PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

PERSPECTIVES

Key challenges for the creation and maintenance of specialist protein resources

Gemma L. Holliday,[1*] Amos Bairoch,[2] Pantelis G. Bagos,[3] Arnaud Chatonnet,[4,5] David J. Craik,[6] Robert D. Finn,[7] Bernard Henrissat,[8,9] David Landsman,[10] Gerard Manning,[11] Nozomi Nagano,[12] Claire O'Donovan,[7] Kim D. Pruitt,[10] Neil D. Rawlings,[7,13] Milton Saier,[14] Ramanathan Sowdhamini,[15] Michael Spedding,[16] Narayanaswamy Srinivasan,[17] Gert Vriend,[18] Patricia C. Babbitt,[1] and Alex Bateman[7]

[1] Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, California 94158

[2] SIB—Swiss Institute of Bioinformatics, University of Geneva, Geneva, Switzerland

[3] Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, 35100, Greece

[4] INRA, Umr866 Dynamique Musculaire Et Métabolisme, Montpellier F-34000, France

[5] Université Montpellier, Montpellier, F-34000, France

[6] Institute for Molecular Bioscience. The University of Queensland, Brisbane, Queensland 4072, Australia

[7] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge Cb10 1SD, United Kingdom

[8] Architecture Et Fonction Des Macromolécules Biologiques, CNRS, Aix-Marseille Université, Marseille, 13288, France

[9] Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia

[10] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20892

[11] Department of Bioinformatics & Computational Biology, Genentech, 1 DNA Way, South San Francisco, California 98010

[12] Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo 135-0064, Japan

[13] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, Cb10 1SD, United Kingdom

Grant sponsor: National Institutes of Health and National Science Foundation; Grant number: NIH R01 GM60595 and NSF DBI1356193 (to GLH and PCB); Grant sponsor: the project "Integration of Data from Multiple Sources" (IntDaMus), which is implemented under the "ARISTEIA II" Action of the "OPERATIONAL PROGRAMME EDUCATION AND LIFELONG LEARNING" and cofunded by the European Social Fund (ESF) and National Resources (to PGB); Grant number: ANR-10-BLAN-1216 (to AC); Grant sponsor: National Health & Medical Research Council (Australia); Grant numbers: APP1026501 and APP1028509; Grant sponsors: Australian Research Council; Grant number: LP130100550 (to DJC); Grant sponsor: European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI; to AB and RDF); Grant sponsor: Agence Nationale de la Recherche (CAZy database)—BIOMINES and BIP:BIP; Grant number: ANR-12-BIME-0006 and ANR-10-BINF-0304; Grant sponsor: Intramural Research Program of the National Institutes of Health, National Library of Medicine (DL Histone Database; KDP RefSeq; work performed at NCBI); Grant sponsors: Grant-in-Aid for Publication of Scientific Research Results; Grant number: 248047 (to NN); Grant sponsor: NN was organized by the Japan Society for the Promotion of Science (JSPS), and also has been supported by Commission for the Development of Artificial Gene Synthesis Technology for Creating Innovative Biomaterial from the Ministry of Economy, Trade and Industry (METI), Japan; Grant sponsor: National Institutes of Health (NIH); Grant number: U41HG007822 (to UniProtKB); Grant sponsor: EMBL-EBI's involvement in UniProtKB comes from European Molecular Biology Laboratory (EMBL), the British Heart Foundation (BHF); Grant number: RG/13/5/30112; Grant sponsor: Parkinson's Disease United Kingdom (PDUK); Grant number: G-1307; Grant sponsor: NIH; Grant number: U41HG02273; Grant sponsor: UniProtKB activities at the SIB are additionally supported by the Swiss Federal Government through the State Secretariat for Education, Research, and Innovation SERI; Grant sponsor: NIH (to PIR's UniProtKB activities); Grant number: R01GM080646, G08 LM010720, and P20GM103446; Grant sponsor: National Science Foundation (NSF); Grant number: DBI-1062520; Grant sponsor: Department of Biotechnology, Government of India; Grant number: BT/01/COE/09/01 (to RS and NS); Grant sponsor: Wellcome Trust; Grant number: WT0077044/Z/05/Z (to Neil Rawlings); Grant sponsor: NIH; Grant number: GM077402 (to TCDB); Grant sponsor: Swiss SERI (Secretariat for Education, Research, and Innovation) through the SIB—Swiss Institute of Bioinformatics; Grant sponsor: Wellcome Trust (to The IUPHAR/BPS Guide to PHARMACOLOGY); Grant number: 099156/Z/12/Z; Grant sponsors: Industry sponsors (Servier), the British Pharmacological Society (BPS), and IUPHAR; Grant sponsor: STW and EU; Grant number: KBBE-2011–5 289350 (to GV).

DL and KDP do not represent the opinion or official policy of the National Institutes of Health in this work.

