Isoform 5a is less abundant than the PAX6 shorter form.

### Developmental stage

Expressed in the developing eye and brain. Expression in the retina peaks at fetal days 51-60. At 6-week old, in the retina, is predominantly detected in the neural layer (at protein level). At 8- and 10-week old, in the retina, the expression is strongest in the inner and middle layer of the neural part (at protein level). 1 Publication

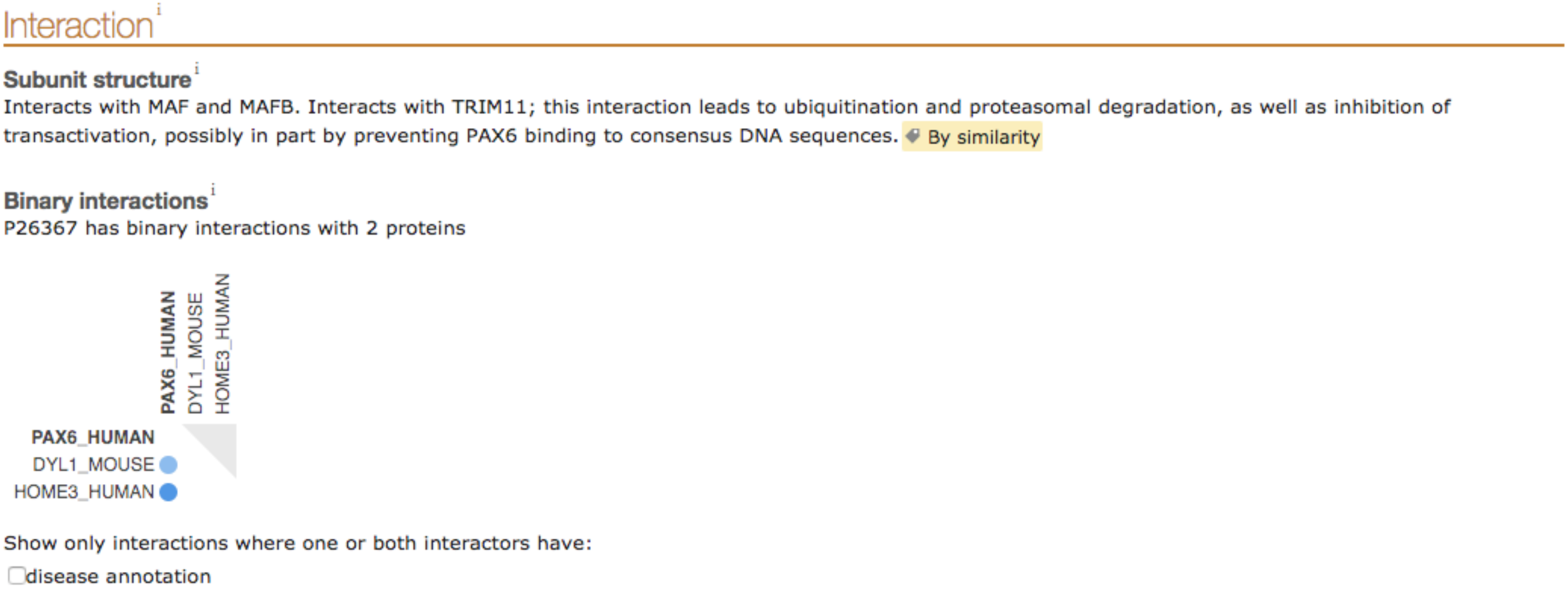
### Gene expression databases

| | |
|---|---|
| Bgee | ENSG00000007372. |
| CleanEx | HS_PAX6. |
| ExpressionAtlas | P26367. baseline and differential. |
| Genevisible | P26367. HS. |

### Organism-specific databases

| | |
|---|---|
| HPA | CAB034143.<br>HPA030775. |

# Uniprot – Example Pax-6 protein



## Interaction

### Subunit structure

Interacts with MAF and MAFB. Interacts with TRIM11; this interaction leads to ubiquitination and proteasomal degradation, as well as inhibition of transactivation, possibly in part by preventing PAX6 binding to consensus DNA sequences. By similarity

