CONCEPT CLEARANCE
NHGRI Advisory Council May 2008
A Centralized Protein Sequence and Function Resource

Purpose

The National Human Genome Research Institute (NHGRI) proposes an RFA to continue the support of a centralized informatics protein sequence and function resource. This resource will serve as a repository of curated protein sequences and will provide high quality annotation of the functional information. This resource will allow many types of queries of the data and will coordinate with other resources containing complementary data to facilitate queries across these resources. This resource is necessary for biologists to translate the enormous amount of data from the Human Genome Project, model organism databases and structural and functional genomics projects to understand human disease.

Background

The completion of genome sequences for many organisms, including human, now allows attention to turn to the identification and function of proteins encoded by these genomes. High-throughput proteomic approaches such as protein microarrays, mass spectrophotometry-based methods, co-immunoprecipitation and yeast two-hybrid assays, have enabled scientists to address in new ways questions about how proteins work and the composition of the molecular machines that perform the functions in a cell. While high-throughput approaches can identify protein function for a large number of proteins in a single experiment, the majority of information about protein function is derived from hypothesis-driven experimental work published in the scientific literature. The curation of this high quality data from the literature is a valuable source of information that needs to be centrally located for access by the scientific community.

With the development of next generation sequencing technologies, the number of identified proteins and associated variants will increase dramatically. Projects using these technologies, such as the NIH Roadmap Human Microbiome Project (HMP) or other metagenomic projects, promise to generate sequence from a wide variety of microorganisms both in the human body and in different environmental sites while the human 1000 Genomes Project will identify variation at the level of 1% frequency in the population. Yet the identification of genes from DNA sequence is only part of the problem since the number of different functional proteins exceeds the number of genes due to the generation of protein isoforms and protein modifications. Therefore, information derived from computational methods to identify protein and their predicted domains and functions also needs to be available for the scientific community.

Recent new technology developments have accelerated the need for high quality informatics resources to help biologists and clinical researchers interpret their experimental results. For example, inexpensive genotyping costs have led to an explosion in the number of Genome Wide Association Studies that successfully pinpoint genomic loci associated with a particular phenotype. Similarly, the promise of the $1000 human genome will generate personal genomic profiles that identify specific genetic variation within an individual's genome. Follow-up studies, guided by existing knowledge about proteins encoded in the genome, are needed to identify and understand how genetic variation interacts with environmental factors to determine a phenotype in humans.

NHGRI funds a variety of informatics resources for the scientific community, such as model organism databases, the Gene Ontology Consortium, and Pathway resources. To provide a centralized protein sequence and functional information resource, NHGRI, with co-funding from NIGMS, supports the UniProt Consortium, a collaboration of SIB (Swiss Institute for Bioinformatics), EBI (European Bioinformatics Institute) and Georgetown University. The UniProt Knowledgebase combined Swiss-Prot, TrEMBL,and PIR (Protein Information Resource) activities into a single, centralized resource. UniProt has been funded for the past 7 years and has made substantial progress towards a unified non-redundant protein resource during this time. However, there has been an increasing expansion of current projects and it is clear that UniProt would like to undertake new projects, which although meritorious, would stretch their resources to breaking point. Given the increasing scale of the problem and the challenges of maintaining a high

quality resource, NHGRI staff think that any renewed resource for protein information will need to be responsive to the needs of the community and will have to provide a clear plan on establishing the priorities for the resource.

On July 9 &10, 2008, the NHGRI will hold a workshop on Protein Sequence Resources to engage the scientific community in discussions about current needs and priorities for protein sequence and function information. The outcome of these community discussions will be used to shape the RFA.

Research Scope and Objectives

At the workshop, the following straw plan will be presented for discussion and critique: This Protein Sequence and Function Resource will be intended as a resource for a wide range of scientists with varying computer skills. For the biologist with limited computer skills who is interested in a single gene, the database should provide a rich resource of information concerning the protein products from this gene. The interface must be simple and easy to understand, while the output should be indexed to allow easy access to different levels of information. This output should include Web-based links to other databases to facilitate the rapid exploration of new data. For the biologist with more advanced computer skills who is interested in many genes, the database should provide tools for complex queries and for retrieval of datasets containing information about the corresponding gene products. These datasets should use a standard data format to enable computational analyses of the information.

A centralized Protein Sequence and Function Resource should have these features:

- The resource should be curated, accurate, stable, and comprehensive.
- The resource should include information on protein sequences, nomenclature, alternatively spliced proteins, homology and paralogy relationships, and family classifications. Additional information on gene function should be included, such as the standardized vocabularies of Gene Ontology (GO) terms, potential protein interactions, expression patterns, and pathways. New data types should be incorporated as they arise.
- Should be easily accessible with multiple methods of querying, including simple web interfaces for common standard queries and tools for more complex queries. The resource should be downloadable so that users independently can acquire and process the data.
- Annotation methods should include computational analyses as well as extraction of information from the literature.
- In consultation with an advisory panel, the resource must be responsible for the establishment of priorities for the types and depth of information to be included. This advisory panel should encourage continuous improvements to the database as methods, data, and needs change with time. A strong emphasis on operating in a cost-effective manner should be established.
- The data should include types of evidence and methods for the annotation along with attribution of their source.
- The quality of the data should be clearly indicated, for both experimental and computational data.
- The resource must coordinate with related databases, including agreeing on controlled vocabularies and common data exchange formats. The output should include links to information in related databases.
- The resource should develop scalable methods to speed up the annotation process both manually and computationally and have the ability to incorporate large datasets.

The resource funded by this RFA will be a stable and enable a broad range of scientists to use the large amount of information becoming available on proteins and their functions.

Mechanism of Support

The mechanism of support probably will be the research resource award (U01), cooperative agreement. The total project period for applications submitted in response to this RFA may be up to three years.

Funds Available

NHGRI and other interested institutes intend to commit up to a total of about $5.0 million per year for each of three years, starting in Fiscal Year 2009 or 2010. It is anticipated that one award will be made.