InterPro Curation: Integrating Predictive Protein Signatures Into Biological Hierarchies

Dr Sarah Burge
Curator, InterPro
swb@ebi.ac.uk

EMBL-EBI

InterPro is a collection of predictive protein signatures

Member databases build sequence-based models to represent biological features such as families, domains, conserved sites

We integrate them (without alteration) into the InterPro database and provide biological context for each signature

- Automatic genome annotation
- Distant relationships between novel sequences
- Streamline analysis
- Protein classification
- Etc...

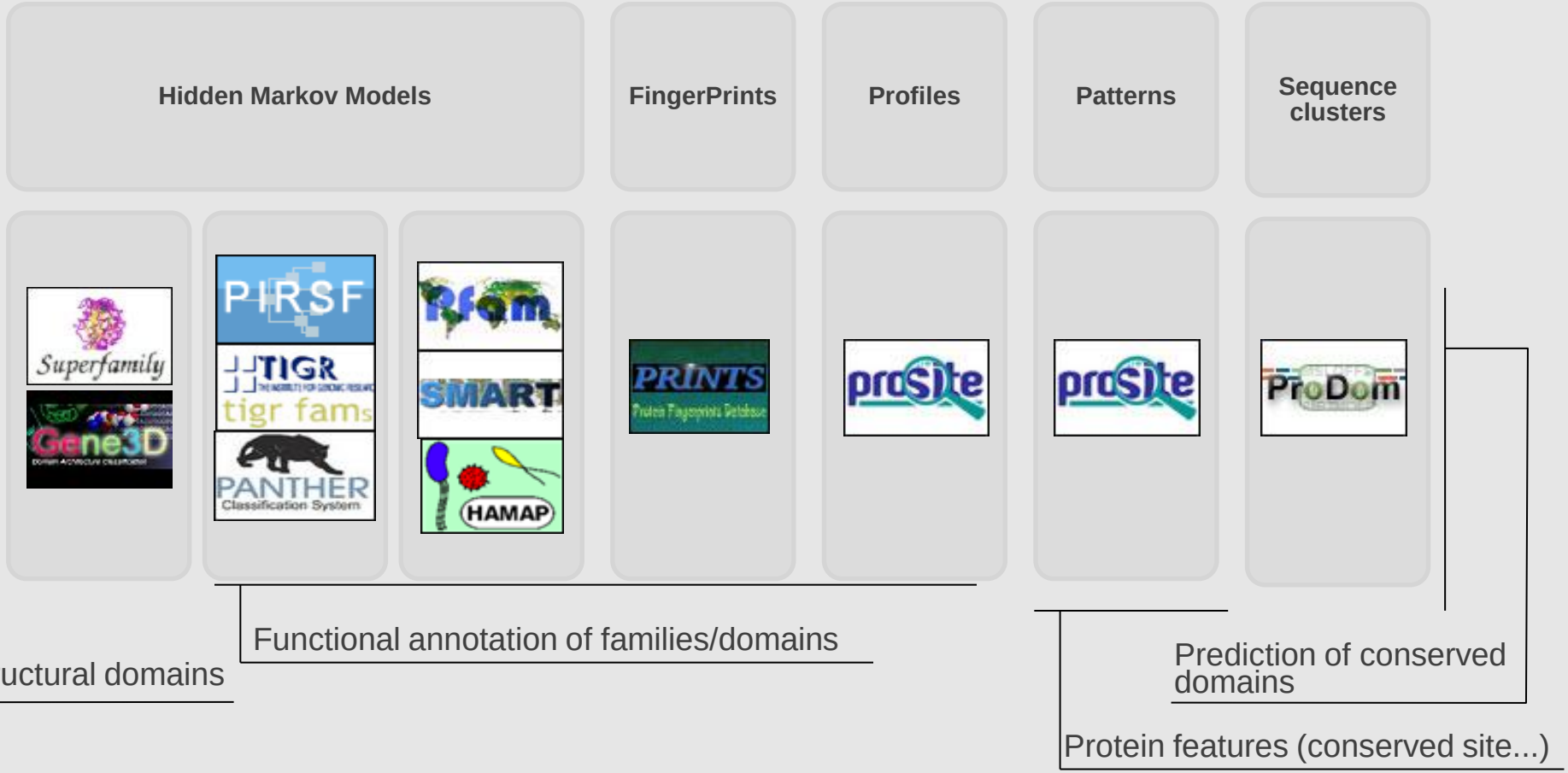
Our member databases all have their particular niche or focus...