*Correspondence to: Gemma L. Holliday; Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA 94158. E-mail: gemma.holliday@ucsf.edu

Received 12 January 2015; Revised 6 March 2015; Accepted 20 March 2015
Published online 27 March 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.24803

[14] Department of Molecular Biology, University of California at San Diego, La Jolla, California 92093

[15] National Centre for Biological Sciences, TIFR, GKVK Campus, Bellary Road, Bangalore 560065, India

[16] Chair NC-IUPHAR, Spedding Research Solutions SARL, 6 Rue Ampere, Le Vesinet 78110, France

[17] Molecular Biophysics Unit, Indian Institute of Science, Bangalore, 560012, India

[18] Centre for Molecular and Biomolecular Informatics (CMBI), Radboud University Medical Center, Geert Grooteplein Zuid 26-28, 6525 GA Nijmegen, The Netherlands

ABSTRACT

As the volume of data relating to proteins increases, researchers rely more and more on the analysis of published data, thus increasing the importance of good access to these data that vary from the supplemental material of individual articles, all the way to major reference databases with professional staff and long-term funding. Specialist protein resources fill an important middle ground, providing interactive web interfaces to their databases for a focused topic or family of proteins, using specialized approaches that are not feasible in the major reference databases. Many are labors of love, run by a single lab with little or no dedicated funding and there are many challenges to building and maintaining them. This perspective arose from a meeting of several specialist protein resources and major reference databases held at the Wellcome Trust Genome Campus (Cambridge, UK) on August 11 and 12, 2014. During this meeting some common key challenges involved in creating and maintaining such resources were discussed, along with various approaches to address them. In laying out these challenges, we aim to inform users about how these issues impact our resources and illustrate ways in which our working together could enhance their accuracy, currency, and overall value.

Proteins 2015; 83:1005–1013.

Key words: specialist protein resource; key challenges; biocuration; longevity; big data; misannotation.

INTRODUCTION

With the advent of the technologies of the omics age, there is far more data to manage, access, and understand than ever before. As the data are far greater than any single group of researchers can hope to ever cope with, repositories for these data are becoming increasingly important. While it is simple to say: "I have data, therefore I shall create a database for it," there are many challenges and hurdles in doing it such that the data can be retrieved and studied effectively and efficiently. Recently, a small group of leaders of web-accessible, knowledge-based, specialist protein resources (SPRs) came together at a retreat sponsored by the Wellcome Trust to discuss the challenges they face. The retreat was held at the Wellcome Trust Genome Campus in Cambridge (UK) on August 11 and 12, 2014. Although each SPR present represented some unique challenges and issues, it became clear that there were some overarching challenges common to all of them. Together, these can be combined into a single question: What makes a database useful? Here, we discuss the top challenges that emerged from this discussion, along with some of the ways that were proposed to address them from the perspective of the researchers at the retreat. The SPRs represented at the retreat covered diverse communities, listed in Table I.

What are SPRs, and why do we need them?

The SPRs represented at the Wellcome Trust meeting are just a tiny proportion of the SPRs available to researchers, but most are designed to perform a similar function: to add value to the data available to researchers. SPRs cover a wide range of different types of protein. Some are general and relate to all types of proteins (*e.g.*, Pfam); others focus on specific types of proteins, for example, transporter proteins (*e.g.*, TCDB), receptor proteins (*e.g.*, GPCRDB), and enzymes (*e.g.*, ExCatDB). In all cases, data are available in many formats, including the primary literature and associated supplementary material, patents, and reference databases (*e.g.*, RefSeq or UniProtKB). An SPR can add value to data in many different ways, from simply collating it into levels of classification, to performing complex data analysis. The most comprehensive list of SPRs can be found in the Nucleic Acids Research Database issue and its associated Molecular Biology Database Collection,[23] published annually in January. In 2014 there were [mt]1500 databases listed in the Molecular Biology Database Collection, which range from comprehensive reference databases to resources that focus on a single protein family, and everything in between. The data types available in SPRs are just as diverse, yet there are commonalities among them. All proteins have a few features in common, namely their amino acid sequence (and often also the associated nucleic acid sequence) and the species from which they come. Thus, most SPRs will contain either a nucleic or amino acid sequence (or both), and at least a minimal amount of metadata. The data that the SPRs add, however, is myriad and varied. Some will annotate the

Table I
The SPRs and Major Databases That Participated in the Wellcome Trust retreat, Their URLs, and Primary References

Database name	URL	Reference
*C*arbohydrate-*A*ctive *E*nzymes database (CAZy)	www.cazy.org	1
ConoServer database for conopeptides	http://www.conoserver.org/	2
CyBase database of cyclic proteins	http://www.cybase.org.au/	3
ESTHER database (*EST*erases and alpha/ beta-*H*ydrolase *E*nzymes and *R*elatives)	http://bioweb.ensam.inra.fr/ESTHER/general?what=index	4
ExTopoDB database of experimentally derived topological models of transmembrane proteins	http://bioinformatics.biol.uoa.gr/ExTopoDB/	5
EzCatDB database of Enzyme Catalytic Mechanisms	http://ezcatdb.cbrc.jp/EzCatDB/	6
GPCRDB (G Protein-Coupled Receptors Database)	http://www.gpcr.org/7tm/	7
gpDB (a database of GPCRs, G-proteins, Effectors and their interactions)	http://bioinformatics.biol.uoa.gr/gpDB/	8
the Histone Database	http://genome.nhgri.nih.gov/histones/	9
the IUPHAR/BPS Guide to pharmacology	http://www.guidetopharmacology.org/	10
Kinase.com	http://kinase.com/	
the KinG database (a database of protein kinases in genomes)	http://megha.garudaindia.in/king/index.jsp	11
the MACiE Database (*M*echanism, *A*nnotation and *C*lassification *i*n *E*nzymes)	http://www.ebi.ac.uk/thornton-srv/databases/MACiE	12
MEROPS (the peptidase database)	http://merops.sanger.ac.uk/	13
neXtProt (knowledge resource on human proteins)	http://www.nextprot.org/	14
OMPdb (a database of β-barrel outer membrane proteins from Gram-negative bacteria)	http://www.ompdb.org/	15
PASS2 database of structure-based sequence alignments of protein structural domain superfamilies	http://caps.ncbs.res.in/pass2/	16
Pfam	http://pfam.xfam.org/	17
Reference Sequence (RefSeq) database	http://www.ncbi.nlm.nih.gov/refseq/	18
the Structure-Function Linkage Database (SFLD)	http://sfld.rbvi.ucsf.edu/	19
Transporter Classification Database (TCDB)	http://www.tcdb.org/	20
TIGRFAMs	http://www.jcvi.org/cgi-bin/tigrfams/index.cgi	21
UniProtKB	http://www.uniprot.org/	22

