



LAWRENCE  
LIVERMORE  
NATIONAL  
LABORATORY

# Practical Applications of Structural Genomics Technologies for Mutagen Research

A. Zemla, B. Segelke

August 12, 2011

39th Annual Meeting of the Environmental Mutagen Society  
Rio Grande, PR, United States  
October 18, 2008 through October 22, 2008

## **Disclaimer**

---

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

**Title: Practical Applications of Structural Genomics Technologies for Mutagen Research**

Adam Zemla<sup>a</sup> and Brent W. Segelke<sup>b\*</sup>

<sup>a</sup>Computation directorate, Lawrence Livermore National Laboratory, P.O. Box 808, Livermore, CA 94551-0808, USA, [zemla1@llnl.gov](mailto:zemla1@llnl.gov)

<sup>b</sup>Physical and Life Sciences directorate, Biosciences and Biotechnology Division, Lawrence Livermore National Laboratory, P.O. Box 808, Livermore, CA 94551-0808, USA, [segelke1@llnl.gov](mailto:segelke1@llnl.gov), 925 424 4752 (phone), 925 422 2282 (FAX)

\*Corresponding author

## **Abstract**

Here we present a perspective on a range of practical uses of structural genomics for mutagen research. Structural genomics is an overloaded term and requires some definition to bound the discussion; we give a brief description of public and private structural genomics endeavors, along with some of their objectives, their activities, their capabilities, and their limitations. We discuss how structural genomics might impact mutagen research in three different scenarios: at a structural genomics center, at a lab with modest resources that also conducts structural biology research, and at a lab that conducting mutagen research absent experimental structural biology. Applications span functional annotation of single genes, to constructing gene networks and pathways, to an integrated systems biology approach. Structural genomics centers can take advantage of systems biology models to target high value targets for structure determination and in turn extend systems models to better understand systems biology diseases or phenomenon. Individual investigator run structural biology laboratories can collaborate with structural genomics centers, but can also take advantage of technical advances and tools developed by structural genomics centers and can employ a structural genomics approach to advancing biological understanding. Individual investigator run non-structural biology laboratories can also collaborate with structural genomics center, possibly influencing targeting decisions, but can also use structure based annotation tools enabled by the growing coverage of protein fold space provided by structural genomics. Better functional annotation can inform pathway and systems biology models.

## **Introduction**

There are a variety of public and private organization engaged in structural genomics with a variety of objectives and a wide range of activities described as structural genomics (appendix A). Structural genomics has been variously described as the structural analog to the human genome project; the experiment based solution to the protein folding problem; or simply high throughput, big science, structural biology. Structural genomics efforts have in common that they start with genomics, or sequence information, in contrast to traditional structural biology efforts that start with extensive biological data (figure 1)[1]. Following the advent of structural genomics a handful of startup companies received startup funds to pursue a variety of business plans[2]. Today, commercial structural genomics efforts are largely focused on drug discovery by structure aided drug design. A number of structural genomics companies have been acquired by pharma. Public structural genomics efforts still cover a wide range of objectives and activities.

The success of the human genome project, the ready accessibility of sequence data and several other advances have made industrial scale structural biology seem feasible. The human genome project demonstrated the success of vertically integrated, multi-disciplinary centers that concurrently pursue technological challenges and scientific objectives. Genome centers also provide the sequence data and gene calls that are the starting point for structural genomics. Several other important advancements that enable structural genomics include recombinant protein expression[3], Nickel affinity chromatography[4], cryo-crystallography[5], selenium-methionine multiwavelength anomalous diffraction (SeMet-MAD) phasing[6], ready accessibility of tunable x-ray sources[7], and fast x-ray detectors[8]. Recombinant protein expression makes it possible to isolate proteins that are not naturally abundant. Cryogenic methods make it possible to collect complete data sets on high intensity x-ray sources from single crystals. SeMet-MAD phasing substantially reduces the uncertainty of obtaining phasing. The accessibility of tunable x-ray sources enables MAD phasing and, along with fast x-ray sources, greatly reduces the time required to collect x-ray diffraction data. With these capabilities in hand or foreseeable on the near horizon, the stage was set, over a decade ago, for the advent of structural genomics.

One of the largest public structural genomics efforts is the National Institutes of Health (NIH) Protein Structure Initiative (PSI), which is approaching the end of the second phase of funding. The first phase of funding established a number of competing vertically integrated, multidisciplinary centers that had as their main objectives accelerating the rate of structure determination and greatly reducing the cost of determining crystal structures, largely by realizing the economy of scale and

industrialization[9]. Phase I PSI centers focused largely on developing structural genomics pipelines (figure 2). Phase II PSI centers continue to develop enabling technologies to increase the throughput and reduce the cost of structure determination, but the large-scale centers now have as their main objective to determine representative structures from each of the large protein sequence families[10,11].

Structural genomics centers and other researchers involved in structural genomics research have made significant advancements that enable industrial scale structural biology, but that also contribute to investigator driven structural biology and other types of biology research. Structural genomics research has made major contribution to number of structure and the coverage of fold space in the protein structure database (PDB) [6]. Structural genomics research has also lead to methods, commercial instrumentation, labware, and reagents for high throughput cloning, expression, and crystallization of proteins[12-16]. New software tools for structure prediction [17, 19], design of experiments, laboratory information management, novel structure based informatics[20-22], such as function prediction from structure, have also been developed [23-25].