### Binary interactions

P26367 has binary interactions with 2 proteins

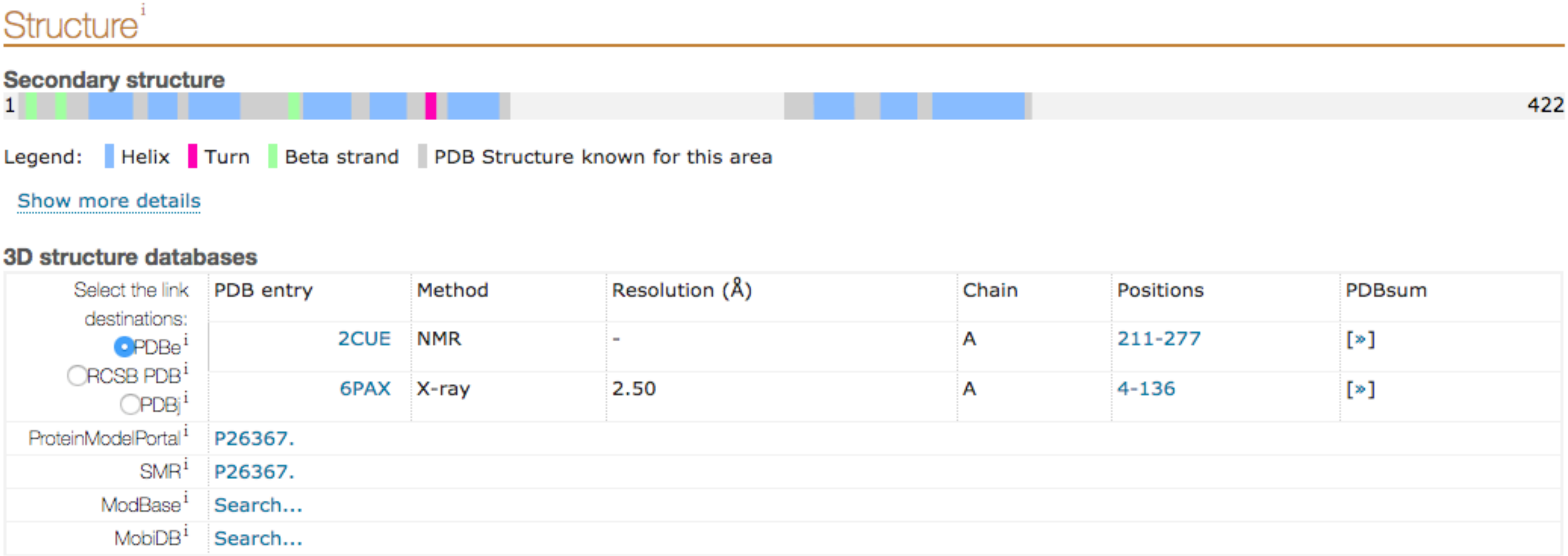
| | **PAX6_HUMAN** | DYL1_MOUSE | HOME3_HUMAN |
|---|---|---|---|
| **PAX6_HUMAN** | | | |
| DYL1_MOUSE | ● | | |
| HOME3_HUMAN | ● | | |

Show only interactions where one or both interactors have:

☐ disease annotation

H3ABioNet

# Uniprot – Example Pax-6 protein



## Structure

### Secondary structure

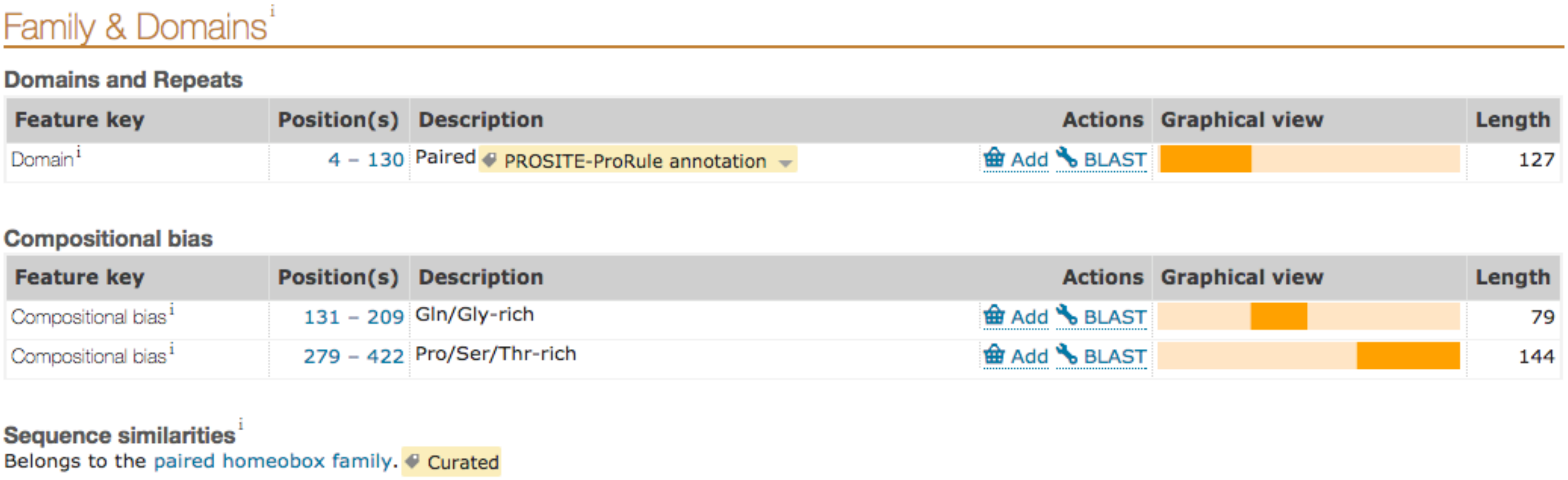
Legend: 🟦 Helix 🟪 Turn 🟩 Beta strand ⬜ PDB Structure known for this area

