



# InterPro and InterProScan 5.0



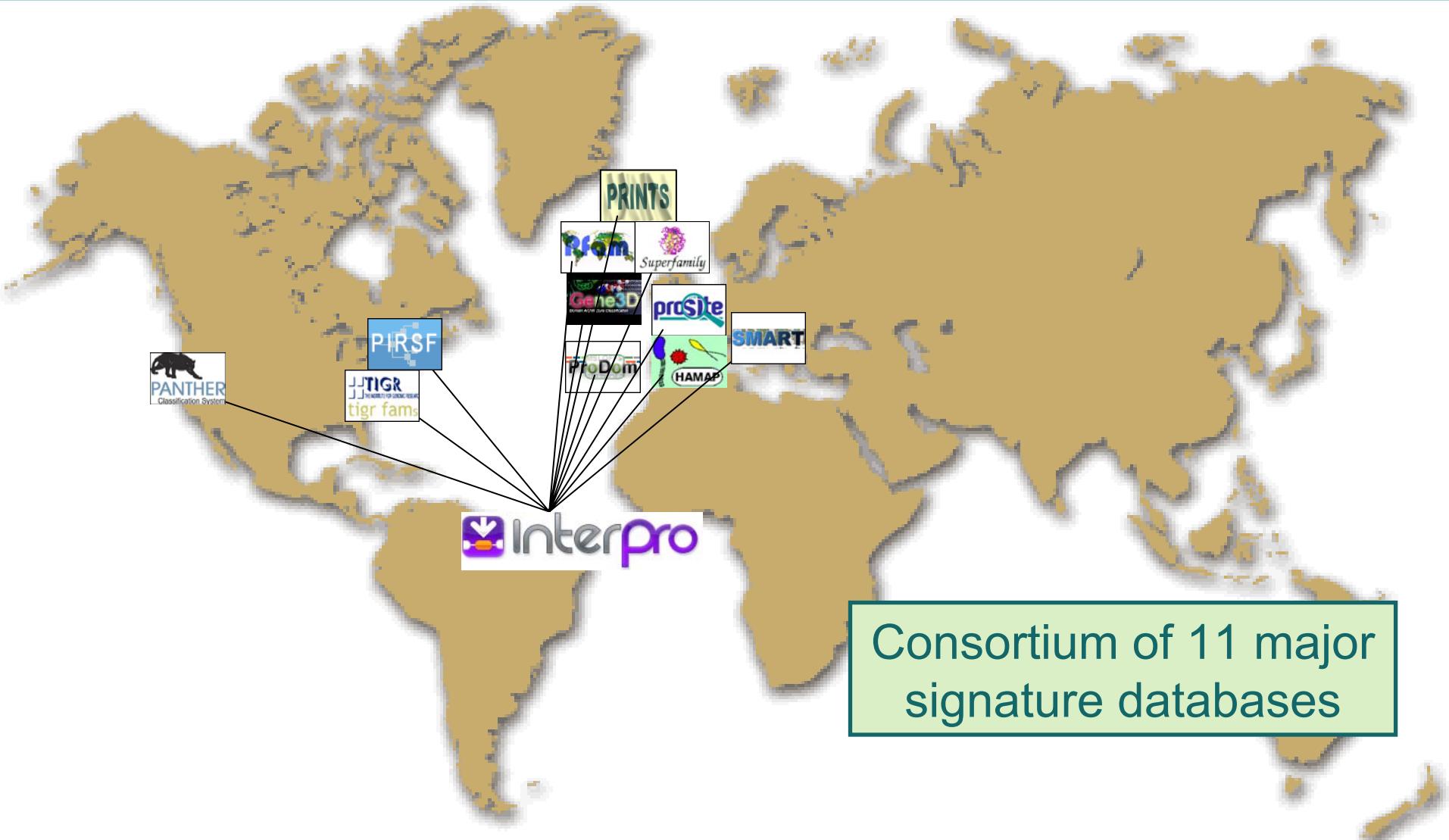
# InterPro

- 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

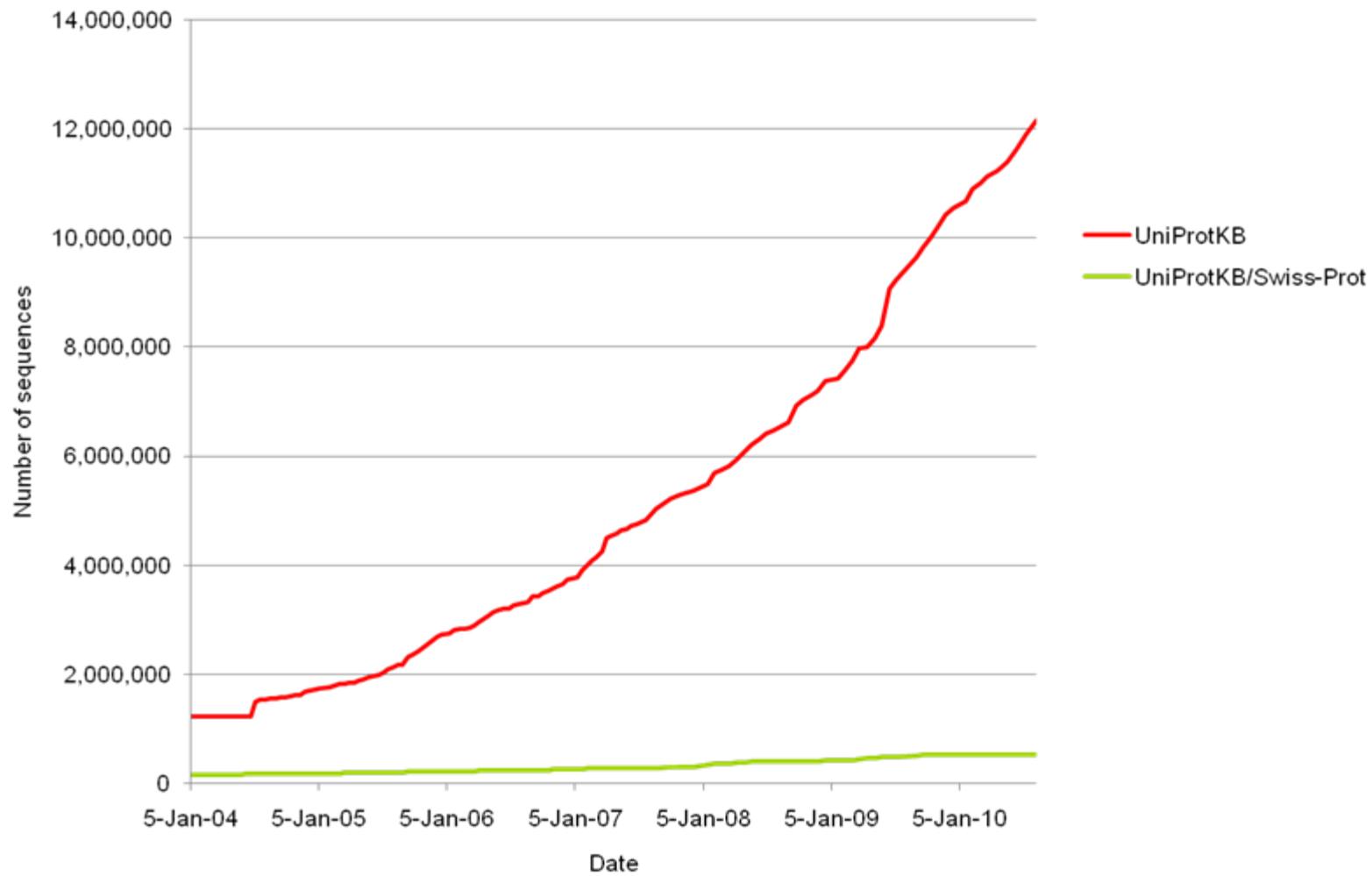
# InterPro Consortium



# 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

ESR1\_XENLA/1-176 MTTHPIPKITIGVTFILQIISSELEIYTRPPIKISLERPIGEMVNWRTGICINYPGEGTYDFAAAAAF...V....YSSA  
ESR1\_CHICK/1-175 MTMTLHTKASGTVLHIOQGTELELTISRFPOLKIPLERSLSDMYVESENKTGVENYPGEGATDGGTTAP...V....YGST  
ESR1\_MOUSE/1-185 MTMTLHTKASCHMALLHQIQNELEPINRPOLKIPMERAIGEVWWVDSKPTWNPFCAYEVNAAAAANASAPIYGQS