Notable examples of methods innovation that enable structural genomics but also impact laboratory scale research are: gateway cloning[13], autoinduction[26], nano-scale crystallization[15, 27], and automated structure determination[28-32]. Gateway cloning facilitates the rapid construction of a variety of expression plasmids from a single entry clone and provides a basis for establishing a more useful clone collection [13]. Autoinduction facilitates parallel expression screening [26]. Since autoinduction obviates the need for monitoring culture density, any number of expression experiments can be run concurrently to completion without the need for investigator intervention. Nanoscale crystallization increases the number of experiments that can be performed with a given amount of often very precious protein stock material, thereby increasing the likelihood of finding a successful crystallization condition [15, 27]. Finally, advances in crystallographic computing, including automated phasing from MAD data and now automated model building, have made crystallography much more accessible[28-32].

As with genomics, the most profound impact structural genomics is likely to have over time is the expansion of the knowledgebase and the new informatics approaches the more comprehensive knowledgebase will inspire. Obvious examples include improved homology based structure prediction[17] as a growing percentage of sequence space has representative structural templates [33]. We will also better understand the size of fold space, its size and complexity and more ancient evolutionary linkages between protein families, as currently unpredictable structural relationships between protein families are revealed [20, 34]. We will have more rigorous fold classification systems and structure based protein ontologies [34, 35]. And perhaps most useful to the greater

biosciences community, we can anticipate new structure aided protein function prediction algorithms [23-25].

We would be remiss in describing structural genomics if we discussed the objectives and advancements and did not discuss current limitations of structural genomics and future challenges. The major challenge ahead for structural genomics is not only to increase structure determination throughput but to increase the structure determination through rate. With current state-of-the art methods, the majority of proteins remain refractory. Even for globular non-membrane proteins, only a small percentage yield their structure (table 1). This threatens the long term objective of the PSI and other structural genomics efforts to obtain representative structures from all large sequence families. New innovative approaches are still needed to increase the target through rate and to realize the ultimate promise of structural genomics.

## **Discussion**

Though structural genomics has its limitations, it could have a significant impact on mutagen research. We discuss three scenarios whereby structural genomics, or methods enabled by structural genomics, could be used in mutagen research. We describe the potential for mutagen research at a structural genomics center, at a lab with modest resources that also conducts structural biology research, and at a lab that conducting mutagen research absent experimental structural biology.

### *The potential for mutagen research at a structural genomics center*

The potential for mutagen research carried out at a structural genomics center is perhaps best exemplified by work recently showcased by the PSI that was carried out at the Northeast Structural Genomics Consortium (NESG). The Northeast Structural Genomics Consortium is one of four PSI II funded large-scale structural genomics centers, as such NESG has considerable resources and can pursue very ambitious projects. According to their website, “the NESG focuses on eukaryotic proteins, particularly Human proteins involved in cancer biology, protein-protein interaction networks, specific biochemical pathways, and proteins implicated in other Human diseases.”

As part of their cancer biology initiative, the NESG took up a structural genomics approach to studying cancer. Cancer has been previously characterized as a “systems biology disease” [36] and NESG investigators assert that structural genomics provides an important approach for understanding systems biology[37]. NESG investigators took advantage of the existing Human Protein Reference Database [38], protein-protein interaction databases [39] and extensive cancer biology literature to build their Human Cancer Pathway Protein Interaction Network (HCPIN) database[37]. The HCPIN includes ~3000 proteins and their interactions [37]. Proteins in HCPIN that did not have

good existing structure templates in the PDB and therefore could not be modeled well, >1000 proteins, were targeted for structure determination.

One of the first proteins from the human cancer pathway protein interaction network to have its structure solved was RBBP9 [40]. RBBP9 was predicted to interact with the retinoblastoma protein, which is known to be important in cell cycle regulation, cell differentiation, and apoptosis [40]. Retinoblastoma protein was the first oncogene to be discovered and retinoblastoma protein mutations are often implicated in cancer. Retinoblastoma protein mutations are thought to circumvent normal cell cycle controls and allow for uncontrolled cell growth. The RBBP9 protein is thought to function by preventing the interaction between retinoblastoma protein and E2F1.

As expected, the structure of RBBP9 revealed the retinoblastoma protein binding motif LxCxE, but unexpectedly, the structure also revealed a serine-histidine-aspartate cluster of residues consistent with a serine protease catalytic triad (figure 3). Investigators have now confirmed RBBP9 protease activity though no catalytic activity was previously known for this protein [41]. Now it is known that in addition to interfering with retinoblastoma protein-E2F1 interactions, RBBP9 functions as a serine protease, acting on a yet-unknown target in the TGF-beta1 signaling pathway[41].

Though the above example is not a direct example of mutagen research, cancer research is closely related to mutagen research and a directly analogous approach could be applied to mutagen research. Many of the signaling pathways implicated in cancer will also be implicated in cellular responses to environmental mutagens as mutagenesis and cancer are often linked as cause and effect. Furthermore, a large-scale effort similar to the one described above could be undertaken by combining the Human Protein Reference Database with array data or other omics data available for specific mutagens of interest to develop a mutagen specific pathway model. This new pathway model could be used to target proteins of particular interest either because of their central role in biological processes or because of the lack of annotation, or both. As demonstrated in the above example, structural genomics can be used to functionally annotate proteins or improve the functional annotation of multifunction proteins.

#### *The potential for mutagen research in a structural biology lab*

Individual investigator labs typically have only a small fraction of the resources and skill sets that structural genomics centers have but individual investigator driven projects can employ some structural genomics approaches at the laboratory scale [42]. Our lab is interested in host-pathogen interactions in bacterial infections. While host-pathogen systems are probably not good surrogates for cellular responses to environmental mutagens the problems are similarly complex—no one gene or regulatory network is responsible for all of the observable phenomena.