Show more details

### 3D structure databases

| Select the link destinations:<br>● PDBe<br>○ RCSB PDB<br>○ PDBj | PDB entry | Method | Resolution (Å) | Chain | Positions | PDBsum |
|---|---|---|---|---|---|---|
| | 2CUE | NMR | - | A | 211-277 | [»] |
| | 6PAX | X-ray | 2.50 | A | 4-136 | [»] |
| ProteinModelPortal | P26367. | | | | | |
| SMR | P26367. | | | | | |
| ModBase | Search... | | | | | |
| MobiDB | Search... | | | | | |

# Uniprot – Example Pax-6 protein



## Family & Domains[i]

### Domains and Repeats

| Feature key | Position(s) | Description | Actions | Graphical view | Length |
|---|---|---|---|---|---|
| Domain[i] | 4 – 130 | Paired PROSITE-ProRule annotation | Add BLAST | | 127 |

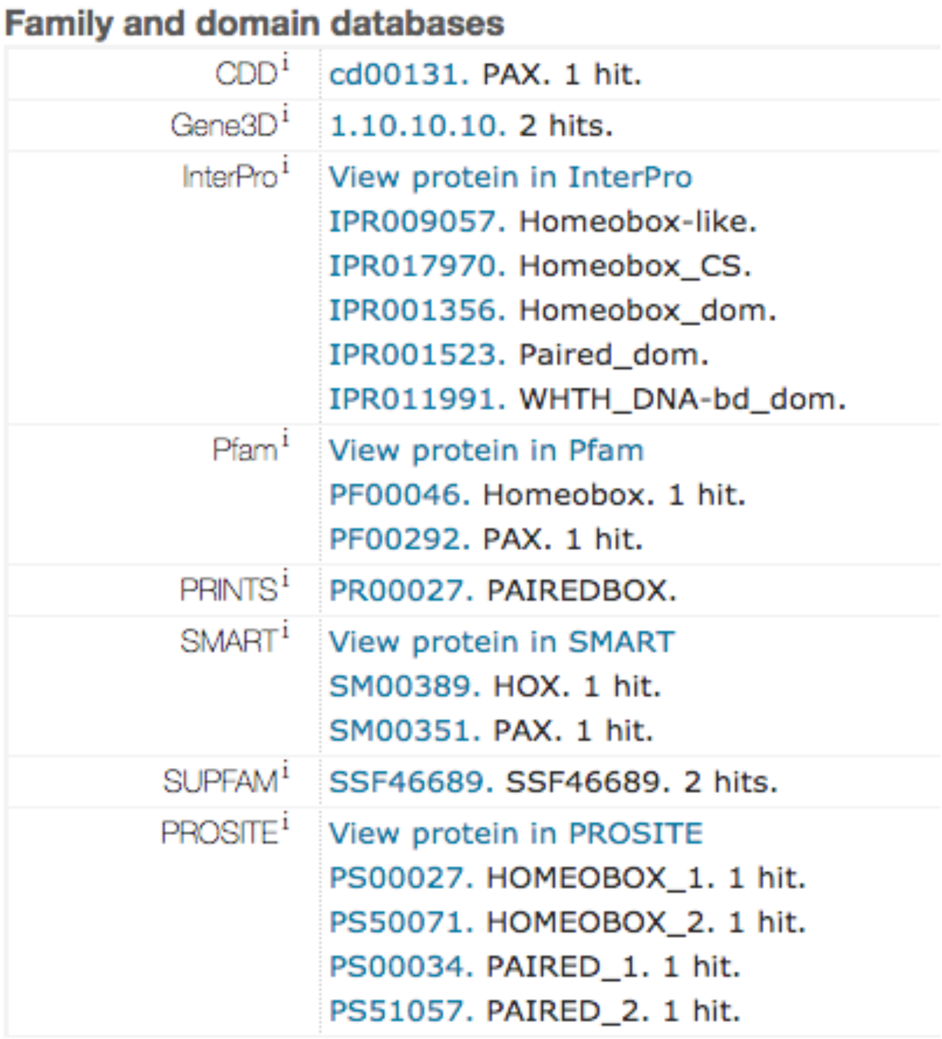
### Compositional bias

| Feature key | Position(s) | Description | Actions | Graphical view | Length |
|---|---|---|---|---|---|
| Compositional bias[i] | 131 – 209 | Gln/Gly-rich | Add BLAST | | 79 |
| Compositional bias[i] | 279 – 422 | Pro/Ser/Thr-rich | Add BLAST | | 144 |

