Responses to Summary Statement (2 U01 HG002712-04)

The National Human Genome Research Institute Special Emphasis Panel (SEP) for 2 U01 HG002712-04 listed a number of points that the SEP considered as weaknesses in the UniProt application 2 U01 HG002712-04. The UniProt Consortium will, in the following document, first address the major points in some detail and add short statements to address important minor points.

The majority of critical remarks were made about: (i) a lack of focus on core areas, (ii) insufficient long-term vision of how UniProt should evolve, (iii) the absence of an advisory board to provide oversight, input on community needs, and guidance on setting short-term and long-term priorities, (iv) insufficient attention paid to the emergence of new types of protein datasets (e.g., networks, pathways and mass spectrometry), and (v) lack of integration of PIR activities with the SIB and EBI activities.

1. Summary Critique (pp.2-3): ***UniProt should be re-focused on the core activities.*** Summary Critique (p.3): ***Research activities should be developed and supported under very separate grant applications.***

> The funding supplied by this NIH grant is indeed being used for the core activities of sequence archiving (with extension to proteins from environmental samples), manual curation of protein sequences (assisted by automated propagation of manually curated rules), the development and maintenance of a user-friendly and unified UniProt web site (with customized information using MyUniProt) and connecting to other protein-related sources (via direct interaction with these sources for expanded cross-references).
> It is important to note that, at each of the sites, the NIH-funded UniProt effort is only a fraction of the work being carried out. Other work, both research and service, is conducted under separate sponsorship and funding. Nevertheless, these other efforts frequently are leveraged to also serve the UniProt project. Thus, new developments have been funded (and will continue to be funded) by separate research grants and incorporated into UniProt only after they have been proven in concept and practice. Examples of separately-funded but complementary projects include the EU-funded TEMBLOR, BioBabel, BioMint, BioSapiens, EMBRACE and FELICS projects at EBI and at SIB, the MRC funded eFamily project and the BBSRC ISPIDER project; at EBI, and the NCI caBIG initiative, NIAID proteomic program and NSF text mining projects at PIR.

2. Summary Critique (p.3): ***the applicants did not present a description of their vision for the future of this resource****...*

> At the heart of the mission of the UniProt consortium is the archiving of sequence-based information, the 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. Outside this core are satellite databases with which UniProt interacts. However, these databases (InterPro,

IntAct, PRIDE, ChEBI, IntEnz, and so on) are separately funded and not part of this proposal.
UniProt is not able to accommodate all types of new data generated by functional genomics, proteomics and systems biology approaches directly in the UniProtKB. For all types of such data we are actively participating in the creation of standards to capture such data in a standardized way that allows the storage of this data in specialized databases. By collaborating closely with these specialized databases we can import in a curator-assisted way the highly reliable data from these databases into UniProt.
We can use the handling of protein-protein interaction data as an example. Such data are imported into UniProtKB from the IntAct protein interaction database. Only highly reliable data will be incorporated into UniProtKB, while we provide cross-references from UniProtKB to the complete IntAct entries. IntAct curation is coordinated by a small core of IntAct curators, but enhanced through intensive collaborations. This approach provides a high synergy effect, as the most time-consuming step in the curation process, being the actual reading of a scientific publication, needs to be done only once. A major milestone in the international collaboration between molecular interaction databases is the agreement that BIND, DIP, IntAct, MINT, and MIPS will regularly exchange all user-submitted data in a manner similar to the DNA data exchange between EMBL, GenBank, and DDBJ. This data exchange will be of great benefit to the scientific community and save global curation effort in the long run. It will eventually lead to a situation where all curated protein-protein interaction data will be available from every partner database. Thus, via IntAct, we can import all reliably curated protein-protein interaction data that has been curated by centers worldwide into UniProtKB. By cross-referencing back to IntAct, users will have ready access to data that is more extensive and with a depth of specialized annotation that is not suitable for inclusion in UniProtKB. UniProt will benefit from the activities of 30 curators worldwide working on protein-protein interaction data while only paying for one UniProt curator liaising with IntAct.
We firmly believe that this concept of focusing on our core activities combined with active liaisons with specialized databases dealing with the emerging new data types in functional genomics, proteomics and systems biology approaches plus active encouragement of the scientific community to engage directly in additional UniProt annotation will be an appropriate approach even beyond the envisaged five-year running time of the grant.

