InterPro and InterProScan 5.0

EMBL-EBI

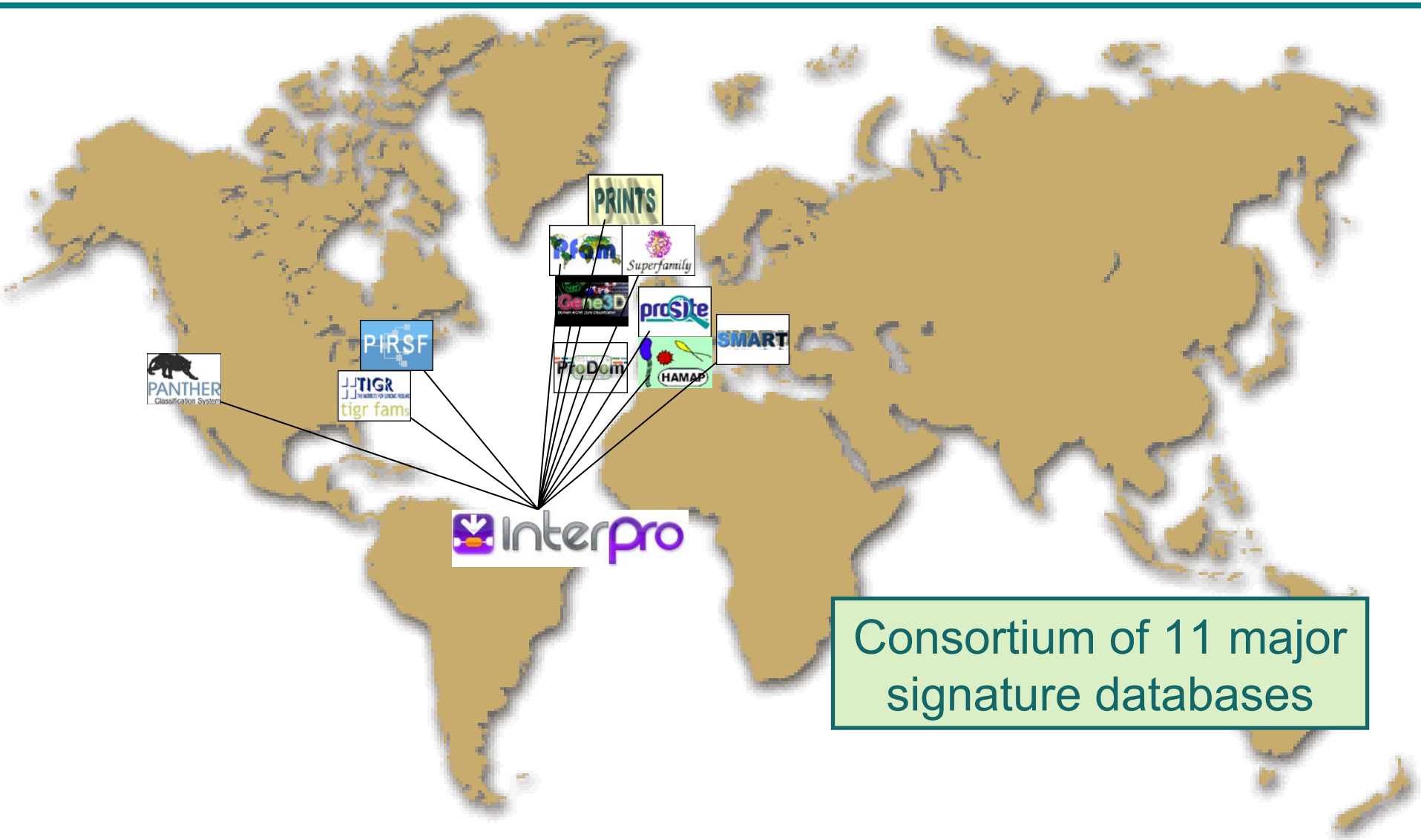
EBI is an Outstation of the European Molecular Biology Laboratory.

InterPro

- is a database that groups predictive protein signatures together
- 11 member databases

- single searchable resource
- provides functional analysis of proteins by classifying them into families and predicting domains and important sites
- Enables whole genome analysis

InterPro Consortium



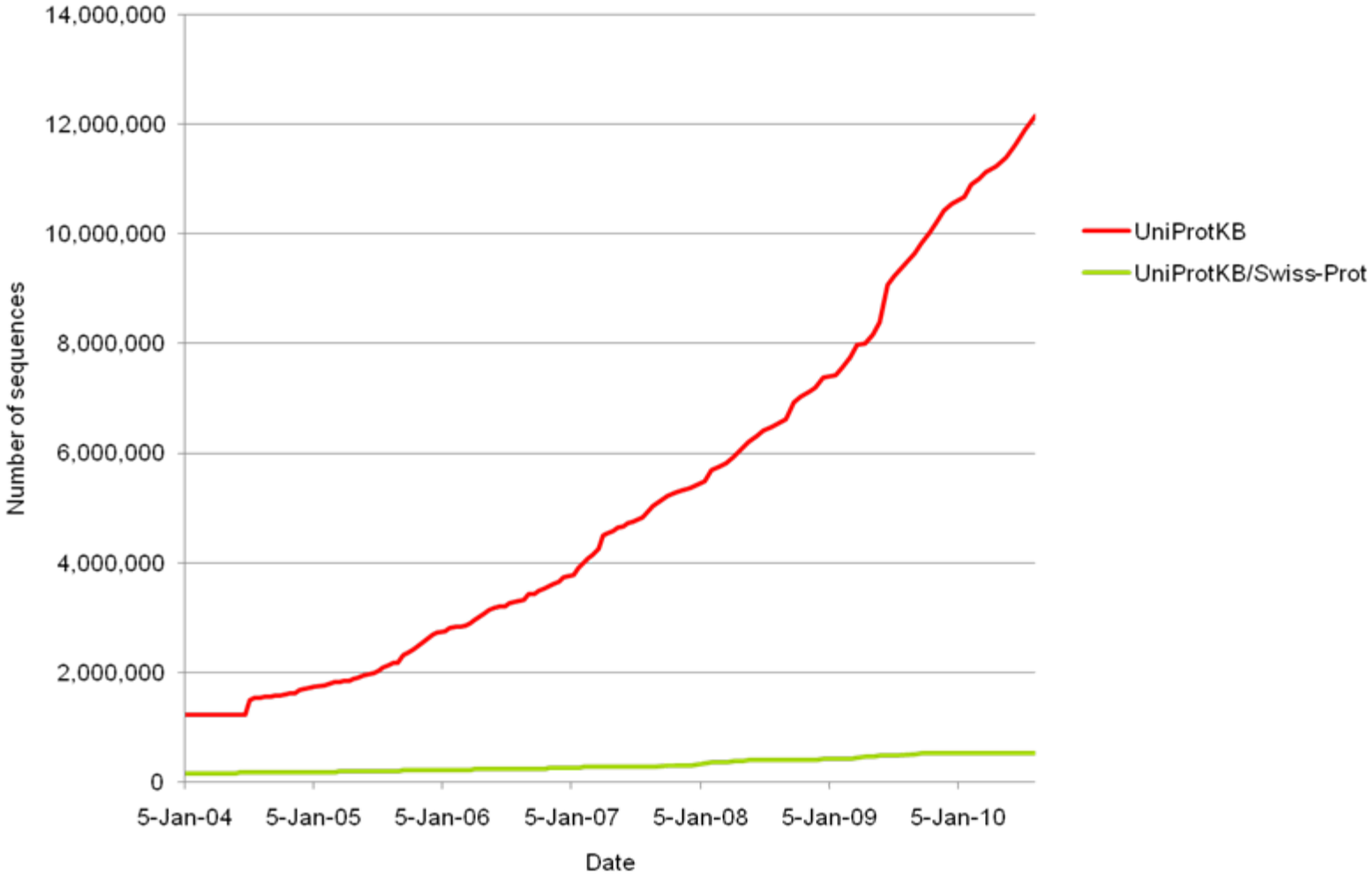
Consortium of 11 major signature databases