### Sequence similarities[i]

Belongs to the paired homeobox family. Curated

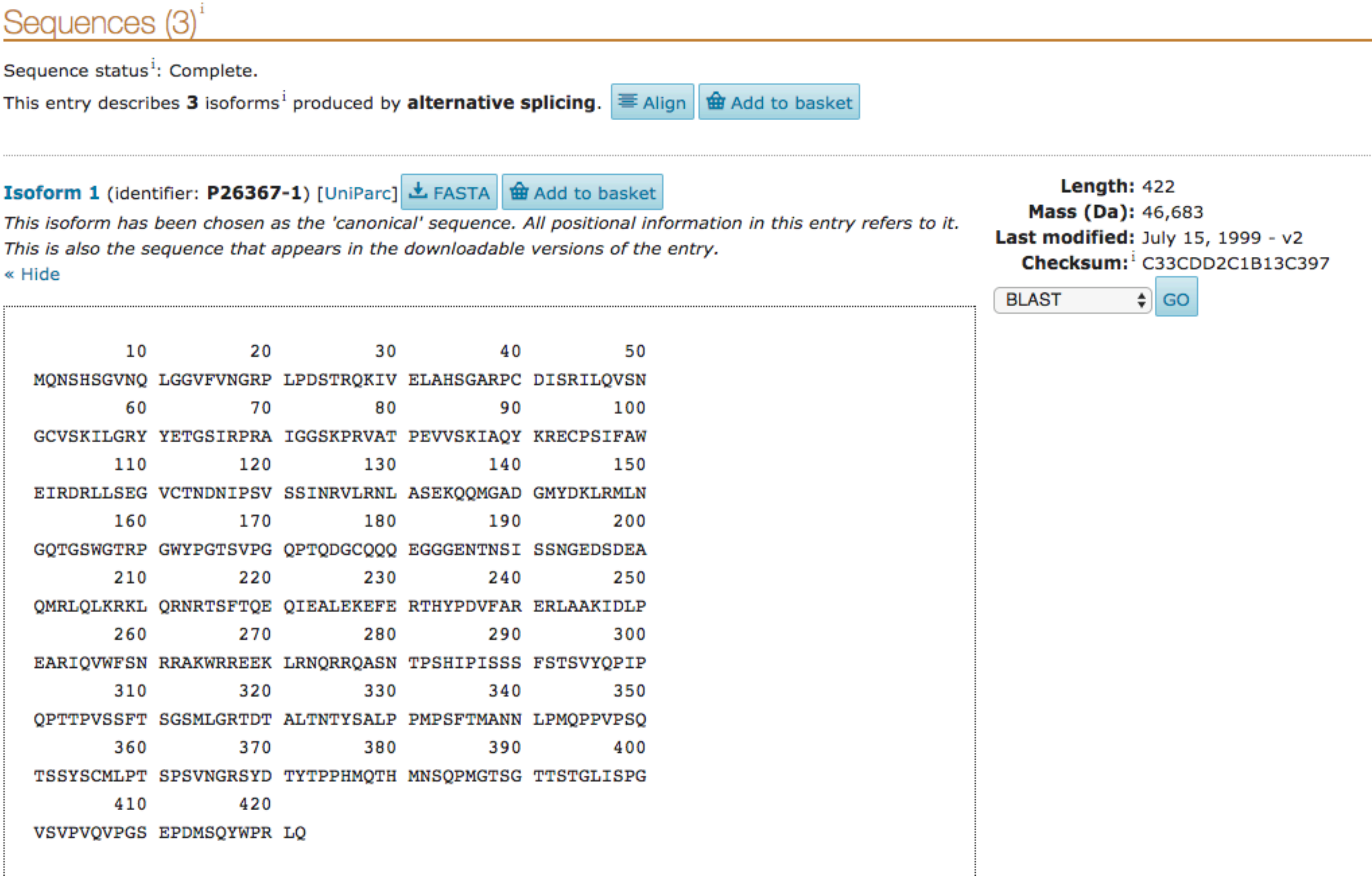
# Uniprot – Example Pax-6 protein



## Family and domain databases

| | |
|---|---|
| CDD[i] | cd00131. PAX. 1 hit. |
| Gene3D[i] | 1.10.10.10. 2 hits. |
| InterPro[i] | View protein in InterPro<br>IPR009057. Homeobox-like.<br>IPR017970. Homeobox_CS.<br>IPR001356. Homeobox_dom.<br>IPR001523. Paired_dom.<br>IPR011991. WHTH_DNA-bd_dom. |
| Pfam[i] | View protein in Pfam<br>PF00046. Homeobox. 1 hit.<br>PF00292. PAX. 1 hit. |
| PRINTS[i] | PR00027. PAIREDBOX. |
| SMART[i] | View protein in SMART<br>SM00389. HOX. 1 hit.<br>SM00351. PAX. 1 hit. |