Starting with gene expression data, we selected ~200 proteins for characterization, including structure determination. We specifically targeted proteins encoded by genes implicated in virulence for which there was little or no functional annotation. Expression array experiments with *Y. pestis*, the etiological agent of bubonic plague, identified a large number of genes that are upregulated in host-mimetic conditions [43]. More than 30% of the implicated genes were annotated as putative or conserved hypothetical, here we will discuss our efforts to characterize one protein, whose gene was annotated as hypothetical conserved protein.

The gene with locus tag YPO0407 is upregulated in host-mimetic conditions [43] and was annotated as a hypothetical conserved protein. Starting with only the gene expression data and the annotated sequence, we targeted the protein encoded by YPO0407 for expression and structure determination. Concurrent with the laboratory scale structural genomics effort, we used traditional informatics tools in an effort to better understand the function of this protein. PSI-blast identified a large family of homologous proteins (figure 4), most of which were also annotated as hypothetical conserved protein, but a few were annotated as antibiotic biosynthesis monooxygenase and one was annotated as LsrG. The high conservation of two residues (E32 and H65) across a diverse family of homologues implicated these residues as functionally important. From PSI-blast alone, we had little additional information about the possible function of YPO0407 than that provided by the genome annotation.

Ultimately, we were able to determine the crystal structure of YPO0407 (pdb 2gff) and we were able to apply structure based informatics approaches to place the protein in a fold superfamily, to partition the superfamily into at least 2 functionally distinct subfamilies, and to hypothesize a catalytic function for the YPO0407 gene product. With the YPO0407 structure in hand STRALCP [20] was used to cluster the protein with its closest structural homologues (figure 5). The clustering results revealed two distinct subfamilies. Some proteins in both subfamilies were annotated as antibiotic biosynthesis monooxygenases, though residues known to be involved in monooxygenase activity from the most well studied protein [44] are conserved in one subfamily and completely absent in the other subfamily. This suggests that the fold is promiscuous (more than one activity assigned to the fold family) and that the annotation for some of the proteins may have been mistakenly borrowed from the ActVA-Orf6 monooxygenase. The function for the nearest homolog to YPO0407 (*E. coli* LsrG) is now known from *in vitro* biochemistry experiments [45] and it is not a monooxygenase. LsrG breaks down phosphorylated auto-inducer 2 molecules, bacterial secreted proteins involved in cell-cell communication. The hypothesized hydrogenase activity is consistent with the known substrate and product identified for LsrG [45].

The above example again is not directly related to mutagen research but again provides an example where, starting with sequence information and minimal biological data, structural genomics and structure based informatics can be used to annotate function. In this case, structure information in combination with subsequently obtained biochemical data could be used to annotate the detailed molecular mechanism and possibly correct the annotation of a large subfamily of proteins.

#### *The potential for mutagen research in a structural biology lab*

There are at least two ways that individual investigator labs can utilize structural genomics in mutagen research, even absent structural biology capabilities within the lab. Non-structural biology labs could use structural genomics methods in mutagen research through the use of structure based informatics tools and by collaboration with structural genomics labs.

Revisiting the YPO0407 example from above, concurrent with the effort to determine the *de novo* crystal structure of homology models of YPO0407 were constructed. Since the experimentally determined x-ray structure was available the homology models were not extensively characterized. *Post facto*, the homology models did provide sufficient detail to accurately assign YPO0407 to a superfamily and to cluster YPO0407 with its subfamily. The putative catalytic triad was also accurately modeled, so the putative catalytic mechanism could also have been inferred. This suggests that even absent a *de novo* structure for proteins of interest, structure based informatics could substantially extend functional annotation of proteins and could be used to check or correct existing annotation.

There are many opportunities for non-structural biology labs to collaborate with structural genomics labs and in particular structural genomics facilities. In fact, as part of the second phase of the PSI the National Institute of General Medical Sciences encourages Biologists to participate in the PSI (<http://www.nigms.nih.gov/News/Results/PSIrelease02122009.htm>) and has mandated that PSI centers have outreach to the greater biosciences community. There is a formal mechanism for nominating targets for structure determination. Information about this process and application forms can be found at <http://cnt.psi-structuralgenomics.org/CNT/targetlogin.jsp>

### **Conclusion**

Structural genomics proceeds from sequence information and limited other biological information to structure determination. There is a large number of vertically integrated, multidisciplinary structural genomics centers that are industrializing structure determination pipelines. Structural genomics centers have been successful innovating approaches to increase structure determination throughput and lower the cost per structure but the long range objectives and ultimate promise of structural genomics are

threatened by limited through rates—approximately 3% of cloned targets are currently yielding structures from PSI structural genomics efforts. Despite the current limitations, structural genomics can be a powerful tool for functional annotation, annotation correction or extension of function annotation to detailed molecular mechanisms. In combination with metabolomics, gene expression, proteomics, structural genomics can also make important contribution to systems biology models even at the laboratory scale.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

