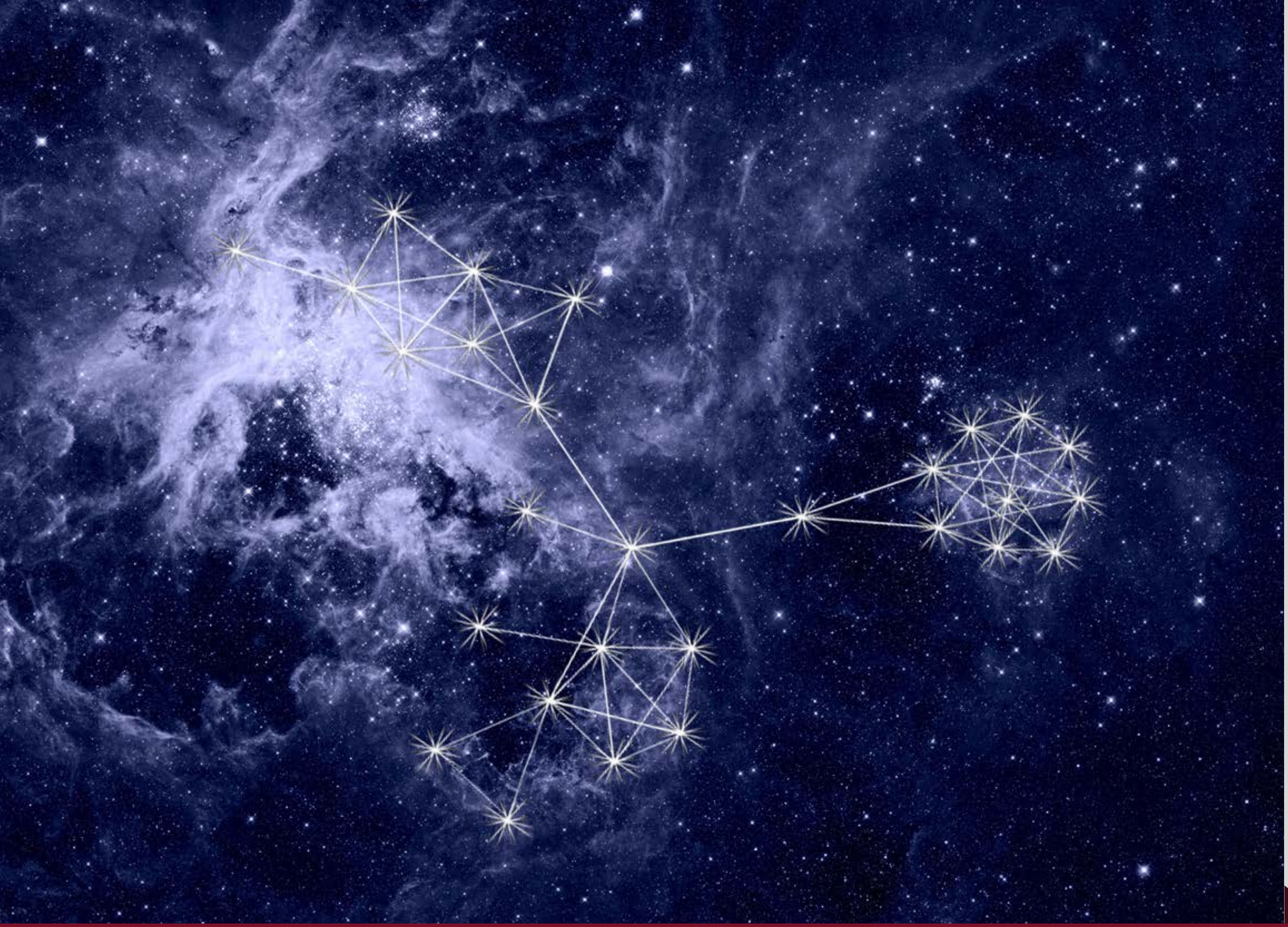


EFIinside

the newsletter for the enzyme function initiative

TABLE OF CONTENTS

EFI ENZYME FUNCTION INITIATIVE

The **Enzyme Function Initiative** (EFI) is developing robust sequence/structure based strategies for facilitating discovery of *in vitro* enzymatic and *in vivo* metabolic/physiological functions of unknown enzymes discovered in genome projects.

DIRECTOR'S NOTE

EFI Director John Gerlt

A Message from the EFI Director, John Gerlt

Looking back at the first EFI Inside newsletter, published in May of 2011, I remarked that the TrEMBL database contained >23,000,000 nonredundant sequences - as of last month, this number is up to >55,000,000. The task set before us continues to grow in magnitude, and yet I remain encouraged by both the scientific discoveries and communal tools that have emerged from the Enzyme Function Initiative. Among the EFI's basic scientific discoveries, I would like to highlight an in-depth study within the Enolase Superfamily, involving the enzymatic characterization of 42 previously unannotated proteins, which has brought into question the historical definition of physiologically relevant catalytic efficiencies (Wichelecki *et al*., p. 11). Additionally, the Microbiology Core has been busy following up the 2PMQ study with metabolomics and transcriptomics that revealed the physiological role of 2PMQ as a salt-dependent switch between catabolism and osmolyte accumulation in *Paracoccus denitrificans* (Kumar *et al*., p. 11).

The past 6 months have been highly transitional as we migrated away from the superfamilies designated in the initial proposal, thus removing constraints and allowing exploration into radical SAM/glycyl radical enzyme-containing anaerobic metabolism and transporter-initiated catabolic pathways (described in more detail on p. 5 and 6). The five superfamilies provided an invaluable training ground with which we honed our computational strategies (Metabolite Docker, Pathway Docking, Covalent Docker, ModBase) and learned how to best utilize sequence similarity networks (SFLD). This expansion into broader territory has been made possible in large part by the maturation of our bioinformatics tools. EFI-Enzyme Similarity Tool, our Web server for generating sequence similarity networks, has allowed researchers, now within and without the EFI, to visualize extremely large data sets of protein sequences and accordingly, tackle increasingly large clusters of unknown enzymes (see p. 4).

Lastly, the EFI is excited to incorporate two pilot projects in Year 5: the Radical SAM Pilot Project, directed by Dr. Squire Booker (Pennsylvania State University) and the Sifting Family (Sfam) Pilot Project, directed by Dr. Katherine Pollard (The Gladstone Institutes). Dr. Booker's contribution is detailed on page 5 of this newsletter. Dr. Pollard will introduce Sfam's ("sifting families", sequence homology-based protein family definitions) into our bioinformatics tools that currently rely on InterPro/Pfam definitions, thus increasing total sequence coverage and enhancing the predictive power of future iterations of the Sequence Similarity Network. The EFI as group has built up substantial momentum over the past four years – I look forward to a highly productive fifth year. ■

NEW METABOLITE DOCKER ROLLED OUT

The 2013 Nobel Prize in Chemistry was awarded to Karplus, Levitt, and Warshel for "Development Of Multiscale Models For Complex Chemical Systems." Acknowledgement of their work to model chemical reactions and systems is a testament to how advanced and pervasive computation has become. The EFI is grounded in the understanding that the only way to make use of the vast amount of sequence data now available is through computational methods. The EFI is fortunate to have an outstanding Computation Core with labs headed by Matthew Jacobson, UCSF Department of Pharmaceutical Chemistry; Andrej Sali, UCSF Department of Bioengineering and Therapeutic Sciences; and Brian Shoichet, University of Toronto Faculty of Pharmacy. Their groups are critical to the EFI because despite the tremendous advances computation has made, modeling and docking are not turnkey techniques. Many parameters have to be taken into account and carefully optimized in order to obtain meaningful results and never is this truer than when attempting to predict enzymatic function.

