SUMMARY STATEMENT
(Privileged Communication)

Release Date: 08/22/2005

Peter Good
301-435-5796
goodp@mail.nih.gov

Application Number: 2 U01 HG002712-04

APWEILER, ROLF PHD
EUROPEAN MOLECULAR BIOLOGY LAB
DEPT OF RES & SERVICE CENTRE
WELLCOME TRUST GENOME CAMPUS

Review Group:	ZHG1 HGR-P (O3) National Human Genome Research Institute Special Emphasis Panel		
Meeting Date:	07/28/2005		
Council:	OCT 2005	*PCC:*	X4PG
Requested Start:	10/01/2005	*Dual PCC:*	NV3
		Dual IC(s):	DE, GM, LM, MH, RR

Project Title: The UniProt protein sequence and function knowledgebase

SRG Action: Priority Score: 185
Human Subjects: 10-No human subjects involved
Animal Subjects: 10-No live vertebrate animals involved for competing appl.

Project Year	Direct Costs Requested	Estimated Total Cost
4	9,759,307	10,000,277
5	10,037,871	10,285,719
6	10,339,009	10,594,292
7	10,649,179	10,912,121
8	10,968,651	11,239,481
TOTAL	51,754,017	53,031,889

ADMINISTRATIVE BUDGET NOTE: The budget shown is the requested budget and has not been adjusted to reflect any recommendations made by reviewers. If an award is planned, the costs will be calculated by Institute grants management staff based on the recommendations outlined below in the COMMITTEE BUDGET RECOMMENDATIONS section.

2U01HG002712-04 Apweiler, R.
COMMITTEE BUDGET RECOMMENDATIONS
FOREIGN INSTITUTION

RESUME AND SUMMARY OF DISCUSSION: In this renewal application the principal investigator (PI) proposes to continue the collaboration involving three major groups that have formed the UniProt Consortium to develop and make available the Universal Protein Resource (UniProt) database. The overall goal of UniProt is to serve as a comprehensive source of information on proteins, including their sequence and function, to provide the research community with ready access to this information, and to link to other databases and information resources that present information about protein structure and function. For the renewal application the applicants propose an extensive list of specific aims including: maintaining and enhancing the UniProt archive, knowledgebase, and literature curation and cross-references databases; perform classification of all protein sequences into families and superfamilies; develop automated annotation procedures for proteins; extend their efforts on evidence attribution to track data entries back to their original source; and enhance the dissemination of UniProt through improvements to user interfaces, additional links to other databases, and providing more tools for queries and searches of datasets of ever increasing size.

This renewal application was received with measured enthusiasm by the members of the Special Emphasis Panel (SEP). The reviewers noted several strengths about this application. The applicants are to be commended for tackling the laudable task of creating and maintaining a comprehensive resource for proteins. The PI and his collaborators have made substantial progress on the challenging task of merging and integrating the three previously independent databases Swiss-Prot, TrEMBL, and PIR-PSD. A major strength of the application is the solid technology that underpins the databases that make up UniProt. The investigative team has substantial previous experience with database and annotation activities and they do this work very well. The reviewers noted that the applicants are highly qualified to carryout many of the centrally important activities critical to the UniProt resource. Finally, it should not be overlooked that UniProt is an invaluable resource to a large number of investigators. As such, the significance of this application was viewed as extremely high by the SEP.

The track records of the investigators working on this resource are excellent. Overall, there was substantial enthusiasm expressed for work that has been done on this resource in the initial funding phase. The current application proposes to continue the multi-institutional structure known as the UniProt Consortium that involves investigators at the European Bioinformatics Institute (EBI), the Protein Information Resource (PIR) at Georgetown University Medical Center, the National Biomedical Research Foundation, and the Swiss Institute of Bioinformatics (SIB). For the most part the consortium members seem to work well together, but the reviewers noted that there appears to be better integration and interaction between EBI and SIB, that tend to take similar approaches to many of the routine aspects of running a database, whereas PIR is not as closely aligned and integrated as the other two groups.

The SEP also identified several weaknesses with the application and expressed concerns about the future directions proposed for UniProt in this application. A major concern voiced by the SEP was the fact that this resource is already large, and it will only increase in size and scope as more and more data become available. The reviewers drew an analogy between UniProt and a very large corporation that finds itself involved in too many disparate activities such that it becomes distracted from what it does best. The plethora of activities also places a premium on making good choices of what activities to pursue in the future since the scale of each activity will represent a substantial investment in resources. Thus, the reviewers expressed the view that UniProt should be re-focused on the core activities that it has done best in the past, and that the research community has come to rely on them to provide; archiving sequence-based information, performing manual curation of protein sequences, and providing a facile interface that can be used by a broad cross-section of researchers to access the information maintained in UniProt, as well as connections to other sources of protein-based information.

A major weakness of the renewal application is that the applicants did not present a description of their vision for the future of this resource. The application contains a long list of extensions and improvements of current on-going activities, as well as some new initiatives, but the proposal was not seen as visionary, and there wasn't a sense that the applicants were attempting to position themselves to meet the anticipated needs of the community five years hence. Related to this point, the SEP noted that the UniProt resource needs more oversight. This should include both an external scientific advisory board (SAB) that would serve the NIH program developers, as well as an SAB set up by the applicants specifically to provide them with advice on both scientific matters and the long-term vision of how UniProt should evolve to meet the needs of the research community. The reviewers noted that the applicants had been advised by the last review panel to establish a SAB and it was regarded as both a disappointment and a major weakness that the applicants have failed to take action on this important issue. The absence of an active SAB, along with a description of an effective plan by which the applicants would interact with the SAB, was a major factor in the SEP's recommendation to reduce the budget to three years of support. Once again, the applicants are strongly advised to establish a SAB, and (among other things) make use of the SAB to help them develop a long-range vision of what UniProt will be in the future.

The future is expected to bring lots of very complex protein datasets from systems biology research activities that focus on identifying gene networks and pathways and protein-protein interactions. The reviewers also envisioned technology that will permit investigators to make quantitative measurements of protein species over time in living organisms. Many of these types of datasets will be produced by small groups of investigators who will make the data available in many (potentially) unrelated "boutique" databases. The reviewers were disappointed that the applicants did not identify these types of activities as more near-term challenges, or set any goals to develop the ability to capture and present these types of data along with appropriate search and retrieval tools that would make them useful and accessible to the research community.

