

# InterPro Curation: Integrating Predictive Protein Signatures Into Biological Hierarchies

Dr Sarah Burge  
Curator, InterPro  
[swb@ebi.ac.uk](mailto:swb@ebi.ac.uk)



# 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...**

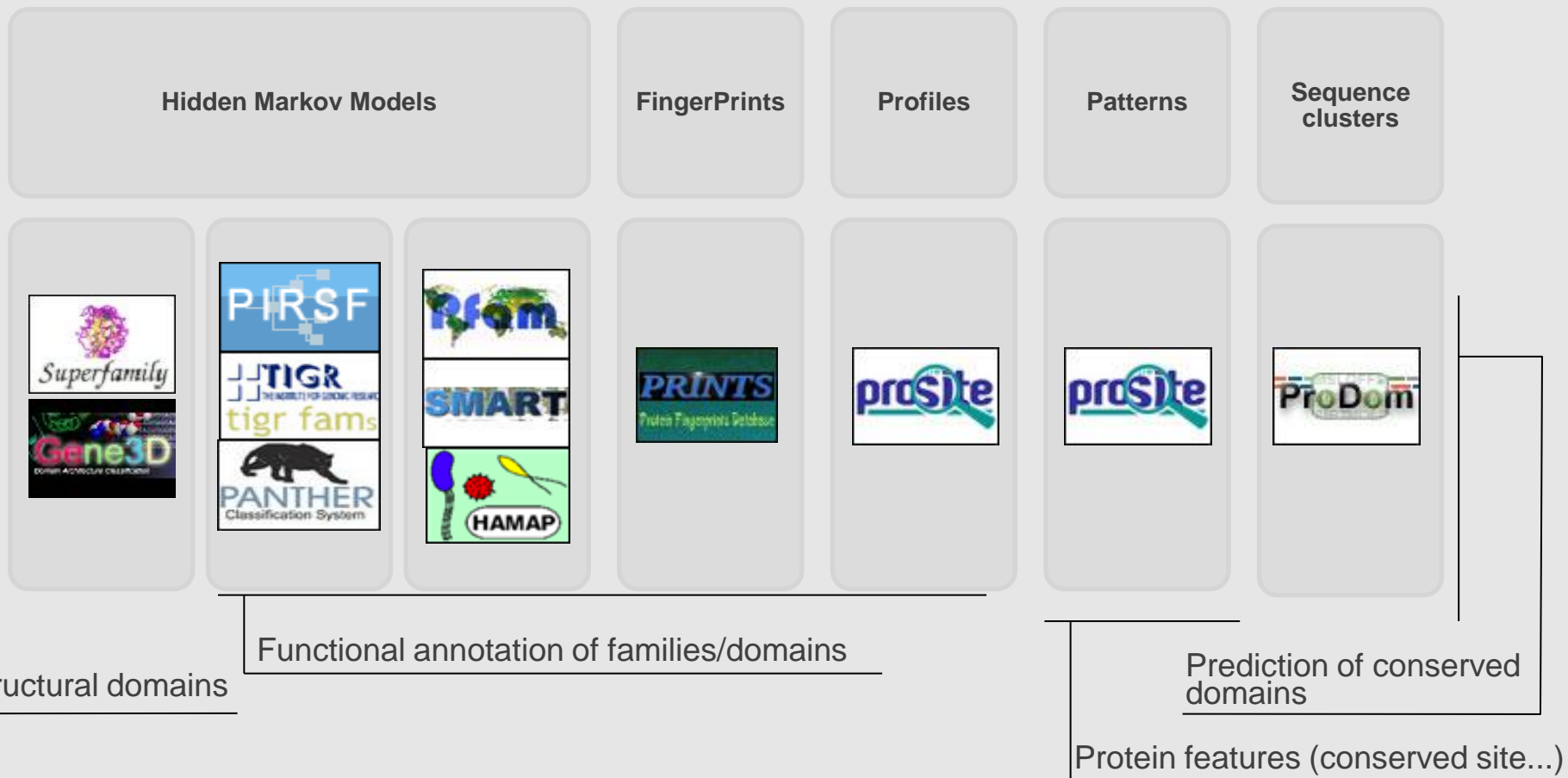
**...but InterPro is a combination of all their areas of expertise!**