## Appendix A

### PSI Centers (<http://www.nigms.nih.gov/Initiatives/PSI/Centers/>)

#### Large-Scale Centers

*Joint Center for Structural Genomics*

<http://www.jcsg.org/>

*Midwest Center for Structural Genomics*

<http://www.mcsg.anl.gov/>

*New York SGX Research Center for Structural Genomics*

<http://www.nysgxrc.org/>

*Northeast Structural Genomics Consortium*

<http://www.nesg.org/>

#### Specialized Centers

*Accelerated Technologies Center for Gene to 3D Structure*

<http://atcg3d.org/default.aspx>

*Center for Eukaryotic Structural Genomics*

<http://www.uwstructuralgenomics.org/>

*Center for High-Throughput Structural Biology*

<http://www.chtsb.org/>

*Center for Structures of Membrane Proteins*

<http://csmp.ucsf.edu/>

*Integrated Center for Structure and Function Innovation*

<http://techcenter.mbi.ucla.edu/>

*New York Consortium on Membrane Protein Structure*

<http://www.nycomps.org/>

#### Homology Modeling Centers

*Joint Center for Molecular Modeling*

<http://jcmm.burnham.org/>

*New Methods for High-Resolution Comparative Modeling*

<http://dunbrack.fccc.edu/nmhrcm/>

#### Non-PSI centers

##### US Centers

*Center for Structural Genomics of Infectious Diseases*

<http://www.csgid.org/csgid/cake/>

*Structural Genomics of Pathogenic Protazoa*

<http://www.sgpp.org/>

*TB Structural Genomics Consortium*

<http://www.doe-mbi.ucla.edu/TB/>

*Seattle Structural Genomics Center for Infectious Disease*

<http://www.ssgcid.org/>

## **International centers**

*Structural Genomics Consortium*

<http://www.thesgc.org/>

*Riken Structural Genomics/Proteomics Initiative*

<http://www.rsgi.riken.go.jp/>

*Protein Structure Factory*

<http://www.proteinstrukturfabrik.de/>

*Israel Structural Proteomics Center*

<http://www.weizmann.ac.il/ISPC/>

## References

- [1] A. Yee, K. Pardee, D. Christendat, A. Savchenko, A.M. Edwards, C.H. Arrowsmith, Structural Proteomics: Toward High-Throughput Structural Biology as a Tool in Functional Genomics, *Acc. Chem. Res.* 36 (2003) 183-189.
- [2] D. Gershon, Structural genomics — from cottage industry to industrial revolution, *Nature* 408 (2000) 273-274.
- [3] D.A. Jackson, R.H. Symons, P. Berg, Biochemical Method for Inserting New Genetic Information into DNA of Simian Virus 40: Circular SV40 DNA Molecules Containing Lambda Phage Genes and The Galactose Operon of Eschericia coli, *PSNA* 69(10) (1972) 2904-2909.
- [4] J. Porath, J. Carlsson, I. Olsson, B. Belfrage, Metal chelate affinity chromatography, a new approach to protein fractionation, *Nature* 258 (1975) 598-599.
- [5] H. Hope, Cryocrystallography of Biological Macromolecules: a Generally Applicable Method, *Acta Crystallogr. B*44 (1988) 22-26.
- [6] W. Yang, W.A. Hendrickson, R.J. Crouch, Y. Satow, Structure of ribonuclease H Phased at 2 Å resolution by MAD Analysis of the Selenomethionyl Protein, *Science* 249(4975) (1990) 1389-1405.
- [7] J.C. Phillips, A. Wlodawer, M.M. Yevitz, K.O. Hodgson, Application of Synchrotron Radiation to Protein Crystallography: Preliminary Results, *PNAS* 73(1) (1976) 128-132.
- [8] M.W. Tate, E.F. Eikenberry, S.L. Barna, M.E. Wall, J.L. Lowrance, S.M. Gruner, A Large-Format High-Resolution Area X-ray Detector Based on a Fiber-Optically Bonded Charge-Coupled Device (CCD), *J. Appl. Cryst.* 28 (1995) 196-205.
- [9] S.A. Lesley, P. Kuhn, A. Godzik, A.M. Deacon, I. Mathews, A. Kreuzsch, G. Spraggon, H.E. Klock, D. McMullan, T. Shin, J. Vincent, A. Robb, L.S. Brinen, M.D. Miller, T.M. McPhillips, M.A. Miller, D. Scheibe, J.M. Canaves, C. Guda, L. Jaroszewski, T.L. Selby, M-A. Elsliger, J. Wooley, S.S. Taylor, K.O. Hodgson, I.A. Wilson, P.G. Schultz, R.C. Stevens, Structural genomics of the *Thermotoga maritima* proteome implemented in a high-throughput structure determination pipeline, *Proc Natl Acad Sci U S A.* 99(18) (2002) 11664–11669.
- [10] R.L Marsden, T.A Lewis, C.A. Orengo, Towards a comprehensive structural coverage of completed genomes: a structural genomics viewpoint, *BMC Bioinformatics* 8 (2007) 86.

- [11] B.H. Dessailly, R. Nair, L. Jaroszewski, J.E. Fajardo, A. Kouranov, D. Lee, A. Fiser, A. Godzik, B. Rost, C. Orengo, PSI-2: Structural Genomics to Cover Protein Domain Family Space, *Structure* 17(6) (2009) 869-881.
- [12] R. Nair, J. Liu, T-T. Soong, T.B. Acton, J.K. Everett, A. Kouranov, A. Fiser, A. Godzik, L. Jaroszewski, C. Orengo, G.T. Montelione, B. Rost, Structural genomics is the largest contributor of novel structural leverage, *J Struct Funct Genomics* 10(2) (2009) 181–191.
- [13] A.J. Walhout, G.F. Temple, M.A. Brasch, J.L. Hartley, M.A. Lorson, S. van den Heuvel, M. Vidal, GATEWAY recombinational cloning: application to the cloning of large numbers of open reading frames or ORFeomes, *Methods Enzymol.* 328 (2000) 575-592.
- [14] H.I. Krupka, B. Rupp, B.W. Segelke, T.P. Legin, D. Wright, H.C. Wu, P. Todd, A. Azarani, The high-speed Hydra-Plus-One system for automated high-throughput protein crystallography, *Acta Crystallogr D Biol Crystallogr.* 58 (2002) 1523-1526.
- [15] C. Hansen, S.R. Quake, Microfluidics in structural biology: smaller, faster... better, *Curr Opin Struct Biol.* 13(5) (2003) 538-544.
- [16] B. Rupp, B. W. Segelke, H. I. Krupka, T. Legin, J. Schäfer, A. Zemla, D. Toppani, G. Snell and T. Earnest, The TB structural genomics consortium crystallization facility: towards automation from protein to electron density, *Acta Cryst. D*58 (2002). 1514-1518.
- [17] B Qian, S Raman, R Das1, P Bradley, A J. McCoy, R J. Read, D. Baker, High resolution protein structure prediction and the crystallographic phase problem, *Nature* 450(7167) (2007) 259–264.
- [18] M.Y. Shen, A. Sali, Statistical potential for assessment and prediction of protein structures *Protein Science*,15(11). (2006) 2507-24.
- [19] D. Petrey, B. Honig, Protein structure prediction: inroads to biology, *Mol. Cell* 20(6) (2005) 811-819.
- [20] A. Zemla, B. Geisbrecht, J. Smith, M. Lam, B. Kirkpatrick, M. Wagner, T. Slezak, C.L.E. Zhou, STRALCP—structure alignment-based clustering of proteins, *Nuc. Acids Res.* 35 (2007) 1–8
- [21] A. Chakicherla, C.L.E. Zhou, M.L. Dang, V. Rodriguez, J.N Hansen, A. Zemla, SpaK/SpaR Two-component System Characterized by a Structure-driven Domain-fusion Method and in Vitro Phosphorylation Studies, *PLoS Comp. Biol.* 5(6) 2009 1-12.