chemistry, such as the enzymatic reaction, cofactors, regulators, and so forth. Others add three-dimensional information such as the PDB structure or active site motifs. Some add disease information, such as disease causing SNPs or polymeric forms; others look at the kinetics of the reaction and small molecule binding. If there is a study or data available in the primary literature and a group of scientists interested in that field, the chances are that there is an associated SPR. Thus, by strengthening and expanding the realm of SPRs, we can provide a richer and more diverse set of resources to the research community, and accelerate the rate at which individual results can be incorporated into interactive databases for greater use.

Misannotation and data integrity

The foremost challenge to most SPRs is the issue of the accuracy of data and its associated annotations, not only in their own resources, but in those of others too. Many different types of error can be found in data resources, all of which present challenges for users, especially those unfamiliar with their specialized content. An analysis done in 2009 on a relatively small set of highly manually curated enzyme superfamilies[24] showed that some major public databases misidentified an average of 5–63% across the six superfamilies studied, usually by "overannotation" of specific function when the evidence only supports annotation of general functional properties. Some errors are relatively easy to identify through automated processes and pipelines (*e.g.*, MisPred,[25] which identifies erroneous protein sequence function predictions in public databases, usually in the form of abnormal, incomplete and incorrect predictions). Others, such as errors in the underlying scientific information (*e.g.*, if the protein sequence has translation errors or the biochemical characterization is incomplete) are much harder to find, especially as our knowledge continues to grow so fast that we often have to move on rather than go back to correct errors.

One example of a problem caused by the growth of knowledge is the enzymatic mechanism for lysozyme. For over 50 years the accepted mechanism involved an ion pair intermediate. It was not until 2001 that new experiments showed that the intermediate was instead a covalent glycosyl enzyme.[26] As researchers are challenged to stay up to date with the scientific literature and SPRs to continually update their information, nonexpert users would be forgiven for thinking the ion-pair mechanism was still the definitive one (especially as this is the mechanism shown in many text books as well). Such examples raise a number of questions for our community and our users: Can we ever say that we know the correct mechanism of an enzyme? Can we ever hope to keep up with the frontier of scientific discovery? Further, even if the new information has been published, will it be incorporated into any database resources and then propagated throughout the many different SPRs? Possibly not—as the key to keeping databases up-to-date requires that curators (or users, or text mining robots) go back over the literature again and again to identify changes, new discoveries, and what information has become obsolete. Ideally, we need an exceptional solution to accurate and automated updating of all relevant databases, even including those that are deeply dependent on specialized knowledge within a field.

Another common error found in protein sequence analysis is the misannotation of a protein due to its modular (or multidomain) structure. For example, the carbohydrate-binding module (CBM) in carbohydrate-active enzymes is frequently found appended to catalytic domains belonging to various families, including domains of unknown function. A best BLAST[27] hit matching only the CBM often leads to erroneous annotation of the adjacent domain. This is because the matched domain is often used to annotate the function of the entire protein, not just the portion found via BLAST. In many cases, such errors can only be identified when researchers go back to carefully examine specific cases in detail. For example, the aminotransferase-related enzyme (UniProtKB: B8NM72) was ultimately found to be involved in synthesis of a ribosomal peptide, rather than acting as a nonribosomal peptide synthetase as previously thought.[28] Such annotation transfer errors can often lead researchers astray and highlights why expert manual annotation is so essential for SPRs.

Although over-prediction, transferring annotation from one annotated protein to another of unknown function using relatively lax parameters, has the advantage of increased data coverage, it can lead to many erroneous function predictions. Such annotation errors can be further compounded by “proof by repetition”; the assumption that the most numerous annotations are the correct ones.[29] Such errors can be protected against by “under annotation,” that is, transferring data only when we have the highest confidence that it is accurate, for example, in requiring not only a high confidence BLAST score, but also in having the active site profile fully matched. These protocols often lead to significantly fewer annotations being assigned, but the quality of the annotation transfer is much better. In both cases, annotation transfer is further complicated by the fact that a protein’s function can be defined as the molecular/chemical role (*e.g.*, a specific serine kinase) or the broad biological process the protein mediates (*e.g.*, mediating the coagulation of blood). Generally, it is quite difficult to decipher the biological role of a protein in the physiological context using computational methods and therefore such predictions should be used with caution. Nowadays a BLAST search of the nonredundant protein database of NCBI (RefSeq) or on UniProtKB often identifies a large number of similar proteins originating almost exclusively from genome sequencing (that is, these proteins have had no characterization performed). Close examination of the names attributed to these proteins shows that they are both heterogeneous and transmitted from one to another via automated processes (creating a mess that is increasingly difficult to discern and fix).

Many protein homologues lack one or more critical residues, making them functionally inactive, another aspect of annotation transfer that may lead to erroneous annotation. These proteins may be biologically relevant but with another function, or on the other hand, the missing residues could be artefacts caused by gene assembly errors. Other typical gene assembly errors lead to the prediction of putative proteins where the wrong initiating methionine has been identified, or where exons have been omitted. Although such errors may be subsequently corrected, finding the time to back-check for these types of errors requires more resources than are available to many SPR curators.

