

# InterPro in 2017—beyond protein family and domain annotations

Robert D. Finn<sup>1,\*</sup>, Teresa K. Attwood<sup>2</sup>, Patricia C. Babbitt<sup>3</sup>, Alex Bateman<sup>1</sup>, Peer Bork<sup>4</sup>, Alan J. Bridge<sup>5</sup>, Hsin-Yu Chang<sup>1</sup>, Zsuzsanna Dosztányi<sup>6</sup>, Sara El-Gebali<sup>1</sup>, Matthew Fraser<sup>1</sup>, Julian Gough<sup>7</sup>, David Haft<sup>8</sup>, Gemma L. Holliday<sup>3</sup>, Hongzhan Huang<sup>9</sup>, Xiaosong Huang<sup>10</sup>, Ivica Letunic<sup>11</sup>, Rodrigo Lopez<sup>1</sup>, Shennan Lu<sup>12</sup>, Aron Marchler-Bauer<sup>12</sup>, Huaiyu Mi<sup>10</sup>, Jaina Mistry<sup>1</sup>, Darren A Natale<sup>13</sup>, Marco Necci<sup>14</sup>, Gift Nuka<sup>1</sup>, Christine A. Orengo<sup>15</sup>, Youngmi Park<sup>1</sup>, Sebastien Pesseat<sup>1</sup>, Damiano Piovesan<sup>14</sup>, Simon C. Potter<sup>1</sup>, Neil D. Rawlings<sup>1</sup>, Nicole Redaschi<sup>5</sup>, Lorna Richardson<sup>1</sup>, Catherine Rivoire<sup>5</sup>, Amaia Sangrador-Vegas<sup>1</sup>, Christian Sigrist<sup>5</sup>, Ian Sillitoe<sup>15</sup>, Ben Smithers<sup>7</sup>, Silvano Squizzato<sup>1</sup>, Granger Sutton<sup>8</sup>, Narmada Thanki<sup>12</sup>, Paul D Thomas<sup>10</sup>, Silvio C. E. Tosatto<sup>14,16</sup>, Cathy H. Wu<sup>9</sup>, Ioannis Xenarios<sup>5</sup>, Lai-Su Yeh<sup>13</sup>, Siew-Yit Young<sup>1</sup> and Alex L. Mitchell<sup>1</sup>

<sup>1</sup>European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, <sup>2</sup>School of Computer Science, University of Manchester, UK, <sup>3</sup>Department of Bioengineering & Therapeutic Sciences, University of California, San Francisco, CA 94143, USA, <sup>4</sup>European Molecular Biology Laboratory, Biocomputing, Meyerhofstasse 1, 69117 Heidelberg, Germany, <sup>5</sup>Swiss-Prot Group, SIB Swiss Institute of Bioinformatics, CMU, 1 rue Michel-Servet, CH-1211 Geneva 4, Switzerland, <sup>6</sup>MTA-ELTE Lendület Bioinformatics Research Group, Department of Biochemistry, Eötvös Loránd University, Pázmány Péter sétány 1/c, Budapest, Hungary, <sup>7</sup>Computer Science department, University of Bristol, Woodland Road, Bristol BS8 1UB, UK, <sup>8</sup>Bioinformatics Department, J. Craig Venter Institute, 9714 Medical Center Drive, Rockville, MD 20850, USA, <sup>9</sup>Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE 19711, USA, <sup>10</sup>Division of Bioinformatics, Department of Preventive Medicine, University of Southern California, Los Angeles, CA 90033, USA, <sup>11</sup>Biobyte Solutions GmbH, Bothestr. 142, 69126 Heidelberg, Germany, <sup>12</sup>National Center for Biotechnology Information, National Library of Medicine, NIH Bldg, 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA, <sup>13</sup>Georgetown University Medical Center, 3300 Whitehaven St, NW, Washington, DC 20007, USA, <sup>14</sup>Department of Biomedical Sciences and CRIBI Biotech Center, University of Padua, via U. Bassi 58/b, 35131 Padua, Italy, <sup>15</sup>Structural and Molecular Biology, University College London, Darwin Building, London WC1E 6BT, UK and <sup>16</sup>CNR Institute of Neuroscience, via U. Bassi 58/b, 35131 Padua, Italy

Received October 24, 2016; Editorial Decision October 25, 2016; Accepted October 27, 2016

## ABSTRACT

InterPro (<http://www.ebi.ac.uk/interpro/>) is a freely available database used to classify protein sequences into families and to predict the presence of important domains and sites. InterProScan is the underlying software that allows both protein and nucleic acid sequences to be searched against InterPro's predictive models, which are provided by its member databases. Here, we report recent developments with InterPro and its associated software, in-

cluding the addition of two new databases (SFLD and CDD), and the functionality to include residue-level annotation and prediction of intrinsic disorder. These developments enrich the annotations provided by InterPro, increase the overall number of residues annotated and allow more specific functional inferences.

## INTRODUCTION

In the post-genomic era, generation of biological sequence data is no longer a scientific barrier; rather, data storage and analysis have become the new bottlenecks in terms of cost

\*To whom correspondence should be addressed. Tel: +44 1223 492 679; Fax: +44 1223 494 46; Email: rdf@ebi.ac.uk

and time (1). With the potential to sequence entire genomes, or to generate hundreds of millions of sequences from environmental samples, the pace of generating sequence data now outstrips the rate of experimental characterisation by many orders of magnitude (2). Consequently, rapid, accurate automatic functional annotation of large numbers of sequences has become a major challenge.

The InterPro database aims to meet this challenge by integrating diverse information about protein families, domains and functional sites. Central to the resource are diagnostic models (profile hidden Markov models (HMMs), profiles, position-specific scoring matrices or regular expressions, collectively known as ‘signatures’), against which protein sequences can be searched to determine their potential functions. The signatures are provided by 14 different member databases: 12 of these are long-standing members of the InterPro Consortium (CATH-Gene3D (3), HAMAP (4), PANTHER (5), Pfam (6), PIRSF (7), PRINTS (8), ProDom (9), PROSITE Patterns (10), PROSITE Profiles (10), SMART (11), SUPERFAMILY (12) and TIGRFAMs (13)); two are new members, the Conserved Domains Database (CDD) (14) and Structure-Function Linkage Database (SFLD) (15) having been added in 2016.

