Adam Mickiewicz University
Institute of Molecular Biology and Biotechnology
Laboratory of Structural Bioinformatics

Doctoral Dissertation of:

Marcin Jakub Domagalski, M.Sc.

Structure determination and functional analysis of isochorismate synthase DhbC from *Bacillus anthracis* using the state of the art SG data management system

Advisor: Janusz M. Bujnicki, Ph.D., D.Habil.

Laboratory of Structural Bioinformatics
Institute of Molecular Biology and Biotechnology
Adam Mickiewicz University
&
Laboratory of Bioinformatics and Protein Engineering
The International Institute of Molecular and Cell Biology in Warsaw

Co-advisor: Władysław Minor, Ph.D.

Department of Molecular Physiology and Biological Physics
University of Virginia, USA

Poznań, 2015

Acknowledgements

I would like to thank my PhD advisors, Professors Wladek Minor and Janusz Bujnicki, who took the time to share their knowledge and appreciation of structural biology, X-ray crystallography, and bioinformatics. I am thankful for the opportunity to work on the exciting scientific projects, chance to attend prestigious conferences, and contribute to excellent manuscripts.

I would also like to thank co-authors of projects, especially Dr. Marek Grabowski and Dr. Matt Zimmermann who supervised my work on data management system, Dr. Maksymilian Chruszcz for teaching me how to collect and process X-ray data, and Dr. Igor Shumilin for teaching me crystallization techniques. Marek, Matt, Maks, and Igor sacrificed many hours of their time helping me in planning my projects and experiments.

I am grateful for productive discussions with Dr. Karolina Tkaczuk, Katarzyna Handing, Dr. Ivan Shabalin, Karolina Majorek, Dr. David Cooper, and Dr. Jing Hou. I need to thank all the members of the Wladek Minor's group who create such a good atmosphere in the lab.

In addition, I would like to thank Dr. David Cooper for proofreading the manuscript and for his valuable comments.

Lastly, I would like to thank my parents and my brother for their encouragement and for allowing me to realize my passion. Most of all, I need to thank my fiancée Maja Buszko, her support and encouragement was in the end what made this dissertation possible.

Peer-reviewed articles and book chapters that resulted from this work

Domagalski, M.J., Tkaczuk, K.L., Chruszcz, M., Skarina, T., Onopriyenko, O., Cymborowski, M., Grabowski, M., Savchenko, A., Minor, W. (2013) Structure of isochorismate synthase DhbC from *Bacillus anthracis*. Acta Cryst F69:956-61

Domagalski, M.J., Zheng, H., Zimmerman, M.D., Dauter, Z., Wlodawer, A., Minor, W. (2014) The Quality and Validation of Structures from Structural Genomics. Methods Mol Biol (Clifton, N.J.) 1091:297-314

Zimmerman, M.D., Grabowski, M., **Domagalski, M.J.**, MacLean, E.M., Chruszcz, M., Minor, W. (2014) Data Management in the Modern Structural Biology and Biomedical Research Environment. Methods Mol Biol (Clifton, N.J.) 1140:1-25

Chruszcz, M., **Domagalski, M.**, Osinski, T., Wlodawer, A., Minor, W. (2010) Unmet challenges of structural genomics. Curr Opin Struct Biol 5:587-97

Table of contents

List of abbreviations

3D	three-dimensional
ADCS	4-amino-4-deoxychorismate synthase
AS	anthranilate synthase
AVA	anthrax vaccine adsorbed
AVP	anthrax vaccine precipitated
BMRB	Biological Magnetic Resonance Data Bank
CSS	Cascading Style Sheets
CM	chorismate mutase
CMM	CheckMyMetal
CPL	chorismate pyruvate-lyase
CSGID	Center for Structural Genomics of Infectious Diseases
CSV	Comma-Separated Values
DHB	2,3-dihydroxybenzoic acid
EFI	Enzyme Function Initiative
FPLC	Fast Protein Liquid Chromatography
FTP	File Transfer Protocol
GPL	General Public License
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
IDE	Integrated Development Environment
ICS	isochorismate synthase
LIMS	Laboratory Information Management System
MBP	Maltose Binding Protein
MCSG	Midwest Center for Structural Genomics
MR	Molecular Replacement
MRSA	Methicillin-resistant Staphylococcus aureus (MRSA)
MVC	Model-View-Controller
NCBI	National Center for Biotechnology Information
NIAID	National Institute of Allergy and Infectious Diseases

NIGMS	National Institute of General Medical Sciences
NMR	Nuclear Magnetic Resonance
NRPS	Nonribosomal Peptide Synthetase
NYSGRC	New York Structural Genomics Research Consortium
OECD	Organisation for Economic Co-operation and Development
ORF	Open Reading Frame
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
PEP	phosphoenolpyruvate
Pfam	Protein FAMily database
PSI	Protein Structure Initiative
PSI-SBKB	PSI Structural Biology Knowledgebase
RCSB	The Research Collaboratory for Structural Bioinformatics
RDBMS	Relational Database Management System
R.M.S.D	Root Mean Square Deviation
SCOP	Structural Classification Of Proteins
SDS-PAGE	Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis
SoC	separation of concerns
SG	structural genomics
SMILES	Simplified Molecular-Input Line-Entry System
SR	synchrotron radiation
SS	salicylate synthase
SSGCID	Seattle Structural Genomics Center for Infectious Disease
SQL	Structured Query Language
TLS	Translation/Libration/Screw
TSV	Tab-Separated Values
UniProt	UNIversal PROTein resource
URL	Uniform Resource Locator
WHO	World Health Organisation
XML	Extensible Markup Language

Abstract

The rapid growth of the number of antibiotic-resistant strains of pathogenic bacteria is becoming a major threat to global public health. In order to limit the number of deaths from simple infections, the development of target specific drugs to replace conventional antibiotic therapies is urgently needed. One of the most promising approaches is based on interrupting iron assimilation in pathogenic bacteria. Isochorismate synthase DhbC from *Bacillus anthracis* is important for the infectivity of this dangerous bacterium because it catalyzes the first step in the pathway for synthesis of the siderophore, bacillibactin. Pathogenic bacteria use siderophores, chelating ferric ions chemical compounds, in order to assimilate scarcely available ferric ions inside of the host organism. The DhbC active site is very similar to the active sites of other chorismate-utilizing enzymes, which suggests the possibility of developing a single inhibitor that targets multiple chorismate-utilizing enzymes. Chorismate-utilizing enzymes are very promising antimicrobial drug targets because of their important role in virulence and in a wide range of bacterial metabolic processes, plus their absence in humans. Therefore, Center of Structural Genomics of Infectious Diseases (CSGID) selected DhbC as a target for structural studies.

Structural genomics (SG) is a relatively new approach to structural biology (first projects started in late 90's) aimed at high-throughput 3D structure determination of macromolecules. Typical SG center consist of specialized laboratories that perform only selected parts of the protein structure determination experimental pipeline. In most of the cases, including CSGID, involved laboratories are located in distant research centers. In order to control the vast amount of data produced by the consortium, the data management system LabDB/UniTrack was developed in Wladek Minor's laboratory at the University of Virginia. The system tracks all experimental work and exchange the data within the lab (group) and between groups involved in the project.

The main scientific objective of my work was to determine the three-dimensional structure of the isochorismate synthase DhbC from *B. anthracis* and subsequently biochemical characterization of this enzyme. The atomic structure of this enzyme will be used for identification of new inhibitors of catecholate siderophore pathways through high-throughput virtual screening approach. The second goal was to develop components of the innovative data management system for structural genomics UniTrack, i.e., the protein target tracking database CSGID-DB, associated knowledge dissemination web portal, target validation tool, and communication protocols with other databases. UniTrack is an important part of CSGID gene-to-structure high-throughput pipeline.

The structure of the apo form of DhbC from *B. anthracis* was solved using single crystal X-ray diffraction at 2.4 Å resolution. DhbC adopts the characteristic fold of other chorismate-utilizing enzymes, and strongly resembles isochorismate synthase EntC from *Escherichia coli*. The enzyme is a homodimer and requires presence of Mg^{2+} ions for its activity. Enzyme kinetics constants were determined using spectrophotometric assay.

The UniTrack system monitors all the experimental work on particular protein targets and provides intuitive workflow between research groups involved in the project. It also reports general progress of the consortium by generating real-time internal reports and statistics as well as XML files, which are used for data submission to external repositories. Moreover, it serves as an information hub for the infectious disease scientific community. In 2011, three other structural genomics consortia, the Midwest Center for Structural Genomics, New York Structural Genomics Research Consortium, and the Enzyme Function Initiative incorporated the UniTrack system for the purpose of data management. To date in CSGID only, over 700 protein structures have been determined with use of the UniTrack and ~7000 protein targets are progressing through the experimental pipeline.

Streszczenie

Gwałtowny wzrost liczby odpornych na antybiotyki szczepów patogennych bakterii staje sie głównym zagrożeniem dla globalnego zdrowia publicznego. W celu ograniczenia liczby zgonów spowodowanych przez proste infekcje, potrzebny jest natychmiastowy rozwój swoistych leków w celu zastąpienia konwencjonalnych terapii antybiotykowych. Jedno z najbardziej obiecujących podejść jest ukierunkowane na uniemożliwienie asymilacji żelaza przez patogenne bakterie. Syntaza izochorizmianu DhbC z *Bacillus anthracis* jest ważnym dla infekcyjności tej groźnej bakterii enzymem katalizującym pierwszy etap w szlaku syntezy sideroforu, bacillobaktyny. Patogenne bakterie używają sideroforów, chelatującyh jony żelazowe związków chemicznych, w celu asymilowania trudno dostępnych jonów żelazowych wewnątrz organizmu gospodarza. Centrum aktywne DhbC jest bardzo podobne do centrów aktywnych innych enzymów wykorzystujących choryzmian, sugerując możliwość opracowania pojedynczego inhibitora dla kilku enzymów wykorzystujących choryzmian. Enzymy wykorzystujące choryzmian są bardzo obiecującymi celami dla leków przeciwdrobnoustrojowych ze względu na ich rolę w wirulencji i w szerokim zakresie bakteryjnych procesów metabolicznych, oraz ich nieobecność u ludzi. Z tych powodów, DhbC została wyselekcjonowana do badań strukturalnych przez Centerum Genomiki Strukturalnej Chorób Infekcyjnych (ang. skrót CSGID).

Genomika strukturalna (skrót: SG) jest nowym podejściem do biologii strukturalnej (pierwsze projekty ruszyły pod koniec lat dziewięćdziesiątych) polegającym na wysokoprzepustowym rozwiązywaniu trójwymiarowych struktur makromolekuł. W skład typowego centrum genomiki strukturalnej wchodzą wyspecjalizowane laboratoria przeprowadzające tylko określone etapy sekwencji eksperymentów prowadzącej do rozwiązania struktury białka. W większości przypadków, również w CSGID, wchodzące w skład centrów laboratoria są ulokowane w odległych ośrodkach naukowych. W celu kontrolowania olbrzymich zasobów danych wyprodukowanych przez konsorcjum, w laboratorium prof. Władysława Minora na University of Virgnia został rozwinięty system zarządzania danymi LabDB/UniTrack. System ten pozwala na śledzenie całości pracy doświadczalnej i wymianę tej informacji w obrębie grupy badawczej oraz pomiędzy grupami zaangażowanymi w projekt.

Głównym celem naukowym mojej pracy było rozwiązanie trójwymiarowej struktury syntazy izochoryzmianu DhbC z *B. anthracis*, a następnie scharakteryzowanie biochemicznych właściwości tego enzymu. Struktura atomowa tego enzymu zostanie wykorzystana do poszukiwań nowych inhibitorów ścieżek metabolicznych siderofrów pirokatechinowych poprzez wysokoprzepustowe badania przesiewowe. Drugim celem było rozwinięcie komponentów innowacyjnego systemu zarządzania danymi dla genomiki strukturalnej UniTrack, tzn. bazy danych monitorującej postęp prac na celami białkowymi CSGID-DB, powiązanego portalu internetowego rozpowszechniającego uzyskaną wiedzę, narzędzia do walidacji celów białywych i protokołów komunikacji z innymi bazami danych. UniTrack jest ważną częścią wyskoprzepustowej sekwencji doświadczalnej "od genu do struktury".

Struktura formy apo DhbC z *B. anthracis* została rozwiązana za pomocą krystalografii rentgenowskiej pojedynczych kryształów makromolekuł do rozdzielczości 2.4 Å. DhbC przybiera zwój charakterystyczny dla innych enzymów wykorzystujących choryzmian i silnie przypomina syntazę izochoryzmianu EntC z *Escherichia coli*. Enzym jest homodimerem i wymaga obecności jonów Mg^{2+} dla swojej aktywności. Stałe kinetyczne dla reakcji katalizowanej przez enzym zostały wyznaczone z użyciem analizy spektrofotometrycznej.

System UniTrack monitoruje pracę doświadczalną nad poszczególnymi celami białkowymi i zapewnia intuicyjny przypływ pracy pomiędzy grupami badawczymi zaangażowanymi w projekt. Monitoruje również ogólny postęp konsorcjum przez generowane w czasie rzeczywistym wewnętrzne raporty i statystyki jak również pliki XML, które są wysyłane do zewnętrznych repozytoriów. Ponadto sluży jako centrum informacyjne dla społeczności naukowej. W 2011, kolejne trzy centra genomiki strukturalnej: Midwest Center for Structural Genomics, New York Structural Genomics Research Consortium i Enzyme Function Initiative zaadoptowały system UniTrack na potrzeby zarzadzania danymi. Do chwili obecnej w samym CSGID, ponad 700 struktur białek zostało rozwiązanych z użyciem systemu UniTrack, a około 7000 celów białkowych znajduje się w fazie badan doświadczalnych.

1. Introduction

1.1 Chorismate-utilizing enzymes as putative drug targets

The shikimate biosynthetic pathway, present solely in bacteria, algae, higher plants, fungi, and Apicomplexa (phylum of parasitic protists), produces chorismate out of D-erythrose 4-phosphate and phosphoenolpyruvate (PEP). Chorismate is anionic form of chorismic acid and serves as intermediate metabolite between the shikimate pathway and the following biosynthetic pathway for aromatic amino acids (i.e., L-phenylalanine, L-tryptophan, and L-tyrosine). Additionally, chorismate is a precursor for biosynthesis of multiple other aromatic compounds such as folate, ubiquinone, phenazines (Dosselaere and Vanderleyden 2001; Kerbarh et al. 2005), and selected siderophores, including enterobactin (O'Brien et al. 1970) and bacillibactin (May et al. 2001). The aforementioned aromatic compounds are essential for bacteria survival and virulence. Because mammals do not possess the above-mentioned pathways, the enzymes have gained attention as potential targets for the development of new antimicrobial drugs (Kerbarh et al. 2005; Ziebart et al. 2010). Up to the present time, seven distinct chorismate-utilizing enzymes have been characterized in bacteria, including chorismate mutase (CM), chorismate pyruvate-lyase (CPL), anthranilate synthase (AS), 4-amino-4-deoxychorismate synthase (ADCS), 2-amino-2-desoxyisochorismate synthase (ADICS), isochorismate synthase (ICS), and salicylate synthase (SS). Five of these enzymes, i.e., ICS, SS, AS, ADICS and ADCS share significant structural similarity (including nearly identical actives), require Mg^{2+} ions for its catalytic activity, and catalyze a similar S_N2' nucleophilic substitution reactions. Thus, it may be possible to develop single compound that will inhibit more than one of those enzymes (Ziebart et al. 2010).

The isochorismate synthase DhbC from *Bacillus anthracis* participates in the bacillibactin biosynthetic pathway. In closely related species *B. cereus*, bacillibactin was recently demonstrated to be crucial for effective virulence through iron acquisition from host ferritin during infection in insects (Segond et al. 2014). Studies on the mechanistic pathways of siderophores may lead to design of small-molecule inhibitors of siderophore biosynthesis and therefore drugs limiting virulence of pathogenic bacteria (Ferreras et al. 2005). Moreover, since bacteria recognize only certain siderophores, it may be possible to use siderophore-mediated iron transport as a 'Trojan horse' for very selective antimicrobial drug delivery (Roosenberg et al. 2000; Wencewicz et al. 2009). Coupling of the siderophore iron-binding groups to an antibiotic should significantly increase effectiveness of the latter one. The drug would be delivered directly to the pathogenic bacteria using microbe specific siderophore.

1.1.1 Antimicrobial resistance – a major threat to public health

Antimicrobial resistance (AMR) is an evolutionarily developed resistance of a pathogen to an antimicrobial drug that was initially effective for treatment of infections caused by the pathogen. Antibiotic resistance refers specifically to resistance of pathogenic bacteria to antibiotics. The main cause of antibiotic resistance is extensive and irresponsible use of antibiotics, which are not only used in medicine, but also in animal feed, plant agriculture, and industry (Barbosa and Levy 2000; Nikaido 2009). Antibiotics are produced at estimated scale of about 100,000 tons annually worldwide. The use of antibiotics creates selective pressure on pathogenic bacteria resulting in the development of resistant strains in humans and livestock animals. Humans spread the resistant bacteria in their families, communities and especially in hospitals and other health care facilities where the most of the infection related deaths occur (CDC 2013). In rare cases resistant bacteria are transmitted to humans from animals via consumption of animal products, contact with animals or by contamination of crops (Hurd et al. 2004; CDC 2013). The emergence of bacterial strains resistant to multiple classes of antibiotics, including most dangerous methicillin-resistant *Staphylococcus aureus* (MRSA), and strains resistant to all clinically relevant drugs like multidrug-resistant *Streptococcus pneumoniae* and multi-drug-resistant *Mycobacterium tuberculosis* is cause for alarm. Outbreaks of multi-drug resistant strains may lead to a global pandemic situation (Choffnes ER 2010). In the United States only, minimum estimates show that antibiotic resistant bacteria are causing ~2 million infections per year, with ~24,000 associated deaths. Additionally, infections caused by *Clostridium difficile,* which usually follow use of antibiotics, result in ~250,000 thousand infections and ~14,000 deaths (CDC 2013). It is important to limit the number of infections through the promotion of good hygiene and sanitation, improvement of the use of antibiotics, and development of new generation of antimicrobial drugs.

On April 30 2014, the World Health Organization released the first global report on antibiotic resistance. The report, 'Antimicrobial resistance: global report on surveillance,' gathers data from 114 countries in all parts of the world which makes it the most complete study on antimicrobial resistance to date. WHO is highlighting the critical actions that should be taken to overcome AMR, i.e., reinforcing global AMR surveillance, monitoring the effectiveness of public health, detecting trends and threads, and most importantly developing a global action plan against AMR (WHO 2014). The WHO report does not leave any doubt that antibiotic resistance has already become a major threat to public health.

1.1.2 Anthrax treatment and antimicrobial resistance

Anthrax is a potentially lethal disease caused by *B. anthracis*, known to humanity since the development of agriculture, and associated with black eschars caused by its cutaneous form (Turnbull 2010). The disease affects wild and domesticated animals (i.e., cattle, sheep, and horses) and occasionally humans. *B. anthracis* forms spores that can be infectious for many years and can be found in soil as well as on hair, wool, and processed skins made from infected animals. Humans working with farm animals and animal products are considered high-risk group for anthrax infection. The most common *B. anthracis* infections are cutaneous, and this anthrax form can be successfully treated using antibiotics. Inhalational infection, on the other hand, has fatality rate of almost 90% (Beierlein and Anderson 2011). Antrax infection has two stages, an intracellular establishment stage in macrophages, and a subsequent extracellular stage that leads to bacteremia, sepsis, and death (Cendrowski et al. 2004). The ability to grow within macrophages and use their trafficking during infection is a distinctive feature of anthrax (Bergman 2011).

Recently, *B. anthracis* gained public attention after its spores were used for bioterror attacks that happened in September 2001 in the USA. Envelopes with spores of the highly virulent Ames strain of *B. anthracis* were mailed to news media offices and U.S. senators, resulting in five lethal and seventeen life-threatening infections.

Currently, military personnel, vulnerable laboratory workers, and livestock workers around the world receive one of two licensed anthrax vaccines: anthrax vaccine adsorbed (AVA) or anthrax vaccine precipitated (AVP). The vaccines are administered in multiple doses over 18 and 8-month periods respectively, and followed by annual booster doses to maintain the immunity (Splino et al. 2005). In case of sudden outbreak of anthrax, vaccines would have limited use because of the slow development of immunity and short period of effective protection (Weiss et al. 2007).

Treatment of anthrax is based on prolonged use of antibiotics and it is effective for some forms of the disease. Similar to treatments for other bacterial infections, it includes large doses of intravenous and oral antibiotics, such as fluoroquinolones, doxycycline, erythromycin, vancomycin, or penicillin (Evans 2002). Typical post-exposure preventative treatment is based on administration of penicillin G, amoxicillin, doxycycline, and ciprofloxacin or ofloxacin given for minimum 60 days (Athamna et al. 2004). It has been showed by multiple *in vitro* studies that prolonged antibiotic treatment might induce resistance to fluoroquinolones (i.e., ciprofloxacin, ciprofloxacin, garenoxacin, levofloxacin and ofloxacin), doxycycline, rifampicin, and β-lactam antibiotics (i.e., amoxicillin, ceftriaxone, penicillin G) in *B. anthracis* (Pomerantsev et al. 1992; Brook et al. 2001; Price et al. 2003; Athamna et al. 2004). Naturally occurring penicillin resistance in *B. anthracis* has been already documented in clinical isolates (Severn 1976; Bradaric and Punda-Polic 1992; Lalitha and Thomas 1997).

1.1.3 Importance of iron for pathogenic bacteria and their host organisms

Iron is an abundant transition metal that is an essential cofactor for the most important cellular processes in practically all forms of life. The iron-dependent processes include photosynthesis, oxygen transport, respiration, the tricarboxylic acid cycle, lipid metabolism, amino acid synthesis, nucleoside synthesis, gene regulation, DNA synthesis, etc. (Cairo et al. 2006). Iron functions as a protein cofactor in the form of mononuclear and binuclear species, as well as more complex iron-sulfur clusters and heme groups (Andrews et al. 2003). Nevertheless, acquisition of iron is a rather challenging problem for organisms living in oxic environment as well as for pathogenic bacteria. In compounds, iron exists predominantly in two oxidation states: iron(II) form called ferrous iron and iron(III) form, referred to as ferric ion. Under aerobic conditions ferrous ions are unstable and react with peroxides forming free radicals which damage DNA, proteins and lipids (Touati 2000). Ferric ions, on the other hand, in aqueous oxic solutions aggregate into very insoluble ferric hydroxides, bringing down the concentration of soluble ferric ions to extremely low levels, i.e., 10^{-18}M in pH ≥ 7.4 (Carrano and Raymond 1978). Moreover, free aqueous Fe^{3+} ion is toxic for the cell. For that reason, the level of free iron in the human body is strictly regulated and kept to a negligible level. In human serum, virtually all iron is either bound to hemoglobin, heme, or iron-storage proteins like ferritins and transferrins or serves as cofactors for various enzymes (Hotta et al. 2010).