While it is widely recognized that computation is complex, there is a strong desire by

NEW METABOLITE DOCKER ROLLED OUT CONT.

the wider scientific community to make use of its power. Lead by Associate Professor John Irwin, the Shoichet lab has been at the forefront of bringing computational tools to individual researchers. Most recently as a result of working with the EFI, they have rolled out a new service dedicated to using molecular docking to computationally predict function. The service, called Metabolite Docker, supports docking of both ground state and high energy intermediate forms of metabolites and commercially available compounds to protein structures.

Metabolite Docker is based on it's successful parent program, **DOCK Blaster**. DOCK Blaster is a free virtual screening server created in 2009 to address the growing need for widely accessible and easy-to-use computational tools for docking proteins and small molecules (1). Metabolite Docker takes the framework of DOCK Blaster and modifies it by providing docking sets that have proven very useful for successful prediction of enzyme function (2). Users have access to several tutorials and practice cases to familiarize themselves with the tool. To use the server with any confidence, the user must provide a protein structure, preferably with a ligand bound. This allows for the pocket to be accurately identified and defined. There are six stages to the server: 1) Preparer, 2) Scrutinizer, 3) Target Prep, 4) Calibration, 5) Docking, and 6) Results. Users can begin with Preparer by providing either a PDB code or uploading a protein structure. The Scrutinizer checks for consistency and ensures that the input data are properly formatted before moving forward. If the data passes scrutiny, then the process moves into the Target Prep phase. This step prepares protein “hot spots” and compound libraries for calibration calculations. During the Calibration stage, the Metabolite Docking server checks the pose fidelity in order to determine the optimum parameters to use when screening a chemical dataset. The results are presented in a color-coded table of sampling (coarser/finer) and scoring (polarized/normal) parameters, which indicate the success of the calibration (e.g. green equates to successful, red does not, and yellow is borderline). A Chimera file is also provided for each binding pocket pose generated. Analysis of these preliminary results helps the user determine the best parameters for moving forward. In the Docking step, the chemical set is selected (e.g. ground state or high energy intermediate). Once docking is initiated, a full screen may take hours to days, depending on the ligand and the size of the database docked.

Like any automated tool, users must take time to explore the system so that they can understand inherent shortcomings and critically evaluate the outcomes. Although not all runs attempted will generate results appropriate for follow-up, Metabolite Docker and DOCK Blaster are both meant to provide non-experts with much-needed access to computational methodology. With appropriate circumspection, such services are powerful tools for addressing unknown function. ■

1. Automated docking screens: a feasibility study. Irwin JJ, Shoichet BK, Mysinger MM, Huang N, Colizzi F, Wassam P, Cao Y (2009) J Med Chem 2009, 52, 5712-20.

2. Predicting substrates by docking high-energy intermediates to enzyme structures. Hermann JC, Ghanem E, Li Y, Raushel FM, Irwin JJ, Shoichet BK (2006) J Am Chem Soc 128, 15882-91.

EFI-EST: PAST, PRESENT, AND FUTURE

QuEST
EFI-EST
EFIESTII
PAST
PRESENT
FUTURE

Above: The evolution of the EFI-Enzyme Similarity Tool logo.

The EFI has championed a novel manner of organizing sequence similarity data for large amounts of protein sequences. The canonical branching tree diagrams that have dominated similarity analyses in the bioinformatics field can become unwieldy when dealing with thousands to tens of thousand of protein sequences. Imagine trees that are too large to visualize on a single computer screen with branch widths and leaf nodes too small to be read by the naked eye. The EFI has adopted the Sequence Similarity Network (SSN), originally popularized by the Babbitt Lab of UCSF (1), which presents interconnected sequences as clusters within 2D space. Two sequences (represented by nodes) are connected via a line (referred to as an edge) if their pair-wise BLAST e-value is below the user-designated cutoff. The SSN is dynamic in that once computed at a particular e-value, the end-user can then adjust that e-value to increasingly more stringent values on the fly, thus abolishing the edges that do not satisfy the cutoff, and re-rendering the network to visualize the remodeling of current clusters and the birth of new clusters that bleb from a former super-cluster. When pertaining to functional discovery, the SSN is of utmost utility when filtered by an e-value that achieves isofunctional fractionation – that is, the network is neither under-fractionated (containing multi-functional sequences connected via similarity) nor over-fractionated (isofunctional clusters now separated into distinct but functionally identical sub-clusters). At this level of clustering, the user can confidently transfer function between members within a single cluster. Unfortunately, there is no universal e-value cutoff capable of achieving isofunctional fractionation for all protein families. This cutoff is predominantly sensitive to sequence length and the degree of sequence divergence observed within the examined dataset. Introducing some “known” variables to the network, such as experimentally characterized functions, quickly helps the user to determine what e-value provides optimal fractionation in a case by case manner.

The EFI is convinced that sequence similarity networks (SSNs) are a powerful tool for analyzing relationships among sequences in protein (super)families and that this method of similarity analysis will be useful for enhancing functional discovery/annotation using strategies developed by the EFI as well as developing hypotheses about structure-function relationships in families and superfamilies. As a result, the EFI is providing “open access” to SSNs.

Past

In October 2014, the **Qu**ick **E**nzyme **S**imilarity **T**ool (QuEST) began as the first Web server to deliver Sequence Similarity Network building capabilities to a broad audience of scientists, where no bioinformatics experience was required. QuEST walked end-users through dataset generation (fetched either by a desired protein family designator or via the BLAST of a single protein sequence), dataset analysis and cutoff selection, and finally network generation and download. Later optimization of the Web server’s back-end computation maximized the quantity of results returned from the BLAST step, and revealed that the cost of thorough sampling of sequence space would be computational time. The process was now not as “quick” as previously advertised. Thus, the re-named and vastly improved, EFI-EST became available to the public in January 2014.

Present

Enzyme Function Initiative - **Enzyme Similarity Tool** (EFI-EST) is currently available online for the generation of networks of 25,000 sequences or less. Users interested in generating larger networks are still welcome to request access to our local computing cluster along with the appropriate Linux commands for executing the program. Via the current Web server, SSN generation generally takes 2 to 3 hours.

Future

The next generation of EFI-EST is nearing completion now (May 2014). In a large step forward, SSNs have been pre-computed for the majority of the 515 Pfam clans and 14,831 Pfam families. The new Web server will allow users to browse networks by clan or family, search all clans and families by entering a protein sequence, filter networks by any e-value cut-off, and finally, download full or representative-node (“rep-node”) networks. This process can be completed in a matter of minutes.

The methods behind the foundation of EFI-EST, as well as their application toward computing similarity networks for the entire “protein universe”, are currently being compiled into a manuscript. ■

1. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. Atkinson HJ, Morris JH, Ferring TE, Babbitt PC. (2009) PLoS One 4, e4345.