The source databases each have their own individual biological focus, method of signature production, and/or signature-match processing. The diversity of approaches helps to ensure that annotations are as comprehensive as possible. For example, related Pfam (profile HMM-based) and Prosite Profiles entries often match subtly different sets of proteins; united however, they match most, if not all, known members of a protein family, while eliminating false-positive annotations. Furthermore, the different databases offer complementary levels of protein classification, from broad-level (e.g. a protein is a member of a superfamily) to more fine-grained assignments (e.g. a protein is a member of a specific family, or possesses a particular type of domain). These different levels of granularity are used by InterPro to produce a hierarchical classification system: one or more-member database signatures are integrated into an InterPro entry, and, where appropriate, relationships are highlighted between different entries, identifying those that represent smaller, functionally specific subsets of a broader entry.

### Database curation

InterPro entries are classified into types (families, domains, repeats or sites) depending on the biological entity they represent. Family and domain entries are placed into distinct, non-overlapping hierarchies: domain entries are able to occur in the same hierarchy as other domains, but not within the same hierarchy as family entries, and vice versa.

Entries are manually annotated with literature-referenced free-text descriptions, explaining the biological information that may be inferred for proteins that match a given signature. Where possible, each entry is also associated with Gene Ontology (GO) (16) terms, which provide a controlled vocabulary to describe protein function, cellular localisation and involvement in wider biological pathways and processes. The granularity of the member database signatures (and hence InterPro entries) determines the specificity of the functional annotation and GO terms that

can be assigned. For example, an InterPro entry representing a small family of functionally conserved enzymes that act on a single substrate, such as the glycerol kinases (IPR005999), is annotated with more specific functional information and terms from the GO hierarchy than an entry representing a more diverse enzyme family acting on a larger class of substrates, such as their parent InterPro entry, the FGGY carbohydrate kinases (IPR000577) (see Figure 1).

At each release of the database, InterPro entries are checked, and updated where necessary, to ensure that the annotations remain accurate. Updates to annotations are typically made in response to changes in signature specificity (e.g. if a signature has been rebuilt by a member database to recognise a tighter functionally-related group of proteins, or to match more distant homologues) or to improved scientific understanding of the function of a protein family (17).

InterPro entries are also automatically annotated with cross-references to a range of relevant databases, including the pathway databases ENZYME (18), MetaCyc (19), UniPathway (20) and KEGG (21), and various 3D structure databases (22,23).

### Use of the resource

InterPro plays a major role in the analysis and annotation of sequence data held in the UniProt Knowledgebase (UniProtKB), the central hub of protein sequences (24). Protein annotations derived from InterPro’s member database signatures are calculated using the InterProScan software package (25) on a monthly basis. These annotations are then used by UniProtKB curators to help annotate Swiss-Prot records and as input to the automated systems that add annotation to UniProtKB/TrEMBL. InterPro’s protein match information is also made available to the public via XML files, and the database’s Web interfaces and services, which can be searched with a protein sequence, a UniProtKB protein identifier, an InterPro or member database identifier, GO term, or free text.

In addition to its use in UniProtKB annotation, InterPro is widely used by the scientific community. Its data feed into a host of annotation pipelines, including ENSEMBL (26), ENSEMBL Genomes (27), PDBe (28), BLAST2GO (29), PhytoPath (30), the digenic diseases database - DIDA (31), and the Endeavour candidate gene prioritisation server (32). The InterProScan Web services are extensively used, processing in excess of 40 million sequence searches per month. InterPro’s data and analysis software are increasingly used in the analysis of metagenomic data: in addition to underpinning EMBL-EBI’s in-house EBI Metagenomics resource (33), the MEGAN metagenomics analysis tool (34) now has the ability to process InterPro-derived data and GO terms.

### New protein family member databases

The constituent databases of InterPro have remained fixed for the last 7 years (see Figure 2), and have changed little in the last decade. Since the last update paper (35), we have been evaluating a range of different databases