- [22] J.A. Capra, R.A. Laskowski, J.M. Thornton, M. Singh, T.A. Funkhouser, Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure, *PLoS Comp. Biol.* 5 (2009)
- [23] O.C. Redfern, B.H. Dessailly, T.J. Dallman, I. Sillitoe, C.A. Orengo, FLORA: a novel method to predict protein function from structure in diverse superfamilies. *PLoS Comp. Biol.* 5 (2009) e1000485
- [24] B.H. Dessailly, O.C. Redfern, A. Cuff, C.A. Orengo, Exploiting structural classifications for function prediction: towards a domain grammar for protein function, *Curr. Op. Struc. Biol.* 19 (2009) 349-56.
- [25] D. Lee, O. Redfern, C. Orengo, Predicting protein function from sequence and structure, *Nat. Rev. Mol. Cell Biol.* 8 (2007) 995-1005.
- [26] F.W. Studier, Protein production by auto-induction in high density shaking cultures, *Prot. Expr. Pur.*,41(1) (2005) 207-234.
- [27] M. Weselak, M.G. Patch, T.L. Selby, G. Knebel, R.C. Stevens, Robotics for automated crystal formation and analysis, *Meth. Enzym.* 368 (2003) 45-76.
- [28] J. Holton, T.B. Alber, TB Automated protein crystal structure determination using ELVES, *PNAS* 101(6) (2004) 1537-1542.
- [29] T.C. Terwilliger, J. Berendzen, Automated MAD and MIR structure solution, *Acta Cryst. D*55 (1999) 849-861.
- [30] T.C. Terwilliger, Maximum likelihood density modification, *Acta Cryst. D*56 (2000) 965-972.
- [31] T.C. Terwilliger, Automated main-chain model building by template matching and iterative fragment extension, *Acta Cryst. D*59 (2003) 38-44.
- [32] A. Perrakis, R. Morris, V.S. Lamzin, Automated protein model building combined with iterative structure refinement. *Nat. Struct. Biol.* 6 (1999) 458-463.
- [33] R. Nair, J. Liu, T.T. Soong, T.B. Acton, J.K. Everett, A. Kouranov, A. Fiser, A. Godzik, L. Jaroszewski, C. Orengo, G.T. Montelione, B. Rost, Structural genomics is the largest contributor of novel structural leverage, *J. Struc. Func. Genom.* 10 (2009) 181-191.
- [34] J. Hou, G.E. Sims, C. Zhang, S.H. Kim, A global representation of the protein fold space,.*PNAS* 100(5) (2003) 2386-2390.