(On the right) Top two images: The 50-linear-foot anaerobic chamber at AECOM showing areas for cell lysis, harvesting, and crystal plate screening. Images courtesy of James Love. Bottom two images: Within an anaerobic chamber at UIUC, culturing obligate anaerobes.

EFI EXPANDS INTO ANAEROBIC TERRITORY

EFI Moves into Anaerobic Protein Production and Microbial Cell Culture

For the past four years, the EFI has focused on five protein superfamilies (Enolase, Haloacid Dehalogenase, Glutathione-S-Transferase, Isoprenoid Synthase, and Amidohydrolase) with the vast majority of protein targets coming from aerobic bacteria. Starting May 2014, the EFI adds the Radical SAM Pilot Project - an experimental effort led by Dr. Squire Booker (Pennsylvania State University), which brings expertise in radical-dependent enzymology, specifically the Radical SAM superfamily, thus expanding the enzymological know-how of this consortium. This particular protein superfamily shares the following enzymatic reaction: an iron-sulfur cluster cleaving S-Adenosyl Methionine (SAM) and producing a radical intermediate. Radical-dependent enzymology allows the execution of a wide range of unusual chemical transformations (C-H bond functionalization, sulfur insertion, C-O bond formation, etc.), thus providing the EFI with new sequence space that will prove to be rich with novel functions. Importantly, Radical SAM enzymes often take the place of enzymes that rely on the cleavage of molecular oxygen for production of potent oxidants – thus, Radical SAM enzymes play a large role in anaerobic metabolism and are accordingly abundant in the portion of the biosphere that lacks access to oxygen.

While the addition of the Radical SAM Pilot Project provides our Bioinformatics and Computational Cores with the challenging task of evolving prediction algorithms to a new enzyme class, the origins of these enzymes, often anaerobic bacteria, will test the flexibility of our Protein, Structure and Microbiology Cores. In preparation, labs at both the Albert Einstein College of Medicine (AECOM) and the University of Illinois at Urbana-Champaign (UIUC) have acquired new anaerobic chambers. At AECOM, the Protein Core has recapitulated their entire high-throughput prokaryotic protein purification and crystallization infrastructure within an anaerobic chamber (i.e., glove box), so as to bring these capabilities to the study of oxygen sensitive macromolecules. In September 2013, Dr. James Love (AECOM) designed and commissioned a novel resource, which includes approximately 50 linear feet of glove box space. This anaerobic chamber houses all elements required for high throughput purification, crystallization and crystal mounting for subsequent X-ray structure determination, functional and mechanistic analysis. Nearly 900 miles away, the Microbiology Core (UIUC) has purchased a 2-person vinyl anaerobic chamber for installation into a Biosafety Level 2 facility for culturing constitutive and facultative anaerobes. The majority of these microbes are members of the human gut microbial community. Any functional characterizations conducted *in vitro* will be corroborated with *in vivo* growth phenotyping, transcriptomics, and metabolomics.

Selection of Radical SAM enzymes for input into the protein production pipeline is already underway. Sequence Similarity Networks (SSNs) are directing researchers toward divergent clusters of protein sequences where no functions have been experimentally verified. This mechanism of prioritization is anticipated to provide the highest likelihood of discovering novel functions in this protein superfamily. Of the 3700 targets identified from the Radical SAM superfamily, approximately 1000 have been successfully cloned and tested for small-scale expression, with actionable amounts of protein being produced for approximately 35% of clones. Proteins have been scaled up and purified successfully via automated methods, reconstituted to yield homogeneous Fe-S clusters, and protein crystals have been grown under strict oxygen free conditions. Concurrently, the Microbiology Core is bolstering the Protein Core's repertoire by purifying additional genomic DNAs from anaerobic microbes such as *Bacteroides spp.* (Bacteroidetes), *Clostridium spp.* (Firmicutes), and *Collinsella aerofaciens* (Actinobacteria). The addition of these bacteria will not only increase the pipeline's biological diversity, but also increase the human health relevance of pipeline results in the form of human gut microbiota-based functional discovery. All Cores look forward to this exciting expansion into anaerobic territory. ■

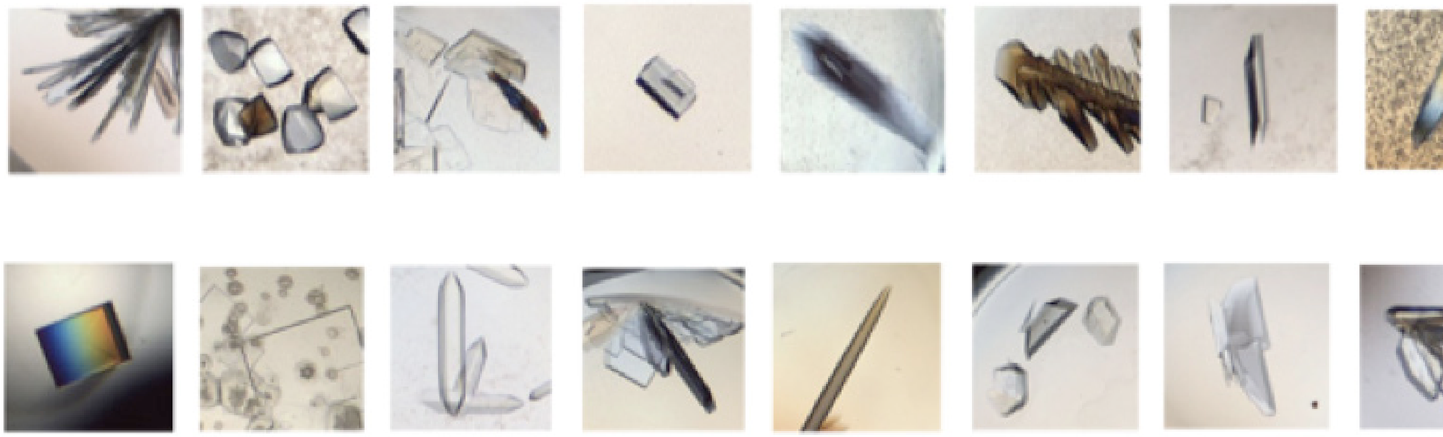
BOOTS*trap*PING FUNCTIONAL DISCOVERY

"Bootstrapping is a database searching technique. One may perform an inexact search (eg. using keywords) and retrieve numerous "hits", some of which will be on-target. When the researcher looks at a relevant document that comes through in the mix, subject headings will be located within the document. The researcher can then execute a new search using authorized subject headings that will yield more focused, pinpointed results."

In Year 4, EFI researchers were inspired to move functional investigations into metabolic pathways up one level to the true beginning of the process: transport of said metabolite into the cell. The adjacent quote regarding "bootstrapping" as defined by the database research community is very appropriate now that the solute binding proteins (SBPs) are being utilized as probes for downstream pathway discovery. SBPs are critical components of three major solute transport systems: The ABC transporters (ATP Binding Cassette), the TRAP transporters (TRipartite ATP-independent Periplasmic transporter), and the TTT transporters (Tripartite tricarboxylate transporter). The SBP components of these transporters have nanomolar to low micromolar affinity for their cognate solute, bringing the solute to an associated transmembrane protein, which passes the solute to the cytoplasm utilizing ATP hydrolysis or an electrochemical gradient. Transport systems are often genomically colocalized or co-regulated with the associated enzymes that utilize that solute, so knowledge of the SBP's cognate ligand assists in functional annotation, and the discovery of downstream enzyme functionalities and metabolic pathways.