The reviewers expressed a concern that UniProt has the appearance of being designed by software developers without sufficient input from the user community, or an adequate assessment of what the community's needs are. The reviewers expressed the view that the activities described in the application have a developer-centric perspective, rather than a user-centric view. Related to this point, the SEP noted that it is important for the applicants to keep in mind that UniProt is a service project that should be focused on addressing the user needs. Research activities should be developed and supported under very separate grant applications.

The applicants could be doing more to stay in touch with the user community to assess future needs. One example would be to hold workshops to establish a public dialogue with the community of investigators who are served by UniProt. To the credit of the investigators, the reviewers were very favorably impressed by the "adopt a protein" plan that would recruit community scientists to update information on specific proteins. This was viewed as a very productive and effective way to bring expertise from the research community to bear on an important problem for any database.

A critically important type of data that is currently being produced, and is expected to be generated in ever increasing amounts, is data from protein mass spectrometry studies. The reviewers noted that the application did not present an adequate plan for how this type of data would be captured and presented to users. In addition, mass spectrometry is an area of expertise that is under-represented on the research team.

The applicants discuss their plans for proteome datasets, but they don't have a satisfactory plan to handle the amount of data for the human or other large mammalian species. The reviewers acknowledged that this would be a large and complex undertaking, but the applicants have not acknowledged the problem or presented adequate attempts to address the challenges they will face.

The applicants appear to have intentionally avoided an in-depth discussion of their interaction with scientific journals. The reviewers regarded an interface and interaction with journals to be an important aspect of the data capture process, but the SEP noted that the applicants almost appear to regard journals as “the enemy” rather than an entity with which they should attempt to establish a partnership. There needs to be a better interaction and better connection between journals and UniProt.

The applicants are to be commended for giving thought to setting up milestones by which they could measure progress and help them to assess whether they were achieving success by following the approaches outlined in the application. However, the reviewers ironically found the 19-page description of the milestones to be so overwrought with detail that they could not determine if the milestones would serve the intended purposes.

In summary, the SEP expressed substantial enthusiasm for the progress the applicants have made during the initial funding phase to forge an effective collaboration among the four funded groups, and to produce a resource that is critically important to a large number of investigators. The investigators were judged to be highly qualified to continue the management and operation of this resource. The reviewers did express several concerns that dampened their overall enthusiasm for the application, and caused them to recommend a reduction in the length of funding to three years. The major weaknesses in the application included: insufficient long-term vision of how UniProt should evolve, the absence of an advisory board to provide oversight, input on community needs, and guidance on setting short-term and long-term priorities, and insufficient attention paid to the emergence of new types of protein datasets (e.g., networks, pathways and mass spectrometry).

The comments in the CRITIQUE section were prepared by the reviewers assigned to this application and are provided without significant modification or editing by staff. The RESUME AND SUMMARY OF DISCUSSION section documents the final outcome of the evaluation by reviewers and is the basis for the assigned priority score.

DESCRIPTION (provided by applicant): Bioinformatics infrastructural activities are crucial to modern biological research. Complete and up-to-date databases of biological knowledge are vital for the increasingly information-dependent biological and biotechnological research. With the recent accumulation of genome sequences for many organisms, most notably the draft human sequence, attention has turned to the identification and function of proteins encoded by these genomes. In the Universal Protein Resource (UniProt) project, funded by the NIH, major European and American protein sequence databases have joined forces and developed a central resource for protein sequences and functions providing a cornerstone for a wide range of scientists active in modern biological research, especially in the field of proteomics.

The broad, long-term objectives of this project are to provide with the Universal Protein Resource a stable and comprehensive resource for information on proteins, their sequences and their functions, to enable scientists to use UniProt to identify and analyze genes and their products and to make queries across databases containing complementary information, and to provide efficient and unencumbered access to the databases produced by the UniProt Consortium.

The specific aims are to maintain and further develop the UniProt Knowledgebase (UniProtKB) as the central database of curated protein sequences with annotations of sequence and functional information, to maintain and further develop the UniProt Archive (UniParc) and create the UniProtKB entry history server to ensure comprehensive coverage of all protein sequences and their annotation history, to maintain and further develop the UniProt Reference Clusters (UniRef) to provide a complete covering of sequence space while hiding redundant sequences (but not their descriptions) from view, to facilitate the use of these databases by providing user-friendly interfaces, tools for simple and complex queries and for retrieval of large datasets, down-loadable database records in defined, parsable format, and user support services; and to provide the flexibility and adaptability needed to be responsive to the changing needs of the scientific community.

These databases produced by the UniProt Consortium will facilitate development of preventive and curative strategies for health maintenance by allowing researchers to integrate the enormous amount of data from the Human Genome Project and other genome projects as well as from structural and functional genomics and proteomics projects to understand the genetic and biological mechanisms causing human disease.

CRITIQUE 1:

Overall this is an excellent proposal from a group of outstanding scientists. Based on the track record of the UniProt group, and on the plans set forth in the application, enthusiasm for the proposal is high.

The major strengths of the proposal are:

(1) Track record of UniProt, especially their demonstrated ability to gather and archive “immutable” and time-stamped sequences. This is the group’s real strength.
(2) Tackling the laudable goal of being a comprehensive resource on protein sequence, a huge undertaking that is of immense importance to the biological research community.
(3) The UniProt database they have proposed is well-designed, and appears to be flexible enough to leave room for the modifications that might be required as science advances.
(4) The UniProt team has assembled a wide network of talented collaborations with other bioinformatics and computational groups including TeraGRID, PrIDe, RefSeq, SEED, International Protein Index, MEROPS, etc.
(5) Their ideas on getting the biology community to “adopt a molecule” to enhance the depth and accuracy of functional annotations is especially provocative.

The weaknesses of the proposal include:

While these stated weaknesses don’t dampen enthusiasm for the concept of the proposal, they emphatically underscore the need for oversight by an advisory board, and by the NIH.