Protein signatures

- More sensitive homology searches

- Each member database creates signatures using different methods and methodologies:
  - manually-created sequence alignments
  - automatic processes with some human input and correction
  - entirely automatically.

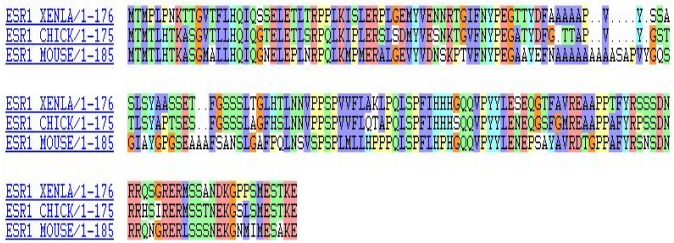
Why do we need predictive annotation tools?



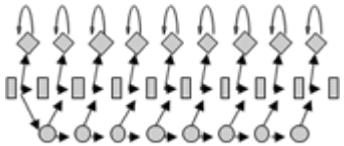
What are protein signatures?

Protein family/domain

Multiple sequence alignment



Build model



Search

it.

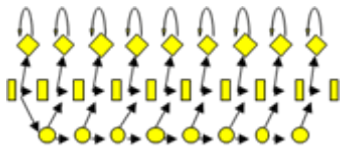
UniProt

Protein analysis

ITWKGPVCGLDGKTYRNECALL

AVPRSPVCGSDDVTYANECELK

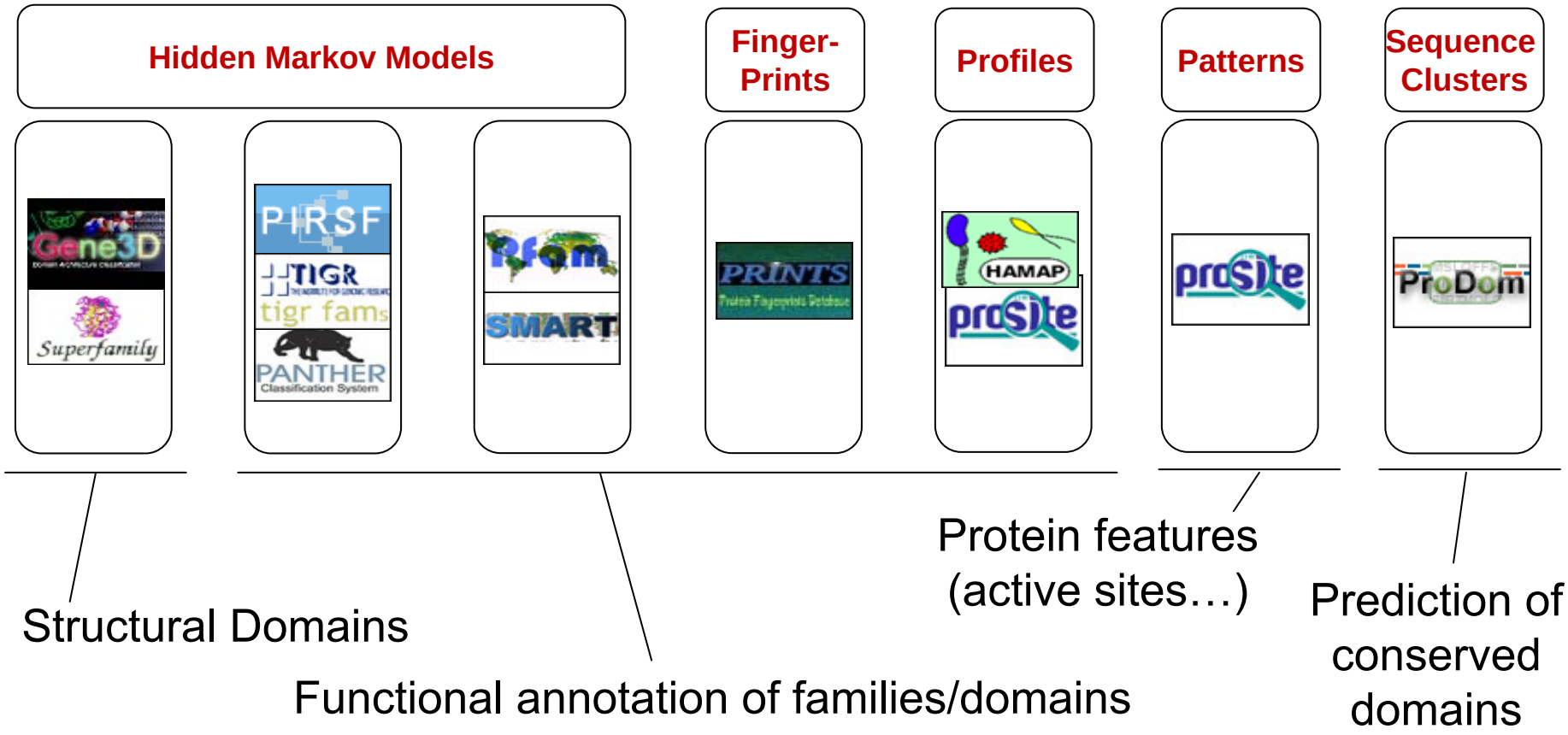
Mature model



Significant match

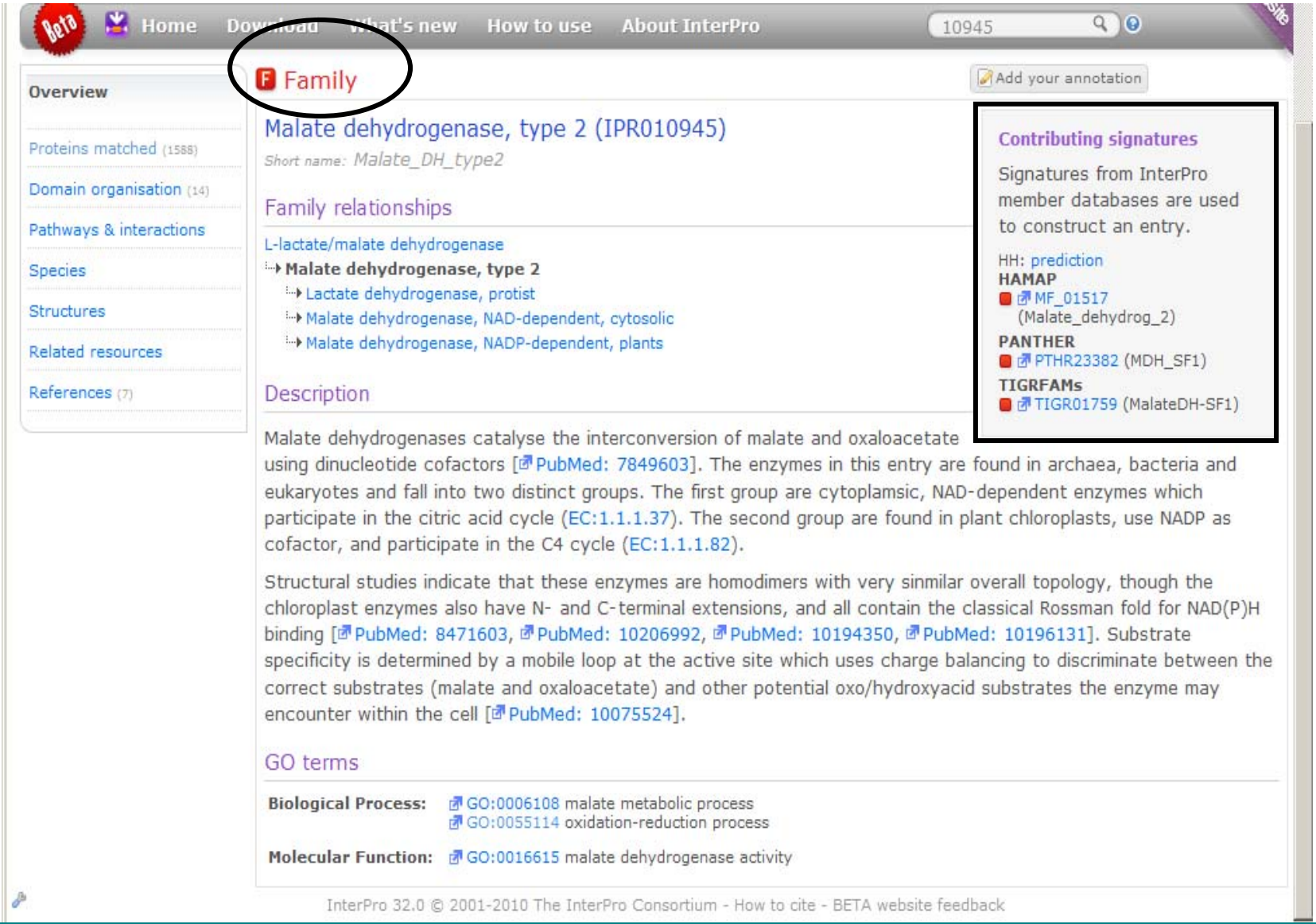
Member databases

METHODS

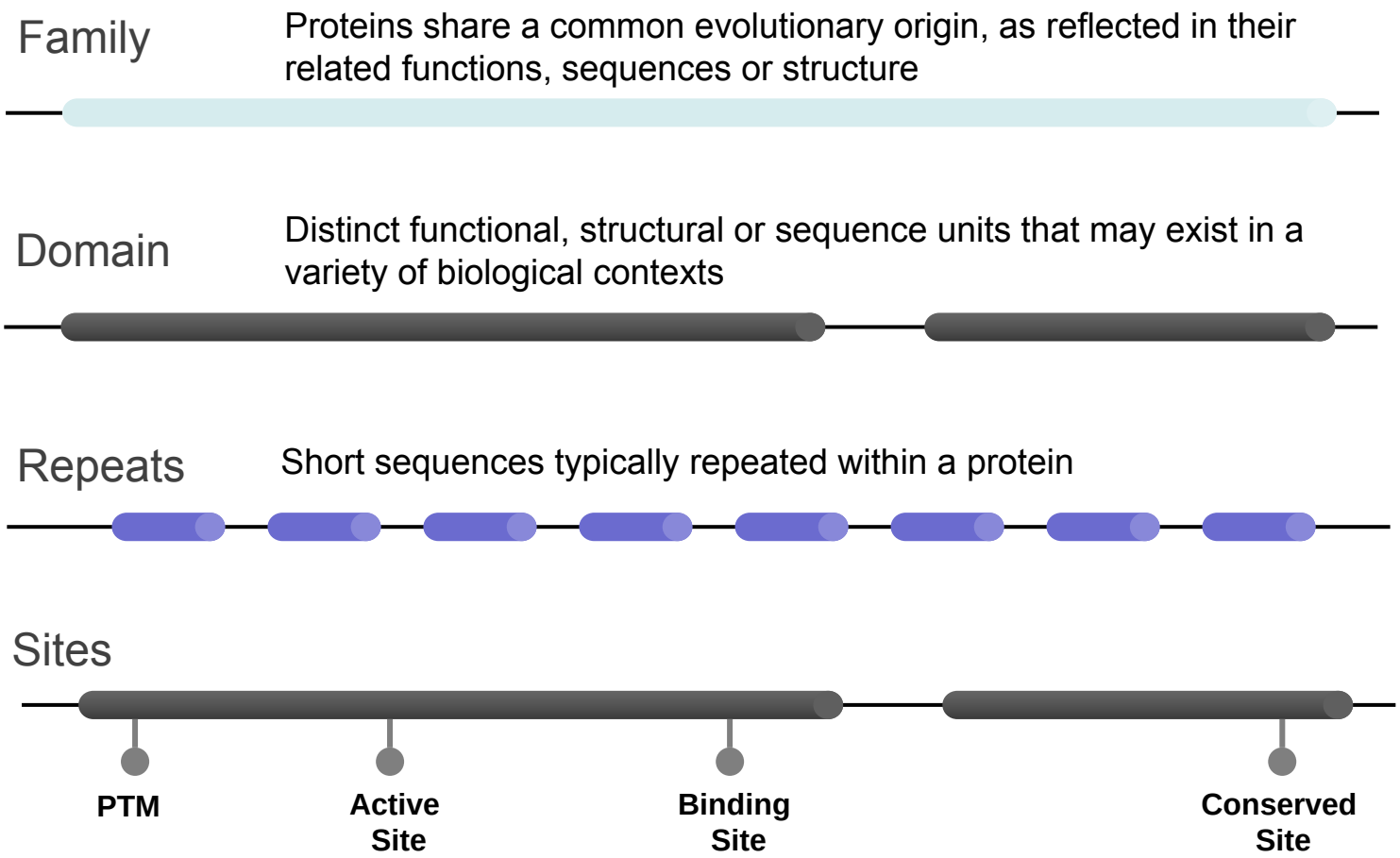


InterPro entry

InterPro entry



The InterPro entry: types



InterPro Entry

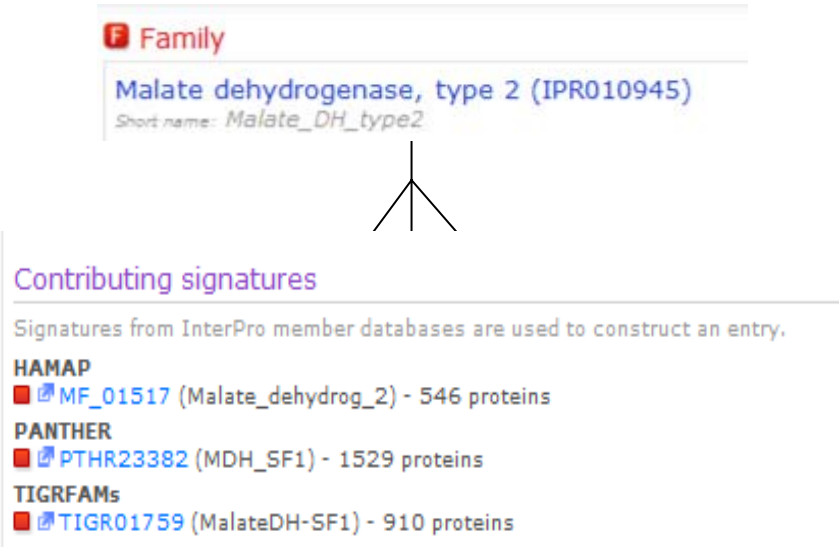
Groups similar signatures together

Adds extensive annotation