# 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

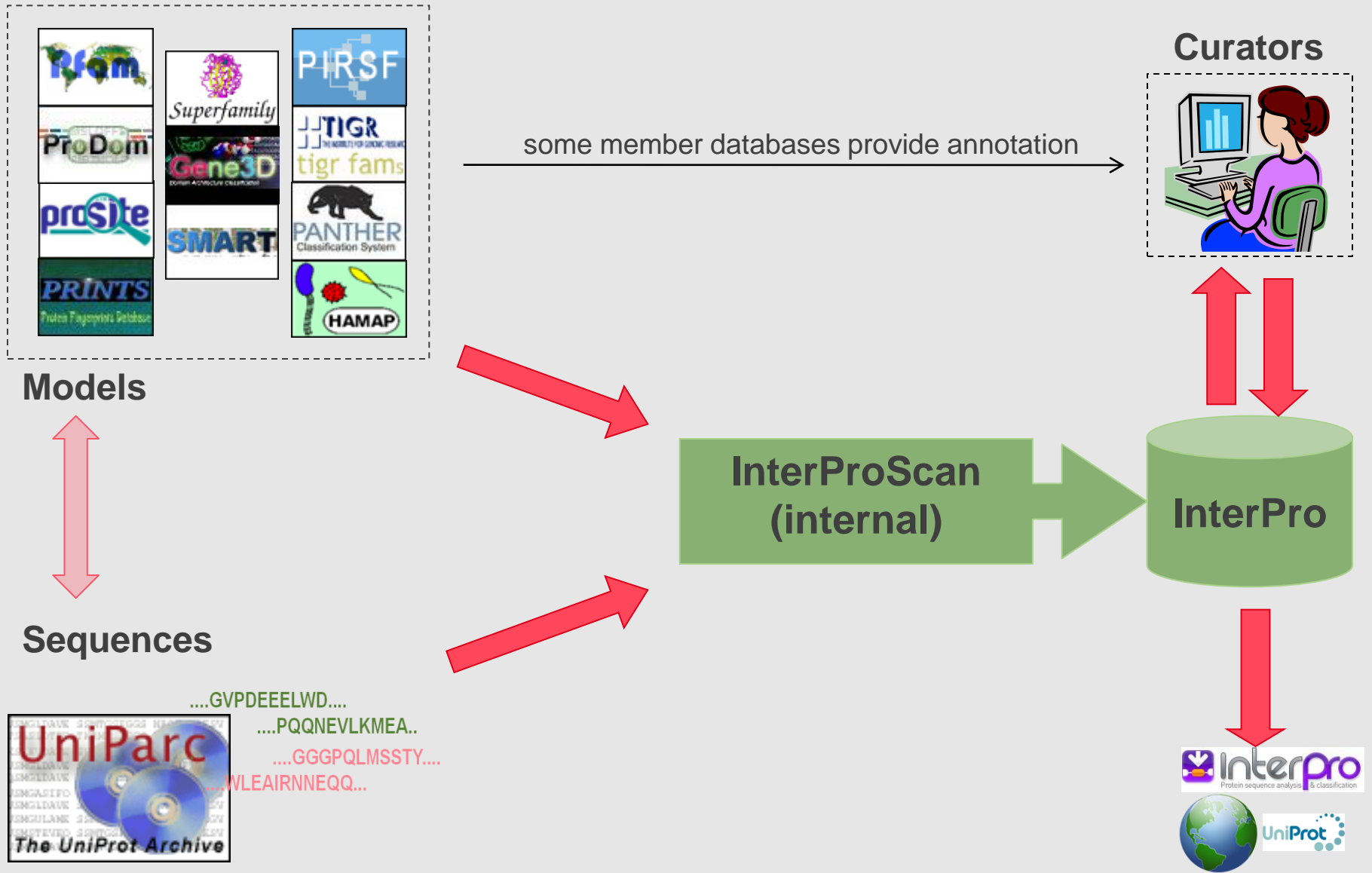
# Different member databases are attempting to describe different things