Iron is equally essential for microbes as it is for higher organisms. The virulence of numerous bacteria including *Escherichia coli* (Bullen et al. 1968), *Klebsiella pneumonia* (Ward et al. 1986), *Listeria monocytogenes* (Martinez et al. 1990), *Salmonella* (Griffiths 1991), *Shigella* (Payne 1989) and other species has been proven to increase with excess of iron. For example, in the case of *Yersinia enterocolitica*, the virulence was enhanced 10 million-fold after the peritoneal injection of ferric desferrioxamine (Bullen et al. 1991). Analogically, bacteriostatic properties of human milk are eliminated by *in vitro* addition of iron (Bullen 1972). Aforementioned studies indicate that strict control of iron availability in mammals is an important element of their protection against bacterial infection (Andrews et al. 2003). The mechanism of protection against microbial infection through active sequestration of nutritional elements is called nutritional immunity (Pishchany 2011). In the absence of highly efficient iron assimilation pathways, pathogenic bacteria would not be able to grow and would be gradually defeated by host's immune system (Ratledge and Dover 2000). Therefore, pathogenic bacteria evolved sophisticated systems for assimilation of iron.

1.1.4 Iron assimilation by *B. anthracis*

The genome of *B. anthracis* contains significantly more iron acquisition systems than genomes of non-pathogenic members of the *Bacillus* genus (Read et al. 2003). It contains 16 ABC uptake systems for iron and iron-complexes and two systems for siderophore biosynthesis (Cendrowski et al. 2004). *B. anthracis* is considered an extracellular pathogen, but it requires a short intracellular phase inside macrophages to initiate the infection (Mock and Fouet 2001). The complex life cycle of *B. anthracis* and ability to infect the host organism through multiple entry points are possible causes of the diversity of iron acquisition systems this bacteria (Skaar et al. 2006). Unfortunately, we still do not fully understand mechanisms of action of iron acquisition systems in the *Bacillus* genus.

In general, inside host organisms pathogenic bacteria acquire iron using multiple different strategies that target specific iron sources. The main approaches are iron acquisition from heme, hemoglobin, iron transport, storage and other heme-containing proteins (i.e., transferrin, lactoferrin, and ferritin) and ferric iron acquisition by small iron-chelating compounds, i.e., siderophores (Caza and Kronstad 2013).

Many Gram-negative as well as Gram-positive bacteria produce siderophores, typically under iron limiting conditions (Krewulak and Vogel 2008). There are three groups of siderophores based on the chemical structure of metal binding site: catecholates, hydroxamates, and hydroxycarboxylates (Raymond 2004). The genus *Bacillus* produces two types of catecholate siderophores, petrobactin (also known as anthrachelin), which contains 3,4-dihydroxybenzoyl moieties and bacillibactin (also known as anthrabactin), which contains 2,3-dihydroxybenzoyl moieties. Biosynthesis of above-mentioned siderophores is performed by *B. anthracis* catechol (bac) and anthrax siderophore biosynthesis (asb) operons, for bacillibactin and anthrachelin respectively (Cendrowski et al. 2004). Bacillibactin has significantly higher affinity for ferric ions ($K_f = 10^{47.6}$) (Dertz et al. 2006) than petrobactin ($K_f = 10^{23}$) (Abergel et al. 2008), but it is being recognized by the immune system protein siderocalin, while petrobactin is able to evade this barrier (Abergel et al. 2006). Petrobactin was shown to be required for bacterial growth in low iron medium and for mouse virulence, while bacillibactin is produced in response to low iron medium but is not required for growth in that medium or for virulence in mice (Cendrowski et al. 2004).

Acquisition of iron from heme sources requires destruction of red blood cells with toxins or hydrolytic enzymes and uptake of heme through secretion of hemophores (heme-binding proteins) (Caza and Kronstad 2013). During the extracellular phase of infection, *B. anthracis* is able to lyse erythrocytes and extract heme from hemoglobin through system known as iron-regulated surface determinant (Isd). Isd protein binds heme and heme-containing proteins through NEAT (NEAr iron Transporter) domains (Gat et al. 2008). The importance of the Isd iron acquisition system for *B. anthracis* virulence was proved by a transcriptome investigation (Carlson et al. 2009). This system is absent in other members of *Bacillus* genus except of *B. cereus* group. *B. anthracis* is part of the *B. cereus* group of bacilli, which also includes *B. cereus*, *B. thuringiensis*, and *B. mycoides* (Dixon et al. 1999). *B. anthracis* can grow on high concentrations of heme in comparison to other bacilli, in spite of toxicity of this compound (Lee et al. 2011). Recent studies on *B. cereus* (Segond et al. 2014) have shown that bacillibactin, in cooperation with the surface ferritin receptor IlsA, is essential for iron acquisition from host ferritin. Lack of the bacillibactin production resulted in a drastic reduction of the ability to acquire iron from ferritin and attenuated virulence in insects. IslA is one of the NEAT proteins and is involved in both ferritin and heme/hemoglobin acquisition. *B. anthracis* has two proteins: BslL, which is nearly identical to last three fourths of IlsA and BslK, which shares similarity with NEAT and SLH domains of IslA. BslK was shown to bind heme and mediate heme delivery to Isd system (Tarlovsky et al. 2010). Unfortunately, involvement of these proteins in iron acquisition from ferritin has not been studied yet (Segond et al. 2014).

Ferric uptake regulator (Fur) controls most of the iron acquisition systems in bacteria, including *Bacillus* genus. Fur is a transcription regulator that binds to DNA in the presence of a co-regulatory Fe^{2+} ion (Bagg and Neilands 1987). The protein is 17 kDa and it functions as a homodimer where each subunit is binding single ferrous ion (Coy and Neilands 1991). Binding of metal ions to a Fur dimer increases its affinity to the DNA-binding site known as Fur box by ~ 1000 fold (Andrews et al. 2003). When iron levels are low, Fur dissociates from Fur boxes derepressing the transcription of various bacterial toxins and virulence factors (Caza and Kronstad 2013). In the *B. cereus* group, Fur regulator controls only biosynthesis of bacillibactin and not petrobactin (Rowland and Taber 1996; Baichoo et al. 2002).

1.1.5 Synthesis of bacillibactin by *B. anthracis*

Isochorismate synthase DhbC from *B. anthracis* is a product of *dhbC* gene, part of the *bac* operon (Bacillus anthracis catechol, BA2368-2372) (Figure 1) which encodes proteins responsible for the synthesis of bacillibactin (Figure 2) (Cendrowski et al. 2004). Biosynthesis of this catechol siderophore has two stages: biosynthesis of 2,3-dihydroxybenzoic acid (DHB) and assembling DHB to a cyclic amino acid core synthesized by multimodular nonribosomal peptide synthetases (NRPS) complex

DhbEBF (May et al. 2001). DhbC catalyzes the first step of DHB biosynthesis, which is conversion of aromatic amino acid precursor, chorismate to isochorismate. The genome of *B. anthracis*, as well as of closely related *B. subtilis*, contain a second isochorismate synthase gene *menF*, located in the biosynthetic operon of respiratory chain component menaquinone. It has been shown that DhbC can compensate for a lack of its isozyme MenF, although depletion of DhbC is not compensated by MenF and results in the absence of DHB (Rowland and Taber 1996). In the second step of DHB biosynthesis, isochorismate is hydrolyzed to 2,3-dihydro-2,3-dihydroxybenzoate and pyruvate by isochorismate lyase (DhbB). Subsequently, 2,3-dihydro-2,3-dihydroxybenzoate is oxidized to DHB by 2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase (EntA) (Hoffmann et al. 2002). DHB is activated in an ATP-dependent reaction by 2,3-dihydroxybenzoate-AMP ligase (DhbE) and transferred to free thiol group of the cofactor phosphopantetheine of the bifunctional isochorismatase/aryl-carrier protein (DhbB) (May et al. 2001). Finally, a dimodular NRPS (DhbF) specifically adenylates threonine and glycine, covalently links these amino acids to corresponding peptidyl carrier domains, amide links the two residues to 2,3-dihydroxybenzoyladenylate and esterifies three of these intermediates to form 2,3-dihydroxybenzoate-glycine-threonine trimeric ester (bacillibactin) (May et al. 2001; Hoffmann et al. 2002). The *bac* operon also contains: a gene encoding an MtbH-like protein whose function is uncertain, but is often associated with NRPS-assisted aryl-containing natural products; a major facilitator superfamily (MFS) efflux transporter (Hotta et al. 2010); the *sfp* gene encoding a 4′-phosphopantetheinyl transferase, essential for proper post-translational activation of DhbB and DhbF (Ollinger et al. 2006); and a homolog of *ubiC*, chorismate pyruvate lyase whose function in bacillibactin synthesis is unclear (Hotta et al. 2010).

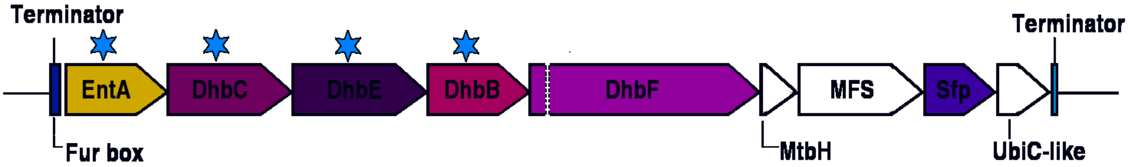


Figure 1 Structure of the bacillibactin biosynthetic operon. Blue asterisks mark genes selected for structure determination by CSGID.

Analogous to *E. coli* enterobactin, the ferric uptake regulator Fur regulates bacillibactin biosynthesis. Bacillibactin is expressed only under iron-limited conditions (Baichoo et al. 2002), regardless of growth aeration (Lee et al. 2011). Availability of iron in concentration of 20 μM is sufficient for nearly complete repression of the accumulation of bacillibactin (Ollinger et al. 2006). The bacillibactin operon is also upregulated by oxidative stress as the highest accumulation of bacillibactin was observed in conditions of low aeration and iron-depletion (Lee et al. 2011). In *B. subtilis* expression of DhbA, DhbB, DhbC, and DhbE is induced by high salinity and corresponding iron limitation (Hoffmann et al. 2002).

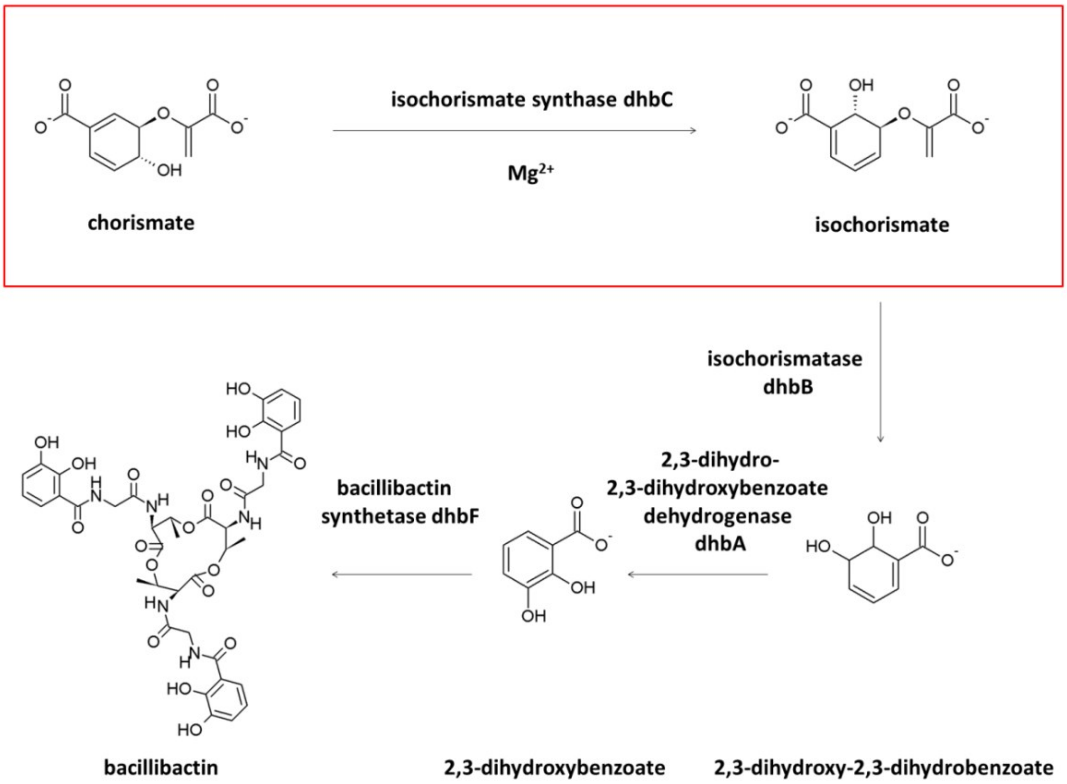

isochorismate synthase dhbC
Mg^{2+}
chorismate
isochorismate
isochorismatase dhbB
2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase dhbA
bacillibactin synthetase dhbF
bacillibactin
2,3-dihydroxybenzoate
2,3-dihydroxy-2,3-dihydrobenzoate

Figure 2 Pathway of bacillibactin biosynthesis in *B. anthracis*. Reaction performed by isochorismate synthase DhbC was outlined with red frame. Image reprinted from an original article (Domagalski et al. 2013)

1.2 Protein X-ray crystallography

Atomic structures of macromolecules are very important for studying their function and way of operation in biological systems. The first three-dimensional structure of protein, the structure of myoglobin, was determined with use of X-ray crystallography by John Kendrew in 1958 (Kendrew et al. 1958). Since then it is continuously been the most commonly used method for structure determination of macromolecules. In essence, to solve a three dimensional protein structure using X-ray diffraction, protein needs to be purified, crystallized and the crystals are subjected to diffraction experiment in an intense X-ray beam. A crystal mounted on a goniometer is gradually rotated on one axes of the goniometer. Some X-rays are diffracted by the electron clouds of crystalline protein atoms resulting in a different two-dimensional diffraction pattern for each angle of rotation. The positions of the reflections are characteristic of the lattice spacing and symmetry of the crystal, but the intensities of the reflections vary based on the contents of the crystal. The three dimensional electron density map can be determined by a crystallographer by sophisticated process of finding phases by SAD, MAD or MR techniques (Drenth 1999). Assuming that we know the polypeptide sequence, in many of the cases model building can be done automatically using modern crystallographic software. Crystallographers need to use his/her own experience in combination with sophisticated validation tools to complete a series of tasks to generate a final model. This includes verifying the correctness of the automatically built model, filling in unmodeled fragments of the polypeptide, modelling ligands incorporated into the crystal (both intentionally as well as unexpectedly), refining stereo-chemical properties of polypeptide bonds, and choosing the most probable and best fitting side chain rotamers of amino acids.

Despite the undeniable advantages, X-ray crystallography is not a trouble free method. The main limitation of X-ray crystallography of macromolecules is a requirement of diffraction-quality protein crystals. Protein crystallization is a difficult process that is different for every protein construct. Small and average-sized globular proteins with rigid structure are more likely to crystallize and form well diffracting crystals. On the other hand, flexible multi-domain proteins very often fail to produce well-ordered crystals. It is common for electron density maps to be absent or difficult to interpret in regions of flexible loops or the N- and C-terminal ends of polypeptides.

Another significant disadvantage of protein crystallization is the need for large quantities of the purified protein of interest. This often necessitates the use of recombinant proteins and their overexpression outside of the source organism.

1.2.1 Protein Crystallization

Protein crystals were studied a long before the discovery of X-rays beams by Röntgen in 1895. The first characterized protein crystals were earthworm hemoglobin described by Hünefeld in 1840. Those crystals were obtained by dehydration of worm's blood between two slides of glass (Hünefeld 1840). The same rationale, slow evaporation of a concentrated protein solution that becomes supersaturated and induces nucleation is a foundation for many current protein crystallization techniques. Until the late 1930s when the first X-ray diffraction images of hemoglobin and chymotrypsin crystals were recorded (Bernal 1938), protein crystallization was used mainly for purification purposes (Luft et al. 2014). Protein crystallization is a critical step for structure determination by X-ray crystallography as only pure, regular and large enough crystals can provide a good quality diffraction data that will allow the determination of high-resolution model of the molecule.

Proteins are usually soluble at physiological conditions, but in a supersaturated solution, the protein concentration exceeds the solubility limit of the protein, resulting in protein precipitation or crystallization. Addition of salt or organic solvents to protein solution can result in precipitation caused by high ionic strength (Drenth 1999). The process of protein precipitation in solution of high ionic strength is called salting out. Protein crystals arise by a repeatable association of protein molecules that interconnect by non-native intermolecular, predominantly hydrophilic, interactions called crystal contacts. Native contacts between protein molecules are referred as biological contacts or oligomeric contacts and usually involve larger surface area with hydrophobic patches (Dasgupta et al. 1997).

There are three common stages during the crystal formation process for both macromolecules and small molecules. These stages are nucleation, crystal growth, and cessation of growth. First two stages occur in supersaturated solutions. Crystal formation begins with the nucleation stage when some critical amount of molecules aggregate in three dimensions creating a thermodynamically stable nucleus. The end of growth is caused by decrease of concentration of free molecules in solution or by build-up of impurities on crystal faces (Russo Krauss et al. 2013).

Solvent is an intrinsic and very important part of protein structure. In contrast to small molecule crystals, protein crystals have high solvent content, in the range of 40 to 60% for most of the cases or 20 to 80% in extreme cases (Trillo-Muyo et al. 2013). This feature causes protein crystals to be very fragile and sensitive to dehydration. Crystal spaces lined with ordered water molecules are called channels (Frey 1994). In addition to ordered solvent molecules, spaces of protein crystal that are filled by unordered water molecules are called cavities. Polar amino acid residues exposed at protein surfaces interact with water molecules, ions and other molecules dissolved in the solvent solution creating the hydration shell of the protein molecule. The hydration shell is mediated by hydrogen and electrostatic bond interactions with neighboring protein molecules (Salemme 1988). Additionally, nonspecific interactions like van der Waals and hydrophobic interactions are also involved in formation of protein-protein contacts. Protein crystallizability and the contribution of specific and nonspecific interactions in crystal contacts varies between proteins and it is dependent on many factors including the identity of the precipitant and its concentration, protein concentration, additives, temperature, buffer identity, crystallization technique, pressure, detergent, magnetic and electric fields, but most importantly pH and the ionic strength (Salemme 1988; Kierzek and Zielenkiewicz 2001; Russo Krauss et al. 2013).

Proteins are large, flexible, and dynamic molecules. Therefore, protein crystals are sensitive to dehydration, change in temperature, pH, or ionic strength. Change in any of these parameters may affect crystal growth. Because proteins are much larger than small molecules, unit cells of protein crystals are bigger and crystals grow slower. Moreover, protein crystals are also smaller and less well ordered. Unfortunately, nowadays protein crystallization is still a process of trial and error.

1.2.2 Diffraction

Diffraction from a three-dimensional periodic structure such as atoms in a crystal is called Bragg diffraction in honor of William Lawrence Bragg and his father Sir William Henry Bragg, who explained this phenomenon (Bragg 1913). Bragg found that a diffraction pattern is a result of reciprocal interference between X-rays that are scattered by parallel crystal planes. The angle of scattered beam is equal to the angle of the incidence beam. If the difference in the path-length of the scattered beam is equal to integer number of wavelengths, then the scattered beam will be subjected to constructive interference (Figure 3). Bragg explained this phenomenon with an equation, which is commonly known as Bragg's law:

$$\mathbf{n\lambda = 2dsin\theta},$$

$\boldsymbol{\lambda}$ is the X-ray wavelength (where $\lambda \leq 2d$), **d** is the distance between crystal planes, $\boldsymbol{\theta}$ is the angle between the incident beam and crystal plane, **n** is the order of the diffracted beam (integer number).

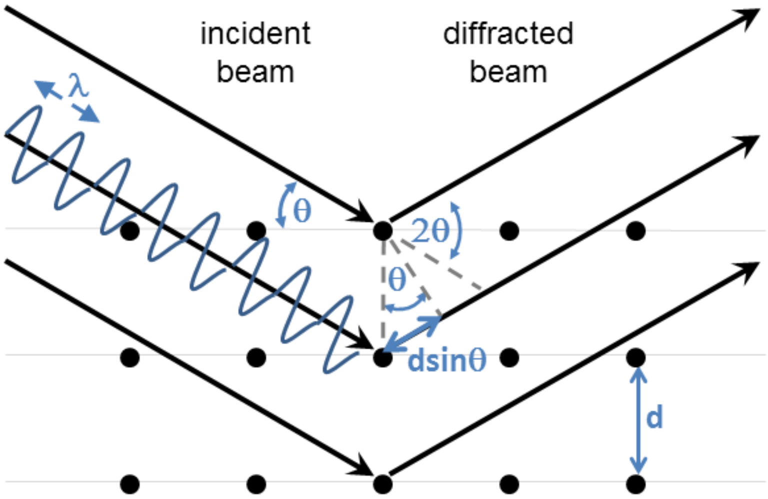


Figure 3 Constructive interference of X-ray waves explained by Bragg's law.

1.2.3 Phase problem

The electron density in a crystal at any position (xyz) can be obtained by calculating the Fourier summation: $\boldsymbol{\rho(xyz) = 1/V \sum |F_{hkl}| \exp(i\alpha_{hkl})\exp(-2\pi ihx+ky+lz)}$, where *hkl* are measured intensities, V is the volume of the unit cell, and α_{hkl} is the phase corresponding to the structure-factor amplitude $|F_{hkl}|$ (Taylor 2003). In an X-ray diffraction experiment, we obtain the intensities of the diffracted X-rays measured at the detector. The amplitude of the wave is proportional to the square root of the intensity, but information about its phase is lost. In macromolecular X-ray crystallography, three approaches are used for recovering the phases:

- isomorphous replacement (SIR / MIR) is recovering phasing information with use of heavy-atoms derivatives of isomorphous crystals,
- anomalous diffraction (SAD / MAD) is based on the presence of sufficiently strong anomalous scattering atoms within the protein crystal,
- molecular replacement (MR) is utilizing phases obtained for the homologous protein or the same protein in a non-isomorphic crystal.

1.2.3.1 Isomorphous replacement (SIR/MIR)

Isomorphous replacement is a method for the phase determination based on determination of the contribution of a heavy atom derivative to structure factors of the sample. Diffraction data for the native and isomorphous, heavy atom soaked crystals are needed in order to calculate contribution of the heavy-atom to each structure factor. The structure factors of heavy-atom derivative crystal are the vector sum of the heavy atom structure factor and native crystal structure factors. The contribution of heavy atoms to each structure factor can be calculated using the Patterson function or direct methods. The method is called single isomorphous replacement (SIR), when a single heavy atom is used and multiple isomorphous replacement (MIR), when multiple heavy atoms are used (Drenth 1999; Taylor 2003).