Links to other databases

Structural information and viewers

- Quality control
- Removes redundancy



InterPro Entry

Groups similar signatures together

Adds extensive annotation

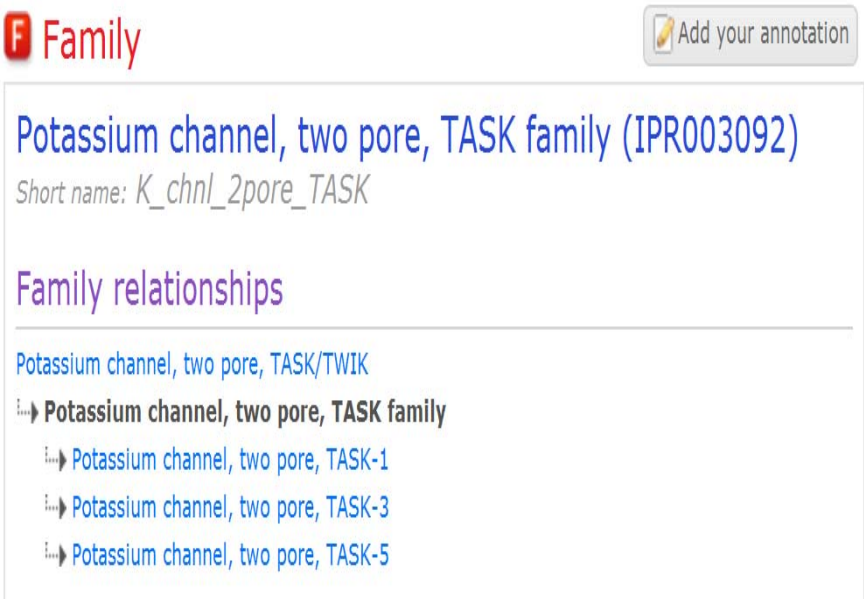
Links to other databases

Structural information and viewers

- Hierarchical classification

Interpro hierarchies: Families

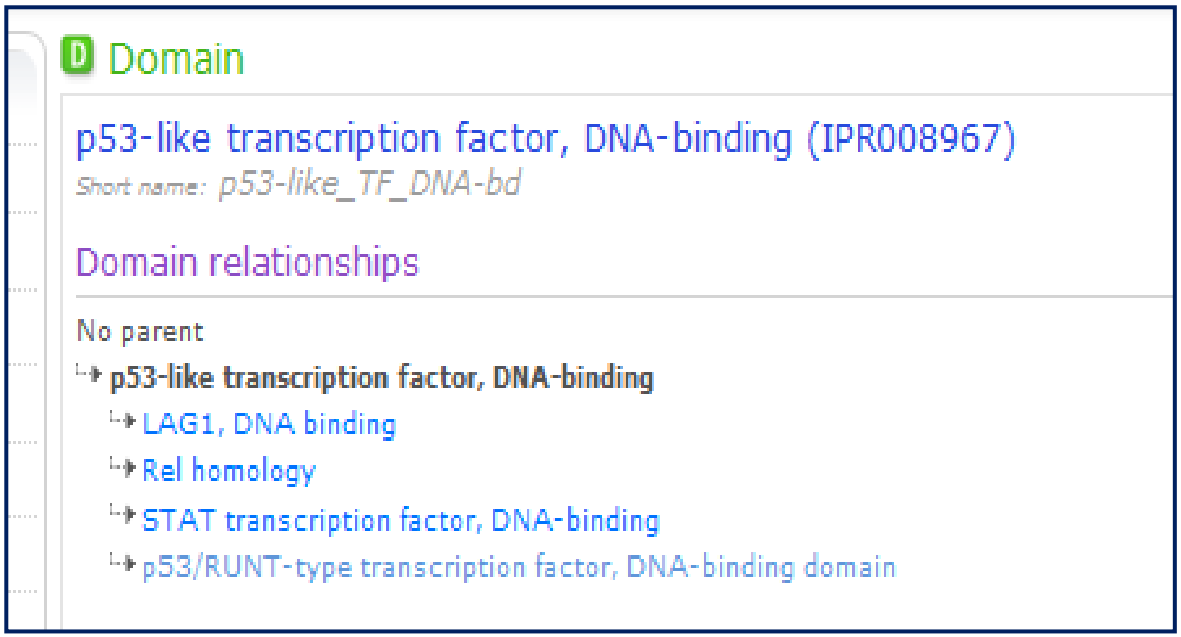
FAMILIES can have **parent/child** relationships with other Families



Parent/Child relationships are based on:

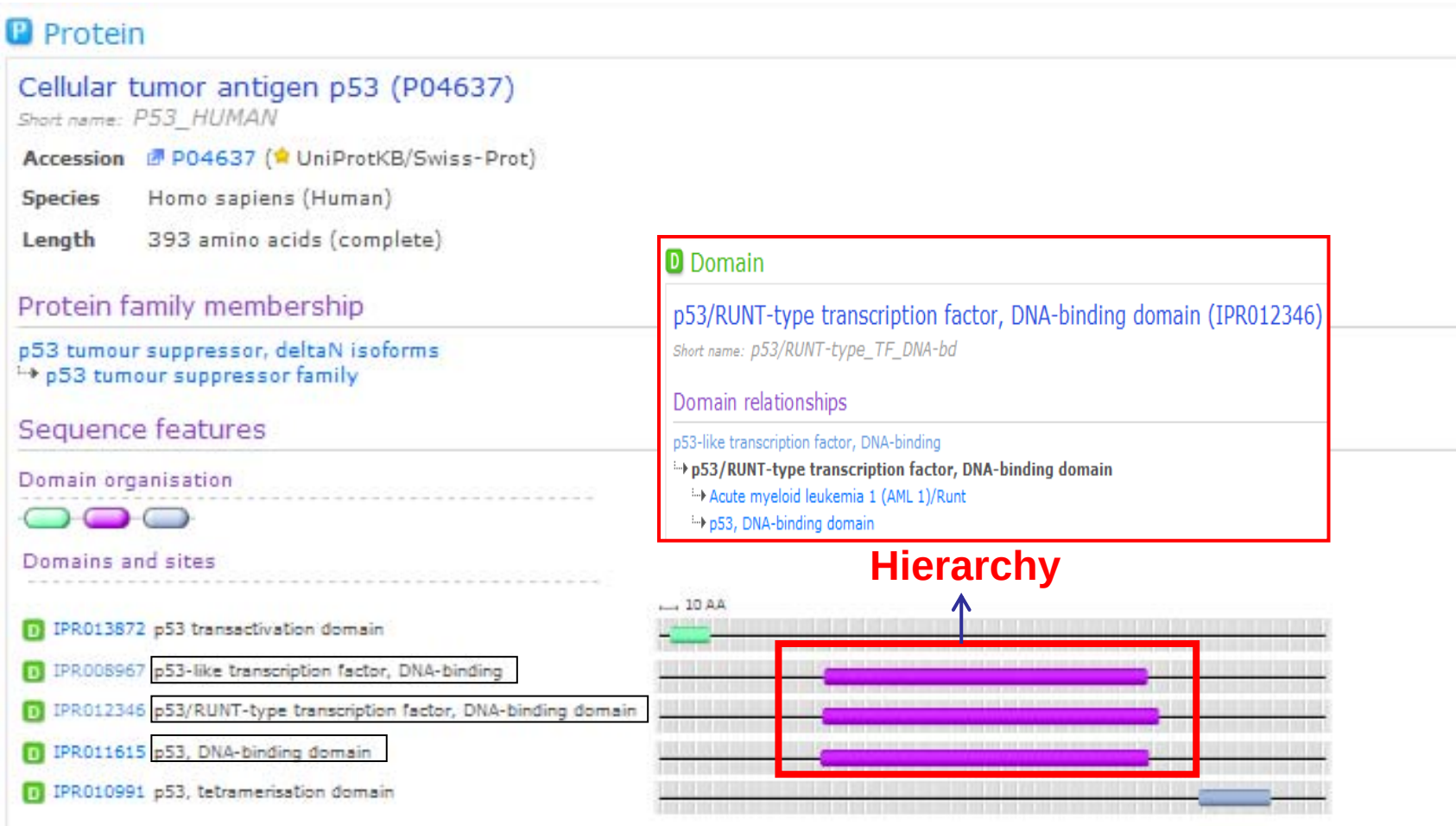
- **Comparison of protein hits**
  - child should be a subset of parent
  - siblings should not have matches in common
- **Existing hierarchies in member databases**
- **Biological knowledge of curators**

InterPro hierarchies: Domains



DOMAINS can have **parent/child** relationships with other domains

Domains and Families may be linked through Domain Organisation



InterPro Entry

Groups similar signatures together

Adds extensive annotation

Links to other databases

Structural information and viewers

Family

Malate dehydrogenase, type 2 (IPR010945)

Short name: *Malate_DH_type2*

Description

Malate dehydrogenases catalyse the interconversion of malate and oxaloacetate using dinucleotide cofactors [PubMed: 7849603]. The enzymes in this entry are found in archaea, bacteria and eukaryotes and fall into two distinct groups. The first group are cytoplamsic, NAD-dependent enzymes which participate in the citric acid cycle (EC:1.1.1.37). The second group are found in plant chloroplasts, use NADP as cofactor, and participate in the C4 cycle (EC:1.1.1.82).

Structural studies indicate that these enzymes are homodimers with very sinmilar overall topology, though the chloroplast enzymes also have N- and C-terminal extensions, and all contain the classical Rossman fold for NAD(P)H binding [PubMed: 8471603, PubMed: 10206992, PubMed: 10194350, PubMed: 10196131]. Substrate specificity is determined by a mobile loop at the active site which uses charge balancing to discriminate between the correct substrates (malate and oxaloacetate) and other potential oxo/hydroxyacid substrates the enzyme may encounter within the cell [PubMed: 10075524].

InterPro Entry

Groups similar signatures together

Adds extensive annotation

Links to other databases

Structural information and viewers

The Gene Ontology project provides a controlled vocabulary of terms for describing gene product characteristics

InterPro Entry

Groups similar signatures together

Adds extensive annotation

Links to other databases

Structural information and viewers

→ UniProt

→ KEGG ... Reactome ... IntAct ...

→ UniProt taxonomy

→ PANDIT ... MEROPS ... Pfam clans ...