As proposed, the UniProt effort is so large and involved that it is almost certain to become unwieldy. Given the importance of this effort to the world, strong and appropriate oversight should be involved (substantially upgraded over what currently exists). There needs to be more integration of the many separate efforts, particularly on nomenclature and problem solving. This proposal looks like one of America’s large corporations that needs to divest individual projects to retain their focus. I felt as if they were describing what should be the national/international plan, rather than their component of the effort.

Although the team has a working history together, it appears as if they are attempting to do too many different things without recruiting additional expertise in new areas (e.g., biology and proteomics). They should have representatives of these fields on staff or on an external advisory boards, if they are to delve into the areas they suggest.

I found the reliance on bioinformatics jargon in this application to be very bad. Even the simplest task, naming and/or numbering genes, is poorly organized. Why does every protein end up with more than

10 accession numbers? This makes it difficult, if not impossible, for the end user “biologist” to extract useful information across databases. The real challenge and problem here is to focus on this key issue and simplify the problem so that it doesn’t need an entirely new language to describe. Since they have linguists on board their extended team, one would hope that this comment would be put to them.

The plan lacks any clear path forward toward integrating the data collected by the proteomics and mass spectrometry community. It is a stated goal, but no solution is provided. This is an extremely challenging problem, one that could be the focus of a separate bioinformatics consortium.

The group is sequence-centric, yet it is now clear that proteins can have very similar folds even with extremely low sequence identity (even less than 15%). There is no discussion of this issue, nor any plan to bring structure-based categorizations into UniProt.

More thought, consideration and national oversight should be given to the issue of manual annotation (induces bias, errors etc).

UniProt is developer-centric, rather than “user-centric.” Consideration should be given to examining and reconsidering some aspects of UniProt from a user’s as opposed to the developer’s perspective. For example, improving the external interface of UniProt to make it more user friendly (e.g., ask up-front “What do you want to do?” and have a list: 1) compare all family members at structural level, 2) compare domain organization, 3) compare tissue or cell expression patterns?

SPECIFIC COMMENTS

The proposal talks about creating proteome sets for mouse and human genomes. This is an incredibly complex task, and they state this as a priority. However, they don’t propose any clear solution to dealing with the complexity. Furthermore, based on the write-up there does not seem to be a clear plan of how to integrate the proteome sets into mass spec search algorithms, or to make it “user friendly” for the proteomics community. They make note of the PrID database for storage of MS data, but the value of this database is at present highly limited and its future maturation is not guaranteed. This is another example of why the team is proposing too much, and that strong oversight of this project is needed to integrate UniProt into other national and international efforts.

Another example of the developer-centric view present in this application is the use of the term “Comment line” in the UniProtKB to list all known properties of the protein (e.g., multiple enzymatic or binding functions). “Comment Line” is clearly programmer language, and doesn’t convey anything descriptive about the information contained in that section. This developer-centric language is found throughout the system and propagates the difficulty with jargon.

The UniProt team plans to use standards adopted by NIH-funded centers like the HUPO for proteomics, STRENDA for reporting enzymology data. This issue it certainly not unique to this proposal, but the standards issue is a substantial problem, if not the major problem with making bioinformatics seamless and usable. The bioinformatics community is essentially developing a language for the “user” who is often a biologist or biochemist. There doesn’t seem to be adequate participation of the biology community, nor is there adequate attention given to simplification. Given UniProt’s skill sets, and their place at the foundation of all of this (i.e., protein sequence), they might be able to take the lead and push for a highly simplified and meaningful nomenclature.

In this regard is it important to bear in mind that UniProt’s position on protein naming seems to be exactly the opposite of what a language is for; to convey meaning. In the application they state that “The protein naming guidelines are written on the premise that a good and stable recommended name for a protein is the name that is as neutral as possible. In general it should not reflect the function or the role of the protein, nor its subcellular localization, domain structure etc." This is a rather strange position to take, and is probably a reflection of the developer-centric view of the problem. What is most important in a name? Is it stability and neutrality, or its use as a descriptor? What they are describing is a number, NOT a name. Although this may seem like a trivial point, it underscores the distinction of a database from the developer's standpoint vs. the user's standpoint, and shows that considerable more thought must go into nomenclature.

One of the superb ideas in the proposal is the "adopt-a-protein" concept. The objective of this effort would be to gain the participation of the "expert" scientific community to annotate proteins. They use the online encyclopedia called Wikipedia as a prototype example. This would be a tremendously valuable resource for the entire community. Importantly though, it is not clear that this idea should be developed in the context of the core UniProt proposal. There are so many things proposed in this application, and so many issues left unmentioned with respect to the basic archiving and nomenclature issues, that the "adopt-a-protein" concept might be better left to a separate proposal.

CRITIQUE 2:

Significance: This proposal addresses a very significant need: that of the protein community to have a central database of sequences, a consistent nomenclature describing them, and a central hub linking the numerous and disparate resources that provide information about proteins. For the protein community to present its information well to the world it is imperative that we have a high-quality central database and this proposal addresses this issue.

Approach: While overall I feel that SwissProt, and later Uniprot, have done a very successful job over the years in performing the services of a robust database, I am concerned that the overall organization of Uniprot is beginning to become unwieldy and unresponsive, very much like a successful company that grows from a start-up into a large multinational through mergers and acquisitions. It is thus imperative for Uniprot to divest of some of its more peripheral activities and focus on its core mission – that is, providing consistent manual curation of protein sequences; a robust platform linking together all the protein resources; an intuitive and easy to use user interface for a broad user base, and a foundation upon which the protein community can collect information.

The particular type of business that I feel that Uniprot proper should avoid conducting is "research." This proposal is supposed to be one of community "service". I believe that the applicants need to clearly separate their efforts on service from research. Research proposals should be sent to the NIH as normal research grants.

In particular, I was not enthusiastic about the data mining and automatic annotation parts of this proposal. This is not to say that I am not enthusiastic about this subject area, which I think is becoming increasingly valuable with the accumulation of more and more information on proteins. However, I see data mining as more of a research area, not a service business. Furthermore, given the "service" format of the proposal, I could not really evaluate the sections of it that dealt with automatic annotation and data mining. There were a few literature references in the sections to papers published by the applicants and others in the field to provide context. In a standard research proposal, one would normally see many references and peer-reviewed publications providing a useful metric with which to evaluate the quality of the proposed research.