...but InterPro is a combination of all their areas of expertise!

Some usage statistics

- InterPro 28.0: **204 145** signatures covering **85.0% of UniProtKB**
- Frequent releases – both protein and method updates
- **45 000** unique visitors per month
- InterProScan: **11 266 969** requests YTD

Different member databases are attempting to describe different things



How we build InterPro

Nature Precedings : doi:10.1038/npre.2010.5330.1 : Posted 25 Nov 2010

Pfam
ProDom
PROSITE
PRINTS
Superfamily
Gene3D
SMART
PIRSF
TIGRFAMs
PANTHER Classification System
HAMAP

Models

Sequences

....GVPDEEELWD....
....PQQNEVLKMEA..
....GGGPQLMSSTY....
....WLEAIRNNEQQ...

UniParc
The UniProt Archive

some member databases provide annotation

Curators

InterProScan (internal)

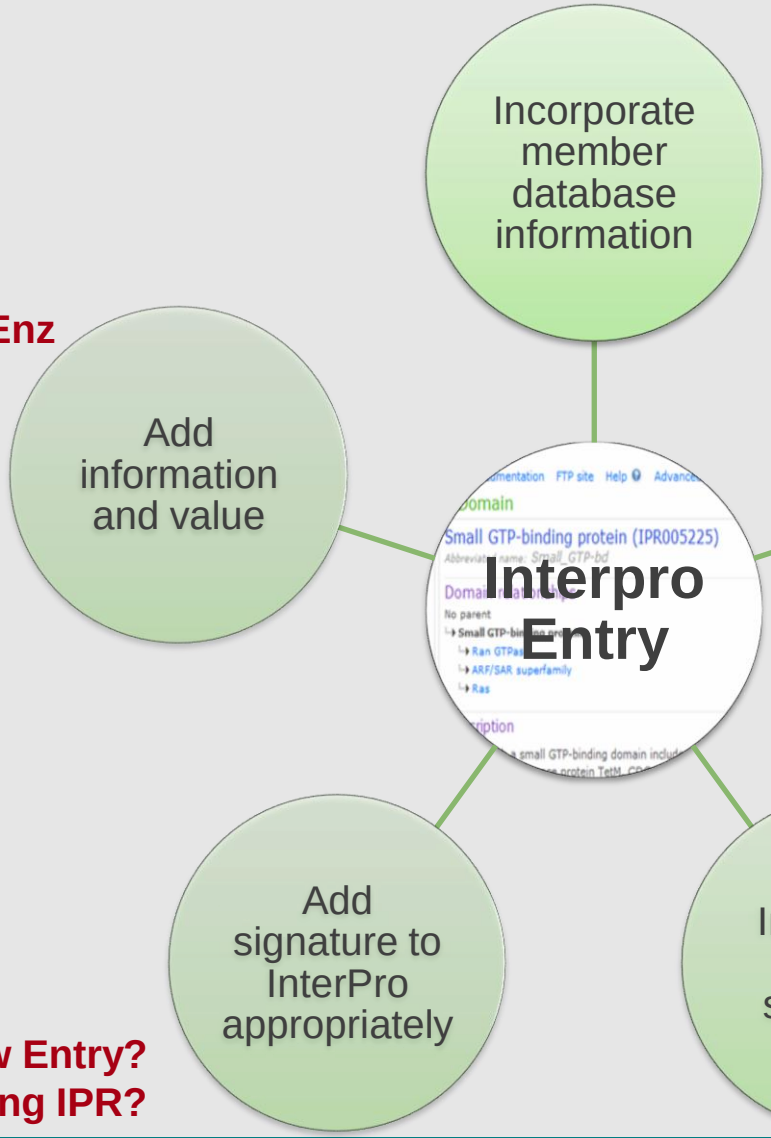
InterPro

InterPro Protein sequence analysis & classification

UniProt

EMBL-EBI

Our job as InterPro Curators



Some db's provide annotation, GO mapping

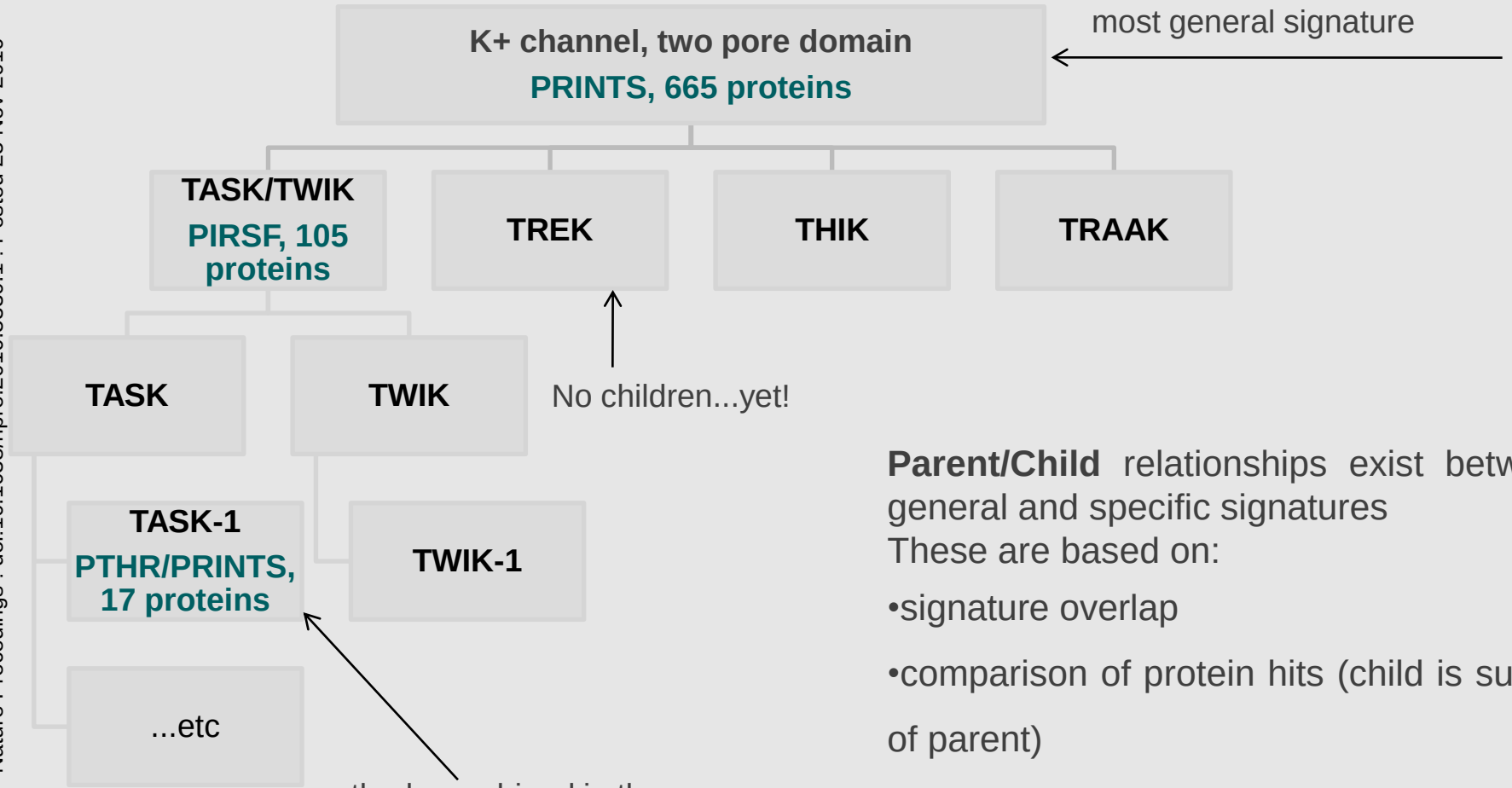
- **Abstract creation**
- **Links to other databases e.g. IntEnz**
- **Hierarchies**
- **GO mapping**

All signatures seen by at least two curators

New Entry?
Add to an existing IPR?