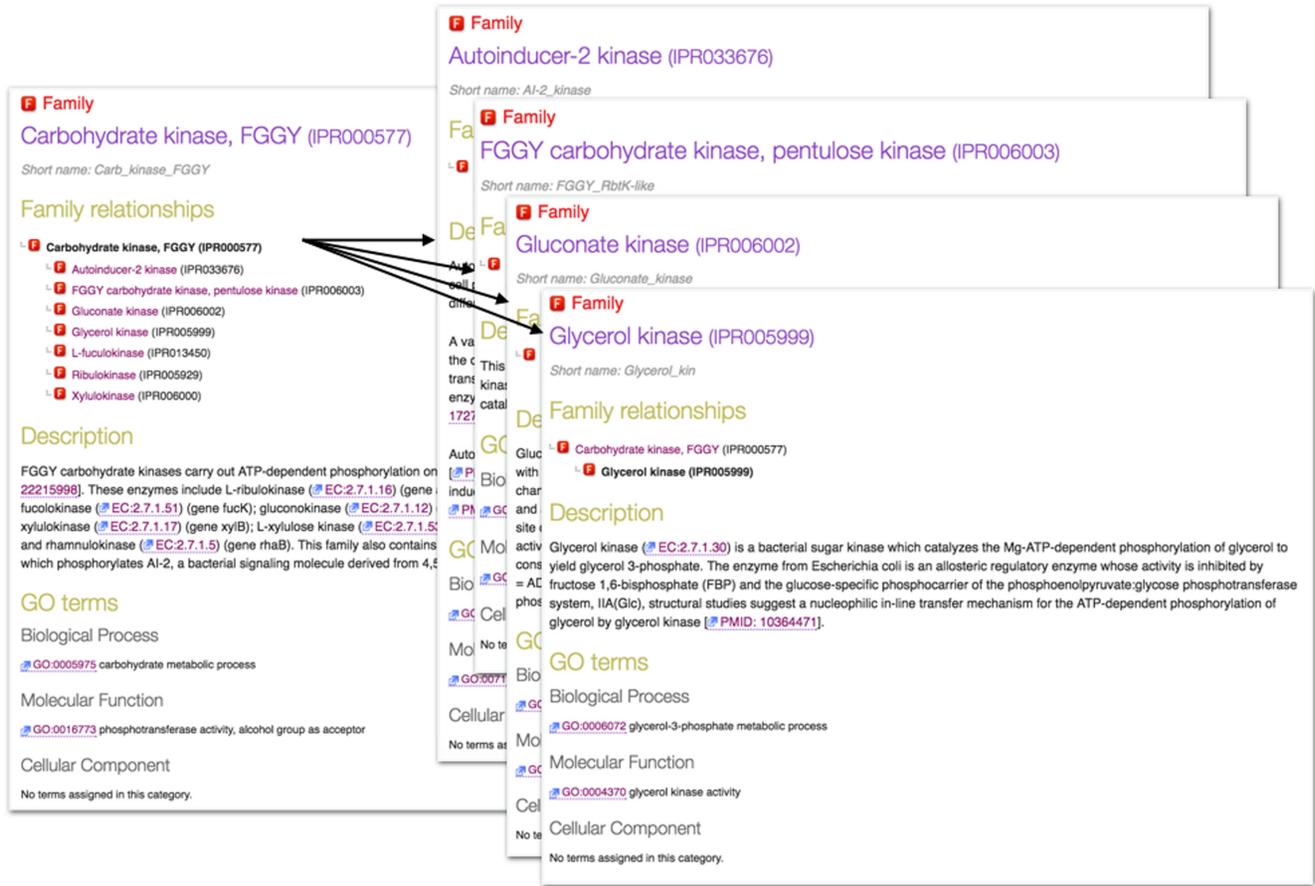


Figure 1. Example of an InterPro family hierarchical relationship. The FGGY carbohydrate kinases entry (IPR000577) provides a parent to a series of child entries that match smaller, more functionally-specific sets of proteins.

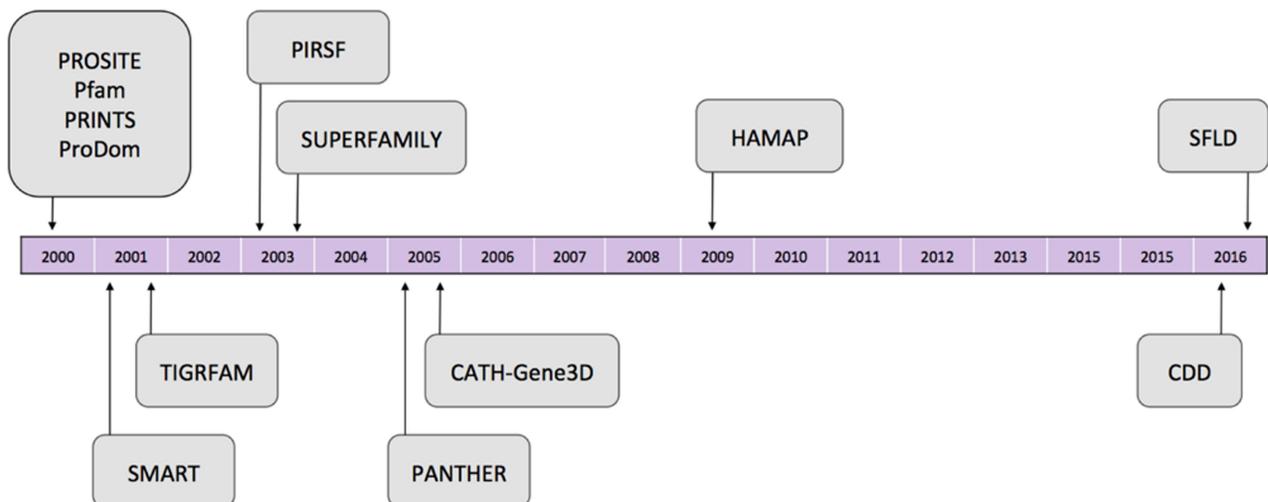


Figure 2. Timeline showing the member databases that have joined InterPro since version 1.0, released in 2000.

that could enhance the information contained within InterPro, in terms both of adding more comprehensive coverage and in providing more fine-grained functional annotations at the individual amino acid level. As a result, two new databases have been added to the resource: the Conserved Domains Database (CDD) and Structure–Function Linkage Database (SFLD).

CDD is a manually curated protein annotation resource representing domain footprints conserved in molecular evolution. Each domain entry is modelled as a multiple sequence alignment, which is also converted into a position-specific scoring matrix (PSSM) that allows fast identification of conserved domains in protein sequences via RPS-BLAST (36). NCBI-curated domains use 3-dimensional structure information to explicitly define the boundaries of known conserved domains, and to provide comprehensive and accurate annotation of protein sequences with the locations and boundaries (footprints) of known conserved domains, including location of conserved/functional sites. CDD content also includes domain models imported from a number of external source databases (Pfam, SMART, COG (37), PRK (38) and TIGRFAMs). Only CDD's own models have been imported into InterPro, mainly owing to the fact that Pfam, SMART and TIGRFAMs are already present in InterPro. Within InterProScan, the RPS-BLAST is substituted by a piece of software called 'rpsbproc', an amendment to the standalone RPS-BLAST program, which allows detailed CD-Search results, including domain superfamily assignments and the predicted locations of conserved sites to be reproduced locally. As well as being bundled into InterProScan, the rpsbproc utility is available from the CDD FTP site at <ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd>. This package ensures that InterPro can faithfully reproduce the results from CDD, thus ensuring consistency of annotations between the CDD and InterPro Web servers for the same sequence.

SFLD is a manually-curated classification resource describing structure-function relationships for functionally diverse enzyme superfamilies. Members within a single superfamily are derived from a common ancestor and mediate a diverse set of related, yet distinct reactions. For example, the enzymes within a superfamily may share active-site features (such as residues in a nucleophile) associated with conserved functional attributes (e.g. part of a reaction mechanism or substrate binding). Consequently, in some annotation resources, such superfamilies are grouped together in a single entry (e.g. the Radical\_SAM entry in Pfam, PF04055), but are inadequately subdivided into their functional groups, leading to imperfect annotations.

To overcome annotation errors and enable transfer of functional features, SFLD provides hierarchical annotations at multiple levels. SFLD subdivides superfamilies into subgroups based upon sequence information, then into families of enzymes known to catalyze the same reaction using the same mechanistic strategy. The family level of the hierarchy defines variations in each set of active site residues that distinguish that family's particular reaction type from other families in the superfamily (Figure 3).

To enable InterPro to perform SFLD annotations, the two resources have developed an approach whereby SFLD produces multiple sequence alignments representing the dif-

ferent superfamilies, subgroups and families. Most of the alignments, particularly those associated with subgroups and families, are annotated with key catalytic residues that are important to the chemical reaction performed by the sequences found within that set. From these alignments, InterPro builds a profile HMM library that is used to annotate sequences. Significant sequence matches are then verified against the key catalytic residues (if present), only those sequences matching all residues being assigned to the family. Thus, the profile HMMs and residue verification steps act as a two-stage assignment criterion, enabling very fine-grained annotations to be made, at rates only marginally slower than searching profiles alone.