Fixing annotation errors and propagating data

Once an error is identified, how do we fix it? Many resources, such as UniProtKB and RefSeq, have mechanisms for users to report problems so that annotation errors can be corrected. Additionally, specialized resources have been developed to help address this issue and provide at least some reannotations (such as PDB_REDO[30] for PDB atomic coordinates). However, many others, such as the Protein Data Bank (PDB),[31] GenBank,[32] the European Nucleotide Archive (ENA),[33] and the DNA Data Bank of Japan (DDBJ),[34] lack these procedures as they are primary repositories that are designed to archive original data. SPRs tend to have their own policies for correcting errors that are relevant to the specific nature of each resource. Many welcome (and need) input from expert users in order to identify and correct data errors.

Identifying the error is only the first step. Once we know an error exists, how do we propagate the fix through all the SPRs that utilize the original entry? The provenance of a datum is often difficult to identify. Have the database curators taken information directly from UniProtKB, or RefSeq, or from the primary literature? Maybe they took it from another resource, but where did that resource’s curators get it from? While such repurposing of data is common place, it is a good way to propagate annotation errors. A better solution might be for all primary data to be stored in a common archive resource so that niche or derived databases could provide pointers to the original information that could then be expanded on demand. However, such an annotation archive would be challenging to implement on a wide scale.

Another promising approach used by some resources (*e.g.*, the Gene Ontology (GO),[35] UniProtKB, and some SPRs) is to use the concept of "evidence," sometimes in combination with the use of the Evidence Code Ontology (ECO),[36] a structured and controlled vocabulary for evidence in biological research. Used in the context of protein function annotation, evidence codes allow for evidence not only to have a type (*e.g.*, inferred from electronic annotation), but to have a source (*e.g.*, a specific resource), providing users with an effective way to judge the confidence with which to judge an annotation. Other SPRs are starting to follow suit although populating a resource with this sort of "back annotation" can be a long and often complicated process as all the data must be cross-checked and back-edited. That being said, if we are ever to propagate annotation "fixes," the ability to follow data back to their source is going to be critical, suggesting the value of using ECO (or something similar) moving forward.

With the ever increasing volume of data, how do we grow and maintain our resources responsibly, especially in the context of the misannotation challenge? Specialist curation (by individuals who are highly trained to a particular resource and/or in a particular field) is always going to be critical because databases that include a high degree of cross-check and human curation provide significant added value over simple repositories or meta-resources/hubs that are especially prone to propagation of misannotation. For example, the IUPHAR/BPS Guide to PHARMACOLOGY uses expert curators for particular protein "receptor" types that are linked to subcommittees of experts who ensure data quality. This approach allows experts to keep their field "clean" and to benefit from highly cited publications[37] rather than using "data trawling" which can lead to misleading information being propagated.

The user's role in expanding SPR data coverage

Users of the SPRs are going to become increasingly important to correcting errors and growing SPRs in the future. For example, expert users are in a position to inform SPRs of errors that they have spotted or to contribute new entries based on their experiments and/or publications. Depending on an evaluation of the evidence provided, the resource can then update the entry. In the experience of the attendees at the retreat, the major hurdle to adopting user based annotation methods is educating users about the benefits of contributing their information to database resources versus the effort of creating the annotation themselves (commonly referred to as the "tragedy of the commons"[38]). One route to achieving more user input (such as that taken by the international crystallographic community) is to require data entry before the results can be published. Without the support and enforcement by the journals, however, it is not practicable to capture functional information efficiently in this manner. On the bright side, the level of detail repositories require of their depositors need not be onerous. For example, including the EC number along with a sequence accession number for an enzyme would represent enormous progress that would allow SPRs and larger resources alike to incorporate research results keyed to those common identifiers. The flip side of encouraging annotation and error correction contributions by users is that the manual incorporation of this information could quickly outgrow a resource's ability to keep up. Again, user submissions enabled via structured information formats supported by the journals would offer progress toward more automated solutions. The International Society for Biocuration (http://www.biocurator.org) is an active proponent in bringing scientists, curators and journals together with a view to enable user submissions. The annual International Biocuration Conference is a great opportunity for these groups come together to discuss the challenges involved.

A different route to maintaining data quality is the use of the Wikipedia model. Rfam[39] and Pfam[17] both utilize this model to populate the respective databases with Wikipedia pages created by authors. Although both of these approaches are promising, general application of this model awaits answers to several basic questions: which resources become the primary repositories of user-contributed data? How do we deal with overlapping resources? When an old resource is retired, who takes on its data? Will one resource become the ultimate one for all protein annotations, which are then used and elaborated upon by SPRs? Will all journals agree to the process? Will the annotation process be both simple and complete enough that authors find benefit to the process? There are no simple answers to these questions, but as the amount of data grows, many aspects of protein research would benefit if SPR developers, users and publishers begin to work together in developing a common plan for moving forward.

Weathering the data deluge

Even the best resource must keep its information up-to-date, and in this omics era possibly the biggest challenge we face is the sheer volume of data currently available, along with its projected growth at a near exponential rate. There is also a constant growth in the number of data-sources. UniProtKB and RefSeq have approximately 89 million and 47 million entries, respectively, as of November 2014, of which just over half a million are manually annotated or reviewed in each database and over 70% of these reviewed proteins are annotated via similarity to a protein of known function. For every bit of information on a single protein that exists, there are even more proteins for which we have no data, save a primary amino acid sequence. SPRs have two important roles to play in weathering this data deluge: one is to provide novel annotation and understanding in their field of expertise and the other is to provide online tools to access those annotations. Additionally, SPRs must determine what information to submit to larger and more general resources, and which to glean from other data resources. These roles, in turn, aid resources such as RefSeq and UniProtKB in extending and improving their predicted annotations. A good SPR should know where its strengths lie and clearly distinguish the primary annotations for which they are the unique source from those data that come from other resources.