ESR1\_XENLA/1-176 SISVAASSET...FGSSSITGHTINNVPPSFVVFIAKIPOLSPFIHHHGQONPYVILESGTFAVREAAPPTFYRSSDN  
ESR1\_CHICK/1-175 TLSVAPTS...FQSSSLAGFHSLINNVPPSFVVFLOTAPOLSPFIHHHSQONPYVILEEGSGCHREALAPPAYVRPSSDN  
ESR1\_MOUSE/1-185 QIAYGPGSAAAPSAANSIDQAFQINSNSPSEIMLLHEPPOLSPFLHPACQONPYVILEENFGSATVARDTGPATFRNSDN

ESR1\_XENLA/1-176 RRQSGRERNSGANDKGPPGNGESTKE  
ESR1\_CHICK/1-175 RRHSIRENNSSTNEKGSIQMESTKE  
ESR1\_MOUSE/1-185 RRQNGRERISSLSSNEKGWNIMESAKE

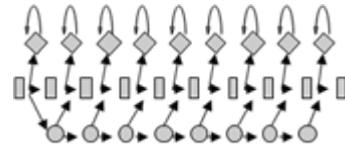
## Significant match



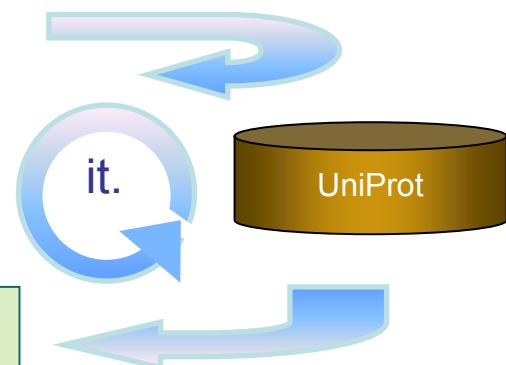
## Protein analysis

ITWKGPVCGLDGKTYRNECALL  
AVPRSPVCGSDDVTYANECELK

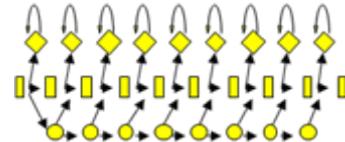
## Build model



## Search



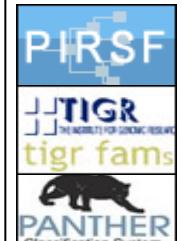
## Mature model



# Member databases

## METHODS

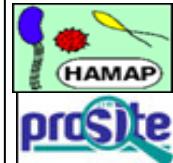
### Hidden Markov Models



### Finger-Prints



### Profiles



### Patterns



### Sequence Clusters



Structural Domains

Functional annotation of families/domains

Protein features  
(active sites...)

Prediction of  
conserved  
domains

# InterPro entry

# InterPro entry

Beta Home Download What's new How to use About InterPro 10945

**Overview**

Proteins matched (1588)

Domain organisation (14)

Pathways & interactions

Species

Structures

Related resources

References (7)

**F Family**

**Malate dehydrogenase, type 2 (IPR010945)**  
Short name: *Malate\_DH\_type2*

**Family relationships**

L-lactate/malate dehydrogenase

- ↳ **Malate dehydrogenase, type 2**
  - ↳ Lactate dehydrogenase, protist
  - ↳ Malate dehydrogenase, NAD-dependent, cytosolic
  - ↳ Malate dehydrogenase, NADP-dependent, plants

**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 cytoplasmic, 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 similar 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].

**GO terms**

**Biological Process:** GO:0006108 malate metabolic process  
 GO:0055114 oxidation-reduction process

**Molecular Function:** GO:0016615 malate dehydrogenase activity

Add your annotation

**Contributing signatures**

Signatures from InterPro member databases are used to construct an entry.

HH: prediction  
HAMAP  
↳ MF\_01517 (Malate\_dehydrog\_2)  
PANTHER  
↳ PTHR23382 (MDH\_SF1)  
TIGRFAMs  
↳ TIGR01759 (MalateDH-SF1)



# The InterPro entry: types

Family	Proteins share a common evolutionary origin, as reflected in their related functions, sequences or structure
Domain	Distinct functional, structural or sequence units that may exist in a variety of biological contexts
Repeats	Short sequences typically repeated within a protein
Sites	<p>The diagram illustrates four types of sites on a protein structure represented by a dark grey horizontal bar. Four small grey circles, each labeled below with its type, are positioned along the bar: 'PTM' on the far left, 'Active Site' in the middle-left, 'Binding Site' in the middle-right, and 'Conserved Site' on the far right.</p>

# InterPro Entry

- Groups similar signatures together
- Adds extensive annotation
- Links to other databases
- Structural information and viewers

- Quality control
- Removes redundancy

## Family

Malate dehydrogenase, type 2 (IPR010945)

Short name: Malate\_DH\_type2



## Contributing signatures

Signatures from InterPro member databases are used to construct an entry.