SBPs that are cloned and purified by the EFI are first screened by thermofluor against a chemical library that is customized for each SBP family. Nawar Al-Obaidi and Matthew W. Vetting (Protein Core and Structure Core, respectively) collaborated extensively on the initial thermofluor setup, ligand selection, and analysis of results. For example, for the TRAP SBPs, whose ligands are predominantly organic acids, the library (184 compounds) was heavily weighted towards carboxylate containing ligands, including mono and di-acid sugars, uronic acids, amino acids, and benzoic acid derivatives. The approximately 300 TRAP SBPs that were put into the EFI pipeline exhibited high solubility rates in small scale expression studies (>70% success) and over 150 targets were screened by thermofluor. Due to the high affinity of SBPs for their ligand and the large surface area buried during an SBP binding event (a "Venus flytrap" closure of two domains), SBPs can exhibit dramatic ligand thermal stabilizations, sometimes in excess of 20 degrees. For the TRAP SBPs >50% of the purified targets exhibited thermal shifts greater than 5 degrees. Specificity has ranged from high (single binding hit) to low (2-6 binding hits). For those targets with multiple hits, the compounds invariably are highly related such that one can designate which moieties, say on a diacid sugar or benzoic acid derivative, contribute to binding. The TRAP thermofluor work has already expanded the number of known ligands for TRAP transporters from 11 to over 30 ligands.

The thermofluor analysis of SBPs is then complemented by 3-dimensional structural studies of liganded SBP complexes, which besides yielding a treasure trove of data on ligand-SBP interactions, also acts as a "sanity" check on ligands suggested by the thermofluor assay. Liganded structures can be used as docking templates by the

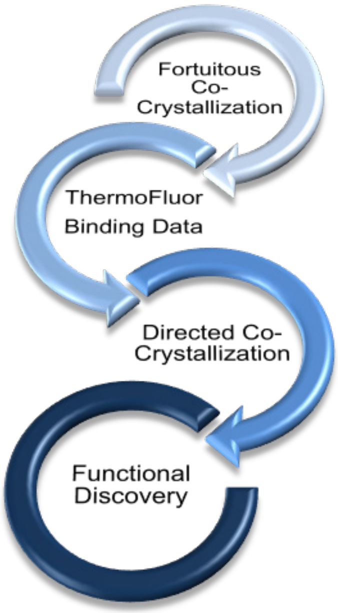


BOOTS*trap*PING FUNCTIONAL DISCOVERY CONT.

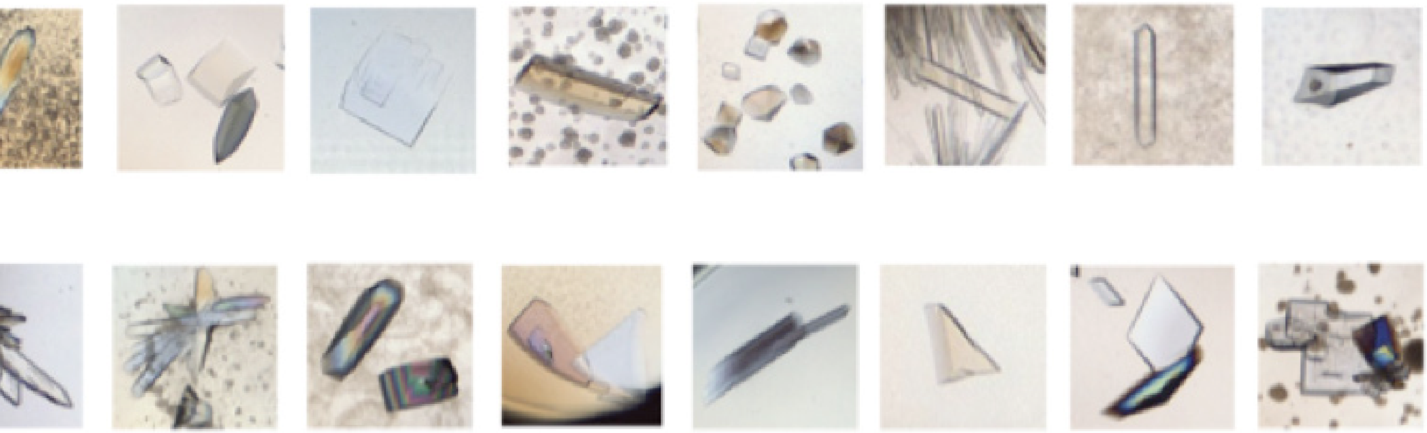
Computation Core to suggest alternative ligands. For example, in one case a rather moderate thermofluor hit on D-Tryptophan (4-5 degrees) was bootstrapped utilizing a D-Trp bound structure, molecular modelling, and additional thermofluor analysis with new ligands to suggest the true ligand was indole-3-acetate (18 degree hit) or indole-3-pyruvate (16 degree hit). Indeed, a later “APO” crystallization experiment was found to contain electron density consistent with indole-3-pyruvate.

Co-purified ligands found in crystal structures have often been described as the “free lunch” when doing high throughput crystallography. This has never been more true than when it came to studying the TRAP SBPs. Those targets with thermofluor stabilizations >5 degrees are co-crystallized with the top hit while those with no hits are set up without ligands (APO). Of those TRAP SBPs setup as APO crystallization experiments, the crystallization rate remained relatively high (53% vs 73% for co-crystallizations), and of the APO structures solved (12 targets), over 2/3 had a ligand that co-purified with the protein. The copurified ligands could be extracted from the protein preparations used for crystallization and positively identified by mass spectrometry. Additional thermofluor analysis with the novel compounds yielded thermofluor stabilizations > 5 degrees. As such the SBPs act as “molecular traps”, extracting whatever ligand is most similar to its native ligand from the E. coli cellular milieu. However, is the copurified ligand relevant to its “natural” ligand? Once again the infrastructure of the EFI can be called upon, where the “validity” of the novel copurified ligand is analyzed utilizing molecular docking, genome neighborhood networks, and regulon analysis. To date, using the SBPs as “molecular traps” have yielded an additional 7 novel TRAP ligands over the regular thermofluor analysis, and as such supplements the difficult process of building an extensive and diverse thermofluor library. As described in the quote at left, an inexact search produces functional “hits”, which reduces the possibilities to a reasonable number for the researcher to begin examining their relevance via more intensive mechanisms.