- **UniProt/NCBI nomenclature**
- **IUPHAR**

Examples of InterPro hierarchies: K+ channel families



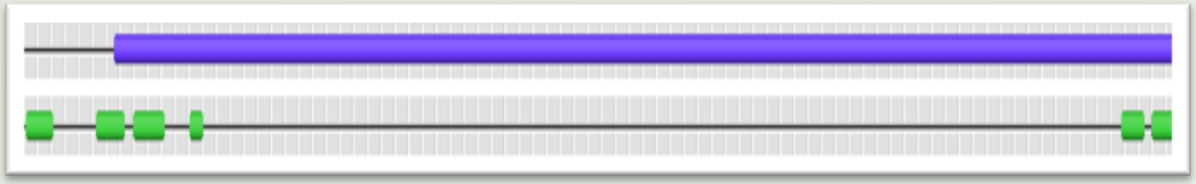
Parent/Child relationships exist between general and specific signatures
These are based on:

- signature overlap
- comparison of protein hits (child is subset of parent)
- Existing hierarchies in member databases
- **Biological knowledge of curators**

Family Parent/Child relationships need curating

Not always possible to use signature overlap to determine how family signatures are related

e.g. High molecular weight glutenins



PF03157 220 protein hits

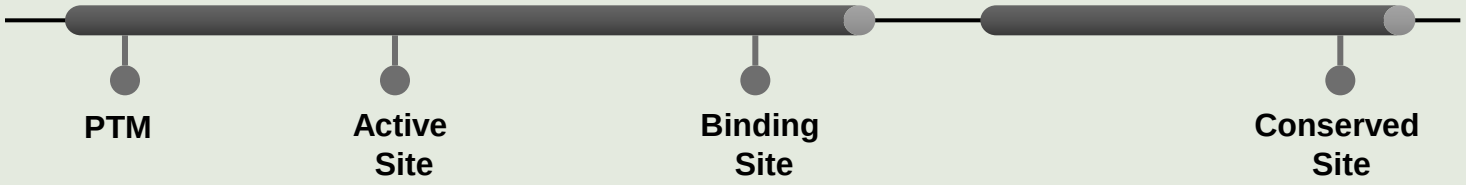
PR00210 232 protein hits

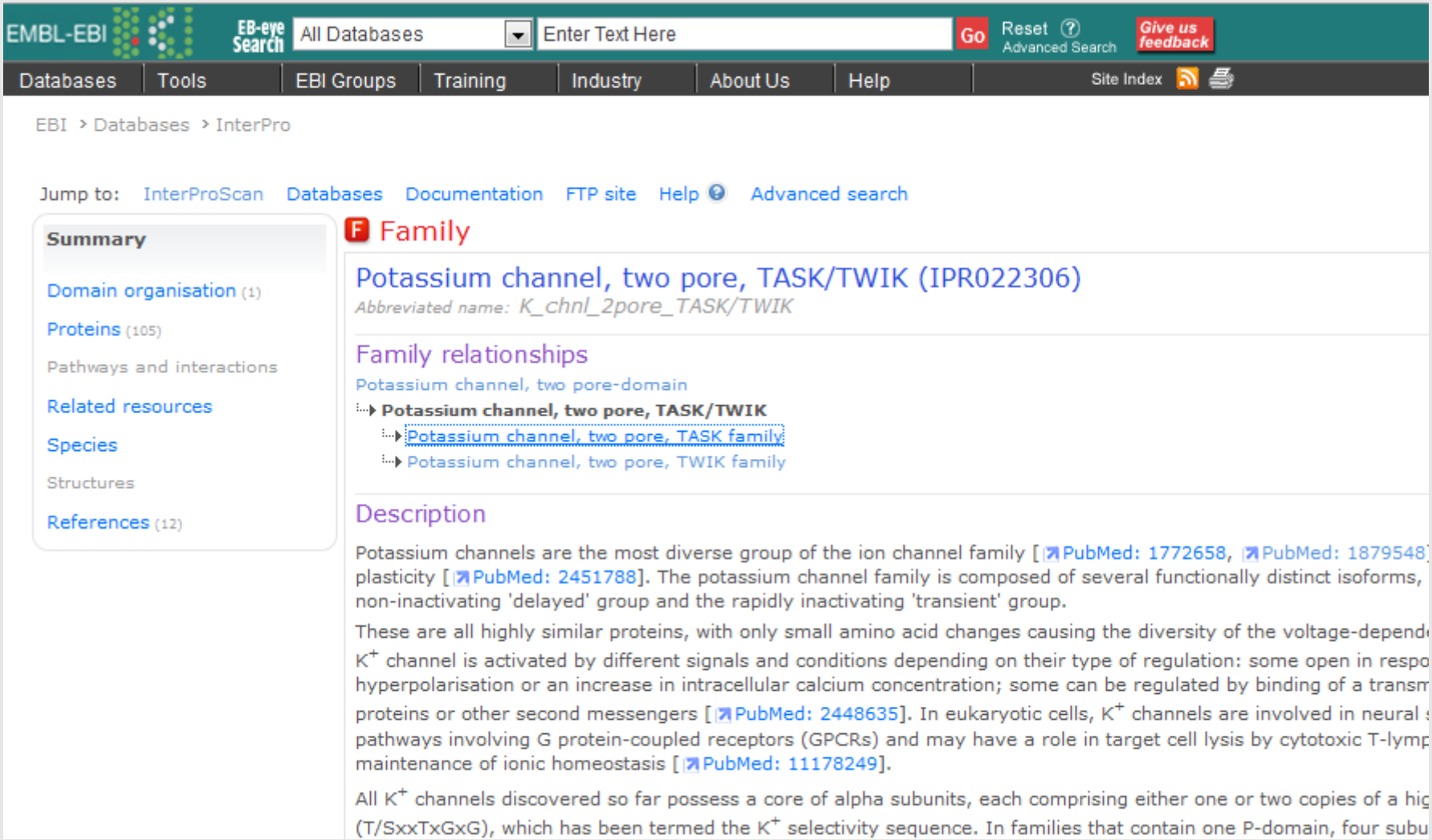
Two very different signatures both describing the same thing!