→ Pubmed

InterPro Entry

Groups similar signatures together

Adds extensive annotation

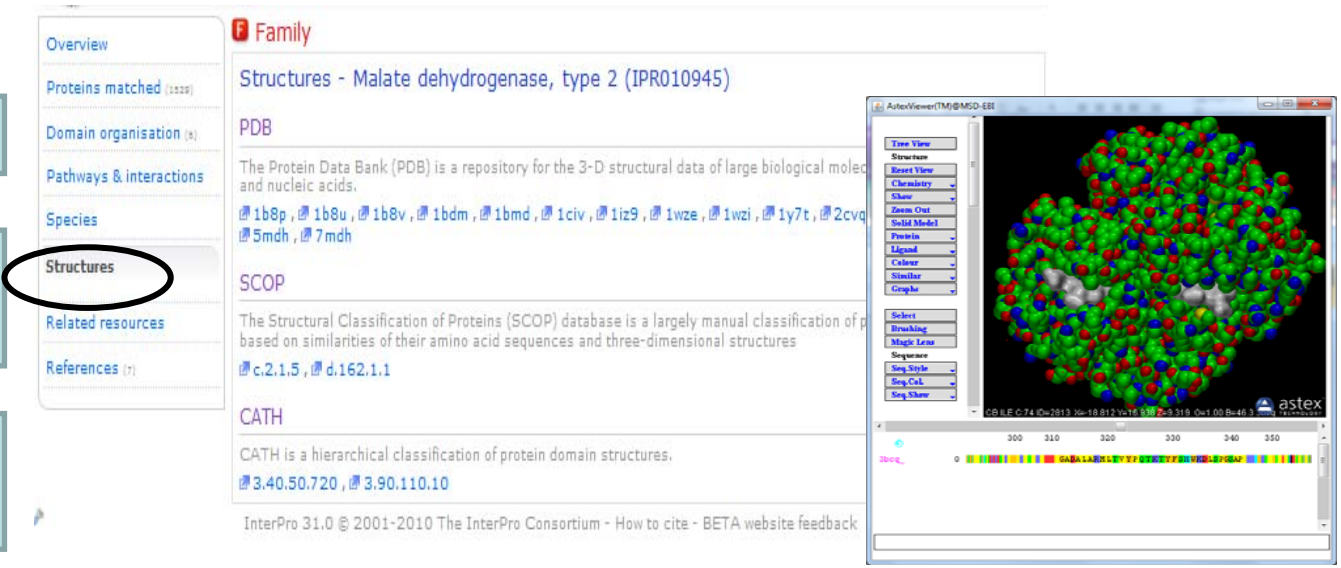
Links to other databases

Structural information and viewers

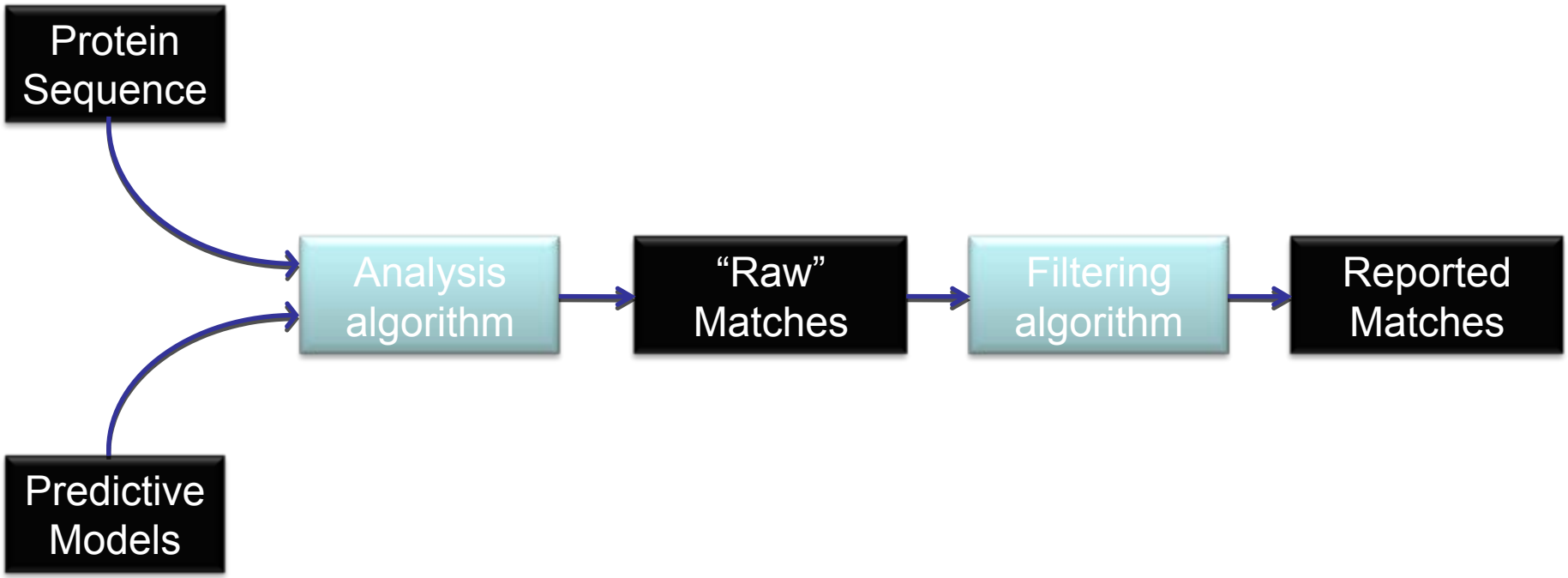
PDB 3-D Structures

SCOP Structural domains

CATH Structural domain

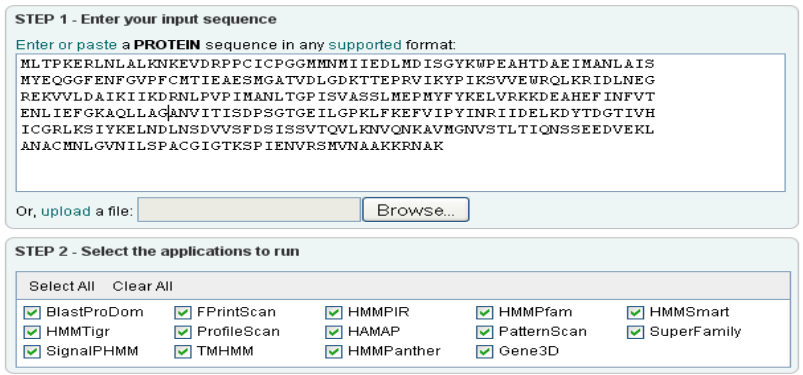


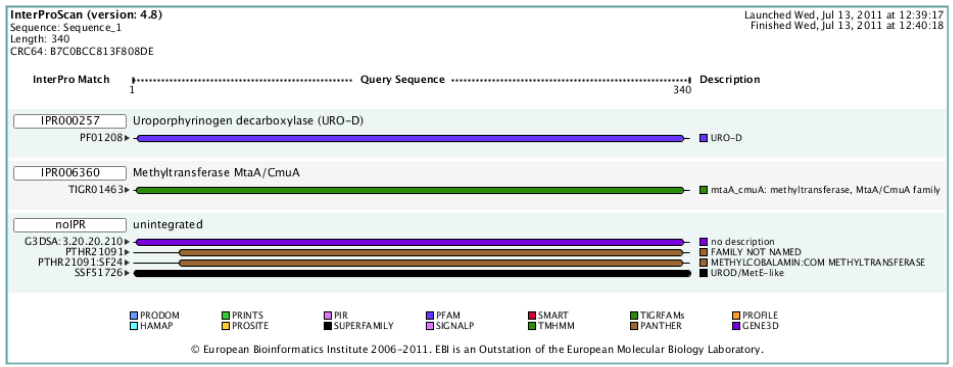
InterProScan



InterProScan access

Interactive:
http://www.ebi.ac.uk/Tools/pfa/iprscan/





Webservice (SOAP and REST):
http://www.ebi.ac.uk/Tools/webservices/services/pfa/iprscan_rest
http://www.ebi.ac.uk/Tools/webservices/services/pfa/iprscan_soap

Downloadable:
ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/

Why redesign InterProScan?