1.2.3.2 Anomalous dispersion (SAD/MAD)

Friedel’s law says that Bragg reflections related by inversion through the origin (i.e., Friedel’s pairs) have equal amplitudes and opposite phase. If the wavelength of the X-rays correspond to the energy of transitions between electron shells of the heavy atom, it will result in phase modification (Drenth 1999). This phase shift results in breaking Friedel’s law and differences between the measured intensities of Friedel pairs. The atomic scattering factor is given by $f + f' + i f''$ where f' and f'' are the real and imaginary parts of the anomalous dispersion correction, and i is a 90° phase shift between these two components. Typically, for SAD/MAD technique, protein methionine residues are substituted with selenomethionine residues and the anomalous scattering is measured in single crystal. The SAD method uses data collected at the peak of anomalous atom diffraction, and the MAD technique additionally uses data collected at inflection point and remote wavelength (Taylor 2003).

1.2.3.3 Molecular Replacement

Molecular replacement (MR) is an approach to solve the phase problem by using a homolog with known structure or even a structure of the same protein in a non-isomorphic crystal. Assuming that r.m.s.d between Cα atoms of the homologous model and the target structure is low, a homologous model can be used for calculation of the initial phases (Taylor 2003). An initial density map can be obtained for a structure using the Patterson function, which discards the phases and using squared amplitudes. The principles of this technique were proposed by Rossman and Blow (Rossman 1962). The first step is to deduce the number of molecules, their orientation, and accurate placement in the target unit cell. Once the MR model is properly oriented and positioned in the unit cell, it can be used to calculate the phases, which in combination with observed structure factors allow calculation of electron densities, and subsequently for building and refinement of the sought structure (Drenth 1999). Structures solved by molecular replacement may contain errors due to the possibility of phase bias. Parts of the model may be wrong, but the map may not show this.

1.3 Structural Genomics (SG)

Structural genomics is a high-throughput (high-output) approach to structural biology, a worldwide effort for determination of three-dimensional structures for all proteins and other gene products that are encoded by complete genomes (Brenner 2001). Pilot SG projects started in late ’90s after sequencing of the first complete genomes. Initially, mapping of the protein universe (Vitkup et al. 2001) and development of high-throughput methods were the primary concerns. The two main aims were to solve a representative set of all proteins that do not show significant sequence similarity to proteins of known structure and provide insight into their function by recognizing homology between proteins that share the same fold regardless of divergent sequences (Brenner and Levitt 2000). Additionally, the novel structures were utilized as templates for homology modeling of millions of protein models. This approach increased the structural coverage of proteins (including reliable homology models) from 30% to 40% (contributing ~50% of the newly characterized families) over the last ten years (Khafizov et al. 2014). Despite the development of novel technologies and thousands of structures, SG projects were criticized for producing large number (i.e., 26% of all structures SG deposited to PDB (Chruszcz et al. 2010)) of structures that are missing functional assignment or their function is referred to as putative. Therefore, the largest structural genomics project, the Protein Structure Initiative, currently named PSI:Biology, shifted its focus to the application of previously developed high-throughput structure determination pipelines via highly organized networks of investigators to research important biological and biomedical problems (SBKB 2015).

1.3.1 Pilot structural genomics projects

The era of SG research started in 1995 with the proposal of the first structural genomics project in Japan. Two years later the pilot project started at the RIKEN institute. The same year in USA, Department of Energy (DOE) and National Institute of General Medical Sciences (NIGMS; one of the National Institutes of Health) started the initial phase of structural genomics in the United States. The New Jersey Initiative in Structural Genomics and Bioinformatics was established. In January 1998, a workshop on Structural Genomics was held at Argonne National Laboratory in USA and initial pilot projects started in Germany, Canada, and USA. In October of 1998, the Structure-Based Functional Genomics meeting took place at Avalon in USA. In June 1999, a call for grant applications for NIGMS/NIH pilot projects was announced. The year 2000 was breakthrough year. In January 2000, OECD Committee for Scientific and Technological Policy (CSTP) proposes to initiate SG studies. The First International Structural Genomics Meeting took place in April in Hinxton, UK. In September, NIGMS started the Protein Structure Initiative, establishing seven SG centers. In

November of 2000, First International Conference on Structural Genomics took place in Yokohama, Japan (MCSG 2014).

1.3.2 Protein Structure Initiative (PSI)

The Protein Structure Initiative (PSI) is the largest ongoing structural biology project established in the year 2000 by NIGMS. Nine pilot centers, i.e., Joint Center for Structural Genomics (JCSG), Midwest Center for Structural Genomics (MCSG), Northeast Structural Genomics Research Consortium (NESGC), New York-Structural GenomiX Research Consortium (NYSGXRC), Center for Eukaryotic Structural Genomics (CESG), Berkeley Structural Genomics Center (BSGC), Southeast Collaboratory for Structural Genomics (SECSG), TB Structural Genomics Consortium (TB), and Structural Genomics of Pathogenic Protozoa SGPP were established during the initial phase of the project. The first phase was dedicated to development of methodology for a subsequent production phase, testing the feasibility of high-throughput structure determination, and solving unique protein structures (Lee et al. 2011). During PSI-1, which lasted from October 2000 to June 2005, PSI centers produced 1416 protein structures, providing the first structure representatives for 355 (2.9% of all) PFAM families (SBKB 2011). PSI-2 lasted from July 2005 to June 2010 and focused on implementing the methods developed in PSI-1, homology modelling and addressing bottlenecks, e.g., modelling membrane proteins (Lee et al. 2011). The number of research centers was increased to 14 and additionally two resource centers were established: the PSI Structural Biology Knowledgebase (SBKB)(Berman et al. 2007) and PSI Materials Repository (PSI-MR). During PSI-2, 3786 structures were solved (SBKB 2011) out of which 561 (4.6% of all PFAM) are the first structural representatives of PFAM families. PSI: Biology, the third and the last phase of PSI started in July 2010 and is focused on utilizing the high-throughput structure determination pipelines to answer broad and challenging biological questions (Montelione 2012). The PSI:Biology research network is organized around 4 centers for high-throughput structure determination, 9 centers for membrane protein structure determination and 15 high-throughput enabled structural biology partnerships (SBKB 2015).

To date, PSI centers solved ~51.5% (as on 8 January 2015; (RCSB 2015)) of all SG structures. PSI researchers developed an impressive number of new technologies, including among others auto-induction media (Studier 2005), a wheat germ cell-free protein production system (Vinarov et al. 2006), and a whole range of methods for improvement of crystallization, i.e., surface entropy reduction (Derewenda and Vekilov 2006), in situ proteolysis (Dong et al. 2007), large-scale reductive methylation of lysine residues (Kim et al. 2008), nanolitre volume crystallization (Gerdts et al. 2008). Aforementioned methods and many other new vectors, expression systems, and experimental protocols decreased time and cost of protein structure determination. PSI also influenced computational modeling projects, i.e., Critical Assessment of Structure Prediction (Moult 2005) and Critical Assessment of Automated Structure Determination by NMR (Rosato et al. 2009) by providing the majority of targets.

1.3.3 Description of selected centers

1.3.3.1 Structural Genomics of Infectious Diseases

The Center for Structural Genomics of Infectious Diseases (CSGID) and Seattle Structural Genomics Center for Infectious Disease (SSGCID) are two consortia that were established by National Institute of Allergy and Infectious Diseases (NIAID) with the common goal of determining three-dimensional structures of proteins from human infectious pathogens (Anderson 2009; Myler et al. 2009). Both centers have their own state-of-the-art high-throughput gene-to-structure pipelines capable of determining the three-dimensional structures of proteins by X-ray crystallography and NMR (Figure 4). CSGID and SSGCID accept structure determination requests from the scientific community and assign to the requested targets the highest priority. Proposed proteins can be drug targets, important enzymes, virulence factors, vaccine candidates, and other proteins with biologically important role (Myler et al. 2009). Both centers target proteins from organisms classified into categories A-C in the NIAID Pathogen Priority List as well as organisms causing emerging and re-emerging diseases, and close homologs of those proteins from closely related organisms (Anderson 2009). The CSGID organisms of interest include members of *Bacilli* genus (i.e., *Bacillus*, *Listeria*, *Staphylococcus*, *Streptococcus*), Gamma-proteobacteria (i.e., *Coxiella*, *Escherichia*, *Francisella*, *Salmonella*, *Shigella*, *Vibrio*, *Yersinia*), Clostridia (*Clostridium*), Epsilon-proteobacteria (*Campylobacter*, *Helicobacter*), dsDNA viruses (*Orthopoxvirus*, *Rhadinovirus*, *Roseolovirus*, *Erythrovirus*), and ssRNA positive-strand viruses (*Calicivirdiae*, *Alphavirus*, *Coronavirus*, *Enterovirus*, *Flavivirus*, *Hepacivirus*, *Hepatovirus*, *Hepevirus*) (Anderson 2009). Other organisms from A-C categories in the NIAID Pathogen Priority List are covered by SSGCID. Targets may include also other human pathogens (with the exception of human immunodeficiency virus) and their phylogenetically related organisms. All structures produced by the consortia are submitted to the Protein Data Bank (PDB), and all materials (clones and protein) generated are publicly available. Experimental procedures and weekly target status reports are submitted to the TargetTrack database. The CSGID and SSGCID experimental results are publicly available through the project websites: http://www.csgid.org/ and http://www.ssgcid.org/, respectively. **The database management system UniTrack that is described in this work was developed specifically for the CSGID.**

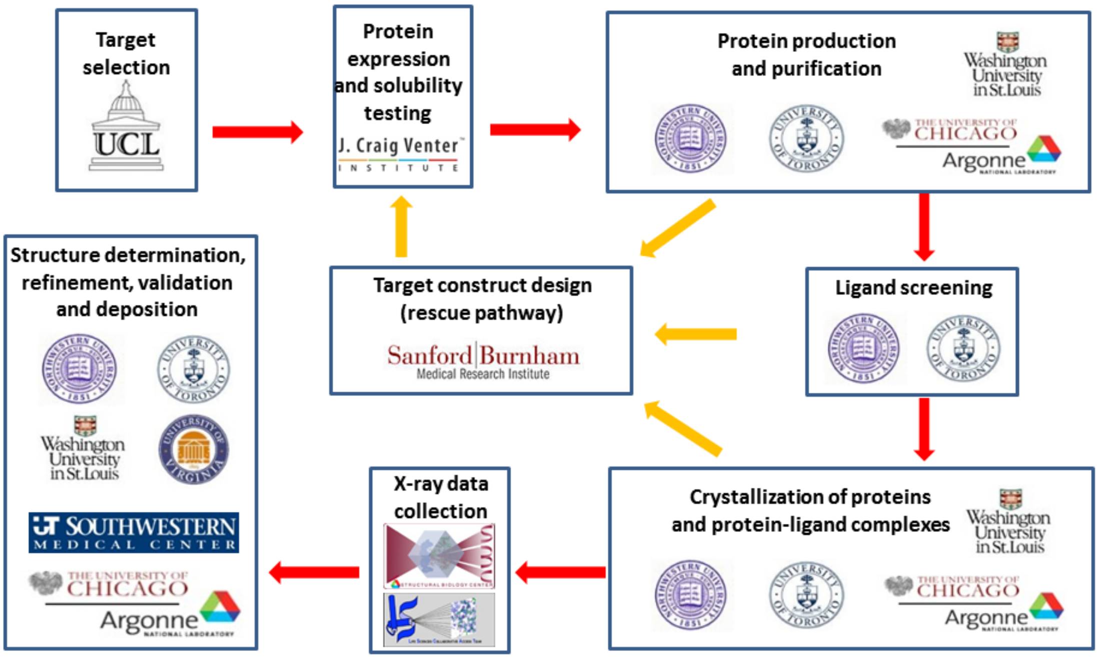

Target selection
UCL
Protein expression and solubility testing
J. Craig Venter INSTITUTE
Protein production and purification
Washington University in St.Louis
THE UNIVERSITY OF CHICAGO
Argonne NATIONAL LABORATORY
Structure determination, refinement, validation and deposition
Target construct design (rescue pathway)
Sanford|Burnham Medical Research Institute
Ligand screening
UT SOUTHWESTERN MEDICAL CENTER
X-ray data collection
Crystallization of proteins and protein-ligand complexes

Figure 4 Diagram of the CSGID structure determination workflow. Red arrows indicate the direction of the standard workflow, while yellow arrows point the alternative/rescue pathways.

1.3.3.2 Midwest Center for Structural Genomics

The Midwest Center for Structural Genomics (MCSG) is a large-scale SG center that was established in the year 2000, during the initial phase of PSI. The main aim of MCSG was to increase the structural coverage of protein superfamilies by the efficient determination of protein structures using X-ray crystallography and advancement in purification, crystallization, data collection, structure solution, and computational methods (MCSG 2014). In result, the center produced over 1000 structures during the first two phases of PSI (Lee et al. 2011). During the current PSI:Biology phase, MCSG is pursuing three scientific programs: proteins associated with virulence in human pathogens, proteins overrepresented and associated with disease in human microbiomes and proteins involved in signaling and transcription regulation (MCSG 2014). Center is organized around seven highly integrated cores: Bioinformatics, Gene Cloning and Protein Expression, Eukaryotic and Viral Proteins Expression, Purification and Crystallization, Data Collection and Analysis, Structure Determination, and Databases and Laboratory Information Management System (LIMS) (MCSG 2014). One of the main considerations of MCSG is data dissemination, which is done through peer-reviewed publications, the PSI-Knowledgebase (PSI:KB), the PSI-Materials Repository (PSI MR) and by maintenance of the database of the experiments and connected knowledge dissemination portal. Since the beginning of the PSI:Biology phase, MCSG is using the UniTrack system for data management. The MCSG data dissemination portal and target tracking database are publicly available through the project website, http://www.mcsg.org/.

1.3.3.4 New York Center for Structural Genomics

The New York Structural Genomics Research Consortium is one of the four large-scale SG centers established during the pilot phase of PSI. During that period the project was based on collaboration of PSI and industrial laboratories and aimed to develop modular technologies that could be utilized in structural biology laboratories in both academia and industry (Bonanno et al. 2005). The main achievement of the first 5 years of NYSGRC was the high-throughput gene to structure pipeline, which to this day produced over 1300 structures deposited in PDB. Throughout the second phase of

PSI, the project was focused on proteins that share less than 30% identity to any protein with known structure. In the current PSI:Biology phase, the project is focused on some high-priority targets including multidomain eukaryotic proteins, multi-component assemblies, secreted proteins, protein phosphatases with the emphasis on human phosphatases, and members of two large protein superfamilies: enolase and amidohydrolase (Almo et al. 2007; Pieper et al. 2009; Sampathkumar et al. 2010). To meet the challenges introduced by new demanding targets, the NYSGRC structure has been reorganized. One of the main changes is a new data management platform based on the LabDB LIMS and a specifically adapted UniTrack system. The NYSRGC experimental data and protocols can be accessed through its web portal: http://kiemlicz.med.virginia.edu/nysgrc/.

1.3.3.5 Enzyme Function Initiative

The enzyme Function Initiative was founded by NIGMS with the main goal to develop large-scale sequence/structure-based strategy for functional assignment of unknown enzymes discovered in genome projects (Gerlt et al. 2011). EFI is not structural genomics center, but its multidisciplinary strategy is being developed and put into practice by specialized scientific cores, including a protein core, structure core, microbiology core, computation core, and data management core. During the first phase of the grant, five bridging projects groups focused on large and functionally diverse protein superfamilies, i.e., amidohydrolases, enolases, glutathione transferases haloacid dehalogenases, and isoprenoid synthases. Each of these superfamilies contains at least 10,000 members. After the first three and half years the research focus was changed to functional discovery in solute binding protein components of transport systems and novel pathways unique for human gut microbiota (Gerlt 2014). The EFI's data management core was established to distribute experimental results to community and most importantly to create data management infrastructure. The data management platform that is used by EFI is based on the LabDB LIMS and UniTrack system. UniTrack derived database of the EFI experimental data can be accessed online: http://kiemlicz.med.virginia.edu/efi/.

1.4 Importance of data management for SG projects

These days when life-sciences research is frequently done by large scale and highly automated scientific organizations, databases and specialized computer software are a prerequisite for efficient experimental data analysis. Design and effective usage of such tools is not an easy task and requires a deep understating of handled data by computer programmers and close cooperation with users during software design and development. The size, complexity, and heterogeneity of data are constantly growing, which makes data management more and more challenging. Structural biology is not an exception from that rule. A single protein project may require many repetitions of the various steps due to difficulties at different levels of the structure determination pipeline. High-throughput techniques are becoming more accessible and even traditional laboratories use crystallization robots that perform large amounts of crystallization trials (Prilusky et al. 2005). Experimental observations may be additionally used for data mining studies that would benefit the success rate of protein production and structure determination experiments. Some of the SG consortia, including Northeast Structural Genomics Consortium (NESGC) and Joint Center for Structural Genomics (JCSG), successfully applied aforementioned approach. NESGC developed a decision tree algorithm for prediction of the protein solubility (Bertone et al. 2001) and JCSG identified features that correlate with protein crystallization and combined them into single score referred to as ‘crystallization feasibility’ (Slabinski et al. 2007).

The data management issues in SG were raised for the first time at the OECD Global Science Forum Workshop on Structural Genomics that was held in Florence in June 2000 (OECD 2000). Scientific delegates of OECD member countries identified three main issues as particularly important for the structural biology projects. In the first place, the delegates pinpointed the need for a stable and permanent funding of databases and for development of bioinformatics tools. Next, they emphasized the necessity to store structural and functional data in publically available data banks. This is also applicable to protocols for cloning, expression, crystallization, and structure determination. Finally, the need for better sharing of structural work between laboratories distributed worldwide was highlighted. The delegates established that it is important to limit the duplication of efforts between the SG centers, despite the fact that a small amount of overlap is beneficial for improving the quality of protein structures (OECD 2000).

The most fundamental roles of data management system, such as documentation, organization, and data sharing, can be done with simple tools like spreadsheets or notebooks. However, large-scale projects need data management systems that not only simply store data, but also provide intuitive, efficient, and secure access to it, allow annotation and modification of its content, allow easy sharing of the data with use of readable hyperlinks and commonly accepted data formats. Ideally, data should be linked to data stored in other databases and repositories. Databases should contain substantial amount of metadata giving complete representation of information. Above all, the logic and design of such system should impose only minimal changes in work organization and current data formats. It should be flexible to adapt to the specific needs of the laboratory and have the possibility to add new functionality. The innovative SG projects evolve during their lifetime requiring from data management system to be easily upgradeable. Finally, a system must be made with main goal to overcome specific needs of its users. Systems developed without cooperation with perspective users may be full of misconceptions or too complicated to use.

Data management in high-throughput structural biology is usually concerned at two distinct levels: the *target tracking* level and *experiment tracking* level (Zimmerman et al. 2014). The experiment tracking includes all of the information and metadata that is typically collected by LIMS's and the target tracking level contains processed data, annotations, and links to external resources. However, it is an arbitrary classification and some of the information is relevant on both levels, in general, a majority of systems is structured in similar manner.

1.4.1 Protein Data Bank

The Protein Data Bank (PDB) is of greatest importance for data management in the field of structural biology. It is a freely and publicly available repository for the three-dimensional structures of macromolecules and related information. The PDB was founded in 1971 at Brookhaven National Laboratories (BNL) (Bernstein et al. 1977) as one of the very first biological data repositories. At the beginning, the data bank contained only seven X-ray structures and nowadays, after forty-three years of existence, it has grown to over 108,000 structures solved by X-crystallography, NMR and electron microscopy. The repository is currently growing at impressive rate of approximately 10,000 structures per year (RCSB 2014) (observation based on data for years 2013 and 2014). Since the early 1990s, the majority of scientific journals followed by some of funding agencies started to require deposition of structure coordinates to PDB for all new structures (Berman et al. 2000), following the guidelines of International Union of Crystallography (IUCr). In 1998, the management of the repository was handed over to the Research Collaboratory for Structural Bioinformatics (RCSB). Five years later, the Worldwide PDB (wwPDB) was established. Since then, depositions, data processing and distribution of structural data are carried out in parallel by RCSB PDB (USA), PDBe (Europe), and PDBj (Japan) that together maintain the single PDB archive (Berman et al. 2003). In 2006 the BMRB (Biological Magnetic Resonance Data Bank) group (USA), responsible for maintenance of data produced by NMR Spectroscopy, joined the wwPDB (Berman et al. 2007). The PDB archive consists of flat files, which are distributed via HTTP and FTP protocols, and each member of the wwPDB provides own view of the data. Model coordinates are distributed in three formats: mmCIF, PDB, and PDBML. PDB is not only repository of 3D structures of macromolecules, but each file with coordinates also contains a description of structure determination experiments and their results. The X-ray diffraction data files contain structure factors - the intensity and phase of the X-ray spots in the diffraction pattern from the structure determination experiment (Berman et al. 2007).

1.4.2 PSI-Nature Structural Biology Knowledgebase

One of the most important outcomes of the second phase of PSI was development of a freely available knowledgebase and data dissemination center called PSI-SBKB. The PSI-SBKB web portal is made in collaboration with Nature Publishing Group (NPG) (Berman et al. 2009). The knowledgebase is centered on the Target Track, a registration database that is monitoring experimental progress and status of protein targets selected for structure determination by PSI and other high-throughput structural biology projects. Initially this was done by two separate tools: TargetDB (Chen et al. 2004) and PepcDB (Kouranov et al. 2006), which were developed for the purpose of a tracking overall experimental status of SG targets and details of purification experiments, respectively. The knowledgebase is also a host for the homology modeling portal, technology portal, and the functional annotation module. In order to promote the advances of PSI and structural biology in general, the system is using several data dissemination routes. One of them is monthly newsletter prepared in collaboration with NPG, which gives the scientific community an insight into the most interesting SG research, new methods, and technologies developed by SG consortia. Set of web tools implemented in PSI-SBKB allows users to check if a protein of interest is being investigated by any of the PSI centers or is one of the protein structure requests. It also searches for related proteins, homology models, and protocols related to the expression, purification, or crystallization of the protein. Additionally, users can check experimental status of protein and availability of DNA clones in PSI:Biology materials repository. Access to the data is not restricted and users can search for the protein using its sequence, macromolecule name, organism, or database identifier, i.e., PDB, PFam etc. It is also possible to filter targets by the experimental progress status, availability of materials and protocols used (Berman et al. 2009; Gabanyi et al. 2011). PSI-SBKB helps to limit the overlap between SG consortia and monitor their progress. Nevertheless, the large-scale SG centers need the data management systems not only for the purpose of providing the data to the scientific community but especially to prioritize targets, effectively manage the vast amounts of experimental data, keep track of experiments, and adjust experimental strategies (e.g., choice of expression vectors, sequence truncation, crystallization conditions, structure determination procedures). These needs require gathering far more data than is required by TargetTrack. Consequently, most of the SG centers developed their own more specialized and comprehensive data management solutions.