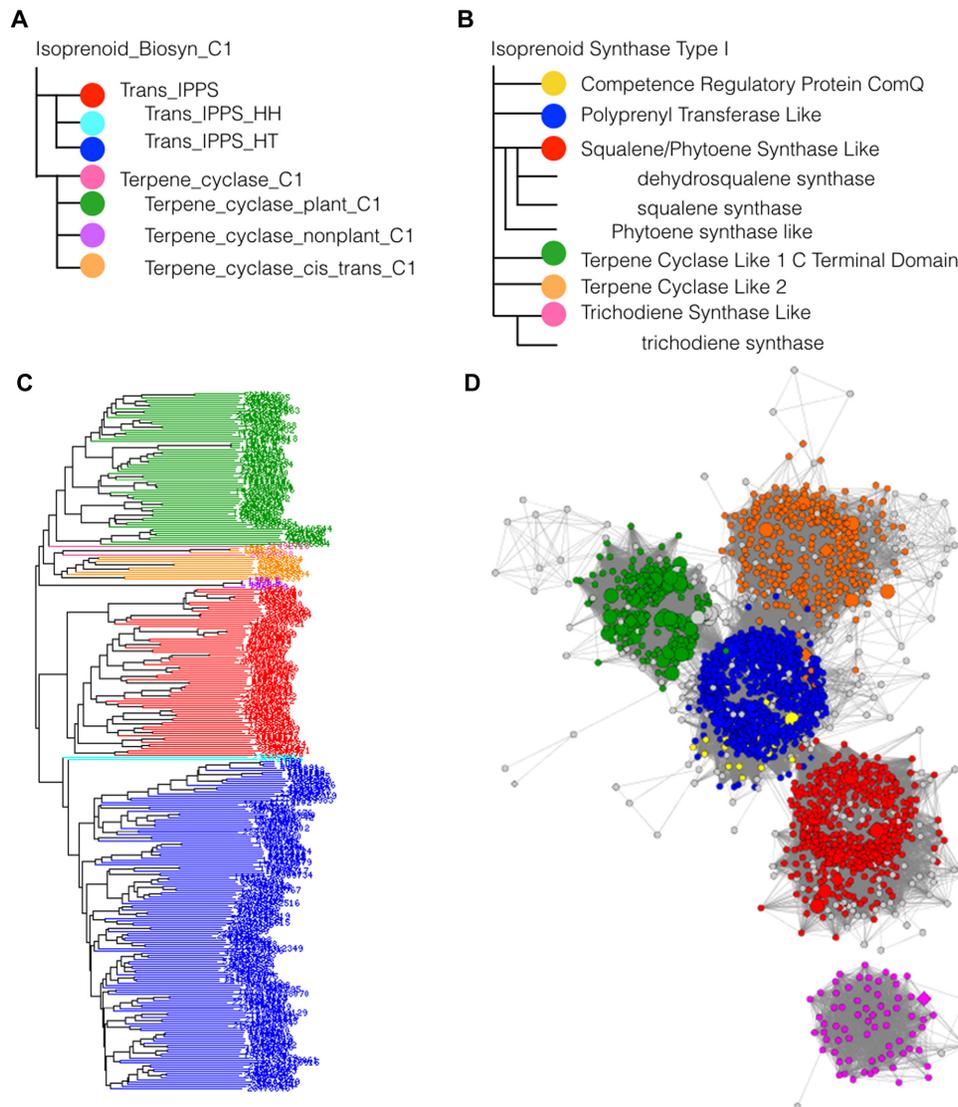
### Integration plans for CDD and SFLD

When a new member database is added to InterPro, the database is added as an entire set of unintegrated signatures, which are then manually annotated and added to InterPro entries, as described above, by the curation team. Thus, CDD and SFLD matches are now available for all signatures, but require a large curation effort to fully integrate each new resource within the InterPro hierarchy. Until integrated, the annotations from these databases are provided at the bottom of each protein page, in the 'Detailed signature matches', under the 'unintegrated signatures' listing. They are provided in a similar way in the InterProScan output. For each member database, it is possible to get the complete listing of signatures from the following pages: [http://www.ebi.ac.uk/interpro/member-database/\(member-database\)](http://www.ebi.ac.uk/interpro/member-database/(member-database)), for example <http://www.ebi.ac.uk/interpro/member-database/CDD> for the CDD database.

As of InterPro release 58.0, CDD version 3.14 added 11 273 signatures, of which 1,005 have now been integrated into InterPro over a period of 4 months. During this time, there has been a dual approach to integrating CDD signatures: (i) adding those that are directly equivalent to other member database signatures that have already been included in InterPro, as this can be done with minimal curation; (ii) adding those that match sequence sets not covered by any other member database, as these add coverage to InterPro.

SFLD, a smaller scale database with a more specific focus, has been added to InterPro more recently. SFLD version 1.0 provided 480 signatures in total, 17 of which were integrated over a 2-week period into the most recent InterPro 60.0 release. Increasing the number of integrated signatures from these databases will be a significant focus for InterPro in the forthcoming year.

Both CDD and SFLD provide hierarchical classifications (see Figure 3), but their hierarchies differ from each other and from that of InterPro. In the case of CDD, the parent signature in the hierarchy does not aim to match all of the sequences matched by its children. Rather, CDD parent entries often provide signatures aiming to cover the relatively few family member sequences that are not matched by the child signatures. As such, one of the most important functions of the parent entry is to provide a root node through which child signatures can be associated. This is significantly different from the hierarchical classification approach used in InterPro, where a parent level entry should



**Figure 3.** Examples of the CDD and SFLD hierarchies (A and B). (A) CDD models for related domains are organized hierarchically, reflecting major events in the domain family's molecular evolution and functional diversification. The hierarchy usually follows a tree structure obtained from (C) phylogenetic analysis of multiply aligned sequences. The relationship between the CDD entries in panel A and the sequences in panel B is indicated by colour. The top 'parent' entry (isoprenoid biosynthesis enzymes, Class 1 superfamily) would be less specific than the 'leaf' node entry (trans-isoprenyl diphosphate synthase, head-to-head). (B) The corresponding superfamily, Isoprenoid Synthase Type I, from SFLD. The specificity relationships between the entries is similarly arranged as in panel A. (D) SFLD network analysis graph showing the sequence identity relationships between the Isoprenoid Synthase Type I superfamily members. The *E*-value threshold for the network is  $1e-10$  and sequences within nodes share 50% or more sequence identity, calculated using CD-HIT. Note, figures C and D are visualizations from the respective source database and are not available from the InterPro website. These figures demonstrate the different approaches for visualizing and defining relationships between families.