Finally, just as any large multinational company needs a strong Board of Directors to represent the interests of its stockholders, a very strong scientific advisory board (SAB) needs to be in place for Uniprot that would represent the interests of the protein community and keep the investigators in this effort focused on their core mission of service.

In evaluating this proposal, I looked carefully at the Uniprot website. The following are some general thoughts on this:

Uniprot is valuable to human users perusing and navigating the site for information, but also progressively as a key database in automated analyses. I am not sure what the balance is in order to optimize for each of these two types of use, but likely until recently the emphasis has been more on human users. I feel that supporting automated analyses is becoming increasingly important.

Users may download the whole UniProt knowledgebase in various formats, including simple FASTA, the main .dat format, and recently as RDF; these downloadable versions constitute the main basis for automated analyses. In general, I feel these are well structured and the documentation explains this structure, the meanings of different line types, etc. However, there is still information available from the web interface that is not available in these downloadable formats which I feel should be added, either directly or as supplemental files. One particular example is the protein properties accessible by the programs, such as ProtParam, ProtScale, etc.

Uniprot has provided a web interface and an ftp download interface for users to access their data. I think it would be more flexible for this project to provide some sort of APIs that users could call remotely to perform complex queries in their own applications.

Innovation: I evaluated this proposal more as a service proposal, focused on providing service to the protein community, than an innovative research one. However, I believe that Uniprot has a number of areas where it can uniquely innovate.

I felt that areas of good innovation included the use of the wiki approach to leverage all the experts in the protein community and enable them to contribute annotation. I also liked the idea of trying to bring in the environmental genomic sequences, which currently do not have a clear taxonomy.

However, I did not feel that the proposal had a clear vision for a 21st-century protein resource. There were some areas that I felt the proposal could have tackled in a more innovative fashion. For example:

There is no real discussion of how Uniprot will deal with journals. Given the weight of this organization it is imperative that they address the issue of the information stored in journals and how that will be linked up with that of databases.

There was also no real innovative way of creating a dialogue with the protein community. In presenting ChEBI, there was no discussion of how to avoid the controversy generated by the NCBI's new service PubChem (in relation to the American Chemical Society).

There was no discussion of new innovations such as LSID for naming proteins.

The file formats that they propose using are somewhat antiquated. I feel that the SwissProt format has now grown too old, and that they should really be moving to the more modern XML type of syntax.

There is a lack of a clear plan on how to integrate proteomics mass spec data well into Uniprot. They definitely need to incorporate more specialists in this area into their effort.

Investigators: The proposed team includes many strong investigators and well-established scientists with a good track record.

Environment: The intellectual environment and computational infrastructure in all three sites in this proposal are outstanding. In particular, since its inception the EBI at Cambridge has established itself as a world-class center in bioinformatics. However, returning to my analogy of the multinational company, here one can definitely see the issue of a number of different businesses, brought together through mergers and acquisitions. One wonders if they really work well together. I did not see the same level of integration between PIR as I saw between the EBI and SIB components.

On the technical side, such lack of integration is apparent in the way the different sites approached database infrastructure. Why are they not all using common database engines? I notice that PIR and EBI are using Oracle, whereas SIB is using something different.

With regard to the website, I find it a little confusing that the three main servers (PIR, EXPASY and EBI) have similar looking homepages but different options and navigation.

Budget: The budget should be cut to no more than three years.

CRITIQUE 3:

It is becoming commonplace for major laboratories to acquire several hundred LC/MS/MS runs per experiment based on ion exchange and RP HPLC separations of proteolytic digests of cellular organelles, fractions or even lysates. These mixtures range from a few hundred proteins to a few thousand usually.

In order to comprehend such large data sets and gain the ability to make comparisons among them with a reasonable investment in expertise in a reasonable timeframe, the proteomics and mass spectrometry communities need to be able to deduce and assign peptide and protein identifications automatically with a high probability of being correct.

Because of the scale of the problem, these tens to hundreds of thousands of sequencing mass spectra(tandem) cannot be interpreted manually and thus a number of search engines have been developed over the past decade by the mass spectrometry community that are designed to search these mass spectral data against gene, protein and EST databases[MASCOT, Prospector, Sequest, etc.] and then employ robust scoring strategies to make correct assignments from all possible matches to any single mass spectrum.

The next phase of complexity will surround the emerging focus on deciphering epigenetic biology that is dependent on the associated assignment of posttranslational occupancies on an increasingly large scale.

UniProt’s accomplishments and its near-term challenges:

Uniprot is the only attempt thus far at the development and maintenance of a comprehensive, non-redundant reference protein database that contains annotated entries. It is, therefore, a vital tool for mass spectrometrists and their collaborating biological and biomedical researchers. Without it, proteomics researchers would be crippled, as the only databases available would be un-annotated nucleotide databases such as genpept, and from these it would be very difficult and time-consuming to convert these identifications into biologically significant information.

Once peptides/proteins have been “identified” by mass spec experiments, it is essential to be able to easily grasp a global view (preliminary understanding) of possible functions of these large sets of proteins, with the possibility of being able to cluster identified proteins together on the basis of common attributes; e.g., cellular function, sub-cellular location, known interaction partners, etc. Uniprot is a central repository for information about given proteins and is heavily referenced and linked to by most bioinformatic databases. This means that even if all the information you want is not in the UniProt database entry, it is easy to get the relevant information from another database using the UniProt accession number.

The Uniprot knowledgebase currently consists of two parts; the manually curated and annotated Swissprot, and the much larger automatically annotated TrEMBL. A major part of the development of the Uniprot knowledgebase is the development and improvement of annotation of TrEMBL entries. This is achieved initially through automated recognition of functional domains, allowing assignment of a protein to a family or superfamily. Protein sequences are then further assessed for homology to proteins of known function to try to identify orthologs and paralogs. Assignment of cellular location on the basis of signal peptide or predicted trans-membrane regions is attempted, and potential ligand-binding sites within protein families are looked for. After all this assignment of potential features is complete a set of extensive rules (RuleBase) is applied that tries to predict the function of a given protein with as high specificity as can be safely assigned. An equally important part of this process is the detection and “weeding out” of false positives, and software is employed specifically for this purpose.