1.4.3 Xtrack

One of the first specialized tools for data management in the field of structural biology was Xtrack (Harris and Jones 2002). The system consists of a relational database with PHP web interface and serves as an electronic notebook that keeps track of protein crystallography projects. The idea behind the software is to replace traditional lab notebook with more permanent and easier to share web-based alternative. The system handles all experiments from protein expression to deposition in PDB. Data is organized around X-ray diffraction dataset, which is referred by the authors as a ‘collection’. Each collection belongs to higher entity called project. Collection is presented on ten separate web pages that contain from 5 to 20 data items and exactly correspond to the data structure of the database. Aforementioned pages contain information about protein chemistry, expression, crystallization, X-ray data collection, data reduction, structure solution, refinement, analysis, and deposition. Useful feature of the program is capability to extract data from log files of the most popular crystallographic programs, i.e., CNS (Brunger et al. 1998), SCALEPACK (Otwinowski and Minor 1997) and REFMAC (Murshudov et al. 2011). Xtrack was not developed to handle large-scale projects and it definitely cannot compete with elaborate data management systems developed for use of SG consortia, but smaller laboratories will appreciate its benefits. It is easy to use and maintain, and it works well for documentation of stretched out in time projects and for sharing results with collaborators.

1.4.4 Sesame

Sesame is a LIMS that was developed specifically for data management in SG projects by the Zsolt Zolnai group at the Center for Eukaryotic Structural Genomics (CESG). The system is composed from a central relational database and set of web based Java applet applications (Zolnai et al. 2003). Sesame, in contrast to traditional LIMS, is not build around a single workflow, but it is focused on objects (e.g., protein, protocol, and screen) that can be linked freely. Users can adjust the system to their own needs by choosing only those modules, which provide functionality relevant to them. The main module handles wet-lab as well as NMR, X-ray crystallography, and cryo-EM structure determination experiments. Additional modules provide support for target requests, lab administration, equipment schedules, metabolomics, crystallization conditions browsing, and cryo-EM and NMR screens. Data can be accessed using set of views, i.e., ORF, protein, solution, crystal, NMR, small molecule, structure deposition, sources (vendors), target submission. Sesame is capable of automatic collection of some types of data, e.g., gel scans, NMR, MS, and UV-VIS data (Zolnai 2014). The system can generate various reports including a weekly TargetTrack report and output data to XML and csv files. Sesame provides tools for managing collaborative projects allows collaborators to have adequate access to data and modify them. The system is also equipped in its own intuitive query system. The authors of the system made it available outside of the consortium for individual investigators worldwide. New York Consortium on Membrane Protein Structure (NYCOMPS) and Promega Corporation are using it, among others.

1.4.5 HalX

HalX is another LIMS based system that was developed for handling SG data at the Yeast Structural Genomics Laboratory, in collaboration with the European SPINE (Structural Proteomics In Europe) project (Prilusky et al. 2005). The system was designed specifically for the high-throughput pipeline used by SPINE. It was developed in PHP scripting language and uses the PostgreSQL database-management system. HalX data model was organized using five categories with database tables that contain information about targets, data from public databases, data relevant to all experiments, experiment-specific data, and core of the data model. The user interface has six views: add experiment, modify experiment, default templates, view experimental results, superuser view, and administrators view (Prilusky et al. 2005). Particularly useful feature is the ‘default templates’, which allows saving the default protocols and using them for faster upload of new records. HalX has a progress page that allows monitoring global progress of all targets. Detailed progress for each target is presented in graph of linked experiments. Clicking on an experiment box displays its experimental details. The system also provides web services for primers design (bestPrimers), choice of the most suitable expression system (suggestES), and verification of cloned sequence (verifyCloning). HalX is distributed under General Public License (GPL) and can be downloaded from http://sourceforge.net/projects/halx/.

1.4.6 SPINE

SPINE (Structural Proteomics In the NorthEast) is a LIMS-based data management system developed to manage protein production pipeline of the Northeast Structural Genomics Consortium (Bertone et al. 2001). The system is a target tracking database and web application that allows tracking of experimental progress using set of data mining features. Moreover, the database content was used as a training set for classification of soluble proteins using machine-learning approach (i.e., decision trees) (Bertone et al. 2001). SPINE is using a MySQL relational database and web interface written in PERL programming language. The SPINE data model (Bertone et al. 2001; Goh et al. 2003) closely mirrors pepcDB data model, with database tables for tracking target, construct, expression, purification, biophysical characterization, X-ray and NMR data, and protein structure data. SPINE data is associated with many local and external resources. Local resources include wiki-based web site, structure gallery, publication page, and target information bulletin board (Goh et al. 2003). Several NESG computational resources including SPINS NMR archival database at Rutgers University, Proteus crystallization database at Columbia University, PartsList and Gene Census databases at Yale University, and University of Toronto LIMS are also connected to SPINE (Goh et al. 2003). The web interface consists of several tools: SpineSearch, SpineStatus, SpineScoreboard, SpineStructuralGalleries, and SpineAlert (Albeck et al. 2006). Data for some of the targets were made publically available after determining the structure. Unfortunately, publically available information provided by SPINE is limited. Currently only experimental status and list of experimental samples are accessible from the web portal.

1.4.7 SPEX Db

SPEX Db (Structural Proteomics EXperimental Database) is the data management system developed for Montreal-Kingston Bacterial Structural Genomics Initiative (M-KBSGI). The system was successfully adapted to serve over 10 structural genomics projects in Canada. SPEX Db was designed to provide both target tracking functionality and LIMS archiving capability. Targets and experiments are accessible using search engine that allows filtering by ids, experimental status, or name of the experimenter. Navigation over the experimental stages of the structure determination pipeline is simplified by introduction of a tree view for a target. The view is a summary table, where columns from left to right correspond to subsequent stages of the pipeline. Table rows contain links to experimental records, which are colored according to experiment status (i.e., in progress, completed, cancelled) and open as new web pages. Each target record is linked with corresponding records in SWISS-PROT/TrEMBL, NCBI, InterPro, and other external databases. Addition of new experimental entries is done manually with use of web forms. The system has homology tools, which checks if any of the targets has homologs in TargetDB or PDB. Users are informed about change of the status for homologous targets by email every week. When sequence identity between target and any protein deposited in PDB is >25%, work on that target might be stopped. Access to the data is controlled by the identification and authority level of a user, which ranges from 0 to 9 (Raymond et al. 2004).

2. Objective

The main goal of this work was to determine the atomic structure of isochorismate synthase DhbC from *B. anthracis* and biochemically characterize this enzyme. The biochemical characterization should prove the function of DhbC and give basis for further investigation on potential inhibitors of chorismate-utilizing enzymes.

A parallel objective was to develop a set of tools that together with other applications developed in Wladek's Minor laboratory will constitute an innovative data management system for the high-throughput structural genomics UniTrack. The tools, the target tracking database for CSGID and a corresponding knowledge dissemination portal, will:

i. improve a data workflow and maximize efficiency of the CSGID high-throughput structure determination pipeline,

ii. allow documentation of experimental work and exchange of this information between groups involved in the project,

iii. satisfy specific needs of four groups of users: community requesters, the scientific community that is not directly involved in the project, CSGID scientists, and an advisory committee,

iv. connect various resources and tools used within the center and supplement those with links to external resources,

v. allow monitoring of work progress on the particular protein targets,

vi. allow monitoring of overall progress of the consortium,

vii. publicly release experimental data including detailed protocols,

viii. serve as information hub for the infectious disease scientific community,

ix. provide numerous a*d hoc* statistics and dashboards for the reports and analysis of the experimental work done within consortium.

3. Materials

3.1 Laboratory equipment

Laboratory experiments, i.e., molecular cloning, expression, purification, and crystallization experiments were performed using the following equipment:

Benchtop centrifuge	Beckman Coulter, Allegra® X-15R
Benchtop shaking incubators	VWR Scientific, 1575A New Brunswick, Innova 4000
Centrifuge	Beckman Coulter, AVANTI J-26XP
Gel imaging instrument	BioRad, Gel Doc™ EZ System
Fast Protein Liquid Chromatography (FPLC) systems	GE Healthcare Life Sciences, ÄKTAprime plus GE Healthcare Life Sciences, ÄKTAFPLC
Incubator	Labnet International Inc., Mini Incubator
Microplate reader	BMG LABTECH, PHERAstar FS
Screen preparation instrument	Emerald BioSystems, Opti Matrix
Shaking incubator	Labnet International Inc., 211DS Shaking incubator
Stereomicroscope	Olympus, SZX16
Ultrasonic Programmable Processor (Sonicator)	Misonix, XL2020
Liquid handling robot	TTP Labtech LTD, Mosquito HTS
Ultracentrifuges	Beckman Coulter, Optima™ L-80 XP Beckman Coulter, Optima™ XL-100K
UV-Visible spectrophotometer	Schimadzu Corp., UV-2450
UV-Visible spectrophotometer (micro-volume)	Thermo Scientific, NanoDrop 2000

3.2 Solutions, buffers and media

LB broth	Research Products International, Miller's LB Broth
M9 selenomethionine growth medium	Shanghai Medicilon, M9 SeMET High Yield
Lysis/Binding buffer	300 mM NaCl, 50 mM HEPES pH 7.5, 5% v/v glycerol, 5 mM imidazole, 0.5 mM phenylmethylsulfonyl fluoride, 1 mM benzamidine
Washing buffer	300 mM NaCl, 50 mM HEPES pH 7.5, 5% v/v glycerol, 30 mM imidazole
Elution buffer	300 mM NaCl, 50 mM HEPES pH 7.5, 5% v/v glycerol, 250 mM imidazole
Dialysis/AKTA buffer	300 mM NaCl, 50 mM HEPES pH 7.5, 0.5 mM TCEP
SDS-PAGE sample buffer	Novex®, Tris-Glycine SDS Sample Buffer
SDS-PAGE running buffer	Novex®, Tris-Glycine SDS Running Buffer
SDS-PAGE protein standards	Bio-Rad, Precision Plus Protein™ Unstained Standards
Coomassie stain	Bio-Rad, The Bio Safe™ Coomassie
Nickel-charged affinity resin	QIAGEN, Ni-NTA Agarose

3.3 Computer equipment

Computer programming, database development and *in silico* analyses were carried out using the following computer workstations owned by Wladek Minor laboratory:

'Anula' - personal workstation

Operating system	Ubuntu 12.04.2 LTS
Processor	Intel® Core™ i7-3770K Processor
Ram memory	16GB
Graphics card	GeForce GTS 250
Disk space	2x1.5TB in RAID-1 array

'Danuska' – CSGID-DB database server / CSGID web portal server

Operating system	Red Hat Enterprise Linux Server release 5.10 (Tikanga)
Processor	2xQuad-core (Intel® Xeon® Processor E5430) @ 2.66GHz
Ram memory	16GB
Graphics card	ATI ES1000
Disk space	2TB + 12 TB in RAID-1 partitioned through LVM

'Soroka' – MetaPDB database server

Operating system	Red Hat Enterprise Linux Server release 5.6 (Tikanga)
Processor	Quad-core Intel® Xeon® Processor 5130 @ 2.00GHz
Ram memory	8GB
Graphics card	ATI ES1000
Disk space	6x750GB drives paired into three RAID-1 arrays

3.4 Software

3.4.1 Experimental data processing and analysis

Gel documentation and analysis	Image Lab (Bio-Rad)
Gel filtration monitoring and analysis	PrimeView (GE Healthcare 2011)
Spectrophotometric enzyme assay	MARS Data Analysis (LABTECH 2011)
Spectrophotometric measurements and kinetics calculation	UVProbe (Shimadzu 1998)
Crystallographic data collection, integration and structure solution	HKL-3000 (Minor et al. 2006) HKL-3000 is integrated with: SHELXC/D/E (Sheldrick 2008), MLPHARE (Otwinowski 1991), DM (Cowtan and Main 1993; Cowtan and Zhang 1999), ARP/wARP (Perrakis et al. 1999), CCP4 (Winn et al. 2011), SOLVE, and RESOLVE (Terwilliger 2004)
Manual model building and validation	COOT (Emsley and Cowtan 2004)
Structure refinement	REFMAC5 (Murshudov et al. 2011)
Structure annotation	ICM Pro (Abagyan 1994), ActiveICM (Raush et al. 2009)
Structure validation	ADIT (Yang et al. 2004), MolProbity (Chen 2010)

Structure visualization	PyMOL (Schrödinger 2010)
Structure similarity search	DALI (Holm and Rosenstrom 2010)
Sequence similarity search	PSI-BLAST (Altschul et al. 1997), HHpred (Soding 2005; Soding et al. 2005)
Structure superposition	SSM (Krissinel and Henrick 2004)
Sequence clustering	CLANS (Frickey and Lupas 2004)

3.4.2 Computer programming and database development

Scripting languages	PHP5 (http://www.php.net/), Python (http://www.python.org/) with following extension packages: BioPython (Cock et al. 2009), NumPy and SciPy (http://www.scipy.org/)
Object-relational Database Management System	PostgreSQL (http://www.postgresql.org/)
Web application framework	CakePHP (http://cakephp.org/)
Integrated Development Environment	Eclipse (http://www.eclipse.org/)
Distributed revision control	Git (http://git-scm.com/)
Interactive tree browser used for visualization of structure determination pipeline	JavaScript InfoVis Toolkit (http://philogb.github.io/infovis/)
Statistics visualization	google charts JavaScript libraries (https://developers.google.com/chart/)

Molecule viewer web applet	Jmol (McMahon and Hanson 2008)
JavaScript toolkits and libraries	Jquery (http://jquery.com/), ExplorerCanvas (http://excanvas.sourceforge.net/), Scriptaculous (http://script.aculo.us/), modalBox (https://code.google.com/p/modalbox/), JSCalendar (http://jscalendar.codeplex.com/)

4. Methods

4.1 Experimental methods

4.1.1 Molecular cloning

X-ray crystallization studies require large quantities of homogenous protein that can only be obtained by overexpression of recombinant protein in an efficient expression host. In order to overexpress the protein of interest, its gene has to be inserted into a proper expression vector using molecular cloning. Jason Stam from CSGID team at J. Craig Venter Institute did the cloning of *dhbC* gene. Detailed protocol for the experiment is available through the CSGID web portal (http://csgid.org/csgid/data/protocols/CSG-003_PCR_and_LIC_v002.pdf). The CSGID high-throughput cloning pipelines use pMCSG7 as the primary expression vector and pMCSG19c, maltose-binding protein (MBP) fusion vector, for a 'salvage' strategy for proteins that show low solubility when expressed in pMCSG7. The open reading frame of *dhbC* was amplified by polymerase chain reaction from *B. anthracis str. Ames* genomic DNA using the forward 5′-TACTTCCAATCCAATGCGATGAATGAATTTACGGCTGTAAA-3′ and reverse 5′-TTATCCACTTCCAATGCTACTTTTCATTAAGTGAACTATC-3′ primers. The gene was cloned into a pMCSG7 plasmid using ligation independent cloning (Aslanidis and Dejong 1990; Haun and Moss 1992). Ligation independent cloning is an alternative for a traditional restriction enzyme cloning that is suitable for high-throughput applications (Eschenfeldt et al. 2009). The pMCSG7 is a fusion expression vector, which encodes N-terminal hexahistidine tag with eight residue spacer followed by the tobacco etch virus protease recognition site (shown as underlined), and an *Ssp*I restriction site (MHHHHHHSSGVDLGTENLYFQ/SN$^{\vee}$IGSG) (Stols et al. 2002). This vector also carries a TVMV protease, which allows *in vivo* his-tag cleavage. Sequencing of the vector revealed cloning artifact, i.e., a single point mutation that resulted in amino acid substitution F24L in the DhbC protein sequence.

4.1.2 Transformation

The pMCSG7-*dhbC* plasmid provided by JCVI was first amplified in a cloning host, *E. coli* XL10-Gold ultracompetent cells (StrataGene). The transformation reaction was performed using a 20 µL aliquot of competent cells and 1 µL of plasmid. For control 20 µL of cells were transformed with 10 pg of pUC18 control DNA. The reactions were incubated on ice for 30 min. Subsequently, each transformation reaction was heat-pulsed for 45 s in a 42°C water bath and instantly chilled by incubation on ice for 2 min. In the next step, 600 µL of LB medium was added to each transformation reaction and the reactions were incubated at 37°C for 1 hour with shaking at 225 rpm. Following transformation, 60 µL samples of the cultures were plated on LB-agar plates with 10 µg/ml ampicillin. The plates were incubated at 37°C overnight. On the next day, for the purpose of purification, 5ml LB medium was aseptically inoculated with single colony picked up from LB-amp-agar plate and incubated for 16 hours at 37°C to OD_{600} of ~4.0. The cells were pelleted by centrifugation at 8000 rpm for 3 min at room temperature. Plasmid was purified using QIAprep Spin Miniprep Kit (Qiagen) and high yield protocol provided with the kit. Purification yielded 40µL of 140 ng/µL of plasmid DNA. The amplified plasmid was transformed into expression host, *E. coli* BL21-CodonPlus (DE3)-RIPL competent cells (StrataGene). The transformation reaction was performed using the same protocol as for the XL10-Gold transformation, but the LB-agar plates and media contained ampicillin (100 µg/mL) and chloramphenicol (25 µg/mL). A glycerol stock of cell culture was made by mixing an overnight culture of transformed cells with equal volume of 50% glycerol, freezing, and storing at -80°C.

4.1.3 Expression

The initial protein expression, purification, and crystallization screening were performed by the CSGID group at the University of Toronto. The 100 µL of glycerol stock of BL21-CodonPlus cells with the pMCSG7-*dhbC* plasmid was revived in 50 ml of M9 SeMET High-Yield growth media containing 100 µg/ml of ampicillin, and grown with shaking at 37°C overnight. The next day, 4x10 mL of overnight culture was used to inoculate a fresh 4x1 L cell cultures. Cells were grown at 37°C to an OD_{600} of approximately 1.2. Then, protein expression was induced by the addition of 0.4 mM isopropyl-β-D-1-thiogalactopyranoside and the cells were grown overnight (~16 hours) with shaking at 20.0°C. Selenomethionine (SeMET) media was used to incorporate selenium in to the protein to permit the use of the SAD technique to determine unbiased crystallographic phases. The total weight of cells was 27.2 g. Harvested cells were flash-frozen in liquid nitrogen and stored overnight at -80 °C for more effective cell lysis.

4.1.4 Cell lysis

Cell disintegration or cell lysis is a process of extraction of intra-cellular components, e.g., an overexpressed protein from expression strain of bacteria, using a mechanical or chemical method. A combination of freeze-thaw with sonication was used for cell lysis of DhbC expression strain because of its high efficiency. Frozen cells were thawed in ice-water bath for about 90 min and suspended in 200 mL of lysis buffer with 4 EDTA-free protease inhibitor cocktail tablets (Roche, cOmplete). Next, thawed cells were dived into 50 mL batches and sonicated. Cells were sonicated on ice for 10 min with 10 s pulses and 10 s pauses at the maximum power of the ultrasonic processor. Sonicated cell lysate was clarified by centrifugation at 35,000 rpm for 45 min at 4°C.

4.1.5 Immobilized metal ion affinity chromatography

The supernatant of clarified cell lysate was applied to a nickel-charged affinity resin at 4°C. Prior to use, the resin was washed with 20x column volume of ddH20 and 10x column volume of washing buffer at room temperature. Resin with bound recombinant protein was washed overnight at 4°C with ~400 ml of washing buffer. The purified protein was eluted at a concentration of 7.9 mg/ml using 10 ml of elution buffer. The protein concentration was calculated from absorbance at 280 nm measured using NanoDrop micro-volume spectrophotometer.

4.1.6 His-tag cleavage

The hexahistidine tag was cleaved from the protein by the addition of 1 mg of recombinant His-tagged TEV protease per 15 mg of eluted protein in the presence of EDTA, TCEP and arginine (final concentrations 1 mM, 0.5 mM, and 200 mM respectively). Arginine was included in the buffer in order to suppress protein aggregation (Tsumoto et al. 2004; Arakawa et al. 2007). The cleavage was performed at 4°C overnight and continued during dialysis to cleavage buffer. Cleaved protein was separated from TEV protease by running over nickel-chelating resin (Domagalski et al. 2013).

4.1.7 Gel filtration chromatography

The gel filtration chromatography is a type of size exclusion chromatography which uses a hydrophilic packing material and an aqueous mobile phase to fractionate macromolecules (Lathe and Ruthven 1956). The eluted protein sample was concentrated to ~20 mg/ml using 10 kDa centrifugal filter (Amicon® Ultra 15 mL) and run in two 2 ml batches through the Superdex G200 column on AKTA FPLC workstation. Both gel filtration buffer and the protein sample were filtered through 0.22 μm membrane before application to the column. The gel filtration flow rate was 1 ml/min and the eluted protein was separated into 2 ml fractions. The sharp protein peak eluted at about 83 ml elution volume (Figure 5). The level of purification was checked using SDS-PAGE gel electrophoresis (Figure 7). The homogenous fractions corresponding to the main AKTA peak were combined together, concentrated to 16 mg/ml using 10 kDa centrifugal filter, divided into 100 μL aliquots and flash-frozen to -80 °C.