The TRAP SBP class is currently being followed down the protein production pipeline by an additional 1200 SBP targets from the ATP-Binding Cassette (ABC) transporter family; specifically two subclasses that are known to be involved in the utilization of polysaccharides. A new thermofluor library (410 compounds) has been constructed with a diverse set of mono-, di-, tri- and tetrasaccharides. In addition, in Year 5, EFI researchers will acquire SBP-associated proteins based on genome neighborhood and regulon analysis in order to test binding hits as substrates of the downstream pathway, as well as confirm the physiological role of predicted SBP ligands via growth phenotyping, transcriptomics, and metabolomics of the respective source organism. We expect the Solute Binding Protein class to be an excellent driver of functional discovery for the following year and beyond. ■

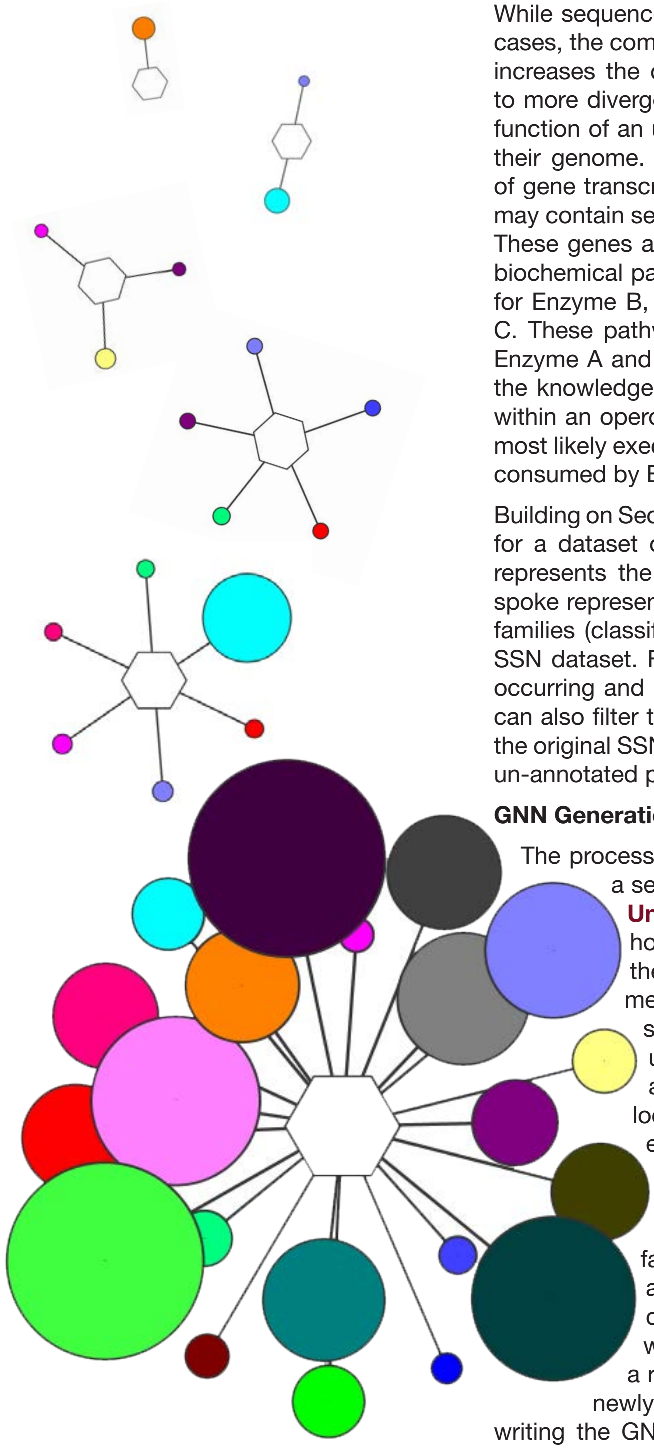


Above: Flow chart of interdependent mechanisms for studying SBPs, ultimately facilitating functional discovery. Below: As of May 2014, a total of 46 SBP targets have structures deposited, with the majority (>80%) having bound ligands derived either through co-crystallizations based on ThermoFluor analysis or ligands that co-purified with the protein preparation utilized for crystallization. Images courtesy of Matthew Vetting.



SNEAK PEAK: EFI-GNT - A WEB SERVER TOOL

Hub-and-spoke clusters of varying occupancy from a Genome Neighborhood Network depict the relationship between Pfam families that are present in the genome neighborhood and a Sequence Similarity Network.



The EFI anticipates making EFI-GNT (Genome Neighborhood Tool) available to the public by mid-Summer 2014. This Web server tool is a companion to EFI-EST (Enzyme Similarity Tool) and operates via input of a Sequence Similarity Network (SSN) and production of a Genome Neighborhood Network (GNN) with its equivalently colored version of the SSN input. The process is extremely efficient, returning up to 20 "neighbor" sequences for each query sequence and rendering the network file in less than a minute for most SSN sizes.

What is a Genome Neighborhood Network?

While sequence homology alone is capable of indicating protein function in many cases, the combination of sequence homology and genome neighborhood analysis increases the confidence of these predictions and expands functional discovery to more divergent proteins. Genome neighborhood analysis can shed light on the function of an unknown protein due to the way bacteria organize the genes within their genome. In order to reduce resources consumed in the turning on and off of gene transcription, bacterial genes are organized into operons. A single operon may contain several genes under the transcriptional regulation of a single promoter. These genes are often related in that their gene products, often enzymes, form a biochemical pathway. For example, the product of Enzyme A is then the substrate for Enzyme B, which produces yet another molecule that is acted on by Enzyme C. These pathways are most often metabolic in nature. Now, if the functions of Enzyme A and Enzyme C are known, but the function of Enzyme B is unknown – the knowledge that Enzyme A and C are close in the genome space, specifically within an operon, gives insight into the possible function of Enzyme B. Enzyme B most likely executes a chemical reaction that bridges the metabolites produced and consumed by Enzyme A and C, respectively.

Building on Sequence Similarity Networks (SSNs), which depict sequence homology for a dataset of protein sequences, the Genome Neighborhood Network (GNN) represents the genome neighborhoods for each query sequence in a hub-and-spoke representation. This network allows a user to quickly identify certain protein families (classified by **Pfam**) that occur within the genome neighborhood of their SSN dataset. From this network, one can quickly distinguish between commonly occurring and rarely occurring protein families in the genome neighborhood. One can also filter this network to examine only the neighbors of specific clusters from the original SSN, in order to quickly assign function to clusters composed entirely of un-annotated protein sequences.

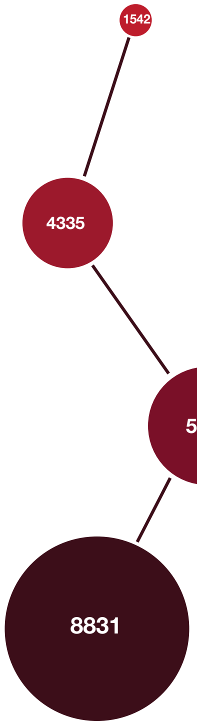
GNN Generation Process

The process of generating a GNN is occurs in three general steps. The input is a sequence similarity network (SSN) of sequences (with their respective **Uniprot** IDs) already fractionated into clusters based on sequence homology at a previously indicated e-value cutoff. In the first step, the network file is parsed for node and edge information, and cluster membership information is associated with each sequence. In the second step, the European Nucleotide Archive (**ENA**) is queried using the Uniprot ID of each sequence. This step is fundamentally a simple, fast, and high-throughput database analysis. Upon location of a query sequence in the ENA, the information from 10 entries proceeding and succeeding the query in the genomic DNA is collected. These 20 entries represent the genome neighborhood. Each "neighbor" entry with a protein-coding sequence (RNAs are discarded) is compared to the Pfam database to determine Pfam family membership. Sequences that do not match to any Pfam family are discarded. Concurrently, neighbor accession IDs are used to query additional databases in order to populate the node attributes with information that is useful to functional discovery. The product is a report that details the relationship between all query sequences and newly retrieved neighborhood information. The third and final step entails writing the GNN network file and coloring the original SSN network. The entire process is extremely fast and computationally inexpensive enough to be carried out on the same machine that hosts the Web server.