Adoption of best practices

Along with annotation input from experimental users, the Retreat discussion also suggested that the SPRs could benefit from best practices that have been developed. However, for any one resource to be useful to another, the language that they both use needs to be standardized. One such exercise in standardization was the EMBRACE project (http://www.embracegrid.info/[40]) which worked to integrate major databases and software tools in bioinformatics, using existing methods and emerging Grid service technologies. Some resources already use the same language, conceptually facilitating the exchange of information between them. For instance, OMPdb uses the commonly accepted family classification system of Pfam. But generally, what one resource means by the term «family» or «superfamily» might not be what another means. For example, the SFLD definition requires that the proteins not only be evolutionarily related, but that they have a conserved chemical aspect to their function. TIGRFAM, on the other hand, only requires evolutionary relatedness.

While we are not advocating that all resources use an identical language (biology is nothing if not messy, so a term in one field will not directly translate to another), there needs to be a way to both establish and translate concepts. Ontologies are certainly the most robust method to do this, and SPRs need to define their language and concepts clearly so that mapping is possible between the different SPRs. Although the Gene Ontology (GO) is probably the most widely known ontology in the field of bioinformatics, a PubMed search for "ontology" in the title of a article yields almost 1500 hits (almost 500 of which involve GO). Especially for some key concepts of biochemistry and biology, relevant to SPRs, the capability exists to link data across resources that share similar data. The ontology repositories, such as BioPortal[41] and the OBO Foundry,[42] offer a good way to find an ontology that will help describe specific types of data by collecting as many biological ontologies as possible into a single location. At the very least, an understanding of the terms used by various resources will allow SPRs to map data between one another, benefitting both our curators and users.

Resource longevity

The 2014 Nucleic Acids Research database issue contained 58 new databases and updates to 123 existing databases, growing the total number of databases represented in the online collection of molecular biology databases (http://www.oxfordjournals.org/our_journals/nar/database/c/) to 1552 in 2014.[23] It is quite easy to create a database, and many databases are created as part of a PhD or Masters projects, but with no plan for future maintenance of the information. Even for more established SPRs, it remains hard to maintain such resources over many years. A 2008 study found that almost 40% of database URLs published in journals were no longer regularly available.[43] "Zombie databases" are not maintained past the original publication for reasons that range from lack of interest or funding to a career move by the creator. Over time, these may become unreliable or even misleading by failing to keep current with the field, including naming conventions, links to other databases, or even browser compatibility, and eventually are taken down. One answer to the problem of longevity is greater integration. An example of such an approach is InterPro,[44] a resource that integrates eleven different protein domain and family resources into a one-stop-shop. The member databases still retain their own identity, data, and role in the wider community while InterPro provides access to their annotations (and expertise) through a single website. The caveat to the inclusion of a new SPR within InterPro is that the source database must have sequence analysis methods that are reproducible and scalable, so is unlikely to be suitable for all SPRs, for example, CAZy, where sequences are annotated on an individual basis. One of the roles performed by InterPro is the provision of the annotations produced by its member databases to UniProtKB on a monthly database, such that annotations are up-to-date with the source member database and all sequences found in UniProtKB. The advantage to the InterPro user is the ability to view all the different annotations in a single resource. To be able to view broad- and fine-grained annotation in a single interface is highly efficient, so in this respect it is arguable that the whole (InterPro) is greater than the sum of its parts (member databases).

Similarly, Pfam (which can be considered both a SPR and a reference database) uses the annotations found in many of the SPRs as either starting points for generating new entries and/or for annotating existing Pfam entries. For example, most of the peptidase families in Pfam have been derived from or annotated using MEROPS (note that Pfam does not contain any of the fine-grained subfamily annotations found in MEROPS). In many ways SPR data integration into Pfam parallels that of data integration of member databases into InterPro; however, the attribution to the source SPR is less obvious than with InterPro. Also, Pfam will derive its own profile hidden Markov model for the entry and possibly supplement the SPR annotation. There is also the risk that smaller SPRs will be subsumed by Pfam and in so doing will reduce traffic to the individual resources. Furthermore, as annotations of both proteins and domains are updated in both in the literature and SPRs, there is no rigorous mechanism in the Pfam production software pipelines to identify and reconcile the differences between Pfam and the SPRs. However, a major advantage of Pfam is its wide use within the scientific community. Moreover, it is a founder member database of InterPro and is used within CDD.[45] Thus, the information in Pfam, both curated by Pfam and derived from the SPRs, is propagated to a broad audience.

There are several other examples of consortia that work toward greater integration of protein resources, and although not formally represented at the inaugural SPR meeting, these have proven to be exceedingly useful. Two such examples are the HUPO Proteomics Standards Initiative[46] and the International Molecular Exchange (IMEx) Consortium of Protein-Protein Interaction databases.[47] The IMEx Consortium is an excellent example of where coordination of a set of related databases has led to standardization and improved interoperability. (UniProtKB was the only member of this consortium represented at the SPR meeting.)