| SUPFAM[i] | SSF46689. SSF46689. 2 hits. |
| PROSITE[i] | View protein in PROSITE<br>PS00027. HOMEOBOX_1. 1 hit.<br>PS50071. HOMEOBOX_2. 1 hit.<br>PS00034. PAIRED_1. 1 hit.<br>PS51057. PAIRED_2. 1 hit. |

# Uniprot – Example Pax-6 protein



## Sequences (3)[i]

Sequence status[i]: Complete.

This entry describes **3** isoforms[i] produced by **alternative splicing**.

**Isoform 1** (identifier: **P26367-1**) [UniParc]

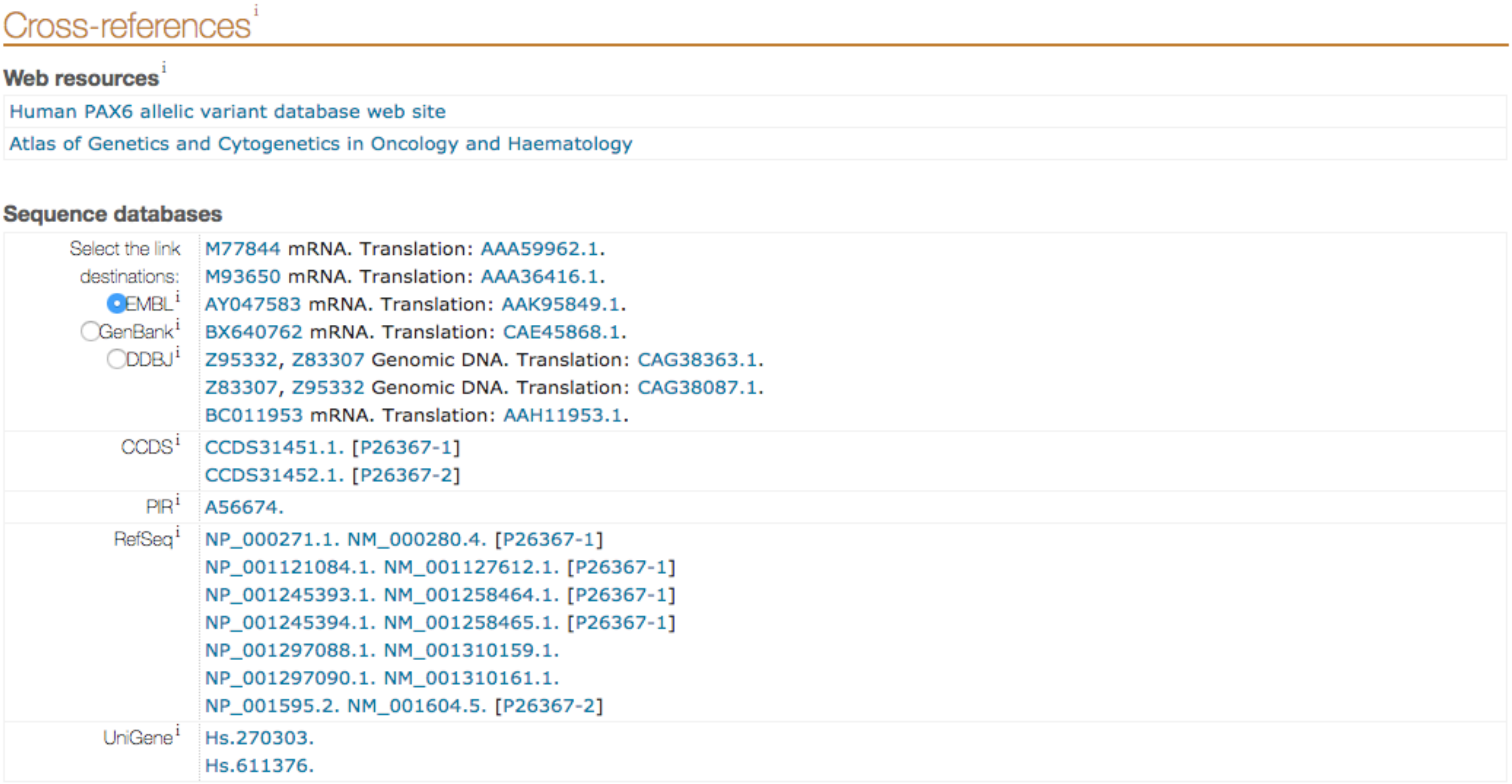
*This isoform has been chosen as the 'canonical' sequence. All positional information in this entry refers to it. This is also the sequence that appears in the downloadable versions of the entry.*