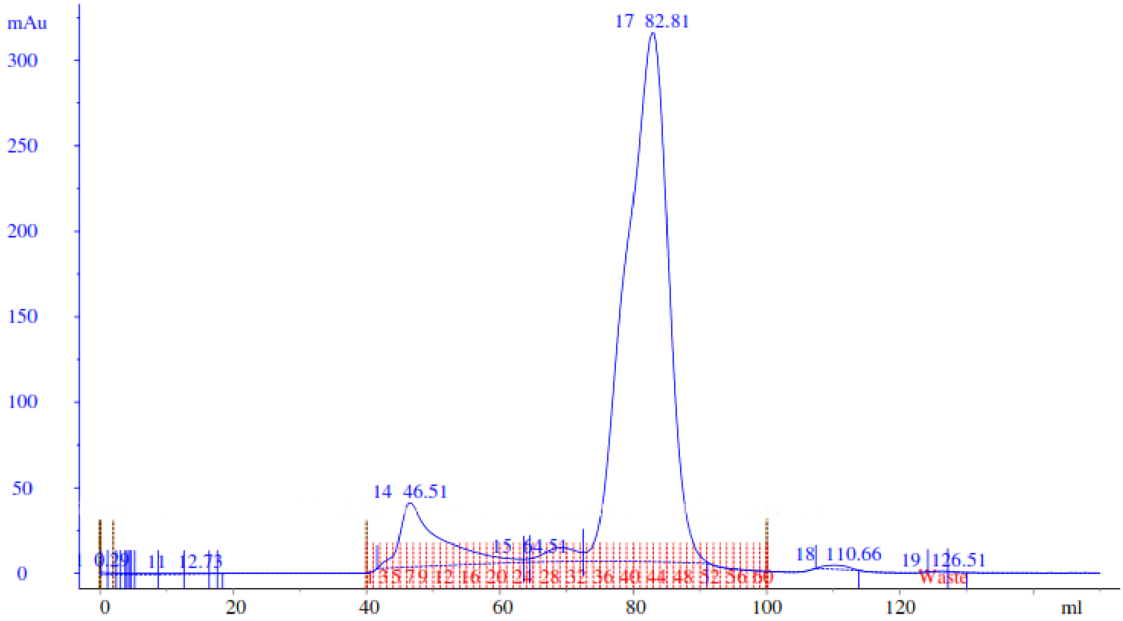


Figure 5 **Gel filtration chromatography elution profile of DhbC. The elution volume is plotted along the horizontal X-axis and absorption at 280 nm is plotted up the vertical y-axis. The red vertical markers on x-axis correspond to elution fractions. Red tick marks correspond to fractions collected for gel electrophoresis check.**

4.1.8 SDS polyacrylamide gel electrophoresis

SDS-PAGE was used to check the purity of protein samples (Figure 6), analyze fractions eluted from the AKTA FPLC (Figure 7), and check the approximate molecular weight of the purified protein. In this technique, the protein is denatured in the presence of an anionic detergent, sodium dodecyl sulfate (SDS) which interacts with hydrophobic amino acids and coats it with negatively charged sulfate groups. Due to the uniform distribution of negative charges, the proteins migrate in electric field towards the positive electrode inside the polyacrylamide gel (Shapiro et al. 1967; Laemmli 1970). The speed of migration, referred as electrophoretic mobility, depends on size, shape, and charge of the molecule. Binding of SDS makes proteins charge-to-mass ratio proportional to their molecular weight, allowing for fractionation based by approximate protein size (Garfin 2003).

Protein samples were prepared by mixing the protein with Laemmli sample buffer in 1:1 ratio, followed by denaturation at 95°C for 5 min and brief centrifugation at a speed of 14,000 rpm. Next, samples were separated on pre-cast mini gel (NuPAGE® Novex®, 4-12% Tris-glycine protein gel) immersed in Tris-glycine SDS running buffer. The electrophoresis was run at 120 V until protein marker reached the foot line of the gel cassette (~ 1 hour). Following electrophoresis, the gels were washed 3 times with water, stained for 1 hour in coomassie stain with slow agitation, and finally slowly washed for 24-48 hours in ultrapure water with slight agitation in order to remove the excess stain. The molecular masses of separated polypeptides were estimated by comparison of the distance traveled relative to the reference bands of the molecular weight protein standard. Finally, gels were documented using a gel imaging instrument, and subsequently dried out between two cellophane membranes using commercial gel drying solution (Novex®, Gel-Dry™).

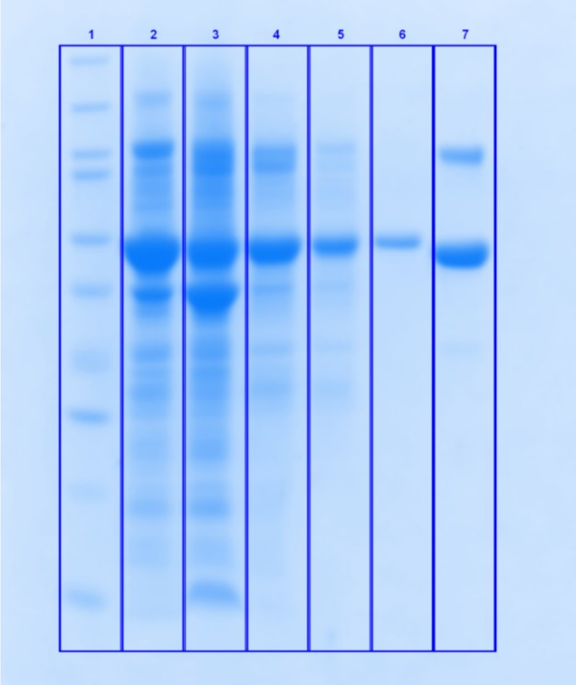


Figure 6 Purification check of DhbC by SDS-PAGE. Lane 1 – protein standards, lane 2 –cell lysate, lane 3 – Ni-NTA flow-through, lane 4 – Ni-NTA washing fractions, lane 5 – Ni-NTA eluted DhbC, lane 6 – DhbC after His-tag cleavage and second Ni-NTA, lane 7 – concentrated DhbC fractions from FPLC

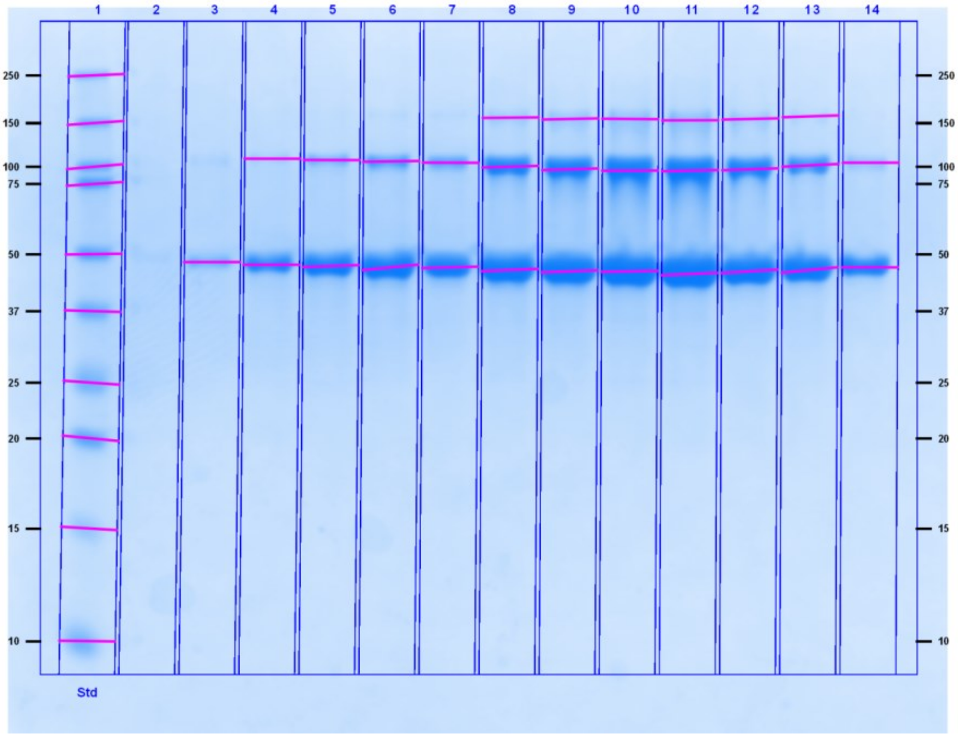


Figure 7 Purity determination by SDS-PAGE of DhbC after purification using fast protein liquid chromatography. Lane 1 contains protein markers, lanes 2-14 contain samples of fractions corresponding to main FPLC peak.

4.1.9 Crystallization

The initial screening of crystallization conditions was done with the Crystal Screen HT kit from Hampton Research and using the sitting-drop vapor diffusion technique. The Crystal Screen HT is a combination of Hampton Research Crystal Screen and Hampton Research Crystal Screen 2 conditions. The screen is a sparse matrix of 96 trial conditions that is biased and selected from known crystallization conditions for macromolecules (Hampton Research 2013).

The crystallization screen was setup using a Mosquito crystallization robot on 96-well, 2-drop-chamber MRC crystallization plate (Swissci, MRC 2 well crystallization plate). The crystallization drops were formed by mixing 400 nL of protein with 400 nL of well solution and equilibrated against 40 µL of well solution. Drops were examined under a stereo microscope immediately after setting up the screen and subsequently once a day for the following two weeks. The crystallization process was monitored and documented using the Xtaldb database system (Zimmerman 2005). The initial protein crystals were observed in the D3 well (100 mM HEPES sodium, 2% v/v PEG400, 2.0 M ammonium sulphate) after 5 days (Figure 8). After crystals were detected in the initial screen, the crystallization conditions were further optimized using the hanging-drop vapor diffusion method. The crystals of selenomethionine-incorporated DhbC used for data collection were grown by hanging drop vapor diffusion method. The well solution consisted of 2M ammonium sulphate, 2% v/v PEG400, 100 mM HEPES pH 7.5. Drops were formed by mixing 2 µL of well solution and 2 µL of 16 mg/mL protein in 300 mM NaCl, 10 mM HEPES pH 7.5, 0.5 mM TCEP. Crystals were grown at room temperature and formed after a week of incubation. Immediately after harvesting, the crystals were transferred into cryoprotectant solution containing 7 % glycerol, 7% sucrose, and 7% ethylene glycol in mother liquor, passed through paratone oil and flash cooled in liquid nitrogen.

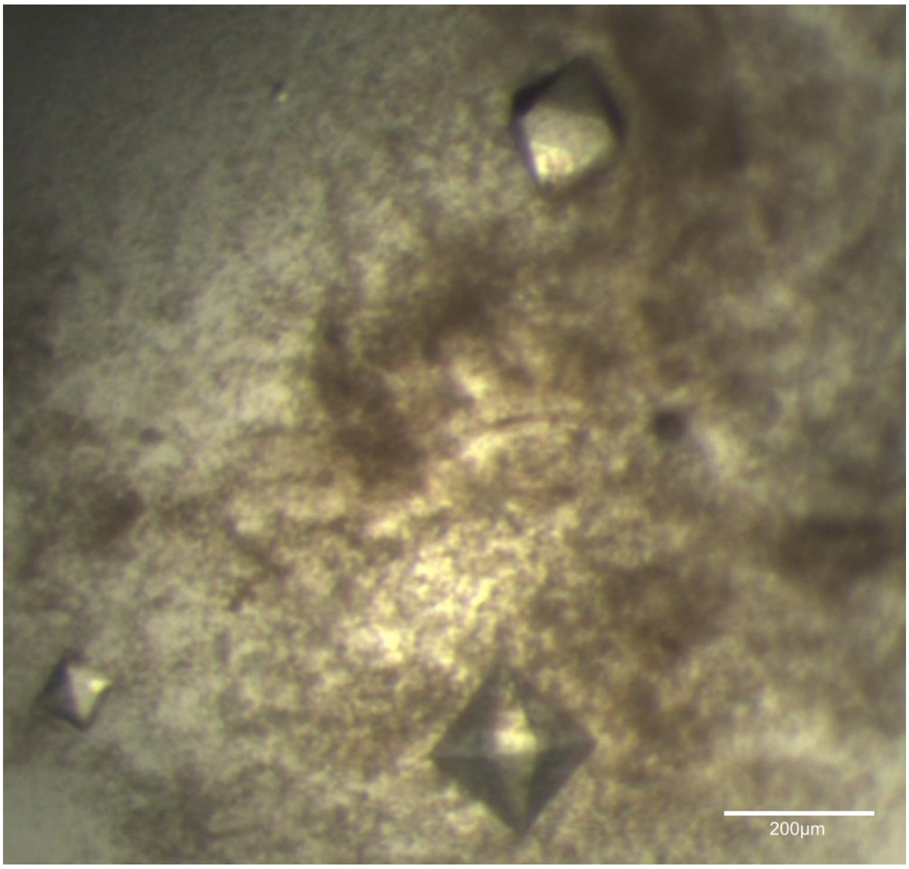


Figure 8 Octahedron shaped crystals of isochorismate synthase DhbC from *B. anthracis*.

4.1.10 X-ray data collection and processing

Diffraction data for the DhbC crystals were collected at 100K at the 19-ID undulator beamline of the Structural Biology Center (Rosenbaum et al. 2006) at the Advanced Photon Source (Argonne National Laboratory, Argonne, Illinois, USA), which is controlled by HKL-3000 (Otwinowski and Minor 1997; Minor et al. 2006). The beamline was operating with standard working energy of 12.660 keV, which corresponds to a wavelength of 0.979 Å, i.e., the K absorption edge of selenium. The 19-ID end station is equipped in ADSC Quantum 315R CCD detector, which was placed at the distance of 300 mm from crystal, resulting in the diffraction limit of ~2.05 Å at the detector edge (Figure 9). The detector was operating in 2x2 binning mode. In the binning readout mode, single pixels are not read individually, but the signal is combined in arrays of four neighboring pixels improving signal to noise ratio. The exposure time was set to 1 s and the attenuation factor was set to 2. In total, 180 still frames were collected with oscillation range of 0.4°. Diffraction data were indexed and integrated with HKL-3000. The resolution cutoff for refinement was determined by commonly used criterion of signal-to-noise ratio, $\langle \overline{I}/\delta(\overline{I}) \rangle$, which should not fall below 2 for the highest resolution shell.

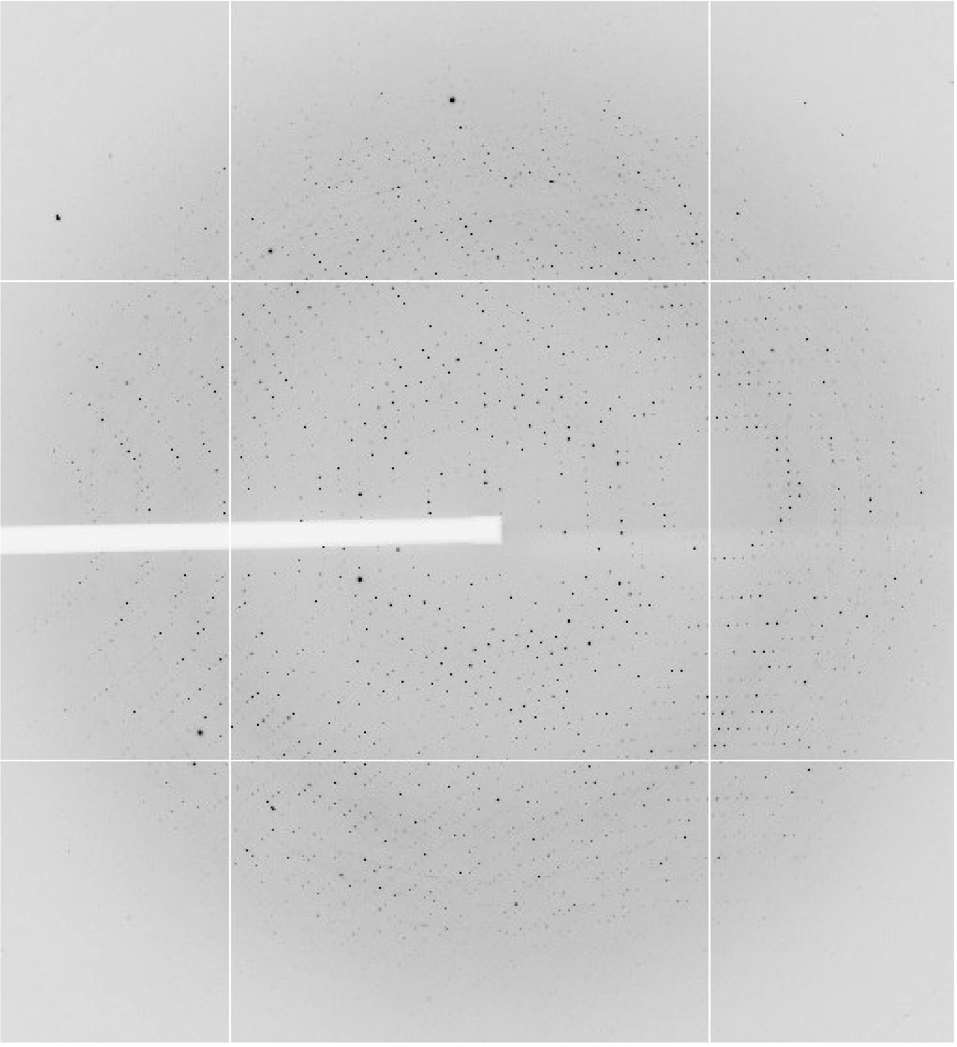


Figure 9 X-ray diffraction patterns obtained from the *P*2$_1$3 DhbC crystal.

The unprocessed diffraction images are publicly available and can be obtained through the CSGID web portal (http://www.csgid.org/csgid/pages/diffraction_images/IDP01205_3os6/). Data collection, structure determination, and refinement statistics are summarized in Table 2.

4.1.11 Structure determination

The next steps after data reduction are determination of the crystallographic phases and model building, a process of construction of a stereochemically accurate atomic model that will correspond to experimentally determined electron density map (Rupp 2010). The structure of the selenomethionine-substituted protein was determined using single-wavelength anomalous diffraction (SAD) phasing, and initial models were built with HKL-3000 coupled with ARP/wARP. The crystals belong to the primitive cubic space group $P2_13$, with unit-cell parameters a,b,c= 201.39 Å. The asymmetric unit consists of 2 homodimers and 28 Se atom sites, 7 per each protomer.

4.1.12 Structure refinement, validation and deposition in PDB

Structure refinement is an optimization step that involves adjusting the initial model coordinates to improve the model's fit to the experimental determined electron density (Figure 10). In essence, it is a process of completing the model and fixing positions of misplaced atoms. The structure of isochorismate synthase DhbC was refined in restrained mode using REFMAC5. TLS refinement (abbreviation for translation, libration, and screw-rotation) is a coarse approximation for the anisotropic vibrations of atom that involves dividing the structure is divided into regions of different isotropic motions. Four TLS groups were used for refinement, one group for each monomer in the asymmetric unit. Water molecules were not included in the TLS groups. Non-crystallographic symmetry (NCS) refinement was used for averaging the density of symmetrical parts of the asymmetric unit. All four subunits of DhbC monomer were restrained in a single NCS group. The molecular modelling program COOT was used for visualization of electron-density maps, model completion, real space refinement, correction of side-chain rotamer conformations, adding solvent molecules and other manual corrections. MOLPROBITY, ADIT, COOT, and HKL3000 were used for structure validation. The coordinates and experimental structure factors were deposited to Protein Data Bank (PDB) with accession code 3OS6.

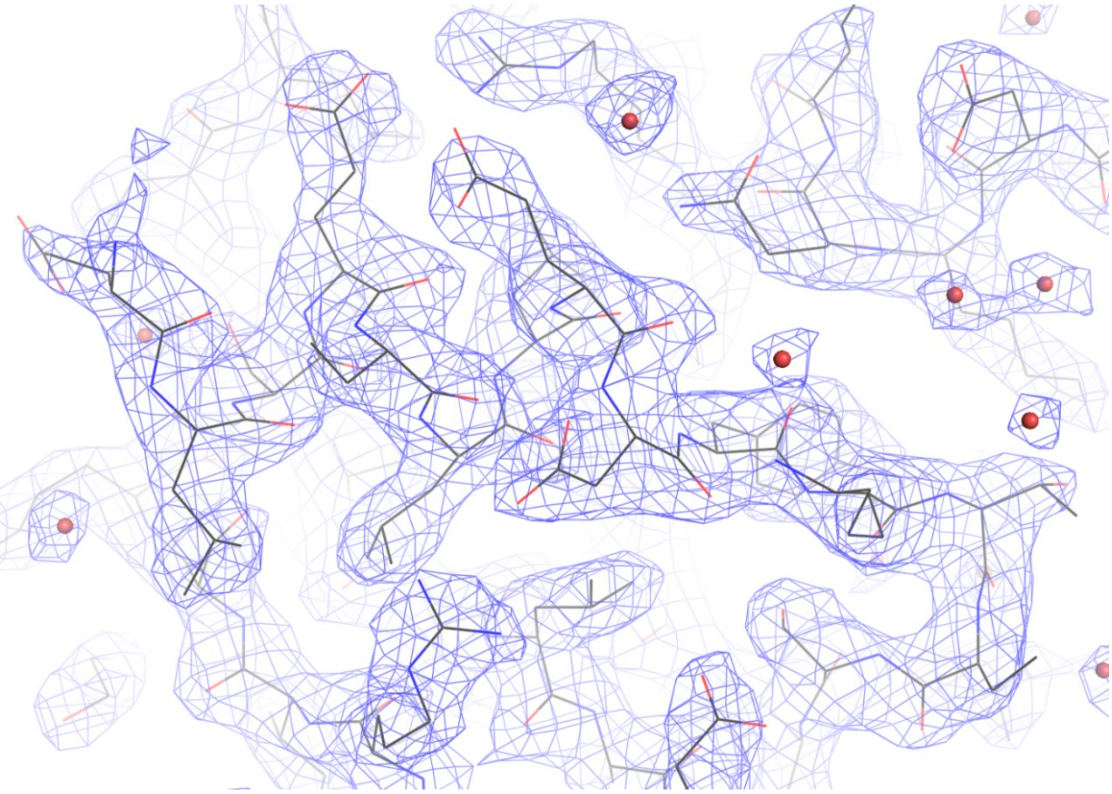


Figure 10 Sample of the 2mFo-DFc electron-density map covering the N-terminal residues of the refined DhbC crystal structure. All oxygen and nitrogen atoms are colored in red and blue, respectively. Red spheres represent oxygen atoms of water molecules. The map was contoured at the 1σ level. Image reprinted from an original article (Domagalski et al. 2013)

4.1.13 Spectrophotometric enzyme activity assay for isochorismate synthase and Michaelis-Menten kinetics

Isochorismate synthase activity and its dependence on Mg^{2+} ions were confirmed using enzyme activity assays. The assay monitors formation of isochorismate by measuring increase of absorbance at 278 nm (He and Toney 2006). Kinetic assays were performed using a PHERAstar FS microplate reader at 30°C for 10 min. 100 µL reaction mixture contained 50 mM HEPES (pH 7.5), 300 mM NaCl, 5 mM MgCl2, 10 µg DhbC, and 1 mM chorismic acid. Samples were incubated for at least 10 min at room temperature prior to addition of chorismate and measurements. The absorbance was monitored every 60 s for 10 minutes of the reaction. The average absorbance changes for three replicates of the experiment were calculated and used for determination of kinetic parameters. The concentration of isochorismate was calculated using the extinction coefficient of isochorismate-chorismate ($\Delta\varepsilon$ = 10211 M^{-1} cm^{-1}) (Domagalski et al. 2013).

Michaelis-Menten kinetic constants were determined by performing the aforementioned spectrophotometric assay with 1 µM of DhbC using a series of substrate concentrations ranging from 5µM to 1 mM chorismate. Measurements were done in three replicates for every substrate concentration. The enzyme was prepared in a Bis-Tris pH 6.5, 50 mM NaCl, 5 mM $MgCl_2$ buffer. Spectrophotometric measurements were performed in temperature of 25°C using Shimadzu UV-2450 UV-Vis spectrophotometer and UVProbe software. The reaction was performed in a black-walled quartz cuvette with 2mm width light slit and 10mm light path. Before the measurements, baseline correction was applied to set the background absorbance to zero to ensure a good reference point before collecting data. Monochromatic absorbance at 278 nm was measured for 300 s with a 0.2 s acquisition rate. Kinetic constants were calculated using GraphPad Prism software.