3. Summary Critique (p.3):... *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*** ...

We agree that UniProt needs an external scientific advisory board. However, we believe that this SAB can both advise UniProt and serve the NIH staff. Please note that each of the UniProt consortium members already has its own advisory boards that meet annually. Also, most of our other non-NIH funded projects have additional advisory boards and reviews. The EBI, for example, has at least two SAB meetings or progress reviews of UniProt-related activities per year by its various funders. It is also important to note that the UniProt consortium named potential members for the UniProt SAB at the start of the UniProt project. The nominees were members of advisory boards of the UniProt consortium. NHGRI received

these nominations early on and was regularly asked by the consortium members to establish an SAB to both advise UniProt and to serve the NIH staff.

4. Summary Critique (pp.3-4): *The future is expected to bring lots of very **complex protein datasets from systems biology research activities** ... A critically important type of data that is currently being produced ... 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 UniProt consortium is mindful of the complex protein datasets from systems biology and proteomic studies and has been actively engaged with these communities in several of our other funded activities to complement the core mission of UniProt. We actively seek the advice of the Proteomics community and especially the mass spectrometry community. In recognition of this proactive engagement, HUPO recently elected Cathy Wu into its Council and Rolf Apweiler as its new President. The EBI has spearheaded the HUPO Proteomics Standards Initiative to develop international standards for capturing and presenting large-scale protein-protein interaction and proteomic/mass spectrometry data. PIR is funded by the NIAID proteomic research program as a member of the Administrative Center that develops a central proteomic database and cyber infrastructure for storing and presenting various systems biology and proteomic data types. Our research expertise is also enhanced by the participation of distinguished researchers on our advisory boards, such as Catherine Fenselau on PIR OSAB. Catherine is the President of US HUPO and a former President of the American Society for Mass Spectrometry, and will be invited to serve on the UniProt SAB.
As previously stated, UniProt is not able to accommodate all types of new data generated by functional genomics, proteomics and systems biology approaches directly in the UniProtKB. For all these types of data we are actively participating in the creation of standards to allow capture in a standardized manner, allowing the storage of this data in specialized databases. By collaborating closely with these specialized databases we can import in a curator-assisted way the highly reliable data into UniProt.
We have described previously the handling of protein-protein interaction data as an example (see above). Mass spectrometry data is just another example of our concept of focusing on our core activities combined with active liaisons with specialized databases dealing with the emerging new data types in functional genomics, proteomics and systems biology.
We will handle mass spectrometry protein identification data in a similar manner to protein-protein interaction data (see above). The EBI has created a **Pr**otein **Ide**ntification database (PRIDE, http://www.ebi.ac.uk/pride/) for storing such data. Mass spectrometry protein identification data suffers from a high rate of false positives, in a similar way to protein interaction data. Therefore, we will import into UniProtKB only the highly reliable subset of PRIDE while cross-referencing back to PRIDE for the data deemed not suitable for inclusion in UniProtKB. Also, with PRIDE we plan to build an international collaboration that will regularly exchange information. In this way, PRIDE would be the central node in a network of freely available, stable, and synchronized proteomics data resources, allowing UniProtKB to import via PRIDE all of the reliable high-quality data that is available worldwide. It is also worth mentioning that we have received very positive feedback from many journals that we

approached with the suggestion of making such submissions mandatory, as soon as the repositories are able to cope with large numbers of submissions expected. An initial meeting of proteomics repositories took place recently. Grant applications to the European Commission and the Wellcome Trust to support both the development and implementation of proteomics standards and the implementation of these standards in PRIDE are in preparation. The EBI has hired staff with experience in mass spectrometry to develop PRIDE and will bring in additional experience as soon as more (non-NIH) funding for PRIDE is secured.