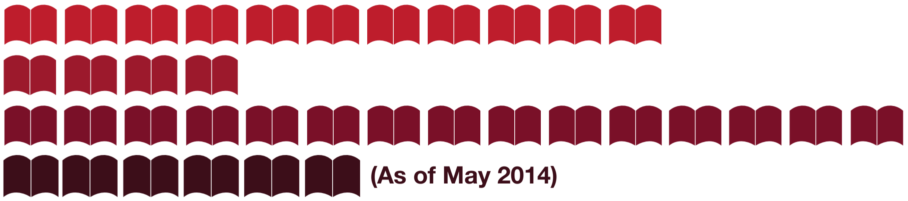
Stay tuned for a beta version of the publically accessible EFI-GNT Web server! ■

AN EFI TIMELINE OF STATS

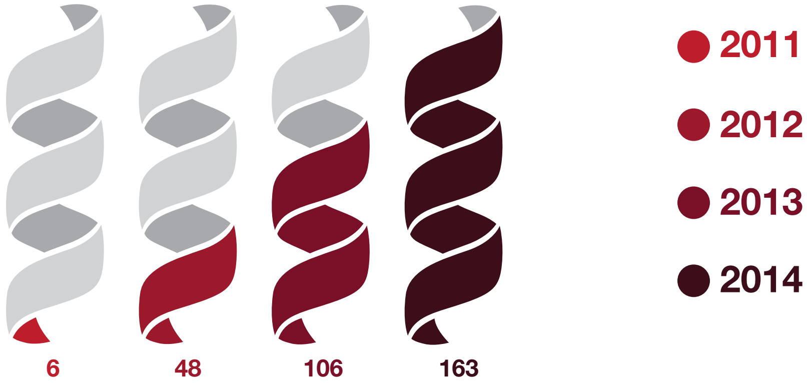
Targets*



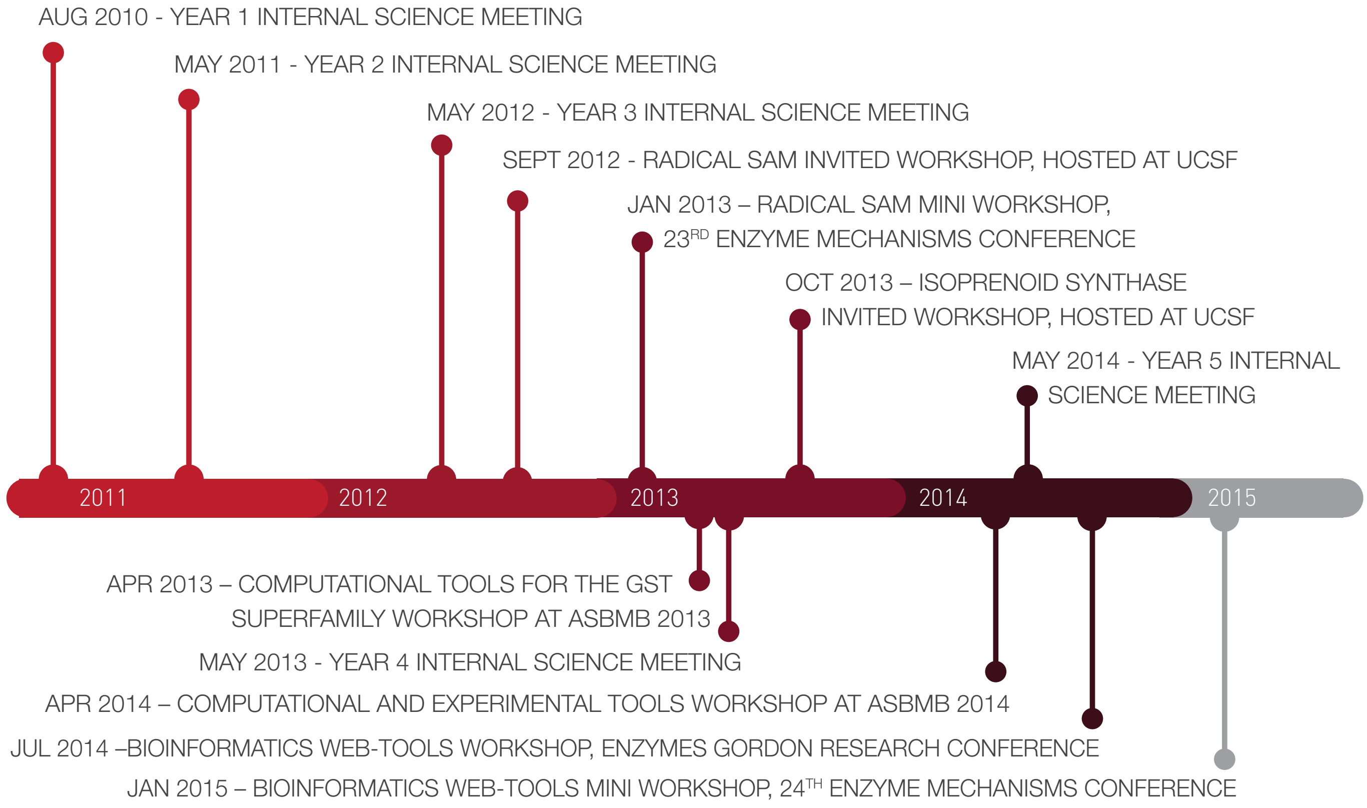
Publications



Structures*



*Target and structure figures were recorded on May of each year. Data available **here.**



AUG 2010 - YEAR 1 INTERNAL SCIENCE MEETING

MAY 2011 - YEAR 2 INTERNAL SCIENCE MEETING

MAY 2012 - YEAR 3 INTERNAL SCIENCE MEETING

SEPT 2012 - RADICAL SAM INVITED WORKSHOP, HOSTED AT UCSF

JAN 2013 – RADICAL SAM MINI WORKSHOP,
23RD ENZYME MECHANISMS CONFERENCE

OCT 2013 – ISOPRENOID SYNTHASE
INVITED WORKSHOP, HOSTED AT UCSF

MAY 2014 - YEAR 5 INTERNAL
SCIENCE MEETING

APR 2013 – COMPUTATIONAL TOOLS FOR THE GST
SUPERFAMILY WORKSHOP AT ASBMB 2013

MAY 2013 - YEAR 4 INTERNAL SCIENCE MEETING

APR 2014 – COMPUTATIONAL AND EXPERIMENTAL TOOLS WORKSHOP AT ASBMB 2014

JUL 2014 –BIOINFORMATICS WEB-TOOLS WORKSHOP, ENZYMES GORDON RESEARCH CONFERENCE

JAN 2015 – BIOINFORMATICS WEB-TOOLS MINI WORKSHOP, 24TH ENZYME MECHANISMS CONFERENCE

EFI PRINCIPAL INVESTIGATORS

Karen Allen
HAD Bridging Project Co-Director, Boston University

Steve Almo
Protein and Structure Cores Director, Albert Einstein College of Medicine