For many small SPRs attending our retreat, a continuing challenge to longevity is obtaining funding. In contrast, this is not a consideration for the Histone Database and the RefSeq Database as these are supported by intramural funds at the National Institutes of Health. Several models currently exist for funding SPRs. These include: Self-funding, that is, they are maintained using funds provided to the research group by the board of directors of the home institute for normal running of the group, for example, GPCRDB and DSSP,[48] grant agency funding, for example, the SFLD is currently supported by NIH and NSF grants, user-based funding (commercial), for example, KEGG,[49] which, due to lack of other funding resources, is now forced to operate via paid licensing fees, and user base "funding" (public), for example, the Little Skate Genome Project,[50] which held many jamborees to annotate the skate genome, minimizing the need for a large curator and bioinformatics staff employed by the resource. It is our job, as a community of SPRs to help one another and to listen to our users. Furthermore, those of us in the SPR community need to work together to minimize duplication of effort, helping one another to maintain quality as well as quantity so that our users have the best possible data from which to work. We also need help from our user communities, without whom our resources cannot hope to thrive. Finally, as funding for the Wellcome Trust Retreat was of necessity limited to a small group of database resources, we would like to encourage any researchers that run their own SPRs to join our mailing list (https://listserver.ebi.ac.uk/mailman/listinfo/sprn) and contribute to further discussion about the issues described in this brief report, including greater data interoperability, standardization, and consolidation.

CONCLUSIONS

SPRs are critical to biological research in this omics age and we can all benefit from sharing the many different technical approaches that have been developed. This is especially true for meeting the challenge of creating user-friendly web interfaces and supporting complex queries, both of which are difficult to develop and maintain. Such technical developments represent a high burden on many SPRs and their requirements can quickly grow beyond an SPR's interest and expertise. Still, many of us have developed similar data structures which could allow us to share methods for developing web interfaces and adoption of external tools. A further useful extension would be an agreement among SPRs to make their data available to other resources in some fairly standard or well defined format, attended by assurance of the quality and provenance of the contributed data/annotations (*e.g.*, when the data were last updated and where they came from). Finally, greater coordination and collaboration between databases is going to be even more important with the ever growing amount of data available. Here, a synergy between SPRs and reference databases will be essential, with the SPRs improving the accuracy and quality of their data and the reference databases integrating these data for broader dissemination to the research community.

ACKNOWLEDGMENTS

The participants of the inaugural SPRN meeting thank the Wellcome Trust for funding the meeting and Alex Bateman and Patricia Babbitt for organizing it.

REFERENCES

1. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 2014;42:D490–D495. [PMC][10.1093/nar/gkt1178] [24270786]
2. Kaas Q, Yu R, Jin AH, Dutertre S, Craik DJ. ConoServer: updated content, knowledge, and discovery tools in the conopeptide database. Nucleic Acids Res 2012;40:D325–D330.