5. Summary Critique (p.2): ***PIR is not as closely aligned and integrated as the other two groups***.

The EBI and SIB have been working together to produce the Swiss-Prot database since 1987, while the collaboration with PIR is less than three years old. Nevertheless, the UniProt group at PIR has already aligned and integrated closely with the EBI and SIB groups, with complementary approaches guided by common goals and perspectives in maintaining high-quality protein sequence databases. Many aspects of the integration are infrastructural with internal exchange mechanisms that may not be evident to end-users. Some of the joint activities are listed below:

- UniProt liaison group (for discussion of high-level issues)
- UniProt web site committee (for unified UniProt web site strategic planning and development)
- Link committee (for working with boutique databases to provide cross-references)
- UniProt help-mail (for answering user inquiries)
- Document committee (for UniProt documentation, tutorials and FAQs)
- UniProt group for XML document changes and maintenance
- UniProt group for routine web site maintenance issues and statistics (for current set of overlapping but non-identical web sites)
- UniProt group for automatic annotation pipeline development and integration
- Manual curation of Swiss-Prot template sequences
- Manual curation of site rules and controlled vocabularies for site features
- Development of rules to be incorporated into EBI's automatic annotation framework
- Development of common protein naming guidelines
- Incorporation of new protein families into InterPro, especially as leaf nodes
- Collaboration with the biweekly update of UniRef, UniParc and UniProtKB databases
- UniProt-PIR staff routinely visit EBI and SIB for extended discussions and exchanges. For example, during August-November 2005, two curators visited the SIB for one week for site rule discussion, two other curators have scheduled a two-week visit to the SIB for Swiss-Prot sequence annotation, one programmer is scheduled for a week-long visit at EBI for web site implementation, and four members will attend the UniProt Consortium annual meeting at SIB.

Below are responses to further critiques addressed in the summary critique.

6. Summary Critique (p.3): *The reviewers expressed the view that the **activities described in the application have a developer-centric perspective, rather than a user-centric view**... **UniProt is a service project** that should be focused on addressing the user needs.*

The UniProt consortium has a well-balanced scientific staff, consisting of an approximately equal proportion of protein scientists/curators and bioinformatics scientists/programmers. The balanced approach between manual curation and automation allows us to provide user-centric, high-quality scientific content with a developer-based framework that considers computational issues such as interoperability and scalability. As a service project, we have set up help-mail services with very rapid response times, and have interacted closely with the user community by giving workshops (such as tri-annual proteomic workshops) and tutorials (such as ISMB demos and visits to Universities). Here we pay special attention to reaching the wet lab biologists to ensure that we serve the whole range of users, from the occasional user in the lab, to the bioinformatics power user. We will continue to extend our interactions with the user community and meet the users' needs.

7. Summary Critique (p.4): *The applicants appear to have intentionally avoided an in-depth discussion of their interaction with scientific journals. ... **There needs to be a better interaction and better connection between journals and UniProt.***

We appreciate the importance of close interaction with scientific journals. The UniProt PIs serve on various Editorial Boards and Editorial Advisory Boards of scientific journals and are actively involved in constant improvements to database-journal interactions. However, the UniProt consortium realizes that these interactions need further strengthening. Therefore we have planned to host a workshop (coinciding with the ISMB-2006 in Brazil) and will invite publishers of major scientific journals to discuss issues regarding connecting journals and UniProt.

Below are responses to individual critiques that are not already addressed in the summary critique.

8. Critique 1 (p.6): *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**. .. UniProt's position on protein naming seems to be exactly the opposite of what a language is for; to convey meaning. ... What is most important in a name? Is it stability and neutrality, or its use as a descriptor?*

The UniProt team adopts common protein naming guidelines designed to provide meaningful and standardized names with stability. The DE (description) lines are being reformatted to serve both as a stable label and as a descriptor with additional functional information. An externally funded project, the NSF-funded BioThesaurus of protein and gene names, provides an additional resource for UniProt, allowing easy search for synonymous and/or ambiguous names during protein naming.

9. Critique 2 (p.7): *In particular, **I was not enthusiastic about the data mining and automatic annotation** parts of this proposal. .. 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.*

As stated in Q2, research activities at UniProt are funded by separate grants, including one previous and one current NSF-funded text-mining project. The major "automatic annotation" efforts of the UniProt project constitutes automatic propagation of manually curated rules based on protein families. The latter includes HAMAP family rules curated at SIB, PIRSF-based name rules,site rules curated at PIR and EBI Rulebase rules, all of which are "service-oriented" curation activities.

10. Critique 2 (p.8): *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.***

All UniProt data has been available in XML format since the first UniProt release.

