Enzyme Function Initiative
206 W. Gregory Drive
Urbana, IL 61801
W: http://enzymefunction.org | E: efi@enzymefunction.org

To: The Office of Science and Technology Policy

From: The Enzyme Function Initiative, a National Institute of General Medical Sciences Large Scale Collaborative Project (U54GM093342)

Response to the Office of Science and Technology Policy's RFI on Building a 21st Century Bioeconomy

December 2011

The Enzyme Function Initiative (EFI) has prepared this statement in response to the OSTP's RFI on the Obama Administration's National Bioeconomy Blueprint. As a large scale collaborative grant addressing a major challenge in contemporary science, deciphering enzyme function, the EFI believes supporting research and development themes that tap into the tremendous information generated by high-throughput genome sequencing will be critical for significant advances in human health and productivity.

The EFI is a Large Scale Collaborative Program (also referred to as a "Glue Grant") supported by the National Institute of General Medical Sciences (NIGMS; U54GM0933342). This recently retired program focused on promoting integrative and collaborative approaches which are increasingly needed to solve complex problems in biomedical science. The EFI was awarded in May 2010 to develop a robust sequence/structure-based strategy for facilitating discovery of *in vitro* enzymatic and *in vivo* metabolic/physiological functions of unknown enzymes discovered in genome projects, a crucial limitation in genomic biology. This goal is being addressed by integrating bioinformatics, structural biology, and computation with enzymology, genetics, and metabolomics.

It is our understanding that the purpose of the RFI is to solicit input on research and investments that will substantially contribute to the US bioeconomy, with a specific directive to comment on multidisciplinary funding efforts that would revolutionize the prediction of protein function for genes. As a dynamic collaboration of researchers devoted to this very goal, the EFI is uniquely qualified to provide perspective on this issue.

In this statement, we recommend OSTP urges the Administration to:

- Support research to develop more sophisticated protein classification algorithms
- Support research aimed at developing platform technologies that would enable accurate large-scale *in silico* docking of metabolites with both experimentally determined crystal structures and homology modeled structures
- Support more multidisciplinary collaborations between experimental groups and computational groups
- Support a comprehensive public database of functional data
- Develop a program promoting and ensuring the success of multidisciplinary collaborations

Current State of Enzyme Functional Annotation

The "genomic age" saw the sequencing of the human genome along with the genomes of hundreds of other eukaryotic and thousands of prokaryotic organisms. As of June 1, 2011, the TrEMBL database contained 16,014,672 protein sequences, up a staggering 47% from 10,867,798 just one year earlier in June 2010. Despite this explosion in genomic knowledge, many of the protein sequences in the databases have uncertain, unknown, or incorrectly annotated functions. Without correctly annotated functions, the tremendous utility that the newly discovered enzymes and associated metabolic pathways could provide to advance medicine, chemistry, and industry will go unrealized.

With improvements in sequencing technology, the cost for sequencing a 4 Mb genome is now ~$10K and decreasing, with the implication that the number of protein sequences will increase indefinitely. For example, deep sequencing technologies are being applied to populations of closely related organisms, including medically important strains of pathogens. As the focus shifts to leveraging these data to discriminate differences between pathogenic and benign strains, the need for reliable functional assignment has become acute, requiring the development of effective approaches for functional assignment of unknown enzymes and pathways.

Challenges to Large Scale Annotation

Existing strategies for functional assignment of unknown proteins utilize clues from sequence similarity analysis, operon/metabolic context, and phenotypic/transcriptional profiling, as well as other approaches. In the case of enzymes, although these approaches may provide functional clues, e.g., the enzyme is a kinase or an aldolase, they rarely define the identities of the substrates and, therefore, the molecular functions. Furthermore, these strategies are by and large extremely low-throughput and efforts to extend this approaches to large scale analyses are either non-existent or in initial stages of development.

New orthogonal approaches for predicting the substrate specificity of unknown enzymes are needed that provide a general, more direct method for functional discovery. To be effective, new approaches must incorporate high throughput predictive methods to focus and enable the more time-consuming experimental assignment of function. However, the necessity of these approaches has come into focus only within the past few years, and the full scope of the functional prediction challenge is just now crystallizing. In essence, progress towards this goal is in it's infancy. Although efforts such as the EFI and attention such as this RFI is moving the issue of prediction of protein function into the public eye, it is clear that considerable investment from the bioinformatic, computational, and biochemical communities is needed with requisite support from funding agencies.

Recommendations

1. Support research to develop more sophisticated protein classification algorithms

Public sequence databases are currently flooded with misinformation. Developing automated algorithms that more accurately classify proteins (and therefore the likelihood of a given protein's physiological function) is critical. Ultimately, classification algorithms should be utilized retrospectively and prospectively on the sequences deposited in public databases to correct past misannotation and avoid future misannotation. Fulfillment would provide the biomedical community with more meaningful upfront estimates of protein function.

2. Support research aimed at developing platform technologies that would enable accurate large-scale *in silico* docking of metabolites with both experimentally determined crystal structures and homology modeled structures

Generation of quality templates for docking and evaluation of results is both computationally intensive and requires careful, skilled, and therefore low-throughput, human intervention. However, to achieve large scale predictive power, support for automated utilities is needed. Ultimately, such utilities would be open source and provide results scored in a format that would allow facile evaluation (e.g. via a graphical output and probability score akin to the E-value). Fulfillment would provide the biomedical community with the ability to generate and evaluate higher quality predictions of protein function.

3. Support more multidisciplinary collaborations between experimental groups and computational groups

Without experimental testing on a subset of computational predictions, the algorithms are unvalidated and their value is nebulous. The importance of pairing experimental and computational groups to develop, test, and refine computational methodologies cannot be overstated. Productive collaborations, especially across disparate disciplines, takes considerable investment by both the researchers who must establish and maintain the partnership and also by the granting institutions who must support such collaborations despite an increase, albeit modest, in the expenses required to effectively carryout the project (e.g. travel, conferencing options, etc). Fulfillment would provide the biomedical community with significantly increased confidence in predictions of protein function.

4. Support a comprehensive public database of functional data

As the number of sequenced genomes has grown, functional information on a fraction of the resulting proteins has also increased. However, there is no comprehensive repository to make the full spectrum of functional data from all experimental, informatic, and computational disciplines accessible. This barrier creates an environment where research is done in isolation with piece-meal information and little large-scale context. Ultimately, a central functional database, ideally linked and/or incorporated into current sequence databases, would be developed. Fulfillment would provide the biomedical community with comprehensive access to evidence of protein functions.

5. Develop a program promoting and ensuring the success of multidisciplinary biomedical collaborations

Progress on any scientific theme of wide scope and complexity demands integration of many disciplines. The EFI's experience is that collaboration on this scale is in itself an experimental science. A centralized program that provides guidance and support for establishment and management of large-scale multidisciplinary collaborations would dramatically increase the efficiency and productivity of such efforts. Fulfillment would ensure the biomedical community receives the maximum benefit from the funding invested.

Concluding Remarks

While the EFI is developing one strategy to predict protein function in functionally diverse enzyme superfamilies, it is only one among the many multidisciplinary efforts needed to address this issue. However, with support for biomedical research increasingly challenged by budgetary constraints at NIH, large-scale grants have been targeted as too costly to maintain, and very few opportunities now exist to solve problems as complex as prediction of protein function. To truly make use of the enormous potential that sequenced genomes hold, quality functional predictions must be generated, validated at least in part, and disseminated. This requires dedication by the scientific community to focus a significant amount of research effort on functional assignment and also requires commitment from funding agencies to invest the requisite resources to accomplish meaningful goals in this area.