3. Wang CK, Kaas Q, Chiche L, Craik DJ. CyBase: a database of cyclic protein sequences and structures, with applications in protein discovery and engineering. Nucleic Acids Res 2008;36:D206–D210.
4. Lenfant N, Hotelier T, Velluet E, Bourne Y, Marchot P, Chatonnet A. ESTHER, the database of the alpha/beta-hydrolase fold superfamily of proteins: tools to explore diversity of functions. Nucleic Acids Res 2013;41:D423–D429.
5. Tsaousis GN, Tsirigos KD, Andrianou XD, Liakopoulos TD, Bagos PG, Hamodrakas SJ. ExTopoDB: a database of experimentally derived topological models of transmembrane proteins. Bioinformatics 2010; 26:2490–2492.
6. Nagano N, Nakayama N, Ikeda K, Fukuie M, Yokota K, Doi T, Kato T, Tomii K. EzCatDB: the enzyme reaction database, 2015 update. Nucleic Acids Res 2015;43:D453–D458.
7. Vroling B, Sanders M, Baakman C, Borrmann A, Verhoeven S, Klomp J, Oliveira L, deVlieg J, Vriend G. GPCRDB: information system for G protein-coupled receptors. Nucleic Acids Res 2011;39: D309–19.
8. Theodoropoulou MC, Bagos PG, Spyropoulos IC, Hamodrakas SJ. gpDB: a database of GPCRs, G-proteins, effectors and their interactions. Bioinformatics 2008;24:1471–1472.
9. Marino-Ramirez L, Levine KM, Morales M, Zhang S, Moreland RT, Baxevanis AD, Landsman D. The Histone Database: an integrated resource for histones and histone fold-containing proteins. Database (Oxford) 2011;2011:bar048.
10. Alexander SP, Benson HE, Faccenda E, Pawson AJ, Sharman JL, McGrath JC, Catterall WA, Spedding M, Peters JA, Harmar AJ, Collaborators C, Abul-Hasn N, Anderson CM, Anderson CM, Araiksinen MS, Arita M, Arthofer E, Barker EL, Barratt C, Barnes NM, Bathgate R, Beart PM, Belelli D, Bennett AJ, Birdsall NJ, Boison D, Bonner TI, Brailsford L, Broer S, Brown P, Calo G, Carter WG, Catterall WA, Chan SL, Chao MV, Chiang N, Christopoulos A, Chun JJ, Cidlowski J, Clapham DE, Cockcroft S, Connor MA, Cox HM, Cuthbert A, Dautzenberg FM, Davenport AP, Dawson PA, Dent G, Dijksterhuis JP, Dollery CT, Dolphin AC, Donowitz M, Dubocovich ML, Eiden L, Eidne K, Evans BA, Fabbro D, Fahlke C, Farndale R, Fitzgerald GA, Fong TM, Fowler CJ, Fry JR, Funk CD, Futerman AH, Ganapathy V, Gaisnier B, Gershengorn MA, Goldin A, Goldman ID, Gundlach AL, Hagenbuch B, Hales TG, Hammond JR, Hamon M, Hancox JC, Hauger RL, Hay DL, Hobbs AJ, Hollenberg MD, Holliday ND, Hoyer D, Hynes NA, Inui KI, Ishii S, Jacobson KA, Jarvis GE, Jarvis MF, Jensen R, Jones CE, Jones RL, Kaibuchi K, Kanai Y, Kennedy C, Kerr ID, Khan AA, Klienz MJ, Kukkonen JP, Lapoint JY, Leurs R, Lingueglia E, Lippiat J, Lolait SJ, Lummis SC, Lynch JW, MacEwan D, Maguire JJ, Marshall IL, May JM, McArdle CA, McGrath J.C, Michel MC, Millar NS, Miller LJ, Mitolo V, Monk PN, Moore PK, Moorhouse AJ, Mouillac B, Murphy PM, Neubig RR, Neumaier J, Niesler B, Obaidat A, Offermanns S, Ohlstein E, Panaro MA, Parsons S, Pwrtwee RG, Petersen J, Pin JP, Poyner DR, Prigent S, Prossnitz ER, Pyne NJ, Pyne S, Quigley JG, Ramachandran R, Richelson EL, Roberts RE, Roskoski R, Ross RA, Roth M, Rudnick G, Ryan RM, Said SI, Schild L, Sanger GJ, Scholich K, Schousboe A, Schulte G, Schulz S, Serhan CN, Sexton PM, Sibley DR, Siegel JM, Singh G, Sitsapesan R, Smart TG, Smith DM, Soga T, Stahl A, Stewart G, Stoddart LA, Summers RJ, Thorens B, Thwaites DT, Toll L, Traynor JR, Usdin TB, Vandenberg RJ, Villalon C, Vore M, Waldman SA, Ward DT, Willars GB, Wonnacott SJ, Wright E, Ye RD, Yonezawa A, Zimmermann M. The Concise Guide to PHARMACOLOGY 2013/14: overview. Br J Pharmacol 2013;170:1449–1458.
11. Krupa A, Abhinandan KR, Srinivasan N. KinG: a database of protein kinases in genomes. Nucleic Acids Res 2004;32:D153–D155.
12. Holliday GL, Andreini C, Fischer JD, Rahman SA, Almonacid DE, Williams ST, Pearson MACiE: exploring the diversity of biochemical reactions. Nucleic Acids Res 2012;40:D783–D789.
13. Rawlings ND, Waller M, Barrett AJ, Bateman A. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res 2014;42:D503–D509.
14. Lane L, Argoud-Puy G, Britan A, Cusin I, Duek PD, Evalet O, Gateau A, Gaudet P, Gleizes A, Masselot A, Zwahlen C, Bairoch A. neXtProt: a knowledge platform for human proteins. Nucleic Acids Res 2012;40:D76–D83.
15. Tsirigos KD, Bagos PG, Hamodrakas SJ. OMPdb: a database of {beta}-barrel outer membrane proteins from Gram-negative bacteria. Nucleic Acids Res 2011;39:D324–D331.
16. Gandhimathi A, Nair AG, Sowdhamini R. PASS2 version 4: an update to the database of structure-based sequence alignments of structural domain superfamilies. Nucleic Acids Res 2012;40:D531–D534.
17. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M. Pfam: the protein families database. Nucleic Acids Res 2014;42:D222–D230.
18. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, Murphy MR, O'Leary NA, Pujar S, Rajput B, Rangwala SH, Riddick LD, Shkeda A, Sun H, Tamez P, Tully RE, Wallin C, Webb D, Weber J, Wu W, DiCuccio M, Kitts P, Maglott DR, Murphy TD, Ostell JM. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res 2014;42:D756–D763.
19. Akiva E, Brown S, Almonacid DE, Barber AE, Custer AF, Hicks MA, Huang CC, Lauck F, Mashiyama ST, Meng EC, Mischel D, Morris JH, Ojha S, Schnoes AM, Stryke D, Yunes JM, Ferrin TE, Holliday GL, Babbitt PC. The structure–function linkage database. Nucleic Acids Res 2014;42:D521–D530.
20. Saier MH, Jr, Reddy VS, Tamang DG, Vastermark A. The transporter classification database. Nucleic Acids Res 2014;42:D251–D258.
21. Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK, Beck E. TIGRFAMs and genome properties in 2013. Nucleic Acids Res 2013; 41:D387–D395.
22. UniProt C. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res 2014;42:D191–D198.
23. Fernandez-Suarez XM, Rigden DJ, Galperin MY. The 2014 Nucleic Acids Research Database issue and an updated NAR online Molecular Biology Database Collection. Nucleic Acids Res 2014;42: D1–D6.
24. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 2009;5:e1000605.
25. Nagy A, Hegyi H, Farkas K, Tordai H, Kozma E, Banyai L, Patthy L. Identification and correction of abnormal, incomplete and mispredicted proteins in public databases. BMC Bioinformatics 2008;9:353.
26. Kirby AJ. The lysozyme mechanism sorted—after 50 years. Nat Struct Biol 2001;8:737–9.
27. Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, Madden TL, Matten WT, McGinnis SD, Merezhuk Y, Raytselis Y, Sayers EW, Tao T, Ye J, Zaretskaya I. BLAST: a more efficient report with usability improvements. Nucleic Acids Res 2013;41:W29–W33.
28. Umemura M, Nagano N, Koike H, Kawano J, Ishii T, Miyamura Y, Kikuchi M, Tamano K, Yu J, Shin-ya K, Machida M. Characterization of the biosynthetic gene cluster for the ribosomally synthesized cyclic peptide ustiloxin B in *Aspergillus flavus*. Fungal Genet Biol 2014;68:23–30.
29. Ursing BM, van Enckevort FH, Leunissen JA, Siezen RJ. EXProt: a database for proteins with an experimentally verified function. Nucleic Acids Res 2002;30:50–51.
30. Joosten RP, Long F, Murshudov GN, Perrakis A. The PDB_REDO server for macromolecular structure model optimization. IUCrJ 2014;1:213–220.
31. Berman HM, Kleywegt GJ, Nakamura H, Markley JL. The Protein Data Bank archive as an open data resource. J Comput Aided Mol Des 2014;28:1009–1014.

32. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res 2014;42: D32–D37.
33. Pakseresht N, Alako B, Amid C, Cerdeno-Tarraga A, Cleland I, Gibson R, Goodgame N, Gur T, Jang M, Kay S, Leinonen R, Li W, Liu X, Lopez R, McWilliam H, Oisel A, Pallreddy S, Plaister S, Radhakrishnan R, Riviere S, Rossello M, Senf A, Silvester N, Smirnov D, Squizzato S, ten Hoopen P, Toribio AL, Vaughan D, Zalunin V, Cochrane G. Assembly information services in the European Nucleotide Archive. Nucleic Acids Res 2014;42:D38–D43.
34. Kosuge T, Mashima J, Kodama Y, Fujisawa T, Kaminuma E, Ogasawara O, Okubo K, Takagi T, Nakamura Y. DDBJ progress report: a new submission system for leading to a correct annotation. Nucleic Acids Res 2014;42:D44–D49.
35. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25–29.
36. Chibucos MC, Mungall CJ, Balakrishnan R, Christie KR, Huntley RP, White O, Blake JA, Lewis SE, Giglio M. Standardized description of scientific evidence using the Evidence Ontology (ECO). Database (Oxford) 2014;2014.
37. Spedding M. Resolution of controversies in drug/receptor interactions by protein structure. Limitations and pharmacological solutions. Neuropharmacology 2011;60:3–6.
38. Hardin G. The tragedy of the commons. The population problem has no technical solution; it requires a fundamental extension in morality. Science 1968;162:1243–1248.
39. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, Finn RD. Rfam 12.0: updates to the RNA families database. Nucleic Acids Res 2015;43: D130–D137.
40. Pettifer S, Ison J, Kalas M, Thorne D, McDermott P, Jonassen I, Liaquat A, Fernandez JM, Rodriguez JM, Partners INB, Pisano DG, Blanchet C, Uludag M, Rice P, Bartaseviciute E, Rapacki K, Hekkelman M, Sand O, Stockinger H, Clegg AB, Bongcam-Rudloff E, Salzemann J, Breton V, Attwood TK, Cameron G, Vriend G. The EMBRACE web service collection. Nucleic Acids Res 2010;38:W683–W688.
41. Grosjean J, Soualmia LF, Bouarech K, Jonquet C, Darmoni SJ. An approach to compare bio-ontologies portals. Stud Health Technol Inform 2014;205: 1008–1012.
42. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Consortium OBI, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007;25:1251–1255.
43. Wren JD. URL decay in MEDLINE–a 4-year follow-up study. Bioinformatics 2008;24:1381–1385.
44. Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S, Sangrador-Vegas A, Scheremetjew M, Rato C, Yong SY, Bateman A, Punta M, Attwood TK, Sigrist CJ, Redaschi N, Rivoire C, Xenarios I, Kahn D, Guyot D, Bork P, Letunic I, Gough J, Oates M, Haft D, Huang H, Natale DA, Wu CH, Orengo C, Sillitoe I, Mi H, Thomas PD, Finn RD. The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res 2015;43:D213–D221.
45. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Bryant SH. CDD: NCBI's conserved domain database. Nucleic Acids Res 2015;43:D222–D226.
46. Orchard S. Data standardization and sharing-the work of the HUPO-PSI. Biochim Biophys Acta 2014;1844:82–87.
47. Orchard S, Kerrien S, Abbani S, Aranda B, Bhate J, Bidwell S, Bridge A, Briganti L, Brinkman FS, Cesareni G, Chatr-aryamontri A, Chautard E, Chen C, Dumousseau M, Goll J, Hancock RE, Hannick LI, Jurisica I, Khadake J, Lynn DJ, Mahadevan U, Perfetto L, Raghunath A, Ricard-Blum S, Roechert B, Salwinski L, Stümpflen V, Tyers M, Uetz P, Xenarios I, Hermjakob H. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. Nat Methods 2012;9:345–350.
48. Joosten RP, teBeek TA, Krieger E, Hekkelman ML, Hooft RW, Schneider R, Sander C, Vriend G. A series of PDB related databases for everyday needs. Nucleic Acids Res 2011;39:D411–D419.
49. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 2014;42:D199–D205.
50. Wang Q, Arighi CN, King BL, Polson SW, Vincent J, Chen C, Huang H, Kingham BF, Page ST, Rendino MF, Thomas WK, Udwary DW, Wu CH. Community annotation and bioinformatics workforce development in concert—Little Skate Genome Annotation Workshops and Jamborees. Database (Oxford) 2012;2012: bar064.