- [35] D.A. Lee, R. Rentzsch, C. Orengo, GeMMA: functional subfamily classification within superfamilies of predicted protein structural domains, *Nuc. Acids Res.* (2009) 1-18.
- [36] J.J. Hornberg, F.J. Bruggeman, H.V. Westerhoff, J. Lankelma, *Cancer: A Systems Biology Disease*, *BioSystems* 83 (2006) 81-90.
- [37] Y.J. Huang, D. Hang, L.J. Lu, L.Tong, M.B. Gerstein, G.T. Montelione, Targeting the Human Cancer Pathway Protein Interaction Network by Structural Genomics, *Mol. Cell. Prot.* 7.10 (2008) 2048-2060.
- [38] S. Peri, J.D. Navarro, R. Amanchy, T.Z. Kristiansen, C.K. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T.K. Gandhi, M. Gronborg, N. Ibarrola, N. Deshpande, K. Shanker, H.N. Shivashankar, B.P. Rashmi, M.A. Ramya, Z. Zhao, K.N. Chandrika, N. Padma, H.C. Harsha, A.J. Yatish, M.P. Kavitha, M. Menezes, D.R. Choudhury, S. Suresh, N. Ghosh, R. Saravana, S. Chandran, S. Krishna, M. Joy, S.K. Anand, V. Madavan, A. Joseph, G.W. Wong, W.P. Schiemann, S.N. Constantinescu, L. Huang, R. Khosravi-Far, H. Steen, M. Tewari, S. Ghaffari, G.C. Blobel, C.V. Dang, J.G. Garcia, J. Pevsner, O.N. Jensen, P. Roepstorff, K.S. Deshpande, A.M. Chinnaiyan, A. Hamosh, A. Chakravarti, A. Pandey, Development of human protein reference database as an initial platform for approaching systems biology in humans, *Gen.Res.* 13 (2003) 2363–2371.
- [39] I. Xenarios, L. Salwinski, X.J. Duan, P. Higney, S.M. Kim, D. Eisenberg, DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 30 (2002) 303–305.
- [40] S.M. Vorobiev, M. Su, J. Seetharaman, Y.J. Huang, C.X. Chen, M. Maglaqui, H. Janjua, M. Proudfoot, A. Yakunin, R. Xiao, T.B. Acton, G.T. Montelione, L. Tong, Crystal structure of human retinoblastoma binding protein 9, *Proteins* 74(2) (2009) 526-529.
- [41] D.J. Shields, S. Niessen, E.A. Murphy, A. Mielgo, J.S. Desgrosellier, S.K.M. Lau, L.A. Barnes, J. Lesperance, M. Bouvet, D. Tarin, B.F. Cravatt, D.A. Cheresch, RBBP9: A tumor-associated serine hydrolase activity required for pancreatic neoplasia, *PNAS* 107(5) (2010) 2189-2194.
- [42] B.W. Segelke, J. Schafer, M.A. Coleman, T.P. Lakin, D. Toppani, K.J. Skowronek, K.A. Kantardjieff, B. Rupp, Laboratory scale structural genomics, *J. Struc. Func. Genomics* 5 (2004) 147-157.
- [43] V.L. Motin, A.M. Georgescu, J.P. Fitch, P.P. Gu, D.O. Nelson, S.L. Mabery, J.B. Garnham, B.A. Sokhansanj, L.L. Ott, M.A. Coleman, J.M. Elliott, L.M. Kegelmeyer, A.J. Wyrobek, T.R. Slezak, R.R. Brubaker, E. Garcia, Temporal Global Changes in Gene

Expression during Temperature Transition in *Yersinia pestis*, *J. Bact.* 186(18) (2004) 6298-6305.

[44] G. Sciara, S.G. Kendrew, A.E. Miele, N.G. Marsh, L. Federici, F. Malatesta, G. Schimperna, C. Savino, B. Vallone, The structure of ActVA-Orf6, a novel type of monooxygenase involved in actinorhodin biosynthesis, *EMBO J.* 22 (2003) 205-215.

[45] K.B. Xavier, S.T. Miller, W. Lu, J.H. Kim, J. Rabinowitz, I. Pelczer, M.F. Semmelhack, B.L. Bassler, Phosphorylation and Processing of the Quorum-Sensing Molecule Autoinducer-2 in Enteric Bacteria, *ACS Chem. Biol.* 2(2) (2007) A-1.

[46] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, T.E. Ferrin, UCSF Chimera--a visualization system for exploratory research and analysis, *J Comput. Chem.* 25(13) 2004 1605-12.

## Tables

**Table 1.** Combined success rate by process step for PSI structural genomics centers

<b>Process Step</b>	<b>Number of Targets*</b>	<b>Cumulative % Success</b>
<b>Cloned</b>	178,000	--
<b>Expressed</b>	128,000	72
<b>Soluble</b>	48,000	27
<b>Purified</b>	44,000	25
<b>Crystallized</b>	15,000	8
<b>Diffraction</b>	6,800	4
<b>Structure determined</b>	5,400	3

\*Approximated numbers as of June 2010 taken from TargetDB (<http://targetdb.pdb.org/TargetDB/>).

## Figures

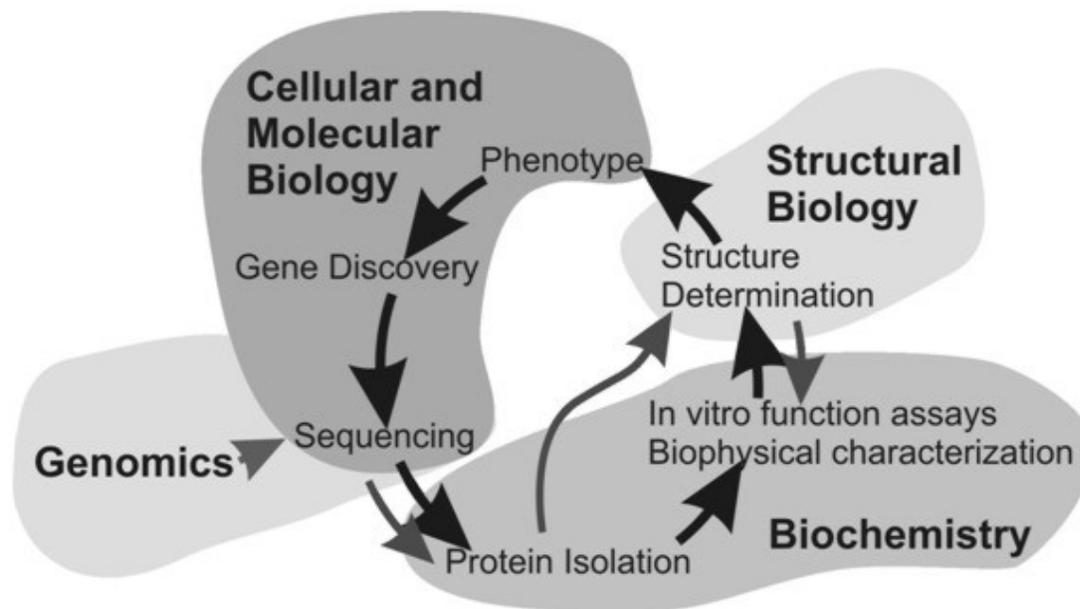


Figure 1. The structural genomics shunt. A traditional structural biology project follows extensive cellular and molecular biology and biochemistry, depicted here by the cycle traced by the black arrows. The function and biophysical properties of a protein are traditionally very well characterized before a structural biology project is initiated. Structural genomics is instead starts with genomic, or gene transcript data, leading to targeting a gene or collection of genes for structure determination. The absence of functional annotation may contribute to targeting and the structure may be the first indication of a proteins function. Function inferred from structure can further motivate biochemistry or cellular biology studies. This figure was adapted from Yee et al. (2003)[1].