4.2 Theoretical methods

4.2.1 Relational Databases

The target tracking database and previously established data mining databases, i.e., MetaPDB, and MetaSG were developed using an open-source relational database management system (RDBMS) PostgreSQL. The relational model is currently one of the most popular *logical models* of database design. In the relational database model, information is divided into small non-redundant entities, which are stored using table data structures and predefined relationships that connect tables. Each relationship belongs to one of three types: one-to-one, one-to-many, and many-to-many. Varied content of the database can be organized using schemas that logically group related tables, procedures, and views. Relational database requires very precise *conceptual design* and understanding of data as it is often very difficult to implement major changes in *logical data model* without breaking already defined relationships and constraints. On the contrary, a well-designed relational database provides high integrity and efficiently reduces redundancy. Among the other most popular types of databases are hierarchical databases, object-relational databases and graph databases. The choice of PostgreSQL was motivated by a fact that it offers high quality standards, stability, security, and reliability, but most importantly, because it is released under a liberal and transparent open source license. PostgreSQL RDBMS is scalable and allows formulation of complex queries using Structured Query Language (SQL). The system is being developed by a vast and active community of developers that provide large number of database design and administration tools. Moreover, PostgreSQL offers an interface for PHP scripting language and it is fully supported by CakePHP framework, which was chosen for development of *database application*, i.e., CSGID portal.

4.2.2 Web application development

The CSGID data dissemination portal was developed using an open-source web application framework CakePHP (http://cakephp.org/). The framework adapts the model-view-controller (MVC) architectural pattern (Figure 11) and it is implemented in PHP5 (http://www.php.net/) scripting language. The concept of the MVC architecture is based on three separate and overlaying data abstraction layers. The base and largest component is the model layer, which is directly interacting with the data. Model is responsible for technology independent interactions between objects, i.e., *business logic*. The view layer handles the actual output of the application and generates forms that allow for user interaction. The controller is the intermediate part that interacts with both model and view layers, handling all of the *application logic*. The MVC design assures clear separation of the *business logic* from the application services and the actual data representation. The complex architecture of MVC framework allows development of very large applications without loss of flexibility. Development in CakePHP is fast and structured. The framework has a large amount of built-in functionality including database access plugins, request handling, web page caching, form validation, authentication components, user sessions, and wide range of security methods. All this components are utilized by CSGID portal and portals of other SG centers managed by UniTrack. The CakePHP community has developed extensive documentation and useful practice guides. Use of this software is regulated by transparent open-source license.

The application models, controllers, and helper components were implemented using the PHP scripting language. The application views are a mix of PHP and HTML. The interactive content of the portal was implemented using JavaScript language. The application layout was set up using the Cascading Style Sheets (CSS) style sheet language.

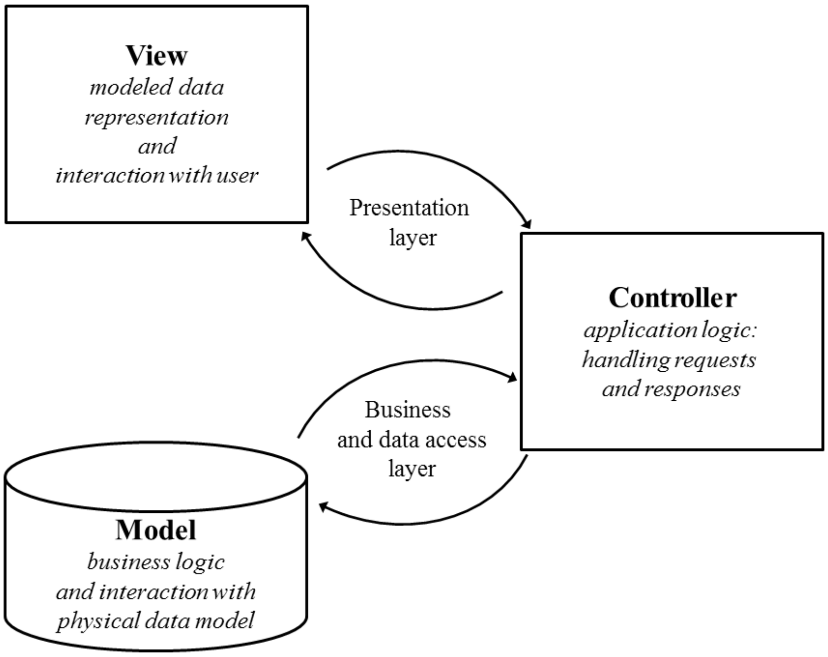

View
modeled data representation and interaction with user
Presentation layer
Controller
application logic: handling requests and responses
Business and data access layer
Model
business logic and interaction with physical data model

Figure 11 Simplistic diagram of CakePHP model-view-controller architectural pattern.

4.2.3 Computer programming

The CSGID web portal was developed using PHP interpreted language (http://www.php.net/). Proposed protein target validation scripts were written using Python interpreted language (http://www.python.org/) with BioPython (Cock et al. 2009), NumPy and SciPy (http://www.scipy.org/) packages. Python was chosen because of the availability of advanced libraries for biological and scientific computation. Development of the CSGID portal was assisted by using the distributed version control system Git. For all of the software development activities, eclipse integrated development environment (IDE) was used.

4.2.4 Graphs and visualizations

Statistics on the CSGID web portal were visualized with use of the Google Charts JavaScript libraries. The experimental pipeline was visualized using JavaScript InfoVis Toolkit. The dynamic content of the CSGID web portal (i.e., pull down menus, calendars, and other) was implemented using Jquery, ExplorerCanvas, Scriptaculous , modalBox, JavaScript InfoVis Toolkit and JSCalendar java script toolkits and libraries. Virtual screening results were presented using Jmol: an open-source Java viewer for chemical structures in 3D. Electronic structure descriptions were generated using the ICM Browser Pro and ActiveICM technology from Molsoft L.L.C.

4.2.5 Bioinformatics analyses

Sequence-based homology searches were conducted with PSI-BLAST (Altschul et al. 1997). Structure-based homology searches were performed with HHpred (Soding 2005; Soding et al. 2005) and DALI (Holm and Rosenstrom 2010). Three-dimensional protein structure alignment was done using SSM (Krissinel and Henrick 2004).

5. Results

5.1 The data management system for Center for Structural Genomics of Infectious Diseases

The CSGID data-management system was developed to serve as a multifunctional tool for monitoring of the progression of protein targets through the high-throughput crystallographic pipeline by documenting the results of the various experiments. The starting point was a system developed for over 7 years for MCSG structural genomics center. In CSGID, a protein is often purified at one site, shipped to another for crystallization screening, and then sent to a third site for structure determination. The unified system keeps track of the location of the samples and collects information about target simultaneously from multiple sources. The database not only contains information about all of the protein samples, but also expression constructs, crystallization drops, crystals, and diffraction datasets. The public view of the experimental pipeline is focused on the status and composition of the experiments, while some metadata is part of the 'super-LIMS' system, LabDB. The central part of the system is a hub database referred as CSGID-DB (Figure 12). The database was created to store the details of all cloning, expression, purification, and structure determination experiments, as well as the results of *in vivo* and *in vitro* analyses as they become publicly available. It means that for great majority of the protein targets, data is available on the day of the experiment. The information is presented in a dynamic, interactive format to allow one for quick browsing through all experimental data. The database content is accessible through the publically accessible CSGID web portal (http://www.csgid.org/). In addition to the information about protein targets and experiments for structure determination, the CSGID portal contains results of virtual screening and annotations in Molsoft's ICM format that were automatically generated for selected protein targets. Other important features of the portal are that it serves as a repository of diffraction images, provides an interactive view of a clustering of the isochorismatase-like hydrolases family, and contains a section of customizable statistics for the database content. Furthermore, the CSGID portal has several other tools including CheckMyMetal (CMM) tool for validation of metal-binding sites (Zheng et al. 2014).

5.1.1 Central role of the target tracking database in the SG data management system UniTrack

The target tracking database (CSGID-DB) plays a central tracking role in the SG data management system UniTrack (Zimmerman et al. 2014) developed in Wladek Minor's laboratory at the University of Virginia. UniTrack is a system that was developed specifically for CSGID with use of some tools developed previously during many years of data management in MCSG. After two years of successful development, variants of the system were applied to three other consortia: MCSG, NYSGRC, and EFI. The target tracking database architecture and set of related support databases and applications is common for the four centers, but the web portals are highly customized for the needs of the particular consortium. Experimental data are incorporated into CSGID-DB using XML files in a predefined format. LIMSs used in participating laboratories regularly update the XML files with complete information about new experiments. On the CSGID-DB side, database update scripts check for changes in those files every 24 hours. Alternatively, LabDB (Zimmerman et al. 2014), a LIMS developed at University of Virginia, communicates with the CSGID-DB via a synchronizing script, which transfers data even more efficiently. The latter approach is a standard for NYSGRC and EFI centers.

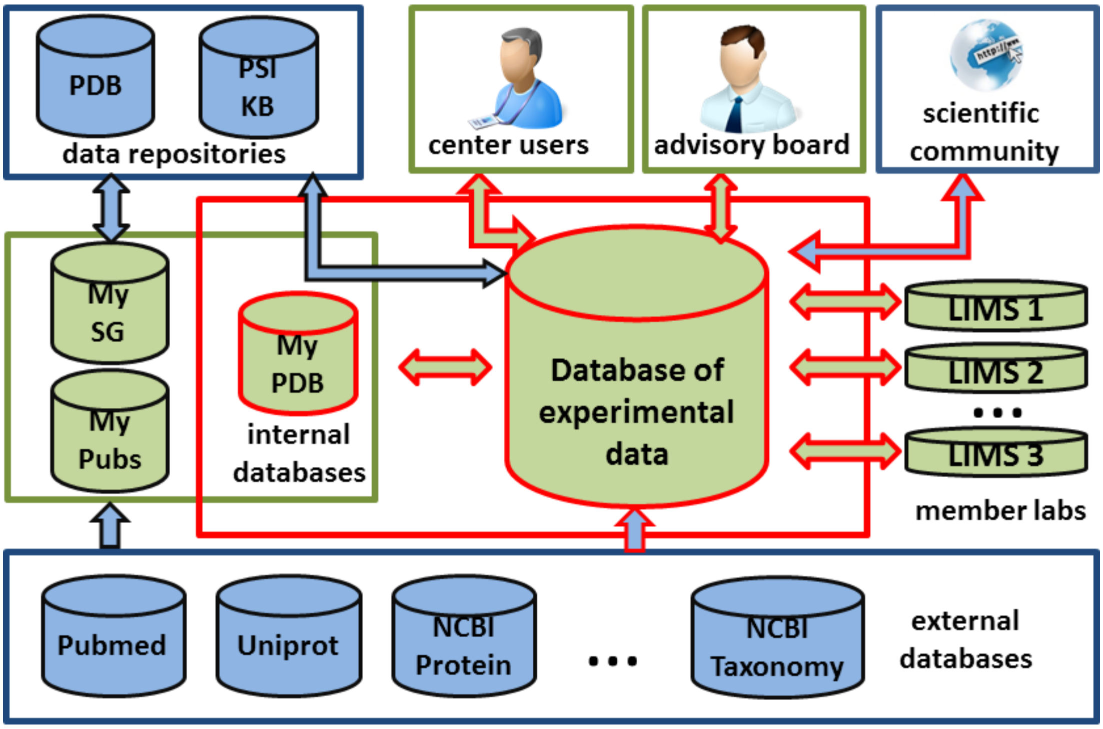


Figure 12 Structure of the SG data management system UniTrack. Publically available data is marked with blue color and internal data is marked with green color. The red line marks the scope of the work described in the thesis.

5.1.2 Relational schema of the database

The target tracking database was designed in a modular way with use of many schemas that allow easy and clear separation of its content. This design allowed adaptation of the database to other SG projects. The CSGID-DB contains following schemas: csgid, experiments, users (Figure 13), uniprot, tigr, targets, taxonomy, synchrotrons, selections, community, drugbank, experiments, homology, hts, ligands, mr, metapdb, metasg, ncbi, nmpdr. Schemas of the target tracking database are divided into three groups:

i. ‘Generic SG center’ schemas, i.e., tables common to any SG center, tables are the same in all instances with the data being different in each center (e.g., users, targets, experiments, and homology).

ii. ‘Center specific’ schemas, i.e., tables present in single center (e.g., csgid.justification_codes, csgid.jcvi_strains).

iii. ‘Mirrored’ schemas, i.e., independent from any data from the SG center, these tables and data will be same in all instances (e.g., ligands, metapdb, metasg, synchrotrons, taxonomy, and tigr).

The schema ‘csgid.experiments’ is a logical core of the database and represents experimental trials of the high-throughput gene-to-structure crystallographic pipeline. The experimental pipeline was divided into 10 stages:

i. (Protein) Target corresponds to the protein from the list of approved CSGID targets. This entity contains information about its gene and common names, protein and DNA sequences, experimental stage, external database identifiers, NCBI annotations, taxonomy, selection phase, justification, and approval date.

ii. Clone corresponds to a molecular cloning experiment and contains information about expression vector, sequence, primers, mutations, experimenter, protocol, status, and date of the experiment.

iii. Expression refers to protein overexpression, expression organism and its strain, media, experimenter’s name, status, and date of the experiment.

iv. Purification is an entity that corresponds to a purified protein sample,

v. Crystallization Drop is a single crystallization trial (a standard screening experiment results in 96 records).

vi. Crystal harvest is a crystal that was harvested with diffraction loop, cryo-frozen, labeled and sent for an X-ray diffraction experiment.

vii. Datasets is a set of diffraction images collected from a crystal.

viii. Structure solution represents processing of the data set. The same dataset can be reprocessed multiple times (e.g., failed attempts with difficult data).

ix. Structure contains final statistics for the model deposited to PDB.

x. Deposit contains information deposition of a final model to PDB, i.e., PDB identification code, title of deposit, list of authors, and dates of deposition and release.

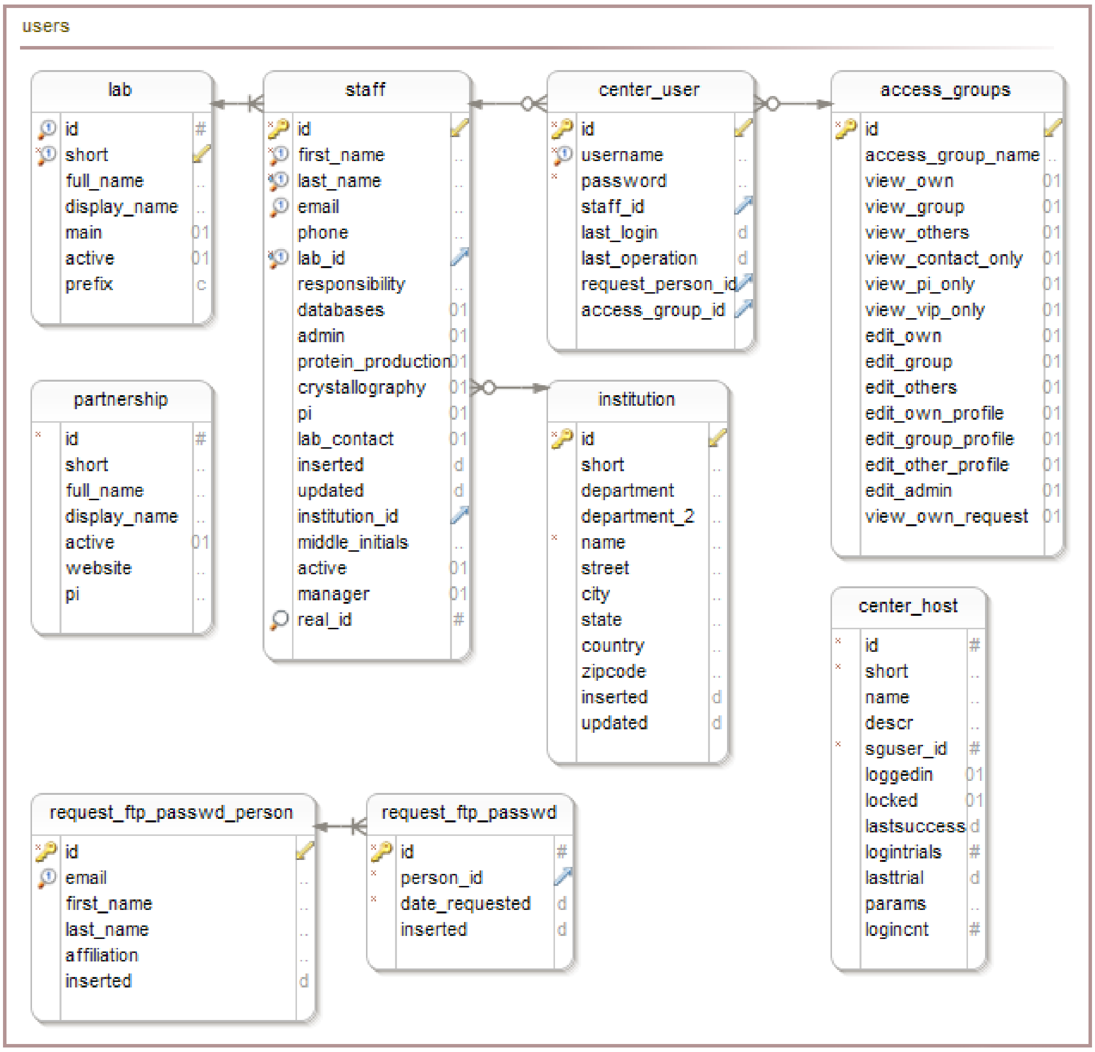


Figure 13 Fragment of entity-relationship diagram for schema ‘users’.

5.1.3 Protein target validation

Before experimental work starts, a new protein target has to pass a validation procedure connected with preceding incorporation of annotations from external databases such as NCBI GenBank (Benson et al. 2013), UniProt (Apweiler et al. 2004), PDB (Bernstein et al. 1977), and the PSI-SBKB (Gabanyi et al. 2011). The validation procedure includes a check of the accuracy of the amino acid and the nucleotide sequences as well as checking if the selected protein does not have homologs with known structure in PDB or among targets already selected by other SG centers. Usually targets are uploaded into the database in large batches of proteins that were proposed based on the same selection rationale, related function, or originating from the same organism. The target validation program was written in the Python programming language with use of BioPython modules and FASTA program. The database update process is done in three discrete stages:

i. File with a batch of accepted targets in a XML, CSV, or TSV format is submitted from a selection person to a database administrator. In CSGID, targets are obtained from UCL target selection database.

ii. A database administrator runs the validation procedure. First, a submitted file is automatically checked for consistency and presence of the obligatory data fields: protein database identifier (NCBI Protein ID or UniProt ID), protein sequence, and NCBI taxonomy identifier. Next, annotations are downloaded from NCBI or UniProt. Finally, raw data, external annotations, and annotated target record are saved into separate tables in the database. A script executes a set of predefined checks and updates validation flags for the annotated record.

iii. Targets that pass all of validation checks get CSGID identifiers. Next, corresponding records are saved to the ‘protein_target’ table and new targets are activated by setting experimental status to ‘selected’. If any targets fail the validation, database administrators identifies the problem and reports it to the person responsible for target selection.

Validation checks for new protein targets include (in the order of execution):

i. Check of sequence coverage between the translation of a nucleotide sequence and the amino acid sequence.

ii. Check of sequence coverage between the amino acid sequence and the sequence reported in NCBI or another reference database.

iii. Check if any of the reference identifiers (i.e., GI number, NCBI accession number, or UniProt ID) is already present in the target database.

iv. Check if the amino acid sequence does not duplicate sequence for an existing target, however, duplication of amino acid sequence is allowed if the nucleotide sequence is different from nucleotide sequence of the existing target.

v. Check if the amino acid sequence does not duplicate any target in a related SG center (i.e., NYSGRC-EFI, or MCSG-CSGID).

vi. Check if the protein target does not have any homologs in PDB, which share more than 30% sequence identity.

5.1.4 Import of experimental data to CSGID-DB

The fundamental mechanism for transferring experiment data from the participating laboratories into the CSGID-DB is by use of XML files, which are published regularly on each site laboratory’s web (httpd) or ftp server. The XML files are generated automatically by specifically adapted LIMS that gather data from connected laboratory equipment. Nevertheless, responsible scientists are required to manually upload scaling log file (Figure 14) and structure coordinates in PDB format with short structure annotation when their target reaches the dataset and structure stages respectively. They can also add or edit information for other experiments using a set of web forms, which are accessible to logged in and privileged users using links in user menu and corresponding experiment views. First authors of PDB deposit are also obligated to upload diffraction images to the CSGID web portal when their structure is released by the PDB. If needed proceeding experimental steps can be also updated manually by privileged users. Thus, data are primarily transferred automatically, but are also curated by responsible scientists.

All experiment records must contain an element or a combination of elements that can serve as unique identifier of the record (i.e., a primary key). Some elements act as foreign keys, which reference records in other XML files as well as relevant protocols, expression vectors, responsible person, or laboratory. These foreign key identifiers must be consistent, referring to existing records in other XML files. Some elements are required and cannot be left empty. Each night a Cron script at the CSGID-DB site first downloads XML files (one file for each type of experiment) from site laboratories and compares them with current files. If there are changes, Cron runs an appropriate PHP script, which updates the database. Scripts process every file that is at the same level or below in experimental hierarchy (for clones it updates all, for crystal_harvest - only crystal_harvests file). When data is saved to the database, a ‘SAVED’ flag for a corresponding record in the XML file is changed to ‘YES’. Additionally, a nightly mirror of each site’s XML files is copied to an FTP directory of the CSGID web portal (ftp://csgid.org/pub/csgid/xml-archive/<site-lab>/). Cron scripts send the e-mails to CSGID addresses with a summary report and a list of errors encountered during the update procedure.