### HAMAP

■ MF\_01517 (Malate\_dehydrog\_2) - 546 proteins

### PANTHER

■ PTHR23382 (MDH\_SF1) - 1529 proteins

### TIGRFAMs

■ TIGR01759 (MalateDH-SF1) - 910 proteins

# InterPro Entry

Groups similar signatures together

Adds extensive annotation

Links to other databases

Structural information and viewers

➤ Hierarchical classification

## F Family

Malate dehydrogenase, type 2 (IPR010945)

*Short name: Malate\_DH\_type2*

### Family relationships

L-lactate/malate dehydrogenase

↳ Malate dehydrogenase, type 2

↳ Lactate dehydrogenase, protist

↳ Malate dehydrogenase, NAD-dependent, cytosolic

↳ Malate dehydrogenase, NADP-dependent, plants

# Interpro hierarchies: Families

FAMILIES can have parent/child relationships with other Families



Potassium channel, two pore, TASK family (IPR003092)

Short name: K\_chnl\_2pore\_TASK

## Family relationships

Potassium channel, two pore, TASK/TWIK

↳ Potassium channel, two pore, TASK family

    ↳ Potassium channel, two pore, TASK-1

    ↳ Potassium channel, two pore, TASK-3

    ↳ Potassium channel, two pore, TASK-5

**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

D Domain

p53-like transcription factor, DNA-binding (IPR008967)  
*Short name: p53-like\_TF\_DNA-bd*

Domain relationships

No parent

- ↳ p53-like transcription factor, DNA-binding
  - ↳ LAG1, DNA binding
  - ↳ Rel homology
  - ↳ STAT transcription factor, DNA-binding
  - ↳ p53/RUNT-type transcription factor, DNA-binding domain

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

# Domains and Families may be linked through Domain Organisation

## P Protein

### Cellular tumor antigen p53 (P04637)

Short name: P53\_HUMAN

Accession: P04637 (UniProtKB/Swiss-Prot)

Species: Homo sapiens (Human)

Length: 393 amino acids (complete)

### Protein family membership

p53 tumour suppressor, deltaN isoforms

↳ p53 tumour suppressor family

### Sequence features

#### Domain organisation



#### Domains and sites

D IPR013872 p53 transactivation domain

D IPR006967 p53-like transcription factor, DNA-binding

D IPR012346 p53/RUNT-type transcription factor, DNA-binding domain

D IPR011615 p53, DNA-binding domain

D IPR010991 p53, tetramerisation domain

## D Domain

### p53/RUNT-type transcription factor, DNA-binding domain (IPR012346)

Short name: p53/RUNT-type\_TF\_DNA-bd

### Domain relationships

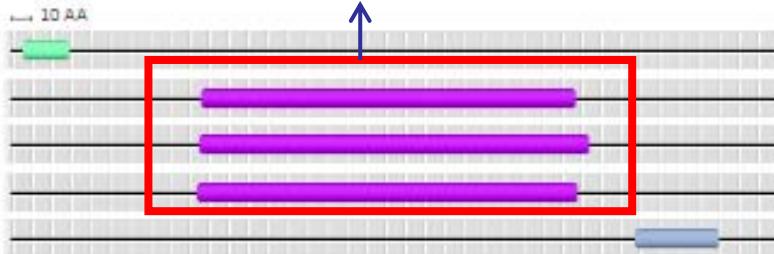
p53-like transcription factor, DNA-binding

↳ p53/RUNT-type transcription factor, DNA-binding domain

↳ Acute myeloid leukemia 1 (AML 1)/Runt

↳ p53, DNA-binding domain

## Hierarchy



# 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 cytoplasmic, 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 similar 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

## GO terms

**Biological Process:**  GO:0006108 malate metabolic process  
 GO:0055114 oxidation-reduction process

**Molecular Function:**  GO:0016615 malate dehydrogenase activity

# InterPro Entry

Groups similar signatures together

Adds extensive annotation

Links to other databases

Structural information and viewers

The screenshot shows the 'Overview' section of the InterPro Entry interface. On the left is a sidebar with the following links:

- Proteins matched (1588) → UniProt
- Domain organisation (14) → KEGG ... Reactome ... IntAct ...
- Pathways & interactions → UniProt taxonomy
- Species → PANDIT ... MEROPS ... Pfam clans ...
- Structures → Pubmed
- Related resources →
- References (7) →

# 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

Overview  
Proteins matched (1525)  
Domain organisation (1)  
Pathways & interactions  
Species  
**Structures**  
Related resources  
References (7)

## Family

Structures - Malate dehydrogenase, type 2 (IPR010945)

### PDB

The Protein Data Bank (PDB) is a repository for the 3-D structural data of large biological molecules and nucleic acids.