The proposal does not present any statistics on the error rates of this software (false positives and false negatives). It mentions that known individual errors in automated assignments can be excluded, but the mechanism to identify these errors is not clear. The performance of the automated annotation software is probably the most important part of the UniProt database, as this is the process that will have the widest impact on the largest number of database entries. It will also be fundamental for their planned creation of a third part of the knowledgebase ‘UniProtKB/ENV’ which is derived from “environmental” samples; i.e., of unknown species origin or function. The doubling rate in size of UniProt seems to be 8-14 months.

A very important part of database entries is consistency and stability in annotation wording. This is required in order to be able to find commonalities between large numbers of proteins identified in proteomics experiments. This is something they are planning to work on, but it is clear that their current basic language is not sufficiently robust to scale as will be required for handling large numbers of post-translationally modified sites, etc. There needs to be a better way of changing accession numbers and tracking these changes – currently finding out what changed is a serendipitous task to say the least. Also, a more sophisticated way of assigning accession numbers is required that can accommodate growing protein sequence complexity (as in histones, for example where each particular sequence with a “name” such as H2B, can have occupancies by different groups even on the same residue (Me or Ac lysine, etc.). Maybe some automated annotation strategy could help provide this, but as they state, manually curated entries (Swissprot) also need to be revised for consistency. This whole topic would benefit tremendously from establishment of a high fidelity dialog with leaders in the mass spec/proteomics community possibly through a new scientific advisory committee structure.

The most labor-intensive part of the database is the manual curation by incorporation of new information published in the literature, and they acknowledge that this cannot be done in a comprehensive fashion, as there is too much literature to review. Hence, they propose focusing on certain topics: proteins previously completely uncharacterized, proteins whose 3D structure is being newly reported, publications reporting post-translational modifications, review papers, and proteins linked to human disease.

A new semi-automated literature curation pipeline is proposed. With the widespread online publication of literature, software has been developed that employs text-mining procedures to automatically search all literature for keywords and phrases and pull out information on when a given protein is discussed, or a modification is reported. Initial results of software looking for references to sites of phosphorylation (which can be extended to look for other post-translational modifications) were very successful. The proposal is to use this software to flag relevant papers that will then be read by a curator to confirm the automated search findings before incorporation into the UniProt knowledgebase. If this process works, this could transform the speed at which information gets incorporated into the knowledgebase, making database entries significantly more extensively annotated and up-to date with the latest information.

It is acknowledged that the mass spectrometry community is creating large amounts of useful information that should be utilized by UniProt. However, at the moment, it is not clear how best to access and use this information, and although the applicants talk about wanting to make use of this data, they do not outline any specific plans on how they plan to implement and achieve this. They have created a repository for storing mass spectrometry data (PrIde), but it is not clear what information should be entered in this database. This is not really entirely their fault though, as the mass spectrometry community itself has not come up with rigid guidelines on requirements for data to be of acceptable reliability. Some mass spectrometry results are more reliable than others (due mostly to the operating performance differences among different types of tandem mass spectrometers), and it is important to be able either only incorporate high fidelity results, or to be able to assign some level of confidence to results. There are now a number of mass spectrometry analysis software packages that can report results with probabilities or expectation values. However, there is no consensus on which software is better than others, and all software is still improving, so the field is not sufficiently mature to adopt the strongest guidelines. Uniprot appears to be actively involved in proteomics community discussions on this topic, but they should add some serious expertise to their leadership team. Otherwise, it is not clear that they will make judicious decisions on what data is appropriate to incorporate or how.

I think the PrIde database will not become extensively populated by the community as there is no incentive to enter data into the database. They should probably try to utilize data from other mass spectrometry data repositories such as PeptideAtlas or GPM (this is something they failed to talk about in the proposal). Data entered into these repositories are analyzed with searching software that reports probabilities or expectation values attached to each assignment, so reliabilities of results can be reported.

The key resource that is currently under-utilized by Uniprot is the scientific community itself. The two mechanisms by which the community can be utilized are either by direct submission of results by the researcher or by submission via a journal publisher. Both of these mechanisms are currently available, but are dramatically under-employed.

The applicants talk about expanding on two strategies to facilitate direct input of data. One is a free text format submission that has the advantage of being very user friendly, but has the potential disadvantage of not necessarily getting all the desired information, meaning data may be incomplete and/or less reliable. The other strategy is standardized submission forms for different information; e.g., a standard form for submission of a PTM. The problem, as the applicant’s acknowledge, is to create an incentive to submit data. In other fields a necessity to submit data as a pre-requisite for publishing has been employed, and the applicants suggest this as a possibility, with a meeting with publishers planned for 2006. Establishment of this dialog is essential and must be given is a high priority. They also propose a few other sensible ideas for how to better utilize community expertise. One is to assign responsibility for the annotation of a particular protein or set of proteins to a world expert on this particular protein; an ‘Adopt-a-Protein’ initiative, which they already employ to a limited extent. This would presumably lead to better annotation of certain protein entries.

As noted above, within the proteomics field the momentum is shifting from simply identifying large numbers of proteins in complex mixtures to global analyses of post-translational modifications and comparative, quantitative analyses of proteins and post-translational modifications in different samples. These are going to be the two data types that Uniprot will have to decide on a strategy to deal with. For the post-translational modification data they are going to need a way to assign a confidence to site assignments. They are also going to be regularly presented with results where one of two or three nearby residues are (for example) phosphorylated, but the exact residue is not determined. Is this information useful enough to be incorporated into the knowledgebase? Protein quantitative data is already incorporated into Uniprot to a limited extent by stating higher expression levels in certain tissue types. In the future they are going to be presented with much more complex information such as a 3-fold increase in protein expression when protein X is phosphorylated on residue Y in cell line Z. Is this information too specific to be in a universal knowledgebase?

Depending on the goal of the experiment, there are two different types of database that researchers would like to be able to search against. If researchers are analyzing a very complex mixture where they are observing relatively low sequence coverage of many proteins, they are not interested in different isoforms of a given protein. Hence, they would ideally like to be able to search against a database that has all highly homologous entries collapsed into a single entry, such as the UniRef90 database that Uniprot creates. One cannot currently search proteomics data against UniRef databases, although all Uniprot entries can be linked to UniRef entries.