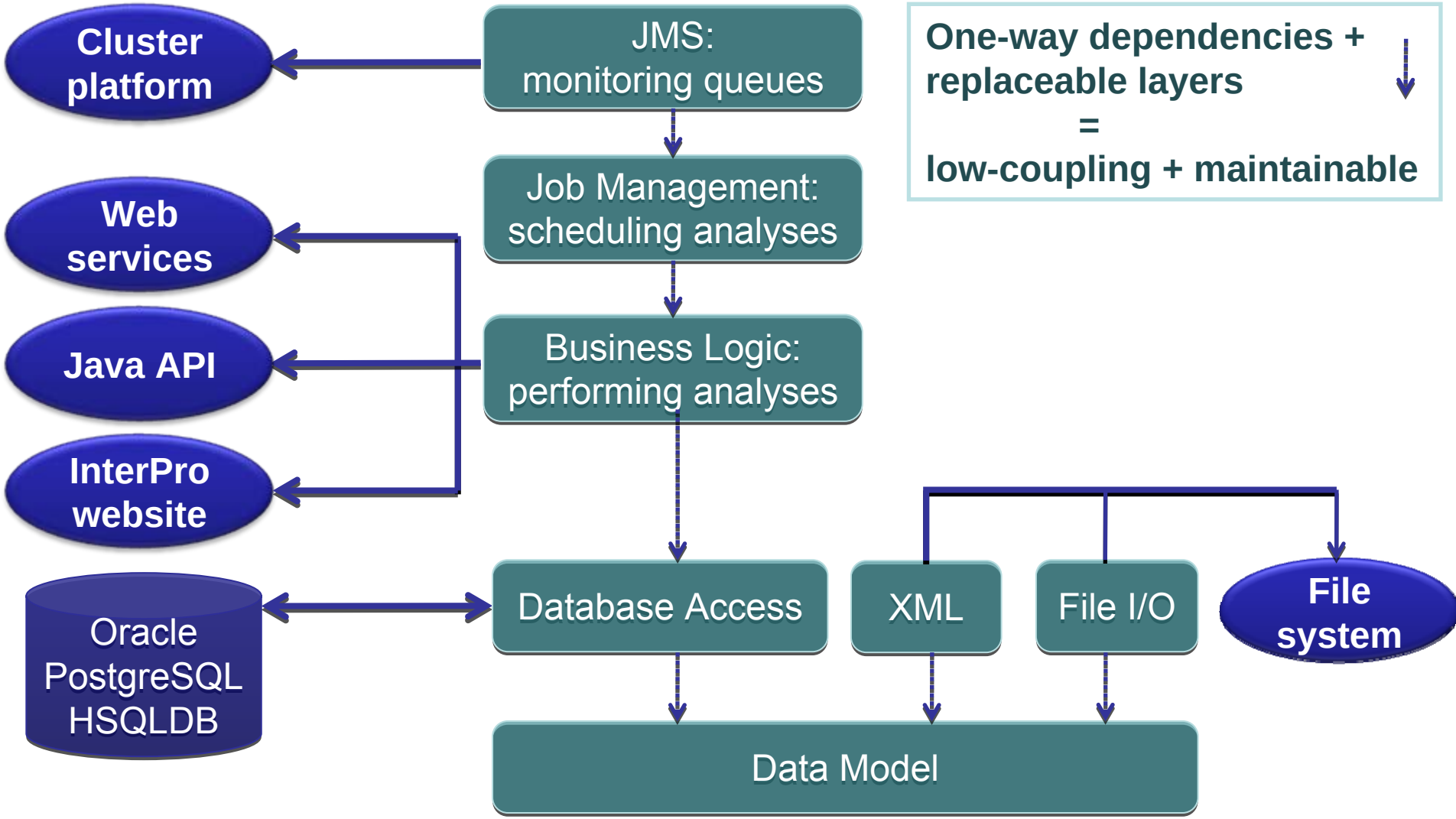
- InterProScan 4
  - complicated installation
  - complicated update
  - limited queuing system
    - Only guaranteed with LSF
  - limited configurability
  - reliability

InterProScan 5.0 aims

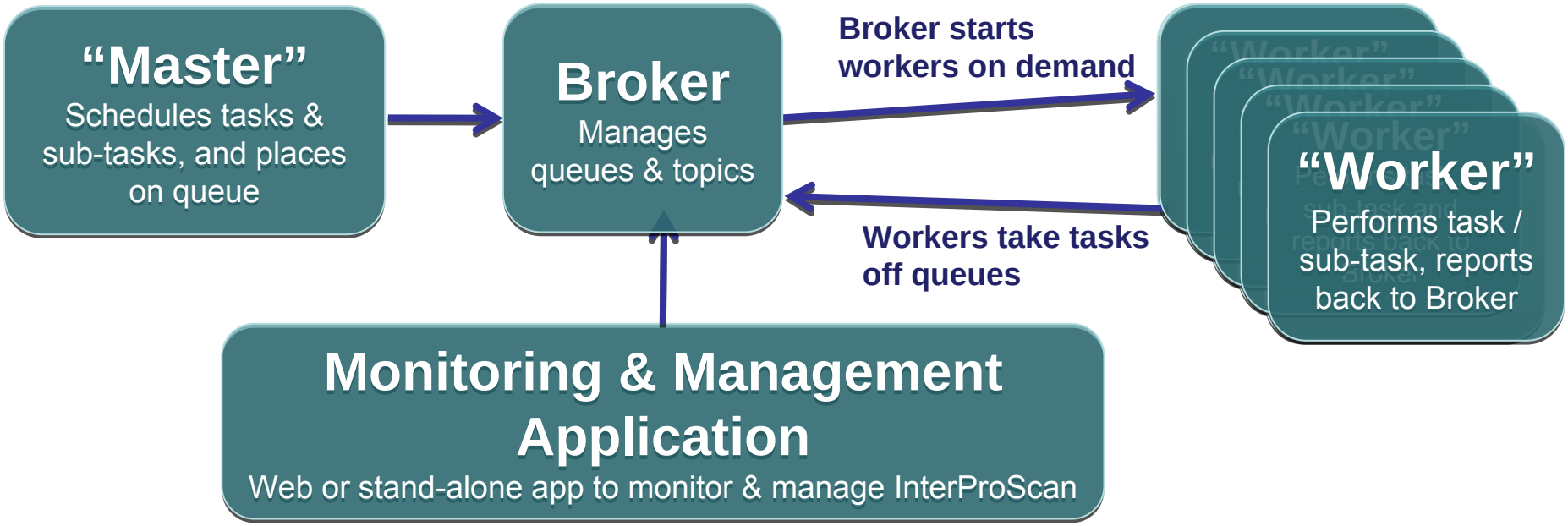
- Easy install and configuration
- Modular
- Expandable
- Easily integrated into existing pipelines
- Incorporate new data model / XML exchange format
- Easy to port on to different architectures:
  - Desktop machine
  - Simple **LAN**
  - **LSF**
  - **PBS**
  - **Sun Grid Engine** ...cloud? GRID?
- Reliablity

InterProScan 5 Technology

Architecture



Java Messaging Service



- Simple and robust programming model
- Mature and stable standard – current JMS version released in 2002
- Guaranteed message delivery to a single worker
- Easy to monitor
- Flexible – easy to implement on multiple platforms

Beta release functionality

Installation

- Requirements
  - Java 1.6
  - Linux
  - Perl
- Installation process

```
wget ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/i5-dist.tar.gz

tar -xzf i5-dist.tar.gz
```

