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

Purpose

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

Background

With the completion of genome sequences for many organisms, including human, attention now turns to now turn to the identification and function of the 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 new questions about how proteins work and about the composition of the molecular machines that perform the functions in a cell. While high-throughput approaches to proteomic analysis are being developed, the majority of information about protein function is derived from hypothesis-driven experimental work published in the scientific literature. The curation of 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, including variants, will increase dramatically. Projects using these technologies, such as the Human Microbiome Project (HMP) and other metagenomic projects will generate sequence from a large number of microbial communities from various sites, and the 1000 Genomes Project that will identify variation at the level of 1% frequency in the human population. At the same time, the identification of genes on the basis of DNA sequence data is only part of the challenge. Protein isoforms and protein modifications mean that the number of different functional proteins vastly exceeds the number of genes and gene products. Furthermore, information derived from computational protein identification methods and the analysis of their predicted domains and functions also needs to be made available to the scientific community.

On July 9 &10, 2008, the NHGRI held 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 proposed RFA is based on this and other community discussions, as well as internal discussions at NHGRI and NIGMS.

Research Scope and Objectives

The outcome of the discussions at the workshop and elsewhere have led NHGRI and NIGMS to the following proposal: A Protein Sequence and Function Resource should be established as a resource for a wide range of scientists with varying computer skills. For the biologists with limited computer skills who are interested in single genes or pathways, the database needs to provide a rich resource of information concerning the protein products from all genes. The interface must be simple and easy to understand, while the output should be indexed to allow easy access to different levels of information. The 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, the database should provide tools for complex queries and for retrieval of large datasets . These datasets should use a standard data format to enable computational analyses of the information. The resource should be stable and enable a broad range of scientists to use the large amount of information becoming available on proteins and their functions.

Such a Protein Sequence and Function Resource should :

- be curated, accurate, stable, and comprehensive.
- 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.
- 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.
- involve annotation methods that include computational analyses as well as extraction of information from the literature.
- set priorities, in consultation with an advisory panel, for the types and depth of information to be included. The 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.
- include types of evidence and methods for the annotation along with attribution of their source.
- clearly indicate the quality of the data , for both experimental and computational data.
- coordinate with related databases, including agreeing on controlled vocabularies and common data exchange formats. The output should include links to information in related databases.
- develop scalable methods to speed up the annotation process both manually and computationally and have the ability to incorporate large datasets.

Mechanism of Support

The mechanism of support will be Cooperative Agreement (U01), . 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.