Nature Precedings .doi:10.1038/npre.2010.5330.1.1. Posted 25 Nov 2010



# How we build InterPro

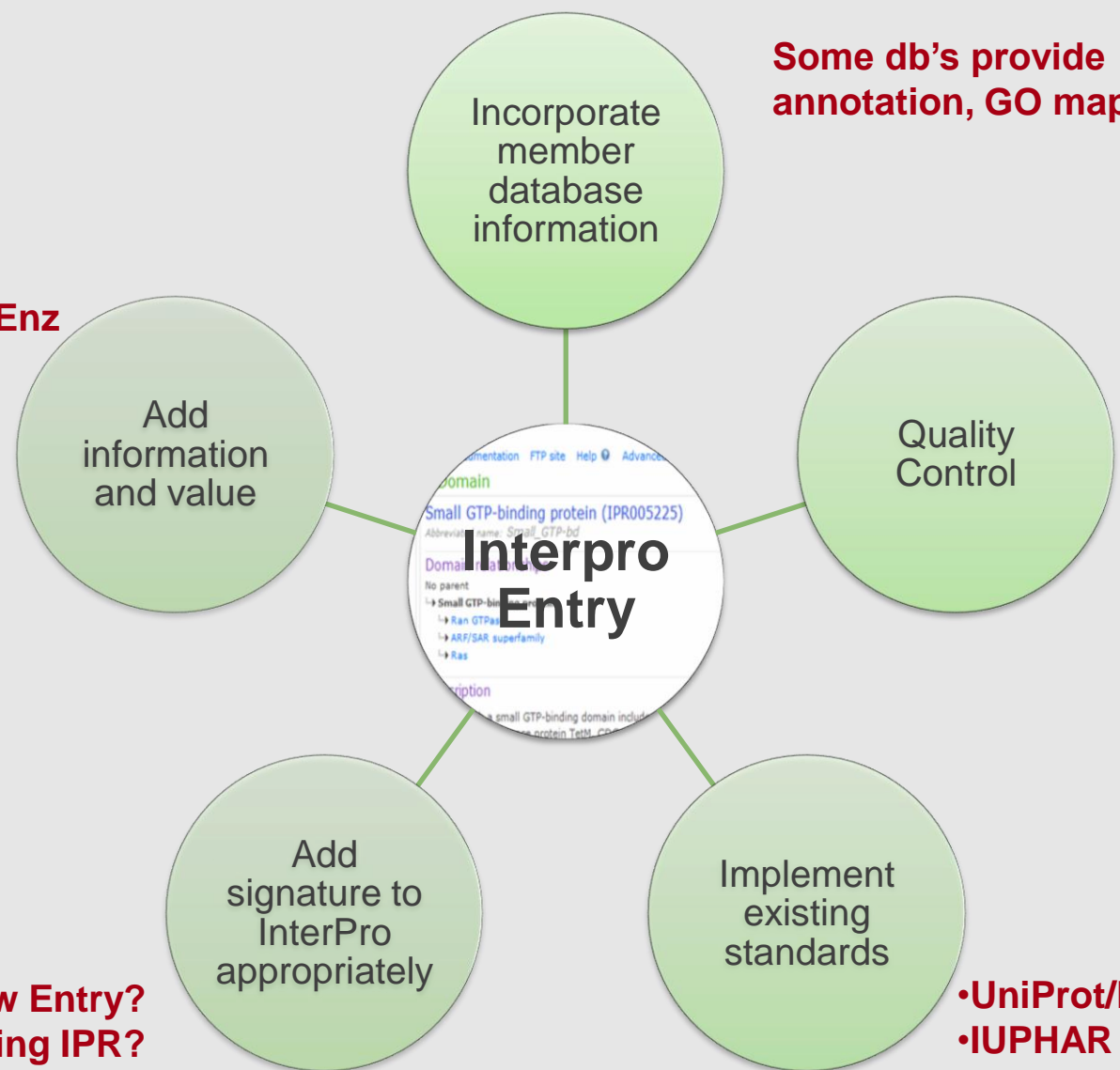
Nature Precedings : doi:10.1038/npre.2010.5330.1 : Posted 25 Nov 2010