(if possible) provide coverage of all proteins matching its child entries. As a result, CDD's higher-level signatures will not be integrated into InterPro. Instead, the aim will be to identify signatures from other member databases that can be substituted in their place, grouping the child entries together.

The SFLD hierarchy is closer to that provided by InterPro, but with some important differences. Superfamily-level entries in SFLD tend to be based upon the common amino acid core of a set of matching sequences. Descending the hierarchy, the child signatures tend to increase in length, pulling in more accessory domains, which provide

functional specificity. As a result, while most SFLD entries correspond to InterPro family-type entries, some superfamilies may appear more like domains (for example, where the functional unit of a sequence family is represented only by a single domain). InterPro aims to follow the SFLD hierarchy as closely as possible, integrating SFLD entries as 'type' Family. Nevertheless, some discrepancies will inevitably occur as we try to merge this hierarchy with InterPro's hierarchy. This is consistent with the current integration strategy for member databases that define families and domains slightly differently from InterPro.

## Updates to data content

InterPro is released publicly every 2 months, with a break in production during August, in harmony with the UniProtKB release cycle. With each InterPro release, one or more of its member databases may have been updated, providing a stream of new entries for integration into the resource. There have been 12 public InterPro releases since the last update paper, with an additional 5,158 signatures being integrated into 3,462 new InterPro entries: 4,035 of these signatures came from existing member databases, and a further 1123 from CDD and SFLD. The latest release (version 60.0) contains 41 925 member database signatures integrated into 29 700 InterPro entries. The member database updates that contributed to recent InterPro releases are shown in Table 1.

The InterPro coverage of sequences in UniProtKB (i.e. the number of proteins with one or more InterPro annotations) is calculated at each release. The signatures integrated into InterPro 60.0 provide matches to 79.8% of the sequences in UniProtKB release 2016.09 (see Table 2), compared to 83.5% in release 48.0. GO terms assigned by the InterPro2GO pipeline (which associates terms with proteins based on their InterPro matches) are cross-referenced more than 130 million times in UniProt 2016.09, representing annotation for 42 million individual proteins. This compares to 168 million terms for almost 50 million proteins in InterPro release 48.0/UniProtKB 2014.07.

The reduction in InterPro's coverage of UniProtKB since our last report may seem counterintuitive, especially as the number of sequences in UniProtKB has decreased (from a peak of 90 million sequences during 2015, to the current level of 71 million), while the number of InterPro entries and associated GO terms has gone up (the latter increasing from ~28 000 in release 48.0 to >32 000 in the current release). However, the reduction in the number of records in UniProtKB has been brought about as a redundancy removal effort, aimed at eliminating close to identical proteomes that are over-represented in the database. As part of this process, UniProtKB records belonging to ~15 000 redundant bacterial proteomes were moved from UniProtKB into UniParc. As InterPro provided high levels of coverage for these proteomes, their removal meant that InterPro's overall coverage of UniProtKB has been disproportionately affected. Increasing coverage of UniProtKB's smaller, but more diverse, sequence set will be an ongoing challenge for InterPro and its member databases in the coming years.

## Per residue annotations

We have been investigating ways to expand the scope of InterPro annotations; specifically, to individual residues that fall within a region defined by a signature. As signature-derived matches are currently based on scores across the entire matched region, the methods can often fail to discriminate between functionally distinct groups. For example, the InterPro entry for the calpain catalytic domain (IPR001300) matches >6000 UniProtKB sequences. While these are undoubtedly derived from a common ancestor, in ~2500 cases the active site residues have been mutated to residues that are no longer capable of performing the proteolytic reaction. Thus, many of the sequences are not

active peptidases, but are likely to perform different functions: e.g. calpamodulin (also known as calpain 6) has the catalytic Cys replaced by Lys, and is a microtubule-stabilising protein, particularly in embryonic muscle, where it has been shown to suppress skeletal muscle differentiation (39). One approach to achieve discrimination between active and inactive forms is to have specific subfamily signatures, e.g. the PANTHER subfamily model for calpain-6: PTHR10183:SF355. However, this requires the production and curation of many models, and there will always be cases where it is hard or impossible to ensure that the signature is capable of separating active and inactive forms, or active forms with (subtly) different mechanisms of action. In such cases, accurately annotating active site residues would help separate active peptidases from inactive homologues.

The mechanism described above is exactly the one employed by the SFLD database to identify specific groups of proteins within entries. Thus, the integration of SFLD into InterPro brings not only increased protein family coverage, but also the annotation of thousands of important residues. Other InterPro databases also provide residue-level annotations: CDD's RPS-BLAST matrices are annotated with a range of per-residue annotations, including active sites, ligand-binding, protein-protein interactions and nucleic acid-protein interactions; Pfam contains active site annotations, based on matches to UniProtKB/Swiss-Prot sequences; PIRSF annotations can also be extended to residues, using the PIRSR resource; and HAMAP and PROSITE annotation rules provide external users and the UniProt automatic annotation pipeline (UniRule) with annotations for functionally important residues (single residues as well as continuous and discontinuous motifs).

To enable the capture of this fine-grained information, we have extended the InterPro data model to deal with per-residue annotations. To date, these annotations have been enabled in InterProScan for SFLD and CDD, with the aim of adding other databases that also provide per-residue annotations in future releases. These developments provide a further tier to the annotations already provided by InterPro. They will also underpin future opportunities to improve annotation granularity. Specifically, the data will allow families to be subdivided into more fine-grained functional groups based on residue patterns, will allow specific annotations to be provided for entries (e.g. identifying the critical functional residues for a given catalytic domain), and will enable the adoption of rule-based approaches (similar to those used by HAMAP, PIRSR and PROSITE) for the assignment of specific functional annotations, such as GO terms.

## Other sequence features annotated by InterPro

For a number of years, InterProScan has included the capability to annotate signal peptides, transmembrane regions and coiled-coils, drawing upon a suite of algorithms to make these annotations (specifically, Coils v2.2, Phobius v1.01, SignalP v4.1 and TMHMM v2.0 in the latest InterProScan release). To complement these, we have integrated a new tool called MobiDB Lite, which provides a consensus prediction of long disordered regions.