« Hide

**Length:** 422
**Mass (Da):** 46,683
**Last modified:** July 15, 1999 - v2
**Checksum:**[i] C33CDD2C1B13C397

```
        10         20         30         40         50
MQNSHSGVNQ LGGVFVNGRP LPDSTRQKIV ELAHSGARPC DISRILQVSN
        60         70         80         90        100
GCVSKILGRY YETGSIRPRA IGGSKPRVAT PEVVSKIAQY KRECPSIFAW
       110        120        130        140        150
EIRDRLLSEG VCTNDNIPSV SSINRVLRNL ASEKQQMGAD GMYDKLRMLN
       160        170        180        190        200
GQTGSWGTRP GWYPGTSVPG QPTQDGCQQQ EGGGENTNSI SSNGEDSDEA
       210        220        230        240        250
QMRLQLKRKL QRNRTSFTQE QIEALEKEFE RTHYPDVFAR ERLAAKIDLP
       260        270        280        290        300
EARIQVWFSN RRAKWRREEK LRNQRRQASN TPSHIPISSS FSTSVYQPIP
       310        320        330        340        350
QPTTPVSSFT SGSMLGRTDT ALTNTYSALP PMPSFTMANN LPMQPPVPSQ
       360        370        380        390        400
TSSYSCMLPT SPSVNGRSYD TYTPPHMQTH MNSQPMGTSG TTSTGLISPG
       410        420
VSVPVQVPGS EPDMSQYWPR LQ
```

# Uniprot – Example Pax-6 protein



## Cross-references

### Web resources

Human PAX6 allelic variant database web site

Atlas of Genetics and Cytogenetics in Oncology and Haematology

### Sequence databases

| | |
|---|---|
| Select the link destinations:<br>● EMBL<br>○ GenBank<br>○ DDBJ | M77844 mRNA. Translation: AAA59962.1.<br>M93650 mRNA. Translation: AAA36416.1.<br>AY047583 mRNA. Translation: AAK95849.1.<br>BX640762 mRNA. Translation: CAE45868.1.<br>Z95332, Z83307 Genomic DNA. Translation: CAG38363.1.<br>Z83307, Z95332 Genomic DNA. Translation: CAG38087.1.<br>BC011953 mRNA. Translation: AAH11953.1. |
| CCDS | CCDS31451.1. [P26367-1]<br>CCDS31452.1. [P26367-2] |
| PIR | A56674. |
| RefSeq | NP_000271.1. NM_000280.4. [P26367-1]<br>NP_001121084.1. NM_001127612.1. [P26367-1]<br>NP_001245393.1. NM_001258464.1. [P26367-1]<br>NP_001245394.1. NM_001258465.1. [P26367-1]<br>NP_001297088.1. NM_001310159.1.<br>NP_001297090.1. NM_001310161.1.<br>NP_001595.2. NM_001604.5. [P26367-2] |
| UniGene | Hs.270303.<br>Hs.611376. |