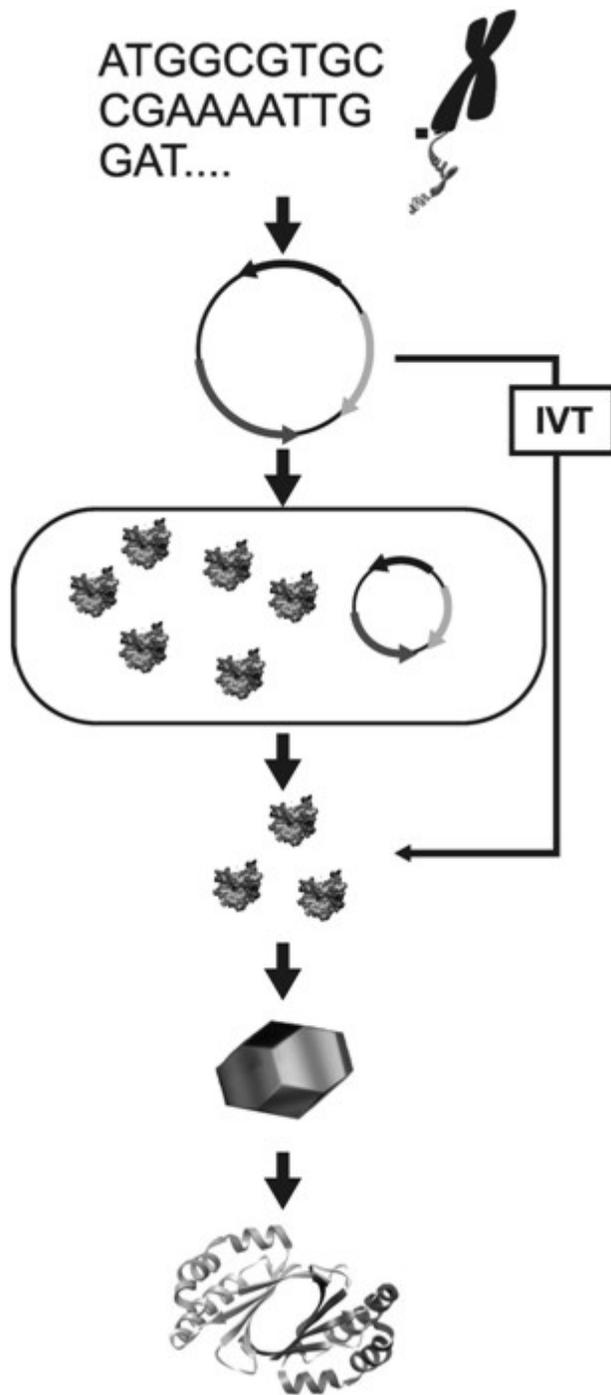


Figure 2. The structural genomics pipeline. Structural genomics largely focused on developing standardized high throughput processes to industrializing structure determination. Starting with just genomic information, e.g. gene sequence, and genetic material a gene is cloned into an expression plasmid, which is transformed into a heterologous expression host. *In vitro* transcription/translation (IVT) can eliminate the need for heterologous expression and can be used to express cytotoxic proteins.

Following transformation, protein is expressed, purified, and crystallized, or used for NMR studies. Crystals are used for x-ray diffraction experiments leading to structure determination. This is a greatly simplified schematic representing only the major steps in the structural genomics pipeline. Most of the early efforts in structural genomics involved automating, parallelizing, and miniaturizing steps in the structural genomics pipeline as well as building information management systems to track progress and data produced in the process.

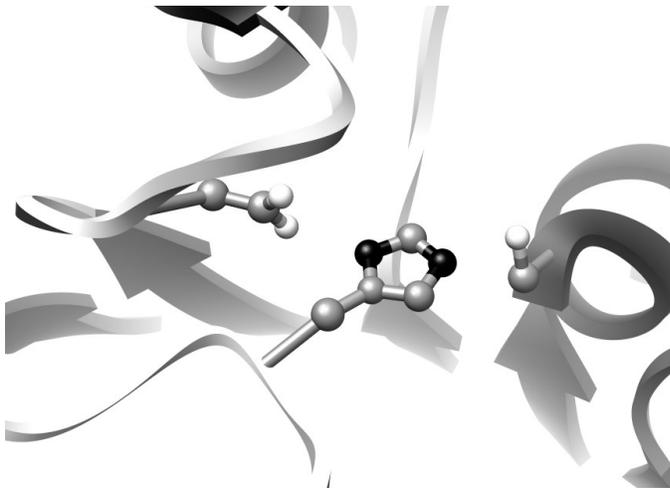


Figure 3. Unexpected serine protease catalytic triad revealed by structural genomics. A region of the crystal structure of RBBP9, a retinoblastoma binding protein, is shown with backbone shown as a ribbon and the sidechains of the catalytic triad shown in ball and stick [41]. Note the three residues of the triad are in close proximity in 3 dimensions, but are brought together from three disparate regions of sequence. Only the 3 dimensional structure reveals the catalytic triad motif. This figure was generated using Chimera[46].