- Children must define a subset of their parent
- Protein function, match count, mDB hierarchies all considered
- Parent must make all the relationships a Child can make
- Siblings should not have matches in common

Parent/child relationships must be based in biology!

Sites and other sequence features are not involved in relationships

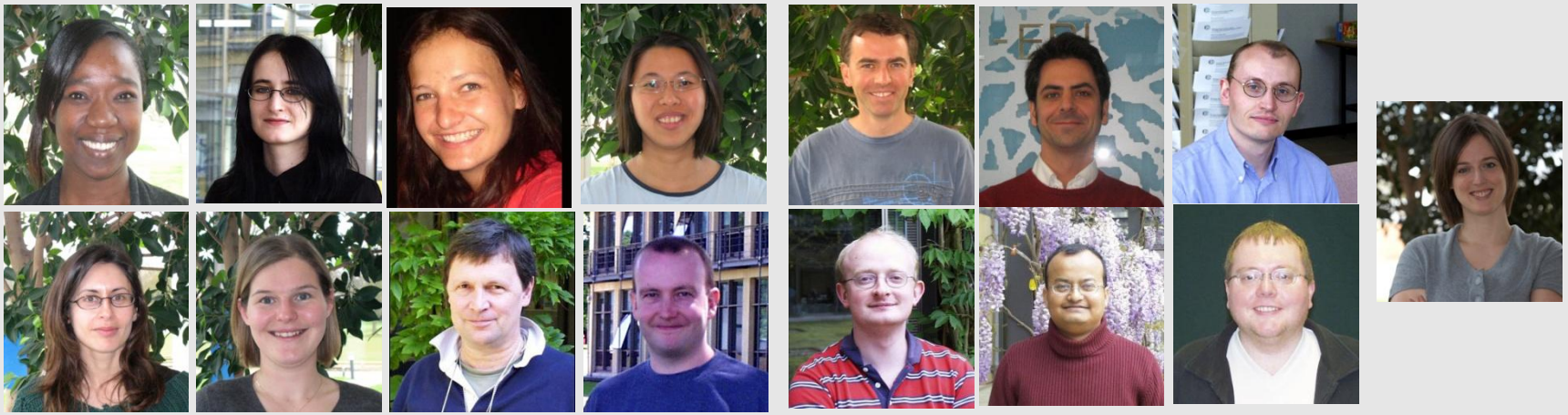




Acknowledgments

Nature Precedings : doi:10.1038/npre.2010.5330.1 : Posted 25 Nov 2010

- InterPro Team



- Member Databases

- Funders