The other extreme is when a researcher is trying to characterize in-depth a single or small number of proteins. In this situation the researcher wants to be able to determine all the isoforms of a protein, and wants to be able to search allowing for known potential amino acid substitutions and sites of post-translational modification. This information is available in UniProt, but it is not possible to annotate a sequence with specific sites of post-translational modifications for searching: it can only be searched in a generic fashion for modifications where every occurrence of a particular amino acid is given equal weighting for potential modification. To achieve this a new coding would be required for each type of modified amino acid. Thus, this could not be encoded in the FASTA format that is currently used for primary sequence storage. This is a topic that needs thought by both UniProt and the mass spec community.

An important part of their role is to act as a common point between many disparate biological databases to try to prevent duplication of work. Their approach to this seems to be to try to exchange as much information as possible between databases, so that new annotations to one database are rapidly incorporated into others. They seem to have this process working well.

CRITIQUE 4:

Summary: This review focuses on Aims D.2-D.4, proposed work on the computational and software development infrastructure and dissemination.

These aims provide a high-level view of how information flows through the database, a description of the software development environment, and the view to users. Several points mentioned as concerns in the original proposal remain problematic, and will be discussed in greater detail below. These include difficulties in the migration from flat-files to a relational database (RDB), the lack of cohesion of the individual efforts, and the dominance of the EBI and SIB groups with a lack of long-term goals or vision for PIR. As noted in the original review, establishing a strong SAB would help in improving these areas.

High-level evidence of the lack of coordination between groups is found in the split, somewhat redundant nature of the bioinformatics efforts. EBI and SIB seem to be working towards integration. PIR seems to be going in a parallel direction, but on a different track. The symptoms are flat-file transfers, different annotation frameworks (InterPro vs. PIRSF), different annotation editors (SIB & EBI vs. PIR), and different CVS repositories. My sense is that EBI & SIB are working well together, providing value to users, and PIR is doing additional bioinformatics analysis that is not central to the mission of UniProt.

D.2: Bioinformatics Framework

Approach: The workflow to generate updates is described as synchronizing UniProtKB; propagating changes to UniParc & UniRef; and distributing a new version every 2 weeks. This is a reasonable update cycle. At this level, the approach is sound.

A schematic overview of the overall approach is provided (p. 328):
SIB is responsible for updates to UniProtKB.
EBI uploads the KB (flatfile?) records, parses them into RDB records, and generates files.
PIR grabs the files by FTP, generates UniRef100/90/50 views.
A brief side note is that the PIR section describes name curation (p. 350). But ... I thought that naming was done by SIB/Swiss-Prot. How do the PIR names feed back to SIB?

EBI's workflow for uploading new sequence information from third parties looks solid. They parse records into an Oracle database.

The transitions from SIB to EBI and EBI to PIR are less well described. The workflow figure shows FTP at the interface, which I take to mean that flat file formats must be defined (or are these table dumps?). EBI and SIB have a specialized workflow for microbial genome data (p. 331). This is essential given the rising rate of production of complete microbial genomes.

I have a hard time understanding the mutable format of data records. The KB, Archive, and Ref portions of UniProt use an RDB. Yet the records that are transferred continue to be in a flat-file format. An example is the new format of the date (DT) field (p. 334). Why not use table dumps or XML to transfer information in a structured way? I find this very puzzling. I would think that an automatic interpreter could keep up with changes in XML. Using flat files, though, a parser must be continually tweaked as the flat files change. I certainly understand the value of keeping legacy software in service. What I would like to understand is how the groups see the trade-off of maintaining flat-file interfaces vs. taking advantage of an RDB infrastructure at each site. Are there any long-term goals to move from flat-file transfers to DB-level interactions? If not, why not? Are the flat-file formats anticipated to continue to be stable despite introduction of new technology such as mass spec?

UniParc provides the ability for a historic snapshot of sequence data, but it is not being designed to track annotation changes (p. 334). I would think that a change log would be very valuable in tracking down reasons for faulty annotations. I suspect that UniProt will find it necessary to build a tool to provide a historical trace backwards based on multiple snapshots. Why not just build it into the database? What is the constraint that trumps a change log?

EBI is using Oracle as its RDB (p. 336). This is a solid choice.
PIR is also using Oracle. They are taking advantage of Oracle "connect by" fields for DAGs, triggers, constraints, etc. SIB is migrating from flat files to an RDB, particularly for the manually-curated Swiss-Prot section that is now in flat files. But they are planning to use PostgreSQL (p. 336). While normally I'm in favor of open-source solutions, I think this is a case where it would be advantageous to switch to Oracle. My understanding is that Postgres doesn't yet perform at the level of Oracle. Also, each RDB has its own quirks, and the knowledge built at EBI won't transfer as easily to Postgres as to Oracle. This leads me to wonder whether the DBA at SIB talks to EBI or PIR.

The EBI and SIB groups are developing annotation editors based on XML. This is a good approach that should work better with changes in the data format (p. 343).

The PIR group seems to be going off in a different direction, using UML diagrams, Java, SOAP-XML,

(p. 346). Similarly, while work at SIB and EBI uses InterPro, the PIR group is developing PIRSF that seems to collect the same type of information (p. 346). This was one of the criticisms of the initial application; that the groups are each good, and the resource is needed, but that there was little in the plan describing how the work would be coordinated. It is worrisome that the groups can't adopt a shared development platform.

In summary: SIB and EBI seem to be working together well, with the exception of SIB's choice to go with Postgres instead of Oracle. PIR seems completely un-integrated, and their contribution seems much less than the other 2 groups. Although the research by PIR is notable, I don't think it adds much value to UniProt users.

D.3 Quality Assurance

EBI uses a standard C++ environment and CVS for version control.
SIB is importing its software into EBI's CVS.
PIR has its own CVS. So PIR and EBI/SIB teams will have to decide which group owns an application at the interface. The software will have to live in one of the two CVS repositories. Again, while this is not a fatal flaw, it points to a lack of cohesive effort.

The quality assurance section describes training good annotators, less on software quality assurance. UniProt's users are experts and will likely catch most software bugs quickly.

D.4 Database Dissemination

The approach is to distribute UniProt free of charge. This is very significant as it permits greater cooperation with other biological databases (NCBI).