**Table 1.** Member database release versions integrated into InterPro since release 48.0

InterPro release	Member database update
49.0	PROSITE patterns (20.105), PROSITE profiles (20.105)
50.0	PIRSF (3.01)
51.0	TIGRFAMs (15.0), HAMAP (201502.04)
52.0	PROSITE patterns (20.113), PROSITE profiles (20.113)
53.0	Pfam (28.0)
54.0	PANTHER (10.0)
55.0	HAMAP (201511.02)
56.0	PROSITE patterns (20.119), PROSITE profiles (20.119)
57.0	Pfam (29.0), SMART (7.1)
58.0	CDD (1.0)*, HAMAP (201605.11)
59.0	Pfam (30.0), SFLD (1.0)*
60.0	MobiDB**

\*New member databases.

\*\*MobiDB is a new non-signature based database that has been integrated into InterPro to provide ID region annotations. See text for details.

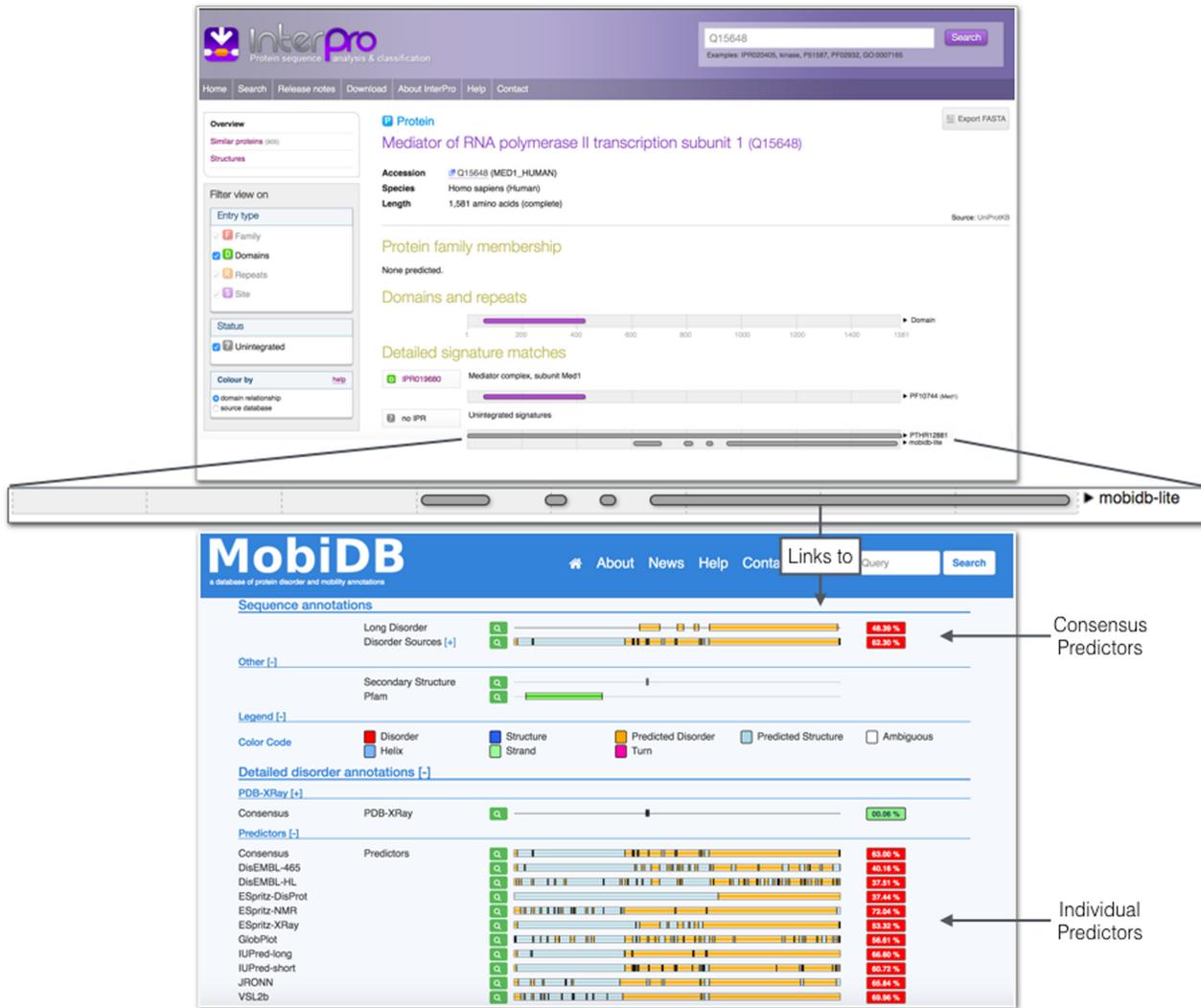
**Table 2.** Coverage of the major sequence databases UniProtKB and UniParc (the non-redundant protein sequence archive) by InterPro signatures

Sequence database	Number of proteins in database	Number of proteins with one or more matches to InterPro
UniProtKB/Swiss-Prot	552 884	533 303 (96.5%)
UniProtKB/TrEMBL	70 656 157	56 310 112 (79.7%)
UniProtKB (total)	71 209 041	56 843 415 (79.8%)
UniParc	132 489 873	103 835 823 (78.4%)

Intrinsically disordered (ID) protein regions, which do not adopt a single well-defined conformation in isolation, rely on a highly flexible state or structural plasticity to carry out their functions (40,41). While ID regions are present in all three domains of cellular life, they often exhibit very little evolutionary conservation and are hence difficult to predict using signature based methods currently employed by InterPro member databases. Information about ID regions largely complements domain and family annotations. At a closer look, ID encompasses different phenomena and different predictors can capture complementary aspects (42,43). MobiDB Lite combines different predictors to generate a consensus focusing on long disordered regions. Eight different algorithms (IUPred-short, IUPred-long (44), GlobPlot (45), DisEMBL-465, DisEMBL-HL (46), Espritz-DisProt, Espritz-NMR and Espritz-X-ray (47)) were chosen for their speed and orthogonality of approaches. The consensus is generated by evaluating the agreement among predictors and smoothing short disorder stretches. A strict agreement threshold of at least 5 out of 8 methods favors precision over inclusiveness and a length cutoff ( $\geq 20$  residues) helps to discriminate functional disorder from ambiguous assignments. MobiDB lite annotations have been enabled in InterProScan, and are available to external users in the InterProScan 5.20–60.0 release, part of InterPro release 60.0. Graphical representations of ID are also to be implemented on InterPro's Web interfaces (see Figure 4). From each ID region, InterPro provides a link to the MobiDB (43) page for the protein of interest where the breakdown of the individual predictions and additional ID annotation may be found. The MobiDB Lite consensus corresponds to the 'Long Disorder' track in MobiDB.