  - ready to use

```
./interproscan.sh -i test_proteins.fasta -o test_proteins.tsv --goterms
```

```
A2YIW7  f927b0d241297dcc9a1c5990b58bf3c4        122     Pfam    PF00085 Thioredoxin     9       112     1.3E-28
T       08-07-2011      IPR013766       Thioredoxin domain      Biological Process:cell redox homeostasis
(GO:0045454)
A2YIW7  f927b0d241297dcc9a1c5990b58bf3c4        122     ProSitePatterns PS00194 Thioredoxin family active site.
32      50      -       T       08-07-2011      IPR017937       Thioredoxin, conserved site     Biological
Process:cell redox homeostasis (GO:0045454)
A2YIW7  f927b0d241297dcc9a1c5990b58bf3c4        122     PIRSF   PIRSF000077     null    4       113
1.50000307E-27  T       08-07-2011      IPR005746       Thioredoxin     Molecular Function:protein disulfide
oxidoreductase activity (GO:0015035), Biological Process:glycerol ether metabolic process (GO:0006662),
Biological Process:cell redox homeostasis (GO:0045454), Molecular Function:electron carrier activity
(GO:0009055)
A2YIW7  f927b0d241297dcc9a1c5990b58bf3c4        122     PRINTS  PR00421 Thioredoxin family signature    39
48      -       T       08-07-2011      IPR005746       Thioredoxin     Molecular Function:protein disulfide
oxidoreductase activity (GO:0015035), Biological Process:glycerol ether metabolic process (GO:0006662),
Biological Process:cell redox homeostasis (GO:0045454), Molecular Function:electron carrier activity
(GO:0009055)
A2YIW7  f927b0d241297dcc9a1c5990b58bf3c4        122     PRINTS  PR00421 Thioredoxin family signature    78
89      -       T       08-07-2011      IPR005746       Thioredoxin     Molecular Function:protein disulfide
oxidoreductase activity (GO:0015035), Biological Process:glycerol ether metabolic process (GO:0006662),
Biological Process:cell redox homeostasis (GO:0045454), Molecular Function:electron carrier activity
(GO:0009055)
A2YIW7  f927b0d241297dcc9a1c5990b58bf3c4        122     PRINTS  PR00421 Thioredoxin family signature    31
39      -       T       08-07-2011      IPR005746       Thioredoxin     Molecular Function:protein disulfide
oxidoreductase activity (GO:0015035), Biological Process:glycerol ether metabolic process (GO:0006662),
Biological Process:cell redox homeostasis (GO:0045454), Molecular Function:electron carrier activity
(GO:0009055)
```

Default tab-separated values output

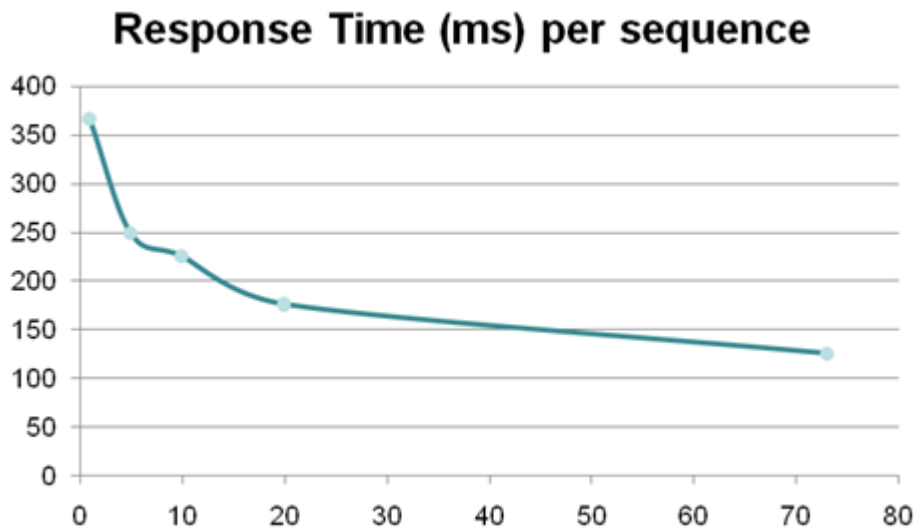
```
./interproscan.sh -i test_proteins.fasta -o test_proteins.xml --goterms -F xml
```

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<protein-matches xmlns="http://www.ebi.ac.uk/schema/interpro">
    <protein>
        <sequence
md5="f927b0d241297dcc9a1c5990b58bf3c4">MAAEEGVVIACHNKDEFDAQMTKAKEAGKVVIIDFTASWCGPCRFIAPVFAEYAKKFPGAVFLKVDVDELKEV
AEKYNVEAMPTFLFIKDGAEADKVVGARKDDLQNTIVKHVGATAASASA</sequence>
        <xref id="A2YIW7"/>
        <matches>
            <fingerprints-match graphscan="III" evalue="2.500000864E-7">
                <signature name="THIOREDOXIN" desc="Thioredoxin family signature" ac="PR00421">
                    <models>
                        <model name="THIOREDOXIN" desc="Thioredoxin family signature" ac="PR00421"/>
                    </models>
                    <signature-library-release version="41.1" library="PRINTS"/>
                </signature>
                <locations>
                    <fingerprints-location score="0.0" pvalue="0.0" motifNumber="3" end="48" start="39"/>
                    <fingerprints-location score="0.0" pvalue="0.0" motifNumber="2" end="89" start="78"/>
                    <fingerprints-location score="0.0" pvalue="0.0" motifNumber="1" end="39" start="31"/>
                </locations>
            </fingerprints-match>
            <hmmer2-match score="100.5" evalue="-INF">
                <signature name="Thioredoxin" ac="PIRSF000077">
                    <models>
                        <model name="Thioredoxin" ac="PIRSF000077"/>
                    </models>
                    <signature-library-release version="2.74" library="PIRSF"/>
                </signature>
                <locations>
                    <hmmer2-location hmm-length="0" hmm-end="108" hmm-start="1" evalue="1.50000307E-27"
score="0.0" end="113" start="4"/>
                </locations>
            </hmmer2-match>
...etc
```

XML output

Pre-calculated match lookup

- BerkeleyDB-backed **REST** web service
- Includes matches for all of UniParc (27 million sequences)
- 250 million matches
- Fast response
- Integrated into i5.



Other functionality

- Increased reliability
- Precalculated match lookup
- Configuration
  - simple properties file
- Nucleotide sequence
  - getOrf
  - map matches to nucleotide coordinates
- Pathway mapping
  - KEGG, Reactome, MetaCyc, Unipathway

Future functionality

- Webservice
- Interact directly with architecture:
  - LAN
  - LSF
  - PBS
  - Sun Grid Engine
- Database persistence
  - Oracle
  - MySQL
  - Postgres
  - etc
- Graphical output
- Other functionality
  - ask!

InterProScan 5 timeline

- Beta release
  - August 2011
  - InterProScan 4 still maintained
- Full release
  - Early 2012
  - InterProScan 4 deprecated

interproscan-5-dev@googlegroups.com

Acknowledgements

Team leader

Sarah Hunter

Developers

Matthew Fraser

Anthony Quinn

Phil Jones

Sebastien Pesseat

Maxim Scheremetjew

Bioinformaticians

Craig McAnulla

Siew-Yit Yong

Curators

Alex Mitchell

Amaia Sangrador

Any Questions → Stand 302

Come and see us at booths 9 and 10!

- Job opportunities
- PhD and postdoc positions
- Training in person and online
- Services
- Industry programme

EMBL-EBI