Going forward, UniProt seems to want to become a portal with MyUniProt. The EBI and PIR sites will become identical (p. 356). MyUniProt will permit sticky preferences, saved query results, recurring queries with notification, which are all good ideas for heavy users.

Nevertheless, to assess the value of MyUniProt, I would like to see the types of users broken out: users who download the database for mirroring or bioinformatic analysis; casual users; and heavy users who would benefit from MyUniProt. For my use, UniProt is far inferior to EnsMart (Ensembl) for downloading huge chunks of data.

APIs will be defined for programmers. This seems to be a PIR effort, based on work with NCI caBIG (p. 361). If this is important here, shouldn't it be equally valuable in Sec. D.2 for replacing the flat-file interfaces? Why can't the UniProt sites use the API mechanism, rather than flat files, to swap information? Wouldn't it be easier to query for the records modified since the previous dump?

There seems to be little motion towards a shared DB infrastructure. Will this get burdensome as the groups attempt to add APIs to the code?

An additional concern with this section is that this work will be a distraction from the key mission of UniProt, providing stable identifiers for proteins.

An important issue that is not discussed is working directly with journals and authors for direct import of annotations from the literature into UniProt. I have heard proposals from Nature and other journals to require XML structure for information presented in papers. I would think that UniProt should be one of the groups agitating for this type of progress, and it is surprising that this is not mentioned.

Budget

The 2004 costs are ~$19M / yr, about 30% covered by NIH
The groups are asking for $10M / yr, which is double their current NIH funding of $5M.
Other DB budgets: DDBJ/EMBL/GenBank $30M
MGD $7M
PDB $10M
The CPU budget looks ok.
Storage equipment seems high: 25TB SAN + switches for $370K. I thought that the going rate for 1 TB was ~$1K, and I would have thought that the sites already had high-end switches. Can this be justified in greater detail?

My sense from the written proposal is that the work at PIR is not central to the mission of UniProt. It seems that much of this work could be spun off into a distinct effort funded under a different mechanism.

UniProt is a resource that is essential, but the proposal as written does not provide the long-term view that is required for a long-term funding commitment. The funding period should be reduced to 2 years, or at most 3 years, to permit UniProt an opportunity to re-think how it is approaching its long-term mission.

CRITIQUE 5:

Significance: The merger of SWISS-PROT and PIR has created a new, integrated database that is now the world's premiere resource for high-quality, annotated protein sequence information. While access to protein sequence data is available through NCBI, the quality of annotation provided by UniProt is unrivaled. In the post-genome era, a perplexing multitude of boutique databases has been established to curate protein-related information, for example protein expression data or protein-protein interactions, but none plays as crucial a role as UniProt. Moreover, with an emphasis on providing cross-references to as many protein information resources as possible, UniProt also acts as a kind of central hub that makes more specialized information in ancillary databases readily accessible. Thus the continued curation and dissemination of UniProt is of highest significance to the life science community.

Approach: In 2002, NIH provided funding to merge the formerly competing SWISS-PROT and PIR databases into a single protein sequence database, UniProt. That integration was a watershed event for the international database community, and technologically and organizationally a non-trivial undertaking. It is reassuring to see that the project team has been largely successful in their endeavor. At the data level, integration appears to be mostly complete though some of the initial goals (e.g. evidence tagging, p.229, or rules-based annotation, p.233) have apparently not progressed as fast as originally anticipated. At the UI level, the collaborators have opted to provide decentralized web access but have to be commended for making it possible for scientists in different countries to access different servers almost unnoticeable.

A major strength of this proposal is the robust technology and curation infrastructure that underlies UniProt. The processes in place to support data acquisition, content creation and curation, and database production are well-established and reflect the long-standing experience of the project team in that space. The underlying technology choices are well-justified and based on industry-standards. The discussion of quality issues displays a keen sense of importance of this topic and is commensurate with the expectations for the world's central resource of protein information. Data dissemination is discussed adequately. UniProt is updated biweekly and distributed in several formats via ftp and the UniProt website. While not discussed in detail, it is likely that the current infrastructure can support continuing growth of the database. Services provided by UniProt include web-based query and analysis tools, which enhance the value of the resource significantly. The outreach component of the proposal, in particular training and support, is excellent. The EBI and PIR are both involved in several collaborations, though the significance of some of them is somewhat unclear.

A major weakness of the proposal is a dearth of new ideas that could take UniProt to the next level. Most of the work proposed to take place under this grant is continuation of, and incremental improvements to, ongoing activities. Descriptions of specific aims for the new funding period are unfortunately not very crisp; the proposal would have been strengthened by distinguishing novel capabilities and non-trivial extensions of current activities more explicitly. New ideas – where presented – are addressed in very general terms. Proposed innovation such as community curation via DAS or Wikipedias and the involvement of young scientists and retired senior scientists as community curators doesn't go beyond the superficial. In several spots there is mentioning of a new UniProtKB division to capture "environmental and other taxonomically unassigned sequences." Specifics are lacking, and there is no cogent explanation why this data set requires a special database division (to deal with the absence of specific species information, one could simply create a "metagenome", "unclassified environmental species" etc., taxonomy code). While existing data structures, file formats, and curation processes are explained in occasionally excruciating detail (e.g., sections 1.5.1 and 1.5.2), such detail is generally absent when new initiatives are described. Among the most compelling ideas presented are changes to the UniProt web site that will offer personalization and user-specific customization. Unfortunately, only 6 pages are devoted to a high-level presentation of quite generic goals, in contrast to – for example – a long discussion of content standardization that earns more than 11 pages.

Even more troublesome is the glaring absence of a long-term vision for UniProt. With funding requested for five years, what will UniProt's role in the database world be at the end of that period? How will UniProt interact with the many post-genome databases that are being developed to capture and curate biological knowledge that goes beyond sequence?

The proposal contains an extensive list of milestones that illustrates excellence in planning and project management. At the same time, the great level of detail makes it impossible to develop a sense of priorities, dependencies, and risks.

Innovation: The amount of innovation in this proposal is low. Most activities are a continuation of ongoing efforts. Potentially exciting new ideas (community curation, improved web site) lack detailed discussion.

Investigators: The co-PIs are recognized leaders in the scientific database arena, with a strong track record of running a major community resource. Their qualification for the operational aspects of this project is beyond any doubt. The addition of a high-caliber Advisory Board would help to strengthen the strategic vision.