## DISCUSSION

Since its inception 17 years ago, InterPro has striven to provide a comprehensive protein classification resource that enables high-quality functional annotation of protein sequences. It has met this aim in collaboration with its member databases, which have provided an invaluable stream of signatures for integration into the resource. As a result, InterPro has grown significantly in terms of coverage and function-annotation specificity, and has developed a substantial worldwide user-base. During this time, most of the member databases have evolved to adopt new algorithms and include new data. For example, eight of the nine profile HMM-based member database now use the significantly faster version of HMMER, version 3.0 (with only SMART using HMMER 2.0). Meanwhile, for calculating sequence matches to HAMAP within InterProScan, the two teams developed a heuristic (using a profile HMM trained on the same alignment used to produce the HAMAP profile and used as pre-filter search), to ensure continued scalability. As indicated in Table 1, 10 of the 14 member databases (CDD, HAMAP, PANTHER, Pfam, PIRSF, Prosite Patterns, Prosite Profiles, SFLD, SMART, TIGRFAM) have been added or updated at least once over the past two years, with CATH-Gene3D awaiting update from version 3.5 to 4.1 in InterPro, demonstrating a steady increase of new data and sustained curation effort from InterPro's member databases. Two new member databases CDD and SFLD further increase coverage and extend the number of resources that annotate discrete functional amino acid residues; this functionality is now available in InterPro for the first time. We will continue to work with our expert member databases, to provide more per-residue annotations, e.g. HAMAP and PROSITE both define annotations dependent on features composed of multiple non-contiguous residues.



**Figure 4.** Integration of MobiDB Lite annotation within InterPro, enabling annotation of intrinsic disordered (ID) regions within proteins. **Top** - InterPro annotations for the Human mediator of RNA polymerase II transcription subunit I protein (UniProtKB accession Q15648). **Middle** - Zoomed in view of the consensus long range ID predictions provided by MobiDB Lite. InterPro only captures the consensus output for each sequence, but the graphical representations of the ID regions link to the source website, MobiDB (**bottom**), where the individual predictions can be viewed.

This additional tier of annotations is a step change for InterPro: it adds a feature that has long been absent from the resource, and, alongside the intrinsic disorder predictions from MobiDB Lite (to complement membrane-topology and coiled-coil prediction), enables the most richly-detailed, informative annotation of protein sequences possible. The integration of similar residue-level annotations from databases like Pfam and PIRSF, coupled with highly specific subfamily-level annotations from resources like PANTHER and PRINTS, extends this functionality even further. Together, these advances will help improve the annotation of proteins in databases like UniProtKB, by adding more discriminatory power to automated annotation systems like UniRule. InterPro's Web interfaces will be expanded to present this additional layer of annotation to users, ensuring both added value and continued usefulness of the resource for the scientific community.

## FUNDING

BBSRC [BB/L024136/1, BB/N00521X/1]; Wellcome Trust [108433/Z/15/Z]; National Sciences Foundation [DBI 1458808]; Intramural Research Program of the National Library of Medicine at National Institutes of Health/DHHS; COST Action [BM1405]; European Molecular Biology Laboratories core funds. Funding for open access charge: Research Councils UK.

*Conflict of interest statement.* None declared.

## REFERENCES

- Mardis, E.R. (2010) The \$1,000 genome, the \$100,000 analysis? *Genome Med.*, **2**, 84.
- Galperin, M.Y. and Koonin, E.V. (2010) From complete genome sequence to 'complete' understanding? *Trends Biotechnol.*, **28**, 398–406.
- Lam, S.D., Dawson, N.L., Das, S., Sillitoe, I., Ashford, P., Lee, D., Lehtinen, S., Orengo, C.A. and Lees, J.G. (2016) Gene3D: expanding