The e-mails are delivered to the following people or groups:

- database administrators (csgid.db@csgid.org) with report summarizing the database update and listing encountered errors,

- the lab that performed the experiment and the database administrators that produced the XML file,

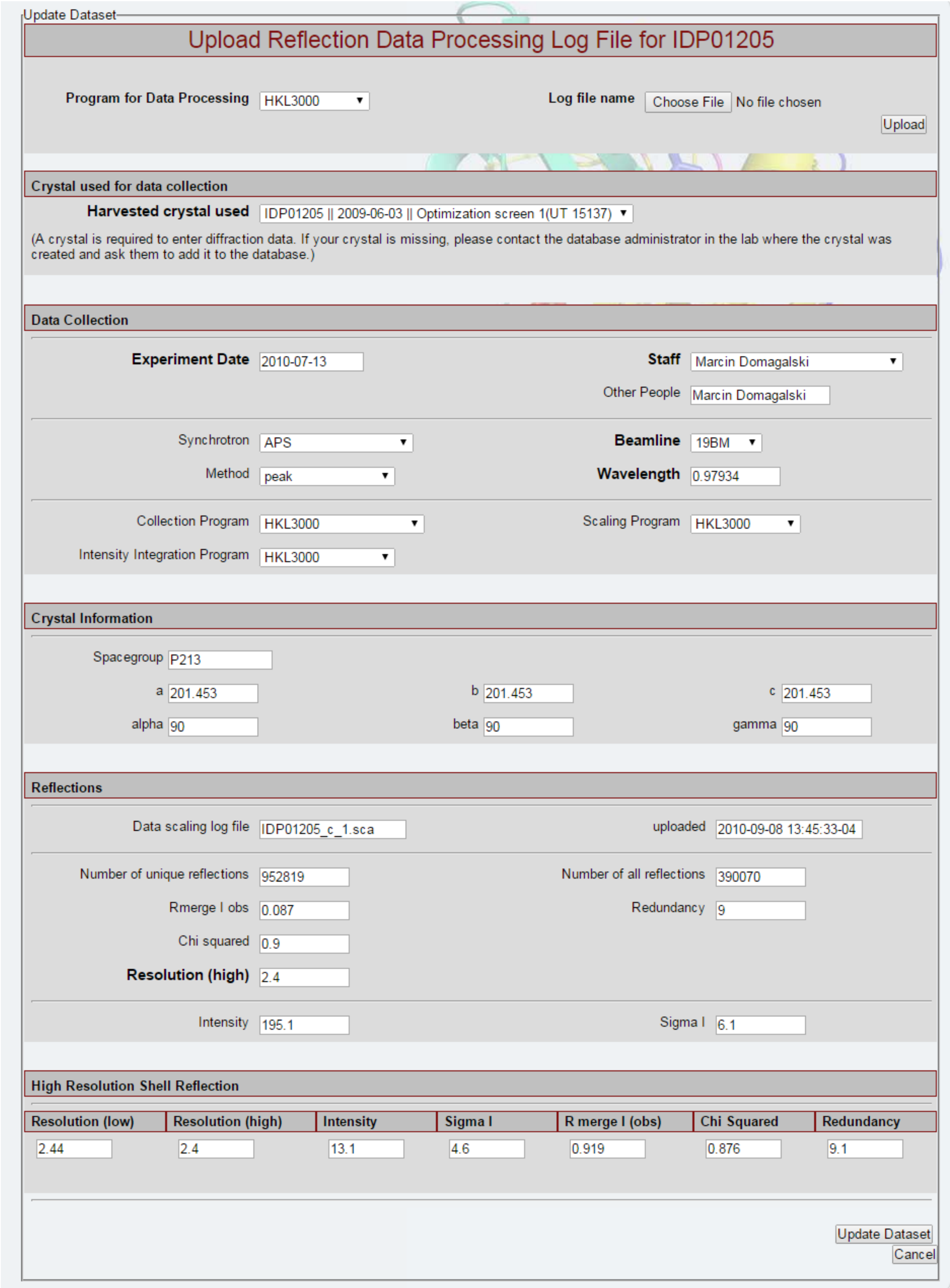
- the responsible laboratory lab, in cases where there was no update in past 2 weeks (e-mails are sent every week after two-week period without update)

5.1.5 Communication layer

The CSGID-DB is a hub database constantly retrieving data from other resources and generating reports for other databases. The aforementioned resources may be classified into three categories: LIMS, internal supporting databases, and external data banks and data repositories. Transfer of the new records between the LabDB LIMS and target tracking database is done by the synchronization script written in Python. The script is automatically executed every hour ensuring that information about new experiment is publically available on the same day. In CSGID, the main update mechanism is based on data transfer through XML files (described in paragraph 5.1.4).

The tracking database employs copies of selected tables from two supporting databases: MetaPDB and MetaSG. MetaPDB is a statistical database, which contains data parsed from the header section of the PDB files. Additionally, MetaPDB stores results of automatic analyses and outputs of many programs, e.g., structure validation report from MolProbity or analysis of surfaces, interfaces, and oligomeric assemblies from PISA. The data is also manually curated for inconsistencies detected in new PDB records. This independent project is a very valuable source of information that was utilized for numerous data mining studies. MetaPDB is also used by a variety of software and databases that were developed in the laboratory. The CSGID database utilizes a copy of multiple MetaPDB tables, which provide data for structure quality statistics, information about homologous structures, and metadata about deposition. The tables are synchronized weekly, a few hours after the RCSB PDB update. MetaSG is a database of information about experimental progress of PSI consortia. The single

MetaSG table utilized by UniTrack contains details of all PSI targets. This information is used for validation of new targets proposals and for statistics.



Update Dataset

Upload Reflection Data Processing Log File for IDP01205

Program for Data Processing HKL3000 ▾ **Log file name** Choose File No file chosen

Upload

Crystal used for data collection

Harvested crystal used IDP01205 || 2009-06-03 || Optimization screen 1(UT 15137) ▾

(A crystal is required to enter diffraction data. If your crystal is missing, please contact the database administrator in the lab where the crystal was created and ask them to add it to the database.)

Data Collection

Experiment Date 2010-07-13 **Staff** Marcin Domagalski ▾

Other People Marcin Domagalski

Synchrotron APS ▾ **Beamline** 19BM ▾

Method peak ▾ **Wavelength** 0.97934

Collection Program HKL3000 ▾ Scaling Program HKL3000 ▾

Intensity Integration Program HKL3000 ▾

Crystal Information

Spacegroup P213

a 201.453 b 201.453 c 201.453

alpha 90 beta 90 gamma 90

Reflections

Data scaling log file IDP01205_c_1.sca uploaded 2010-09-08 13:45:33-04

Number of unique reflections 952819 Number of all reflections 390070

Rmerge I obs 0.087 Redundancy 9

Chi squared 0.9

Resolution (high) 2.4

Intensity 195.1 Sigma I 6.1

High Resolution Shell Reflection

Resolution (low)	Resolution (high)	Intensity	Sigma I	R merge I (obs)	Chi Squared	Redundancy
2.44	2.4	13.1	4.6	0.919	0.876	9.1

Update Dataset

Cancel

Figure 14 View of the dataset upload form containing information parsed from IDP01205 (*B. anthracis* DhbC) SAD dataset log file.

During validation of proposed targets, the target validating script collects information from many external resources. The script requires an identifier of at least one of the two large protein data banks, i.e., NCBI Protein and UniProt. Information parsed from the data banks includes identifiers, DNA and amino acid sequences, annotations, GO terms, and taxonomic information. The system also updates the protein functional annotations from TIGR. UniTrack generates weekly performance reports for TargetTrack, which are available through the CSGID FTP server. The reports include updates of experimental status for each target and protocols.

5.2 CSGID web portal

The web interface of the CSGID-DB is implemented using the Model–View–Controller (MVC) architecture, with separate layers for data representation model (model), application logic (controller), and web page rendering (view). This modular organization allows easier maintenance and development. The web portal was designed to fulfill the needs of four groups of users:

i. Community requesters, researchers from outside of the consortium who submitted a request for protein structure determination,

ii. Scientific community that is interested in the CSGID research, but not directly involved in the project,

iii. Researchers working for the consortium,

iv. Members of the CSGID advisory committee, who monitor the progress of the consortium and provide valuable advice on further strategy development.

In order to achieve the desired flexibility of the interface, set of roles with different access levels were introduced for registered users. Unauthorized users can access publically available data, but only registered users assigned to a particular group are able to read and modify data belonging to that group. A complete list of predefined user roles and their access privileges is presented in Table 1.

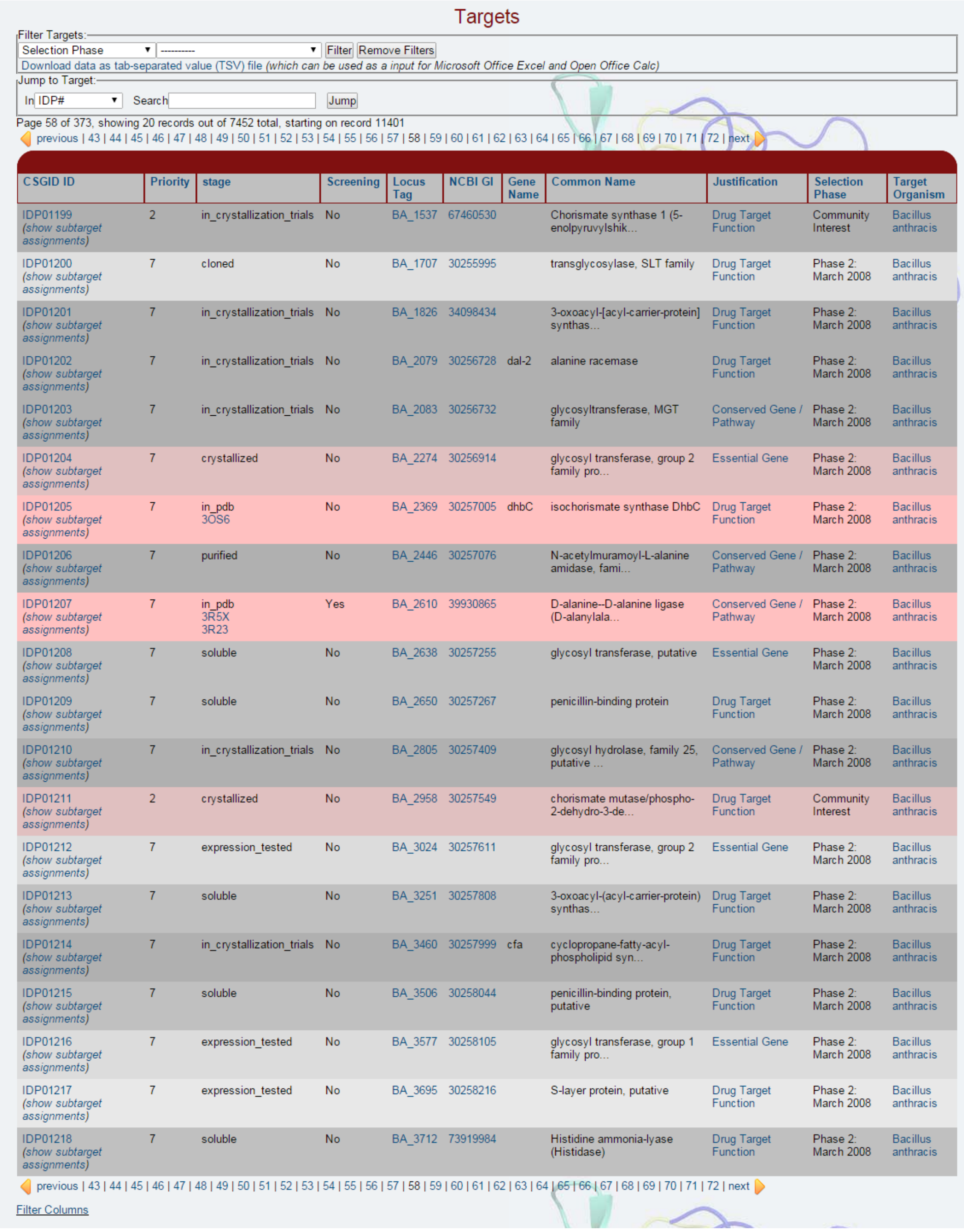
Implementation of the CSGID web portal contains over 50,000 lines of source code. The source code was used as a base for development of sibling web portals for MCSG, NYSGRC, and EFI. Data management in abovementioned centers is driven by adapted instances of UniTrack.

Table 1 User roles and access privileges in CSGID web portal.

	Admin	Staff	Crystallographer	Lab contact	PI	VIP
view own	x	x	x	x	x	x
view group	x	x	x	x	x	
view others	x	x	x	x	x	x
view contact	x			x	x	x
view PI only	x				x	x
view VIP only	x					x
edit own	x	x	x	x	x	
edit group	x	x	x	x	x	
edit others	x		x	x	x	
edit own profile	x	x	x	x	x	x
edit group profile	x			x	x	
edit other profile	x				x	
edit admin	x					

5.2.1 Target search engine

A target search engine is an access gateway for information about the CSGID protein targets. This tool can be accessed by clicking on 'target list' in the top navigation menu of the web portal. The view contains a full list of proteins that were accepted by the selection committee. Every row on the list contains the CSGID ID, priority, stage, locus tag, organism, gene name, protein name and other basic information about a target. Each target on that list is highlighted in a color that corresponds to its experimental status (Figure 15). The colors range from white, which means 'work stopped', through multiple shades of gray that correspond to statuses from 'selected' to 'in crystallization trials' and four shades of pink that mark the final stages from 'crystallized' to 'in PDB'. By clicking on any target on that list, a user is redirected to the interactive browser of experiments linked to the project. Users can search for protein targets using a wide range of cumulative filters which include: selection phase, organism, species, keyword, experimental stage, laboratory (by clone location), TIGR category, presence of virtual screening results, focus area, ligand studies, functional follow up, and priority. Selection of any of the aforementioned filters brings up pull-down menu with a list of possible options. Users can also use CSGID identifier, NCBI GI, or locus tag to search for the particular target of interest. By using the ‘filter columns’ at the bottom of the table, a user can not only decide which targets they want to display, but also which information will be reported.



Download data as tab-separated value (TSV) file *(which can be used as a input for Microsoft Office Excel and Open Office Calc)*

Page 58 of 373, showing 20 records out of 7452 total, starting on record 11401

previous | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 | 61 | 62 | 63 | 64 | 65 | 66 | 67 | 68 | 69 | 70 | 71 | 72 | next

CSGID ID	Priority	stage	Screening	Locus Tag	NCBI GI	Gene Name	Common Name	Justification	Selection Phase	Target Organism
IDP01199 *(show subtarget assignments)*	2	in_crystallization_trials	No	BA_1537	67460530		Chorismate synthase 1 (5-enolpyruvylshik...	Drug Target Function	Community Interest	Bacillus anthracis
IDP01200 *(show subtarget assignments)*	7	cloned	No	BA_1707	30255995		transglycosylase, SLT family	Drug Target Function	Phase 2: March 2008	Bacillus anthracis
IDP01201 *(show subtarget assignments)*	7	in_crystallization_trials	No	BA_1826	34098434		3-oxoacyl-[acyl-carrier-protein] synthas...	Drug Target Function	Phase 2: March 2008	Bacillus anthracis
IDP01202 *(show subtarget assignments)*	7	in_crystallization_trials	No	BA_2079	30256728	dal-2	alanine racemase	Drug Target Function	Phase 2: March 2008	Bacillus anthracis
IDP01203 *(show subtarget assignments)*	7	in_crystallization_trials	No	BA_2083	30256732		glycosyltransferase, MGT family	Conserved Gene / Pathway	Phase 2: March 2008	Bacillus anthracis
IDP01204 *(show subtarget assignments)*	7	crystallized	No	BA_2274	30256914		glycosyl transferase, group 2 family pro...	Essential Gene	Phase 2: March 2008	Bacillus anthracis
IDP01205 *(show subtarget assignments)*	7	in_pdb 3OS6	No	BA_2369	30257005	dhbC	isochorismate synthase DhbC	Drug Target Function	Phase 2: March 2008	Bacillus anthracis
IDP01206 *(show subtarget assignments)*	7	purified	No	BA_2446	30257076		N-acetylmuramoyl-L-alanine amidase, fami...	Conserved Gene / Pathway	Phase 2: March 2008	Bacillus anthracis
IDP01207 *(show subtarget assignments)*	7	in_pdb 3R5X 3R23	Yes	BA_2610	39930865		D-alanine--D-alanine ligase (D-alanylala...	Conserved Gene / Pathway	Phase 2: March 2008	Bacillus anthracis
IDP01208 *(show subtarget assignments)*	7	soluble	No	BA_2638	30257255		glycosyl transferase, putative	Essential Gene	Phase 2: March 2008	Bacillus anthracis
IDP01209 *(show subtarget assignments)*	7	soluble	No	BA_2650	30257267		penicillin-binding protein	Drug Target Function	Phase 2: March 2008	Bacillus anthracis
IDP01210 *(show subtarget assignments)*	7	in_crystallization_trials	No	BA_2805	30257409		glycosyl hydrolase, family 25, putative ...	Conserved Gene / Pathway	Phase 2: March 2008	Bacillus anthracis
IDP01211 *(show subtarget assignments)*	2	crystallized	No	BA_2958	30257549		chorismate mutase/phospho-2-dehydro-3-de...	Drug Target Function	Community Interest	Bacillus anthracis
IDP01212 *(show subtarget assignments)*	7	expression_tested	No	BA_3024	30257611		glycosyl transferase, group 2 family pro...	Essential Gene	Phase 2: March 2008	Bacillus anthracis
IDP01213 *(show subtarget assignments)*	7	soluble	No	BA_3251	30257808		3-oxoacyl-(acyl-carrier-protein) synthas...	Drug Target Function	Phase 2: March 2008	Bacillus anthracis
IDP01214 *(show subtarget assignments)*	7	in_crystallization_trials	No	BA_3460	30257999	cfa	cyclopropane-fatty-acyl-phospholipid syn...	Drug Target Function	Phase 2: March 2008	Bacillus anthracis
IDP01215 *(show subtarget assignments)*	7	soluble	No	BA_3506	30258044		penicillin-binding protein, putative	Drug Target Function	Phase 2: March 2008	Bacillus anthracis
IDP01216 *(show subtarget assignments)*	7	expression_tested	No	BA_3577	30258105		glycosyl transferase, group 1 family pro...	Essential Gene	Phase 2: March 2008	Bacillus anthracis
IDP01217 *(show subtarget assignments)*	7	expression_tested	No	BA_3695	30258216		S-layer protein, putative	Drug Target Function	Phase 2: March 2008	Bacillus anthracis
IDP01218 *(show subtarget assignments)*	7	soluble	No	BA_3712	73919984		Histidine ammonia-lyase (Histidase)	Drug Target Function	Phase 2: March 2008	Bacillus anthracis

previous | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 | 61 | 62 | 63 | 64 | 65 | 66 | 67 | 68 | 69 | 70 | 71 | 72 | next

Figure 15 View of the CSGID target search engine.

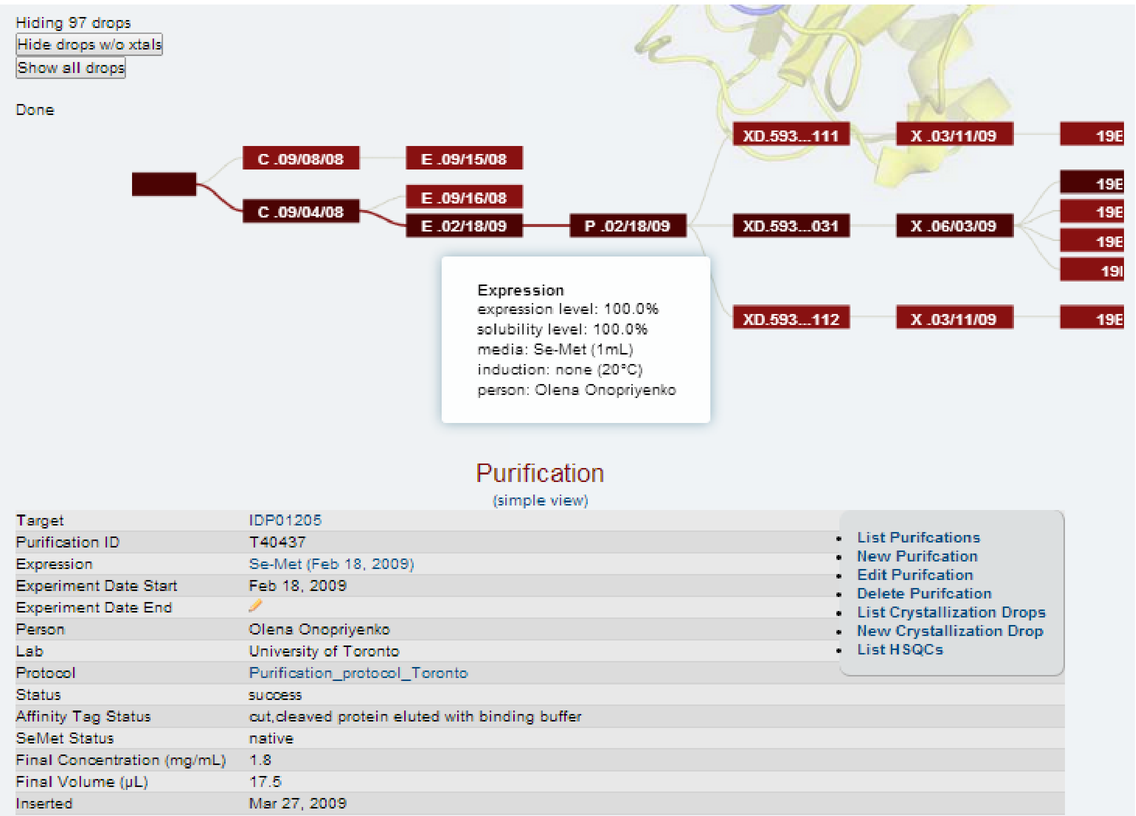
A complete list of over 7000 (7388 as on April 12, 2015), targets is divided into many pages, but it is possible to download all the data (or a filtered list) in CSV or TSV formats for use with spreadsheet application, e.g., Microsoft Excel or Open Office Calc. The equivalent search views for particular experiments, i.e., clones, expressions, purifications, crystallizations, crystal harvests, and datasets, are accessible from the left hand navigation menu under link ‘Experiments’.

5.2.2 Implementation of the crystal structure determination pipeline

The CSGID web portal provides a very intuitive way of navigation through the specifics of the experimental process. Each protein target selected from the target search engine opens inside an interactive node-link tree browser where the target record is a root node and consecutive experiments occupy subsequent child nodes of the diagram (Figure 16). Nodes are represented as boxes labeled in an experiment type dependent manner. All paths reaching the furthest experimental stage of the tree are highlighted with a darker color for easier navigation. While hovering a mouse over any of the tree nodes, the most important details of the experiment are shown in balloon pop-up. After clicking on that node, information that is more complete appears under the tree. Information about each experiment include protocol, experiment date, responsible person, information specific for type of experiment, e.g., media type, growth temperature, or buffers composition, links to external databases plus researchers’ comments. Clicking on some of the identifiers on the page will link directly to other databases. This structure is very convenient when one wants to compare different cloning, expression, purification, or crystallization trials. The tree not only contains successful, but also failed experiments, which can help identify the problems and adapt the experimental strategy for targets that are problematic. The system collects up to 400 parameters for the full pipeline (cloning to structure determination and/or functional characterization).