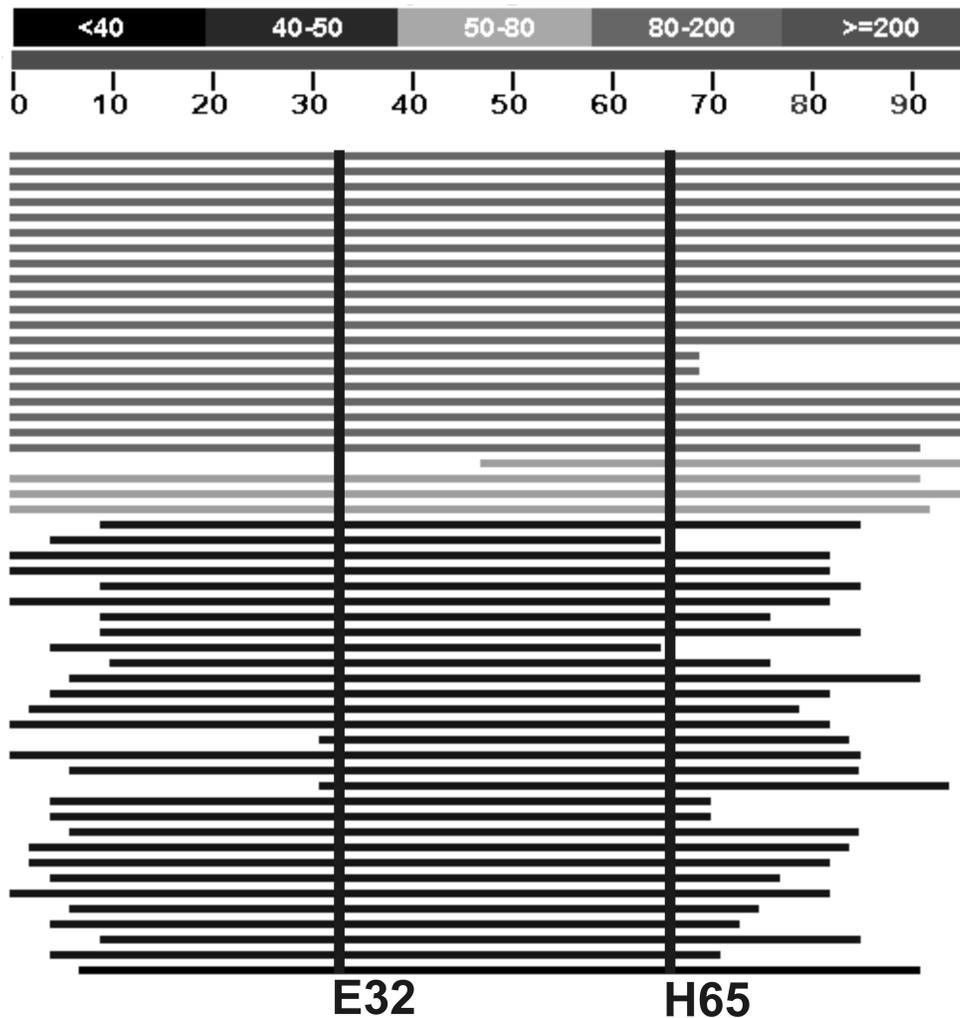


Figure 4. PSI blast comparison between YPO0407 encoded protein and homologues. PSI-blast initially identified a number of close homologues for *Salmonella*, *E. coli* and other species closely related to *Y. pestis* (upper medium grey block of horizontal bars)—all of these homologues were annotated as hypothetical conserved protein, except one salmonella protein that was annotated as *IsrG*. A large number of more distantly related proteins were identified (middle light grey block and lower block of black horizontal lines), the majority of which were annotated as hypothetical conserved protein, though some were annotated as antibiotic biosynthesis monooxygenases. Two residues were highly conserved across the whole family of proteins, E32 and H65 (using YPO0407 numbering). The high conservation of these residues across a diverse family of homologues implicates these residues as functionally important.



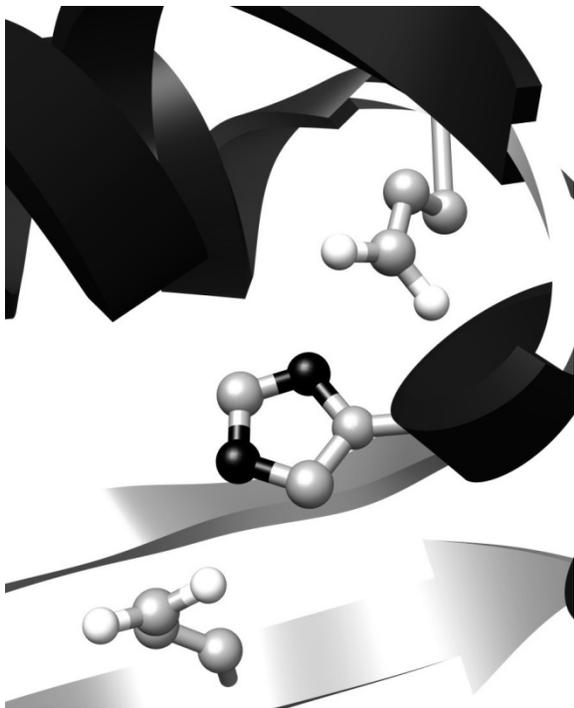


Figure 6. Putative hydrolase triad of the YPO0407 gene product. The conserved residues of the subfamily the YPO0407 gene product clusters with are arranged in a suggestive 3D motif. This arrangement of residues is consistent with hydrolase activity and not monooxygenases. Hydrogenase activity is consistent with the know substrate and product for IsrG[45]. This figure was generated with Chimera[46].