[1b8p](#), [1b8u](#), [1b8v](#), [1bcd](#), [1bmd](#), [1civ](#), [1iz9](#), [1wze](#), [1wzi](#), [1y7t](#), [2cvq](#), [5mdh](#), [7mdh](#)

### SCOP

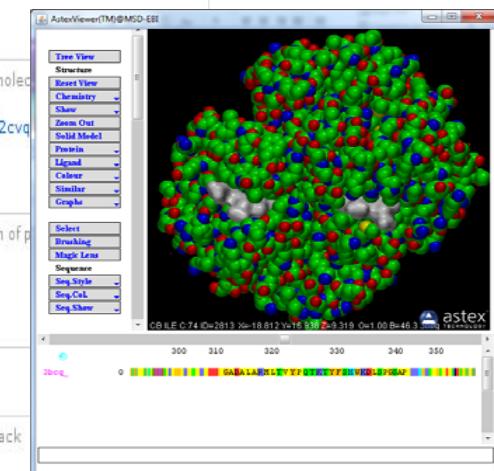
The Structural Classification of Proteins (SCOP) database is a largely manual classification of proteins based on similarities of their amino acid sequences and three-dimensional structures.

[c.2.1.5](#), [d.162.1.1](#)

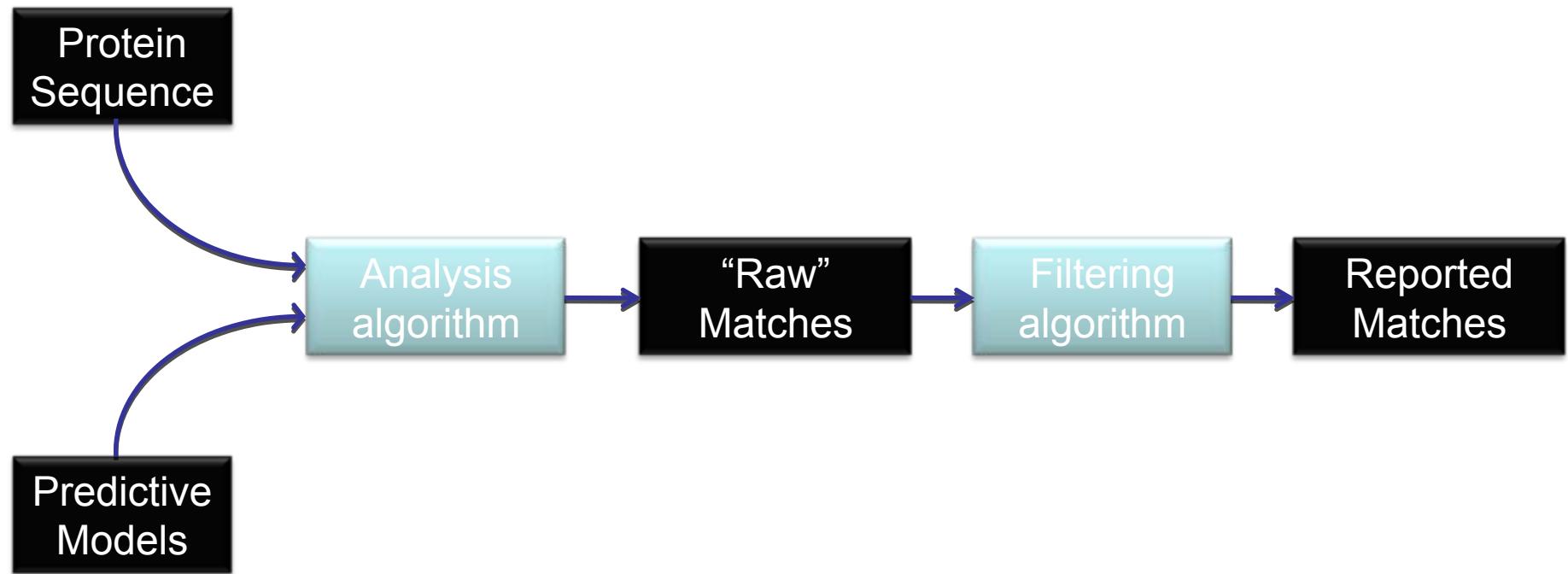
### CATH

CATH is a hierarchical classification of protein domain structures.