Squire Booker
Radical SAM Pilot Project Director, Pennsylvania State University

John Cronan
Microbiology/Metabolomics Core Director, University of Illinois at Urbana-Champaign

Debra Dunaway-Mariano
HAD Bridging Project Co-Director, University of New Mexico

John Gerlt
EFI Director and EN Bridging Project Director, University of Illinois at Urbana-Champaign

Matt Jacobson
Computation Core Director, University of California, San Francisco

Wladek Minor
Data and Dissemination Core Co-Director, University of Virginia

Katherine Pollard
Sfam Pilot Project Director, The J. David Gladstone Institutes

Dale Poulter
IS Bridging Project Director, University of Utah

Andrej Sali
Computation Core Co-PI, University of California, San Francisco

Brian Shoichet
Computation Core Co-PI, University of California, San Francisco

Jonathan Sweedler
Microbiology/Metabolomics Core Co-PI, University of Illinois Urbana-Champaign ■

PAST AND CURRENT EFI GRADUATE STUDENTS, POSTDOCS, AND STAFF

Eyal Akiva, Ricki Alford, Daniel Almonacid, Nawar Al-Obaidi, Jon Attonito, Joshua Baker-Lepain, Radhika Banu, Sarah Barelier, Alan Barber, David Barkan, Jeff Bonanno, Jason Bouvier, Dan Brown, Shoshana Brown, Sarah Calhoun, Backy Chen, Kyuil Cho, Jeng Yeong Chow, Sukanya Chowhurdy, Jennifer Cummings, Ashley Custer, Marcin Domagalski, Guangqiang Dong, Olga Esakova, Brad Evans, Hao Fan, Jeremiah Farelli, Ryan Foti, Salehe Ghasempur, Swapnil Ghodge, Scott Glenn, Marek Grabowski, Anthony Grizzi, Alissa Goble, Marek Grabowski, Brandan Hillerich, Daniel Hitchcock, Eric Hobbs, Gemma Holliday, Gabriel Horton, Jong Hou, Hua Huang, Heidi Imker, John Irwin, Kevin Jagessar, Amy Jones, Jeff LaFluer, Karol Langner, Florian Lauck, Lili Li, Chunliang Liu, Nir London, James Love, Lingqi Luo, Chakrapani (CK) Kalyanaraman, Siddhesh Kamat, Jungwook Kim, Peter Kolb, Magdalena Korczynska, Ritesh Kumar, Tiit Lukk, Merced Malabanan, Susan Mashiyama, Aldo Massimi, Fiona Mills-Groninger, David Mischel, Taka Miyairi, Matt O'Meara, Sunil Ojha, Jian-Jung (JJ) Pan, Chetanya Pandya, Yury Patskovsky, Ursula Pieper, Przemek Porebski, Udupi Ramagopal, Gurusankar (Guru) Ramamoorthy, Swarnamali Indumathie (Indu) Rupassara, Boris Sadkhin, Ayano Sakai, Brian San Francisco, Avner Schlessinger, Alexandra Schnoes, Ron Seidel, David Slater, Carla Smith, Hilary Smith, Jose Solbiati, Mark Stead, Teague Sterling, Doug Stryke, Lin Sun, Boxue Tian, Sarah Toews-Keating, Rafael Toro, Matt Vetting, Megan Wadington (Branch), Frank Wallrapp, Min Wang, Patrick Weinkam, Katie Whalen, Dan Wichelecki, Kamila Wojciechowska, McKay Wood, Jeff Yunes, Dao Feng Xiang, Wendy Zencheck, Xianshuai Zhang, Zhi Zhang, Suwen Zhao, Li Zheng, Matt Zimmerman, Lucas Zimney ■

RECENT EFI PUBLICATIONS

The structure-function linkage database, Akiva E, Brown S, Almonacid DE, Barber AE 2nd, Custer AF, Hicks MA, Huang CC, Lauck F, Mashiyama ST, Meng EC, Mischel D, Morris JH, Ojha S, Schnoes AM, Stryke D, Yunes JM, Ferrin TE, Holliday GL, Babbitt PC. (2014) **Nucleic Acids Res** 42, D521-30. PMID: 24271399 SUPERFAMILY/ GENOME CORE

ModBase, a database of annotated comparative protein structure models and associated resources, Pieper U, Webb BM, Dong GQ, Schneidman-Duhovny D, Fan H, Kim SJ, Khuri N, Spill YG, Weinkam P, Hammel M, Tainer JA, Nilges M, Sali A. (2014) **Nucleic Acids Res** 42, D336-46. PMID: 24271400 COMPUTATION CORE

Prediction and biochemical demonstration of a catabolic pathway for the osmoprotectant proline betaine, Kumar R, Zhao S, Vetting MW, Wood BM, Sakai A, Cho K, Solbiati J, Almo SC, Sweedler JV, Jacobson MP, Gerlt JA, Cronan JE. (2014) **MBio** 5, e00933-13. PMID: 24520058 MICROBIOLOGY CORE - COMPUTATION CORE - PROTEIN CORE - STRUCTURE CORE - EN BRIDGING PROJECT

Protein production from the structural genomics perspective: achievements and future needs, Almo SC, Garforth SJ, Hillerich BS, Love JD, Seidel RD, Burley SK. (2013) **Curr Opin Struct Biol** 23, 335-44. PMID: 23642905 PROTEIN CORE

Galactaro δ-lactone isomerase: lactone isomerization by a member of the amidohydrolase superfamily, Bouvier JT, Groninger-Poe FP, Vetting M, Almo SC, Gerlt JA. (2014) **Biochemistry** 53, 614-6. PMID: 24450804 EN BRIDGING PROJECT - STRUCTURE CORE

Prospecting for unannotated enzymes: discovery of a 3',5'-nucleotide bisphosphate phosphatase within the amidohydrolase superfamily, Cummings JA, Vetting M, Ghodge SV, Xu C, Hillerich B, Seidel RD, Almo SC, Raushel FM. (2014) **Biochemistry** 53, 591-600. PMID: 24401123 AH BRIDGING PROJECT - PROTEIN CORE - STRUCTURE CORE

Data management in the modern structural biology and biomedical research environment, Zimmerman MD, Grabowski M, Domagalski MJ, Maclean EM, Chruszcz M, Minor W. (2014) **Methods Mol Biol** 1140, 1-25. PMID: 24590705 DATA AND DISSEMINATION CORE