# Our job as InterPro Curators

Nature Precedings : doi:10.1038/npre.2010.5330111 posted 25 Nov 2010

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



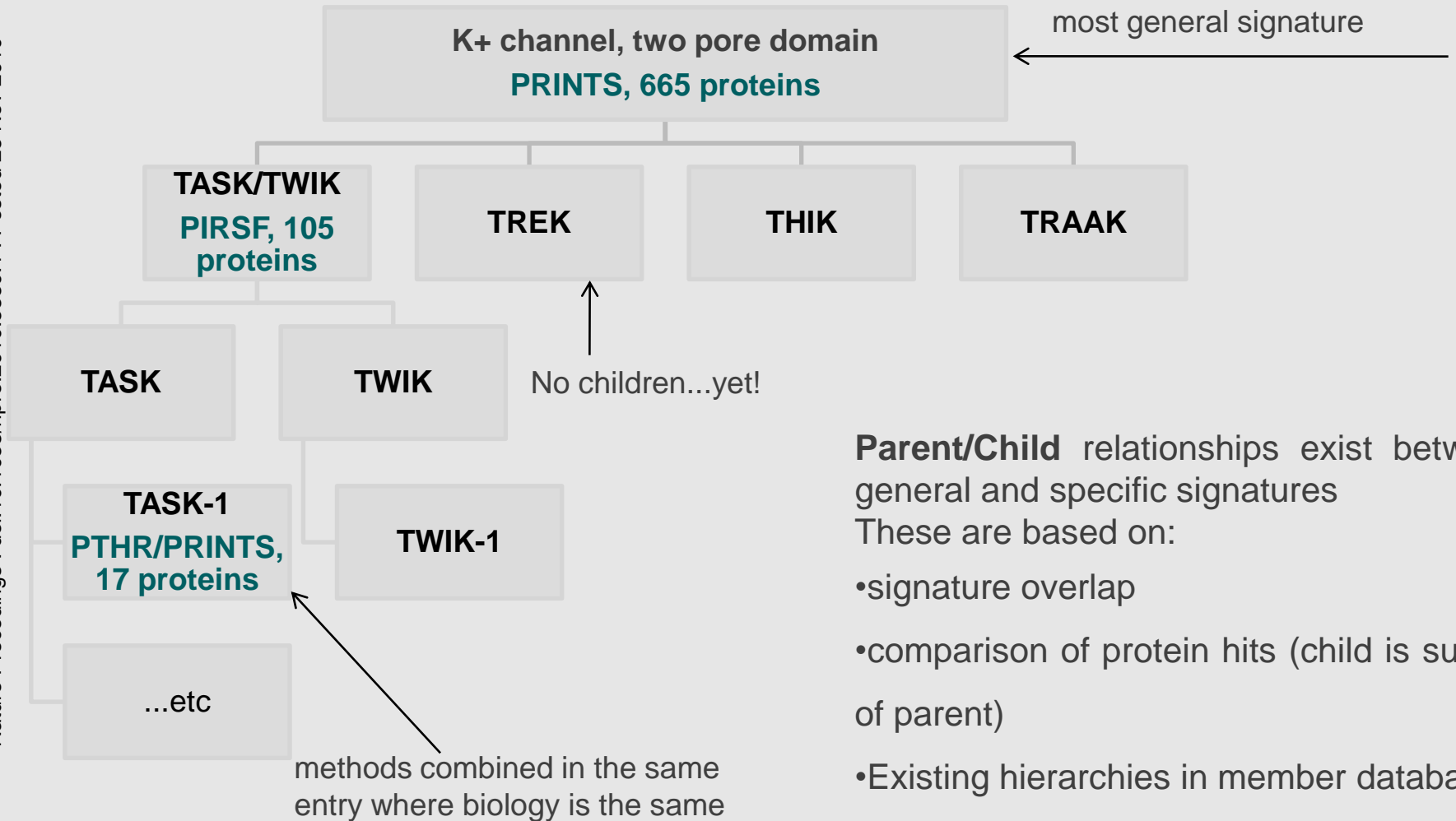
Some db's provide annotation, GO mapping

All signatures seen by at least two curators

New Entry?  
Add to an existing IPR?

- UniProt/NCBI nomenclature
- IUPHAR

# Examples of InterPro hierarchies: K<sup>+</sup> channel families



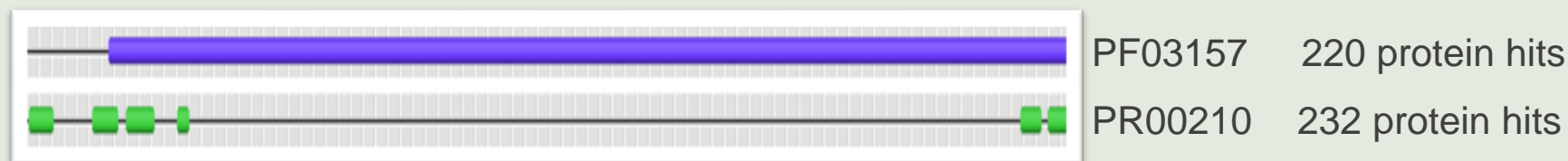
**Parent/Child** relationships exist between general and specific signatures  
These are based on:

- 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

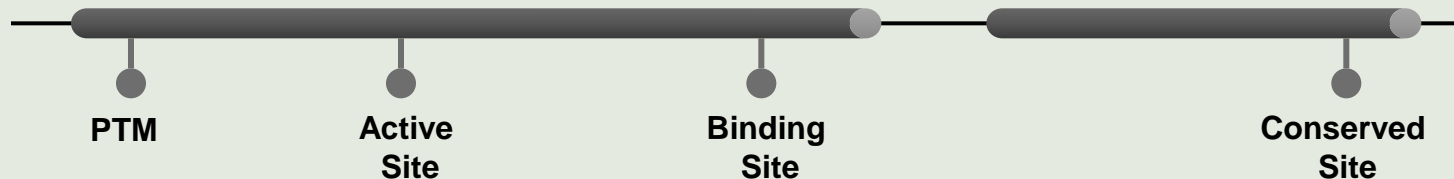


Two very different signatures both describing the same thing!