- the utility of domain assignments. *Nucleic Acids Res.*, **44**, D404–D409.
4. Pedruzzi, I., Rivoire, C., Auchincloss, A.H., Coudert, E., Keller, G., de Castro, E., Baratin, D., Cucho, B.A., Bougueleret, L., Poux, S. *et al.* (2015) HAMAP in 2015: updates to the protein family classification and annotation system. *Nucleic Acids Res.*, **43**, D1064–D1070.
  5. Mi, H., Poudel, S., Muruganujan, A., Casagrande, J.T. and Thomas, P.D. (2016) PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res.*, **44**, D336–D342.
  6. Finn, R.D., Coghill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
  7. Wu, C.H., Nikolskaya, A., Huang, H., Yeh, L.-S.L., Natale, D.A., Vinayaka, C.R., Hu, Z.-Z., Mazumder, R., Kumar, S., Kourtesis, P. *et al.* (2004) PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Res.*, **32**, D112–D114.
  8. Attwood, T.K., Coletta, A., Muirhead, G., Pavlopoulou, A., Philippou, P.B., Popov, I., Romá-Mateo, C., Theodosiou, A. and Mitchell, A.L. (2012) The PRINTS database: a fine-grained protein sequence annotation and analysis resource—its status in 2012. *Database (Oxford)*, **2012**, bas019.
  9. Bru, C., Courcelle, E., Carrère, S., Beausse, Y., Dalmar, S. and Kahn, D. (2005) The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res.*, **33**, D212–D215.
  10. Sigrist, C.J.A., de Castro, E., Cerutti, L., Cucho, B.A., Hulo, N., Bridge, A., Bougueleret, L. and Xenarios, I. (2013) New and continuing developments at PROSITE. *Nucleic Acids Res.*, **41**, D344–D347.
  11. Letunic, I., Doerks, T. and Bork, P. (2015) SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res.*, **43**, D257–D260.
  12. Oates, M.E., Stahlhacke, J., Vavoulis, D.V., Smithers, B., Rackham, O.J.L., Sardar, A.J., Zaucha, J., Thurlby, N., Fang, H. and Gough, J. (2015) The SUPERFAMILY 1.75 database in 2014: a doubling of data. *Nucleic Acids Res.*, **43**, D227–D233.
  13. Haft, D.H., Selengut, J.D., Richter, R.A., Harkins, D., Basu, M.K. and Beck, E. (2013) TIGRFAMs and genome properties in 2013. *Nucleic Acids Res.*, **41**, D387–D395.
  14. Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I. *et al.* (2015) CDD: NCBI's conserved domain database. *Nucleic Acids Res.*, **43**, D222–D226.
  15. Akiva, E., Brown, S., Almonacid, D.E., Barber, A.E., Custer, A.F., Hicks, M.A., Huang, C.C., Lauck, F., Mashiyama, S.T., Meng, E.C. *et al.* (2014) The structure-function linkage database. *Nucleic Acids Res.*, **42**, D521–D530.
  16. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
  17. Sangrador-Vegas, A., Mitchell, A.L., Chang, H.-Y., Yong, S.-Y. and Finn, R.D. (2016) GO annotation in InterPro: why stability does not indicate accuracy in a sea of changing annotations. *Database (Oxford)*, **2016**, baw027.
  18. Bairoch, A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.
  19. Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C.A., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A. *et al.* (2016) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **44**, D471–D480.
  20. Morgat, A., Coissac, E., Coudert, E., Axelsen, K.B., Keller, G., Bairoch, A., Bridge, A., Bougueleret, L., Xenarios, I. and Viari, A. (2012) UniPathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Res.*, **40**, D761–D769.
  21. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
  22. Andreeva, A., Howorth, D., Chandonia, J.-M., Brenner, S.E., Hubbard, T.J.P., Chothia, C. and Murzin, A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
  23. Sillitoe, I., Lewis, T.E., Cuff, A., Das, S., Ashford, P., Dawson, N.L., Furnham, N., Laskowski, R.A., Lee, D., Lees, J.G. *et al.* (2015) CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res.*, **43**, D376–D381.
  24. UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
  25. Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
  26. Aken, B.L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., García Girón, C., Hourlier, T. *et al.* (2016) The ensemble gene annotation system. *Database (Oxford)*, **2016**, baw093.
  27. Kersey, P.J., Allen, J.E., Armean, I., Boddus, S., Bolt, B.J., Carvalho-Silva, D., Christensen, M., Davis, P., Falin, L.J., Grabmueller, C. *et al.* (2016) Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Res.*, **44**, D574–D580.
  28. Velankar, S., van Ginkel, G., Alhroub, Y., Battle, G.M., Berrisford, J.M., Conroy, M.J., Dana, J.M., Gore, S.P., Gutmanas, A., Haslam, P. *et al.* (2016) PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic Acids Res.*, **44**, D385–D395.
  29. Götz, S., García-Gómez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M., Talón, M., Dopazo, J. and Conesa, A. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.*, **36**, 3420–3435.
  30. Pedro, H., Maheswari, U., Urban, M., Irvine, A.G., Czuzick, A., McDowall, M.D., Staines, D.M., Kulesha, E., Hammond-Kosack, K.E. and Kersey, P.J. (2016) PhytoPath: an integrative resource for plant pathogen genomics. *Nucleic Acids Res.*, **44**, D688–D693.
  31. Gazzo, A.M., Daneels, D., Cilia, E., Bonduelle, M., Abramowicz, M., Van Dooren, S., Smits, G. and Lenaerts, T. (2016) DIDA: A curated and annotated digenic diseases database. *Nucleic Acids Res.*, **44**, D900–D907.
  32. Tranchevent, L.-C., Ardeshirdavani, A., ElShal, S., Alcaide, D., Aerts, J., Aubouef, D. and Moreau, Y. (2016) Candidate gene prioritization with Endeavour. *Nucleic Acids Res.*, **44**, W117–W121.
  33. Mitchell, A., Buccini, F., Cochrane, G., Denise, H., ten Hoopen, P., Fraser, M., Pesseat, S., Potter, S., Scheremetjew, M., Sterk, P. *et al.* (2016) EBI metagenomics in 2016—an expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res.*, **44**, D595–D603.
  34. Huson, D.H., Beier, S., Flade, I., Górská, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.-J. and Tappu, R. (2016) MEGAN community edition - interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Comput. Biol.*, **12**, e1004957.
  35. Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S. *et al.* (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.*, **43**, D213–D221.
  36. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
  37. Galperin, M.Y., Makarova, K.S., Wolf, Y.I. and Koonin, E.V. (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.*, **43**, D261–D269.
  38. Klimke, W., Agarwala, R., Badretkin, A., Chetvernin, S., Ciufu, S., Fedorov, B., Kiryutin, B., O'Neill, K., Resch, W., Resenchuk, S. *et al.* (2016) The national center for biotechnology information's protein clusters database. *Nucleic Acids Res.*, **37**, D216–D223.
  39. Tonami, K., Hata, S., Ojima, K., Ono, Y., Kurihara, Y., Amano, T., Sato, T., Kawamura, Y., Kurihara, H. and Sorimachi, H. (2013) Calpain-6 deficiency promotes skeletal muscle development and regeneration. *PLoS Genet.*, **9**, e1003668.
  40. Wright, P.E. and Dyson, H.J. (2015) Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.*, **16**, 18–29.
  41. Necci, M., Piovesan, D. and Tosatto, S.C.E. (2016) Large-scale analysis of intrinsic disorder flavors and associated functions in the protein sequence universe. *Protein Sci.*, doi:10.1002/pro.3041.
  42. van der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., Gough, J., Gsponer, J.,

- Jones, D.T. *et al.* (2014) Classification of intrinsically disordered regions and proteins. *Chem. Rev.*, **114**, 6589–6631.
43. Potenza, E., Di Domenico, T., Walsh, I. and Tosatto, S.C.E. (2015) MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. *Nucleic Acids Res.*, **43**, D315–D320.
44. Dosztányi, Z., Csizmók, V., Tompa, P. and Simon, I. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.
45. Linding, R., Russell, R.B., Neduva, V. and Gibson, T.J. (2003) GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, **31**, 3701–3708.
46. Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J. and Russell, R.B. (2003) Protein disorder prediction: implications for structural proteomics. *Structure*, **11**, 1453–1459.
47. Walsh, I., Martin, A.J.M., Di Domenico, T. and Tosatto, S.C.E. (2012) ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics*, **28**, 503–509.