Environment/Organizational Structure: The environment at each site is outstanding. Coordinating three separate groups across international borders is non-trivial. The model deployed for managing the UniProt consortium is that of largely independent groups, with a shallow layer of cross-coordination. While there may be concerns about the lack of tighter project management, this specific approach appears to have served the consortium well. Of greater concern is the absence of a Scientific Advisory Board, which was proposed for the current funding period but apparently has not been installed. Each site has a separate Advisory Panel. Information about their activities, e.g. minutes of panel meetings, would have been very useful.

Overall Summary: This proposal is very strong in many areas. The need for the UniProt database to continue to provide this resource on protein information to the research community is blatantly obvious. The co-PIs have an outstanding reputation and are uniquely qualified to manage this project. The integration of SWISS-PROT and PIR has been, for all intents and purposes, a resounding success. UniProt is run as a highly professional operation, with processes and technologies well suited for producing a top-tier community database.

What limits enthusiasm for this proposal is its myopic focus on operational issues to the detriment of a strategic vision for UniProt in the post-genome era. Maybe rightly so, the applicants are emphasizing the status quo, and most activities proposed under this renewal application are straightforward extensions of current efforts. A couple of innovative ideas contained in here are not discussed at the level of detail they deserve. More importantly, however, the PIs have missed an opportunity to take a step back and develop a cogent, engaging vision for the protein database of the future. In this era of “systems biology”, the scientific community will be inundated with new types of protein-related knowledge. UniProt, at its core, is a sequence database. Everything in the post-genome era will not relate back easily to sequence, and the primary mechanism for linking UniProt to other databases (via cross-references) is likely to prove inadequate to preserve the richness of scientific complexity. It would have been of great interest to learn more about the PI’s vision for the role UniProt should play in systems biology and pathway research.

A serious concern is the absence of a Scientific Advisory Board, which was proposed to be in place for the current funding period but apparently has not been instituted. The absence of such an oversight committee, which can provide help and guidance with respect to planning and execution, goals, vision and priorities, and interactions with the scientific user community at large, is unacceptable.

BUDGET: The EBI hardware request seems excessive.

THE FOLLOWING RESUME SECTIONS WERE PREPARED BY THE SCIENTIFIC REVIEW ADMINISTRATOR TO SUMMARIZE THE OUTCOME OF DISCUSSIONS OF THE REVIEW COMMITTEE ON THE FOLLOWING ISSUES:

COMMITTEE BUDGET RECOMMENDATIONS: The SEP recommended the requested level of funding for the first three years, but recommended that funding for the fourth and fifth years be eliminated. The applicants are encouraged to refocus their efforts on the important task of archiving protein sequence data, performing protein annotation, and with the assistance of a Scientific Advisory Board developing a long-term vision of how the UniProt resource should evolve to serve the needs of the biomedical research community as new datasets of information on proteins are developed.

FOREIGN INSTITUTION: Two of the four institutions requesting funding (the European Molecular Biology Laboratory (EBI) and the Swiss Institute of Bioinformatics) are foreign institutions. The SEP noted that there is unique expertise at EBI and SIB in the form of a large number of experienced annotators that represent a decade of experience providing high quality entries for the database. While there are several US investigators who would be considered qualified to direct this resource, it is also the case that it would take years and significant funds to duplicate the hardware and human resource infrastructures that exist at EBI and SIB in order to support the UniProt resource at a US site.

ROP

NOTICE: The NIH has modified its policy regarding the receipt of amended applications. Detailed information can be found by accessing the following URL address: http://grants.nih.gov/grants/policy/amendedapps.htm

NIH announced implementation of Modular Research Grants in the December 18, 1998 issue of the NIH Guide to Grants and Contracts. The main feature of this concept is that grant applications (R01, R03, R21, R15) will request direct costs in $25,000 modules, without budget detail for individual categories. Further information can be obtained from the Modular Grants Web site at http://grants.nih.gov/grants/funding/modular/modular.htm

MEETING ROSTER

National Human Genome Research Institute Special Emphasis Panel
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
ZHG1 HGR-P (O3) 1
July 28, 2005

CHAIRPERSON

NADEAU, JOSEPH H., PHD
CHAIR AND PROFESSOR
DEPARTMENT OF GENETICS
SCHOOL OF MEDICINE
CASE WESTERN RESERVE UNIVERSITY
CLEVELAND, OH 441064955

MEMBERS

BADER, JOEL S., PHD
ASSISTANT PROFESSOR
DEPARTMENT OF BIOMEDICAL ENGINEERING
JOHNS HOPKINS UNIVERSITY
BALTIMORE, MD 21218

BURLINGAME, ALMA L, PHD
PROFESSOR
DEPARTMENTS OF CHEMISTRY AND
PHARMACEUTICAL CHEMISTRY
SCHOOL OF PHARMACY
UNIVERSITY OF CALIFORNIA
SAN FRANCISCO, CA 941430446

FUCHS, RAINER T., PHD
VICE PRESIDENT
BIOGEN IDEC, INC.
RESEARCH INFORMATICS
CAMBRIDGE, MA 02142

GERSTEIN, MARK , PHD
ASSOCIATE PROFESSOR
MOLECULAR BIOPHYSICS AND BIOCHEMISTRY
YALE UNIVERSITY
NEW HAVEN, CT 06520

SMITH, JEFFREY W, PHD
DIRECTOR & PROFESSOR
THE BURNHAM INSTITUTE
CENTER ON PROTEOLYTIC PATHWAYS
LA JOLLA, CA 92037

SCIENTIFIC REVIEW ADMINISTRATOR

POZZATTI, RUDY O., PHD
SCIENTIFIC REVIEW ADMINISTRATOR
SCIENTIFIC REVIEW BRANCH
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
NATIONAL INSTITUTES OF HEALTH
BETHESA, MD 20892

GRANTS TECHNICAL ASSISTANT

WILLIAMS-BEY, DIANE D
GRANTS TECHNICAL ASSISTANT
SCIENTIFIC REVIEW BRANCH
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
BETHESDA, MD 20892

Consultants are required to absent themselves from the room during the review of any application if their presence would constitute or appear to constitute a conflict of interest.