[3.40.50.720](#), [3.90.110.10](#)



# InterProScan



# InterProScan access

Interactive:

<http://www.ebi.ac.uk/Tools/pfa/iprscan/>

STEP 1 - Enter your input sequence

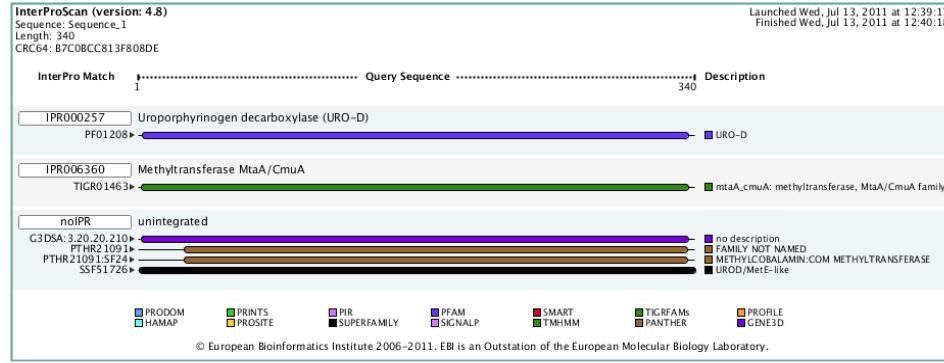
Enter or paste a **PROTEIN** sequence in any [supported](#) format:  
MLTPKERLNLALKNKEVDRPPCICPGGMNNMIIEDLMDISGYKWKPEAHTDAEIMANLAIS  
MYEQGGFENFGVPPFCMTIEAESMGATVDLGDKTTEPRVIKYPIKSVEWROLKRIDLNEG  
REKVVLDAIKIICKRNLPVPIMANLTGPISVASSLMEPHFYKELVRKKDEAHFINFV  
ENLIEFGKAQLLAGANVITISDPSGTGEILGPKLKFKEPVIPYINRIIDELKDYTDGTVH  
ICGRLEKSIYKELNDLNSDVVSPDSISSTVQVLKNVQNKAVMGNVSTLTIQNSSEEDVEKL  
ANACMNLGVNLSPACIGTKSPIENVRSMVNAAKKRNAK

Or, [upload a file:](#)  [Browse...](#)

STEP 2 - Select the applications to run

Select All Clear All

<input checked="" type="checkbox"/> BlastProDom	<input checked="" type="checkbox"/> FPrintScan	<input checked="" type="checkbox"/> HHMPiR	<input checked="" type="checkbox"/> HMMpFam	<input checked="" type="checkbox"/> HMMSmart
<input checked="" type="checkbox"/> HMMTigr	<input checked="" type="checkbox"/> ProfileScan	<input checked="" type="checkbox"/> HAMAP	<input checked="" type="checkbox"/> PatternScan	<input checked="" type="checkbox"/> SuperFamily
<input checked="" type="checkbox"/> SignalPHMM	<input checked="" type="checkbox"/> TMHMM	<input checked="" type="checkbox"/> HHMPPanther	<input checked="" type="checkbox"/> Gene3D	



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_rest)

[http://www.ebi.ac.uk/Tools/webservices/services/pfa/iprscan\\_soap](http://www.ebi.ac.uk/Tools/webservices/services/pfa/iprscan_soap)

Downloadable:

<ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/>



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

EMBL-EBI

# Why redesign InterProScan?

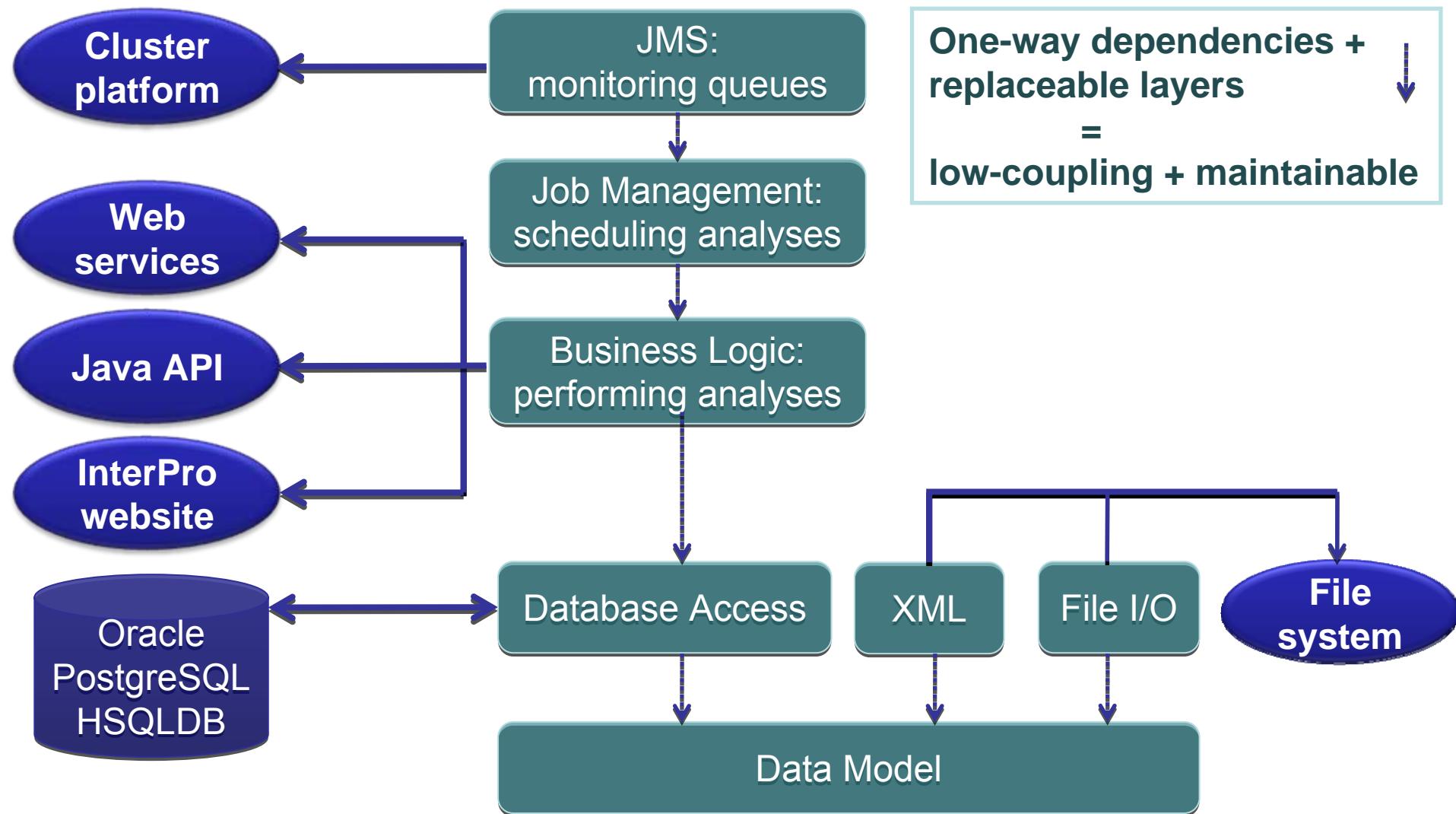
- InterProScan 4
  - complicated installation
  - complicated update
  - limited queuing system
    - Only guaranteed with LSF
  - limited configurability
  - reliability