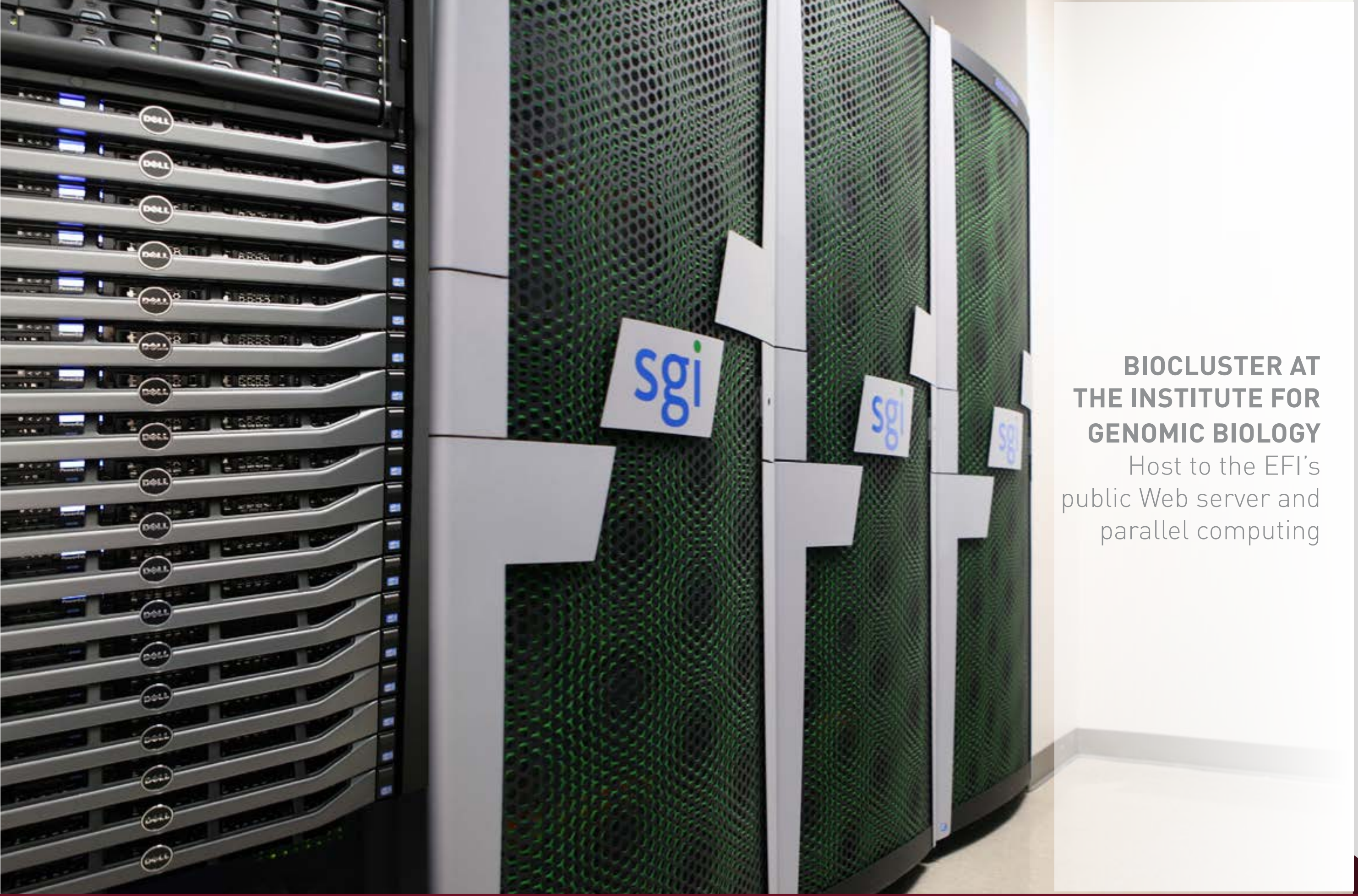
Discovery of function in the enolase superfamily: D-mannonate and D-gluconate dehydratases in the D-mannonate dehydratase subgroup, Wichelecki DJ, Balthazor BM, Chau AC, Vetting MW, Fedorov AA, Fedorov EV, Lukk T, Patskovsky Y, Stead MB, Hillerich BS, Seidel RD, Almo SC, and Gerlt JA. (2014) **Biochemistry** 53, 2722-31. PMID: 24697546 EN BRIDGING PROJECT - STRUCTURE CORE

Enzymatic and structural characterization of rTSγ provides insights into the function of rTSβ, Wichelecki DJ, Froese DS, Kopec J, Muniz JRC, Yue WW, Gerlt JA. (2014) **Biochemistry** 53, 2732-8. PMID: 24697329 EN BRIDGING PROJECT

Avoidable errors in deposited macromolecular structures: an impediment to efficient data mining, Dauter Z, Wlodawer A, Minor W, Jaskolski M, Rupp B (2014) **IUCrJ** 1, 179-193. PubMed in Process DATA AND DISSEMINATION CORE ■

NEXT ISSUE OF EFI INSIDE

Look for a comprehensive analysis of the successes and setbacks of the first 5 years of the Enzyme Function Initiative Large-Scale Collaborative Project.



BIOCLUSTER AT THE INSTITUTE FOR GENOMIC BIOLOGY
Host to the EFI's public Web server and parallel computing

EFIinside

visit us online at: http://enzymefunction.org/

The *EFIinside* provides information on the EFI's organization and progress towards realizing large-scale functional assignment strategies. The EFI is funded by the National Institute of General Medical Science of the National Institutes of Health (U54GM093342).

Questions or comments may be directed to efi@enzymefunction.org.

EFI Mailing Address:
Room 3107, MC 195
Institute for Genomic Biology
University of Illinois
1206 W. Gregory Dr.
Urbana, IL 61801

Editor:
Katie Whalen, Heidi Imker

Design:
Nicholas Vasi, Ricki Alford

Photography:
Ricki Alford

Contributers:
Matthew Vetting, James Love

On the cover: Mapping the protein universe via homology and function - superposition of a cluster from the Sequence Similarity Network of the Sugar/Inositol Transporter family (IPR003663) on a photograph of the Tarantula Nebula taken by the Hubble Space Telescope (April 17, 2012, NASA).

Workshop on Computational and Experimental Tools at ASBMB

A workshop focused on the EFI's computational and experimental tools was held at the Annual ASBMB Meeting April 29th, 2014. Approximately 100 researchers attended this workshop which featured an introduction to the sequence similarity network building tool, EFI-EST (John Gerlt - UIUC), an overview of docking tools pioneered by the EFI (Brian Shoichet, UCSF), and an in-depth description of the EFI's protein production and crystallization capacities (Steven Almo, AECOM). Slides are available on the **EFI Workshops page**.

Workshop on Sequence Similarity and Genome Neighborhood Web Tools at the GRC Enzymes Conference

A workshop is planned for July 14th at the 2014 Gordon Research Conference on Enzymes, Coenzymes, and Metabolic Pathways. EFI director, John Gerlt, and assistant director, Katie Whalen, will present background information, example use cases, and a step-by-step tutorial for use of the EFI's Web tools: EFI-EST and EFI-GNT. Participants are encouraged to bring a laptop in order to follow along and receive one-on-one assistance generating and interpreting networks for the protein family of their choice. More information **here**.

BOSTON UNIVERSITY
EINSTEIN Albert Einstein College of Medicine of Yeshiva University
GLADSTONE INSTITUTES
PENNSTATE
UCSF University of California San Francisco
ILLINOIS
UNM
THE UNIVERSITY OF UTAH
UNIVERSITY of VIRGINIA
NIH National Institute of General Medical Sciences
Basic Discoveries for Better Health