11. Critique 2 (p.8): *I feel that **supporting automated analyses** is becoming increasingly important. ...However, there is still information available from the web interface that is not available in these downloadable formats. ... 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. ... There was no discussion of new innovations such as **LSID for naming proteins**.*

The UniProt team is developing APIs for accessing UniProt databases in separately funded grants. In particular, the grid-enabled API and web services developed at PIR as part of the NCI caBIG initiative will be ported to UniProt as an additional access mechanism (i.e., programmatic data access) for UniProt databases. Other innovations, such as LSID for naming proteins, are also being developed in the caBIG initiative and will be adopted by UniProt when they become more broadly accepted standards/technologies.

12. Critique 2 (p.9): *I notice that PIR and EBI are using Oracle, whereas **SIB is using something different.***

The SIB is using Oracle infrastructure at the EBI. Only some SIB specific components are currently maintained in other Database management systems, but even these are in the process of being ported to Oracle, either at the EBI or SIB.

13. Critique 2 (p.9): *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**.*

The UniProt Web site, as originally developed, was actually three distinct sites, largely unified in overall look, feel, and layout, but nonetheless with slight differences owing either to specific tailoring to user expectations or to the use of different software. One deliverable of the renewal grant will be to develop a completely unified UniProt web site so that every

user will be provided with a consistent site layout, entry view, and retrieval format. This will, in addition, allow us to better serve the user community by providing expanded services and customizable interfaces. For example, there is a growing demand for programmatic access to the underlying database engine through web services or related technologies. The needs of different types of users will also be taken into account by the MyUniProt concept, which will provide each user with the ability to customize the site according to individual needs.

14. Critique 4 (pp.12-13): *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….**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.*

This was described in the progress report part of the proposal. The appropriate reference was cited, but unfortunately we omitted it from the references section:
Wieser D., Kretschmann E., Apweiler R.; Filtering Erroneous Protein Annotation. Bioinformatics 20:i342-i347(2004).

14. Critique 4 (pp.12-13): ***This review focuses on Aims D.2-D.4**, proposed work on the computational and software development infrastructure and dissemination. ...High-level evidence of the **lack of coordination between groups** is found in the split, somewhat **redundant nature of the bioinformatics efforts**. 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 ... **PIR is doing additional bioinformatics analysis that is not central to the mission of UniProt.***

As stated above and detailed in the grant proposal, PIR efforts (along with EBI and SIB efforts) are fully discussed among the three UniProt consortium members and are broadly integrated into different aspects of the UniProt projects. The scientific justifications for the complementary (not redundant) approaches employed by PIR, SIB and EBI were outlined in sections outside the description of the computational and software aspects that this reviewer focused on. For example, while InterPro is an umbrella for many curated family databases, PIRSF is an underlying InterPro member database that provides finer granularity for functional annotation based on homeomorphic (rather than domain) families. The PIRSF family-driven, rule-based annotation enabled by this annotation framework feeds directly into the UniProt core activities in that protein classification provides the scientific basis for many individual protein annotations, whether manually curated or automatically propagated. Further, only by direct comparison of sequences can manually-curated propagation rules be derived.

PIR, in addition, supports several UniProt mission-critical activities, including the production of UniRef100/90/50 databases and the development of a unified UniProt website. In addition, infrastructure funded by separate grants – such as NCI-funded caBIG grid-enablement ,web services and NSF-funded iProClass for ID mapping also contribute directly to the public access and utility of the UniProt databases.

15. Critique 4 (p.13): *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?*

PIRSF classification serves as a basis for the curation of name rules and site rules that are used for automatic propagation of protein names and feature sites in TrEMBL entries via EBI automatic annotation pipeline and for propagation of sites in Swiss-Prot entries via logfiles. In addition to protein naming of TrEMBL entries, PIR communicates discrepant protein names with SIB via emails and a web interface, and co-develops UniProt protein naming guidelines with SIB.

16. Critique 4 (pp.12-13): ... ***the records that are transferred continue to be in a flat-file format.*** *...* ***Why not use table dumps or XML*** *to transfer information in a structured way?*

XML is the primary format for data exchange within the UniProt consortium. Flat-file format data files are transferred from EBI to PIR for distribution to users in the FTP site.