# InterProScan 5.0 aims

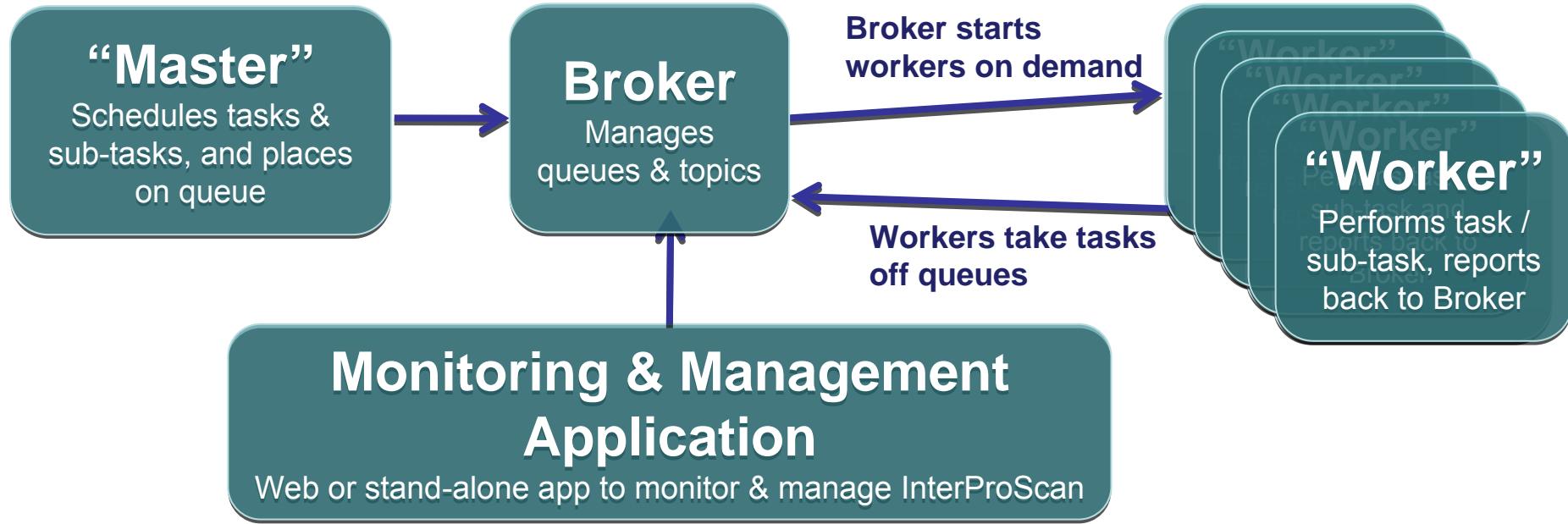
- Easy install and configuration
- Modular
- Expandable
- Easily integrated into existing pipelines
- Incorporate new data model / XML exchange format
- Easy to port on to different architectures:
  - Desktop machine
  - Simple **LAN**
  - **LSF**
  - **PBS**
  - **Sun Grid Engine** ...cloud? GRID?
-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

# 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			
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
```

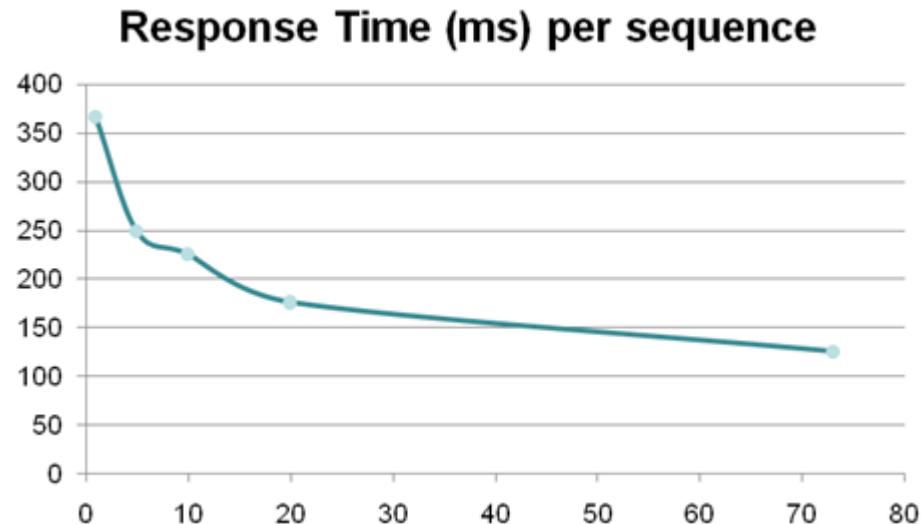
---

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<protein-matches xmlns="http://www.ebi.ac.uk/schema/interpro">
    <protein>
        <sequence
            md5="f927b0d241297dcc9a1c5990b58bf3c4">MAAEEGVVIACHNKDEFDAQMTKAKEAGKVIIDFTASWCGPCRFIAPVFAEYAKKFPGAVFLKVDVDELKEV
AEKYNEAMPTFLFIKDGAEADKVVGARKDDLQNTIVKHVGATAASASA</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](mailto:interproscan-5-dev@googlegroups.com)

# Acknowledgements



Team leader



Sarah  
Hunter



Matthew  
Fraser



Anthony  
Quinn



Phil  
Jones

Developers



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