- 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





**Summary**[Domain organisation](#) (1)[Proteins](#) (105)[Pathways and interactions](#)[Related resources](#)[Species](#)[Structures](#)[References](#) (12)**F Family****Potassium channel, two pore, TASK/TWIK (IPR022306)***Abbreviated name: K\_chnl\_2pore\_TASK/TWIK***Family relationships**[Potassium channel, two pore-domain](#)↳ **Potassium channel, two pore, TASK/TWIK**↳ [Potassium channel, two pore, TASK family](#)↳ [Potassium channel, two pore, TWIK family](#)**Description**

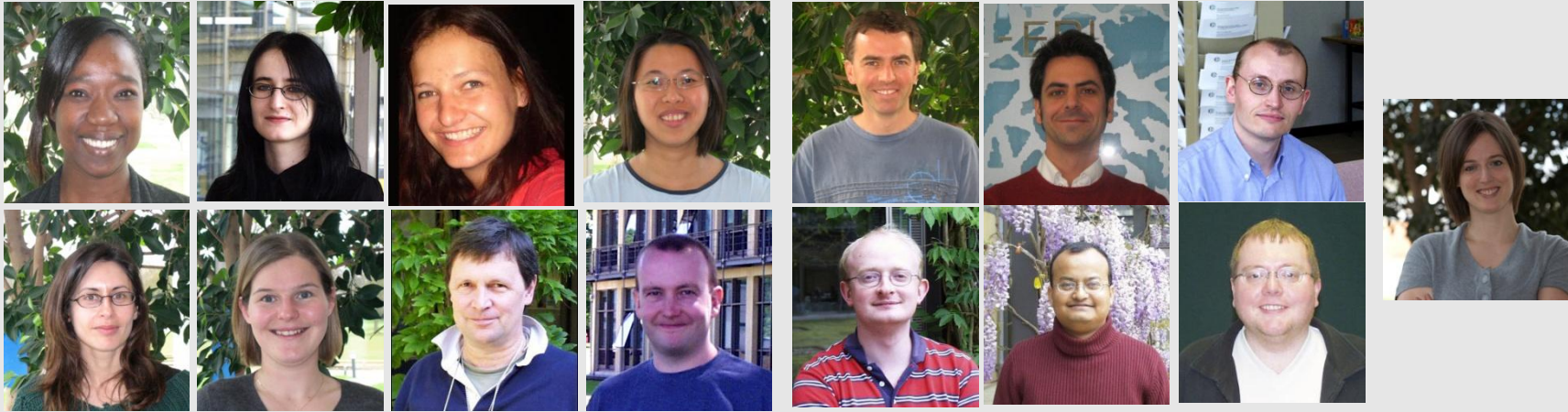
Potassium channels are the most diverse group of the ion channel family [[PubMed: 1772658](#), [PubMed: 1879548](#)]. The potassium channel family is composed of several functionally distinct isoforms, non-inactivating 'delayed' group and the rapidly inactivating 'transient' group.

These are all highly similar proteins, with only small amino acid changes causing the diversity of the voltage-dependent  $K^+$  channel is activated by different signals and conditions depending on their type of regulation: some open in response to hyperpolarisation or an increase in intracellular calcium concentration; some can be regulated by binding of a transmembrane protein or other second messengers [[PubMed: 2448635](#)]. In eukaryotic cells,  $K^+$  channels are involved in neural signalling pathways involving G protein-coupled receptors (GPCRs) and may have a role in target cell lysis by cytotoxic T-lymphocytes. Maintenance of ionic homeostasis [[PubMed: 11178249](#)].

All  $K^+$  channels discovered so far possess a core of alpha subunits, each comprising either one or two copies of a highly conserved sequence (T/SxxTxGxG), which has been termed the  $K^+$  selectivity sequence. In families that contain one P-domain, four subunits

# Acknowledgments

- InterPro Team



- Member Databases



- Funders