Subsequent experiment nodes of the crystallographic pipeline are labelled in the following manner:

i. Protein target: 'T' + CSGID identifier
ii. Clone: 'C' + date of the experiment
iii. Protein expression: 'E' + date of the experiment
iv. Protein purification: 'P' + date of the experiment
v. Crystallization drop: 'XD' + crystallization drop identifier
vi. Crystal harvest: 'X'+ date of harvest
vii. Dataset: beamline name
viii. Structure solution: 'Sol:' + phasing method
ix. Refinement: R-factor
x. PDB deposit: PDB ID



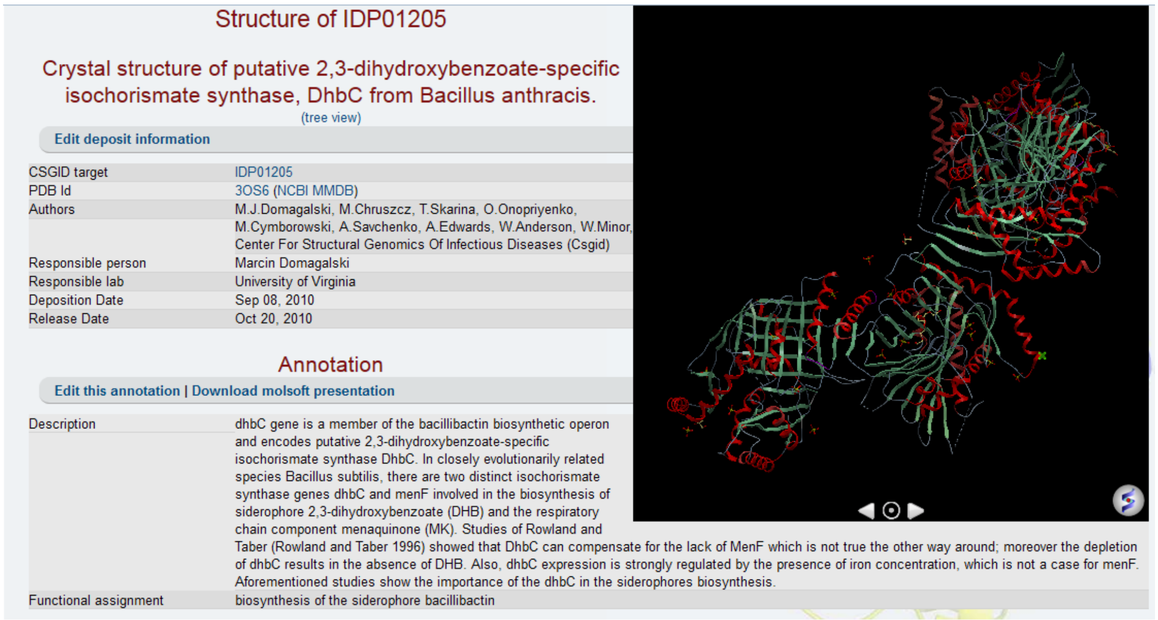
Expression
expression level: 100.0%
solubility level: 100.0%
media: Se-Met (1mL)
induction: none (20°C)
person: Olena Onopriyenko

Target	IDP01205
Purification ID	T40437
Expression	Se-Met (Feb 18, 2009)
Experiment Date Start	Feb 18, 2009
Experiment Date End	
Person	Olena Onopriyenko
Lab	University of Toronto
Protocol	Purification_protocol_Toronto
Status	success
Affinity Tag Status	cut,cleaved protein eluted with binding buffer
SeMet Status	native
Final Concentration (mg/mL)	1.8
Final Volume (µL)	17.5
Inserted	Mar 27, 2009

Figure 16 View of the experimental trials for target IDP01205 (*B. anthracis* DhbC), displayed inside the node-link tree browser. Complete experimental details are displayed under the tree after clicking on the corresponding node. For easier navigation, the most important details are displayed in clouds visible when hovering the mouse pointer over the corresponding box. The longest track on the tree is marked with darker color to distinguish it from unsuccessful experiments.

5.2.3 Electronic structure description

One of the key goals of the CSGID is to convert its structural data into useful information that can be used by the scientific community. For selected protein structures solved as part of the CSGID, interactive 3D presentations are available (Figure 17). Slides and animations are generated using ICM software developed by Molsoft LLC. Interactive content is embedded directly on the structure description web pages and can be accessed after installation of a freely available ActiveICM plugin (Raush et al. 2009). Users can rotate and manipulate structures to view structural units, ligands, oligomerization states, and B-factor distributions. When available, annotations are also provided, giving a functional context to specific structures. Additionally, presentations can be downloaded and edited using ICM Browser, Browser Pro, or ICM Pro (Abagyan 1994). ActiveICM format is being accepted for scientific publishing of enhanced versions of articles (Raush et al. 2009) by journals such as PLoS ONE (Qiu and Dhe-Paganon 2011) and Nature (Li et al. 2011).



Structure of IDP01205

Crystal structure of putative 2,3-dihydroxybenzoate-specific isochorismate synthase, DhbC from Bacillus anthracis.

(tree view)

Edit deposit information

CSGID target	IDP01205
PDB Id	3OS6 (NCBI MMDB)
Authors	M.J.Domagalski, M.Chruszcz, T.Skarina, O.Onopriyenko, M.Cymborowski, A.Savchenko, A.Edwards, W.Anderson, W.Minor, Center For Structural Genomics Of Infectious Diseases (Csgid)
Responsible person	Marcin Domagalski
Responsible lab	University of Virginia
Deposition Date	Sep 08, 2010
Release Date	Oct 20, 2010

Annotation

Edit this annotation | Download molsoft presentation

Description	dhbC gene is a member of the bacillibactin biosynthetic operon and encodes putative 2,3-dihydroxybenzoate-specific isochorismate synthase DhbC. In closely evolutionarily related species Bacillus subtilis, there are two distinct isochorismate synthase genes dhbC and menF involved in the biosynthesis of siderophore 2,3-dihydroxybenzoate (DHB) and the respiratory chain component menaquinone (MK). Studies of Rowland and Taber (Rowland and Taber 1996) showed that DhbC can compensate for the lack of MenF which is not true the other way around; moreover the depletion of dhbC results in the absence of DHB. Also, dhbC expression is strongly regulated by the presence of iron concentration, which is not a case for menF. Aforementioned studies show the importance of the dhbC in the siderophores biosynthesis.
Functional assignment	biosynthesis of the siderophore bacillibactin

Figure 17 Automatically generated electronic structure description for the DhbC structure embedded inside the CSGID web portal.

5.2.4 Homology searches

In order to limit the efforts overlap within consortium and with other SG projects, the CSGID does not accept structure determination requests for proteins that share more than 80% sequence identity to any structure in PDB or any targets of other SG centers. This condition also applies to targets, which are already in the structure determination pipeline. When a structure of homologous protein is released by PDB, experimental work on the target is stopped unless it is the final stages of the pipeline. The CSGID portal contains a tool for the detection of sequence homology. A link to the homology search is located on each target page adjacent to protein sequence. The search is done on CSGID and PDB sequence databases using the Fasta program (Pearson and Lipman 1988). The automated homolog search is run every week by a Crontab script, which sends the results by email to all participating laboratories. The script reports all active targets that share more than 85% relative sequence identity (as determined by FASTA) to any protein deposited in PDB. Targets deposited in PDB, stopped (priority 0), flagged as ligand studies or functional studies are omitted from this list. The results of this search are available in form of XML files, which are located on CSGID FTP server.

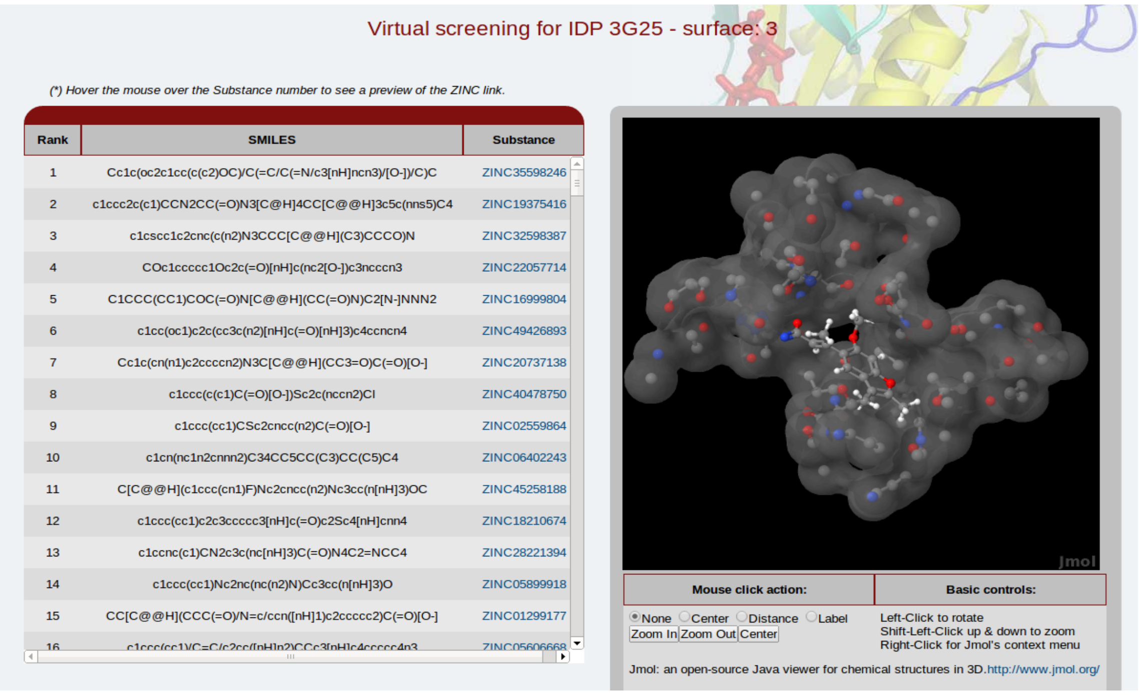
5.2.5 Protocols

Among the many details listed for each experiment accessed through the experiment browser, users can find links to complete, detailed protocols used during the experiments. The CSGID web portal contains a repository of all of the experimental protocols used in the consortium. A complete listing of 72 CSGID protocols can be accessed through the following URL: http://csgid.org/protocols/. Protocols are categorized according to the experiment type, i.e., selection, cloning, expression, purification, crystallization, crystal harvest, NMR and by source laboratory, i.e., ANL, Collaborators, JCVI, NIAID, NU, SBI, UCL, UT, UTSMC, UVA, WU. Every protocol contains detailed description of the experimental procedure.

5.2.6 Implementation of virtual screening results

The CSGID puts a special pressure on determination of protein complexes with biologically significant ligands or protein partners. As a part of this approach, Andrew Binkowski at the University of Chicago developed a single hierarchical pipeline that combines a series of protein analysis, docking and molecular dynamics software packages that allow for an exhaustive investigation of protein-ligand interactions. The APPLIED (Analysis Pipeline for Protein-Ligand Interactions and Experimental Determination) pipeline was designed to predict ligand interactions and provide insights on protein function (Binkowski et al. 2014).

When available, results of the virtual screening are accessible through the CSGID web portal (http://csgid.org/screenings/). The virtual screenings web page contains table listing all of the experiments and provide information about the docking template (e.g., PDB ID, polypeptide chain ID, surface ID), experiment (date, person, run ID, compound library used) and results (five of the top hits). A mouse click on any row in the table redirects to the experiment view, which contains a table with 1000 compounds listed in docking rank order and interactive Jmol applet window that visualizes binding pocket and docked compounds (Figure 18). The table contains information about rank, SMILE code and the ZINC compound identifier linked to the ZINC database. Clicking on any row in the table displays the fit of the compound into the pocket. User can rotate, zoom and use options available from the Jmol context menu, e.g., change representation, display bonds, surface etc. Alternatively, a virtual screening results page can also be accessed by clicking on the icon of the protein molecule that is displayed on top of target tree browser when the results for particular target are available.



Virtual screening for IDP 3G25 - surface: 3

(*) Hover the mouse over the Substance number to see a preview of the ZINC link.

Rank	SMILES	Substance
1	Cc1c(oc2c1cc(c(c2)OC)/C(=C/C(=N/c3[nH]ncn3)/[O-])/C)C	ZINC35598246
2	c1ccc2c(c1)CCN2CC(=O)N3[C@H]4CC[C@@H]3c5c(nns5)C4	ZINC19375416
3	c1cscc1c2cnc(c(n2)N3CCC[C@@H](C3)CCCO)N	ZINC32598387
4	COc1ccccc1Oc2c(=O)[nH]c(nc2[O-])c3ncccn3	ZINC22057714
5	C1CCC(CC1)COC(=O)N[C@@H](CC(=O)N)C2[N-]NNN2	ZINC16999804
6	c1cc(oc1)c2c(cc3c(n2)[nH]c(=O)[nH]3)c4ccncn4	ZINC49426893
7	Cc1c(cn(n1)c2ccccn2)N3C[C@@H](CC3=O)C(=O)[O-]	ZINC20737138
8	c1ccc(c(c1)C(=O)[O-])Sc2c(nccn2)Cl	ZINC40478750
9	c1ccc(cc1)CSc2cncc(n2)C(=O)[O-]	ZINC02559864
10	c1cn(nc1n2cnnn2)C34CC5CC(C3)CC(C5)C4	ZINC06402243
11	C[C@@H](c1ccc(cn1)F)Nc2cncc(n2)Nc3cc(n[nH]3)OC	ZINC45258188
12	c1ccc(cc1)c2c3ccccc3[nH]c(=O)c2Sc4[nH]cnn4	ZINC18210674
13	c1ccnc(c1)CN2c3c(nc[nH]3)C(=O)N4C2=NCC4	ZINC28221394
14	c1ccc(cc1)Nc2nc(nc(n2)N)Cc3cc(n[nH]3)O	ZINC05899918
15	CC[C@@H](CCC(=O)/N=c/ccn([nH]1)c2ccccc2)C(=O)[O-]	ZINC01299177
16	c1ccc(cc1)/C=C/c2cc([nH]n2)CCc3[nH]c4ccccc4n3	ZINC05606668

Mouse click action:	Basic controls:
● None ○ Center ○ Distance ○ Label Zoom In Zoom Out Center	Left-Click to rotate Shift-Left-Click up & down to zoom Right-Click for Jmol's context menu

Jmol: an open-source Java viewer for chemical structures in 3D.http://www.jmol.org/

Figure 18 Interactive view of virtual screening results.

5.2.7 Structure gallery

The structure gallery is a visual presentation of all structures solved by CSGID. This view consists of two separate pages for X-ray and NMR structures, which can be accessed using the links on the left hand navigation menu of the web portal. Each entry contains information about a target ID, deposition details, (i.e., PDB ID, title, authors' names, deposition, and release dates), links to external databases, an image with ribbon representation of the 3D structure, and a link to diffraction images, which can be downloaded by logged in users. Users can access complete target and deposit records using the links on target ID and deposit title. The listing can be sorted and filtered by particular organism, selection phase, target, PDB deposit, keyword, or deposition dates.

5.2.8 Statistics and reports

The CSGID-DB web portal provides numerous statistics divided into two categories: public statistics and performance reports. These are convenient data mining tools, and are very important for control of experimental work. A wide range of statistics enables researchers to oversee general progress of the project and assess bottlenecks.

5.2.8.1 Public statistics

The CSGID dissemination portal has a public statistics section, which is accessible from the header navigation bar of the website. The statistics section is further divided into several overviews that can be switched using the menu on the right side of the page. Those overviews focus on the progress of protein production and evaluation of the X-ray structure quality. All statistics in this section are generated in real-time and some of them can be further adjusted using one of the predefined filters, e.g., by (protein) superfamily, species, or organism. The section contains eight dashboards: 'summary statistics', 'organism distribution: targets and structures', 'organism distribution: pipeline', 'structure statistics', 'ligands in deposits', 'infectious pathogens', 'homologs in PDB', 'detailed target report'.

The ‘summary statistics’ page provides the information about the number of targets that achieved certain stages of the experimental pipeline. Data is presented using a table and column chart. Subsequent stages of the pipeline include: ‘target selected’, ‘target cloned’, ‘protein expressed’, ‘soluble protein expressed’, ‘protein purified’, ‘protein crystallized’, ‘diffraction data collected’ (or ‘HSQC obtained’ in the case of NMR experiment), and finally ‘submitted to PDB’. Using the pull-down menu on top of the page results can be limited to certain organism, species, or selection phase.

The ‘organism distribution: targets and structures’ page contains two tables and two pie charts (Figure 19A), which presents the taxonomic distribution of CSGID targets and structures. The data is grouped by organism or species level and can be further limited to a particular selection phase.

The ‘organism distribution: pipeline’ page presents the efficiency of the structure determination pipeline for particular organisms. The overview contains a large table listing all source organisms and numbers of targets from those organisms that reached certain stages in the crystallographic pipeline. An additional last column of the table contains the percentage of deposits per clone. The data can be grouped by species and filtered by a selection phase.

The ‘structures statistics’ page is an overview of the structure quality. The quality metrics for the deposition of X-ray structures to PDB were discussed on the programmatic meeting held between CSGID and SSGCID in November 2008 (Chicago, IL). At this meeting, researchers defined the ‘refined structure criteria’ that need to be fulfilled in order to deposit a structure to PDB. The criteria relate to model resolution, R-value and R-free, geometry, completeness, percent of overloads, and I/σ value in the high-resolution shell. First two bar charts of the dashboard plot the R and R-free distribution in 0.2 Å resolution bins for all CSGID structures (Figure 19B). Two scatter plot charts indicate R and R-free values versus resolution for all structures. Two other scatter plots represent MolProbity clashscore percentile and MolProbity score versus resolution. The last scatter plot shows the number of water molecules per polypeptide residue versus resolution for every structure. Dots on all scatter plot charts have a color scheme that reflects the structure solution method. All scatter plot charts are interactive, a mouse click on any of the chart dots redirects to a corresponding row in the table placed below the charts. The table contains target id, information about the deposit (i.e., PDB id, first author), structure solution method, quality parameters (resolution, R-value, R-free, MolProbity clashscore percentile, MolProbity score) and a link to diffraction data if available.

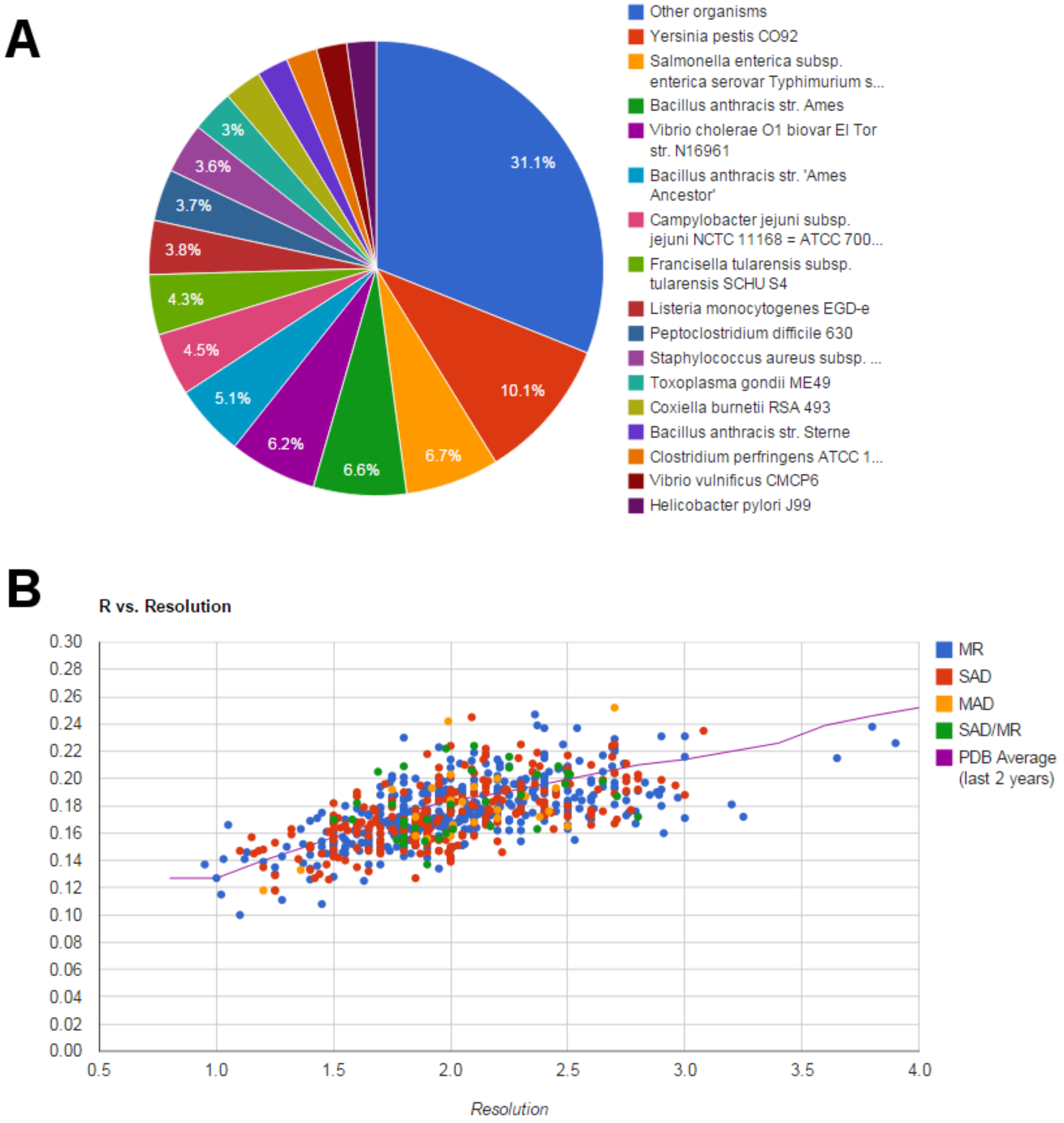


Figure 19 Selected charts from the CSGID portal public statistics section: The organism distribution of all CSGID targets (Fig A) and R-value distribution vs. resolution for all structures solved by CSGID (Fig B).

The 'ligands in deposits' contains a listing of all ligands identified in CSGID structures. Each line describes one ligand and contains information about PDB component identifier, name, number of targets and deposits that contain the ligand, and ligand category. Ligands are grouped into three categories: biological (i.e., biologically relevant), crystallization (crystallization artifacts), and other (i.e., unclassified). Users assign the categories when they upload the structure and fill in structure information form.

The ‘infectious pathogens’ dashboard juxtaposes numbers of protein structures from selected pathogens solved by CSGID with total number of deposits from these species. The presented information includes taxonomic assignment (i.e., classification, genus, species), number of PDB deposits before start of the CSGID, current number of PDB deposits, number of structures deposited by the CSGID researchers, percentage of the CSGID structures among all structures, and percentage of CSGID structures among new structures.

The ‘homologs in PDB’ is a list of all active targets that share more than 85% sequence identity with any structure in PDB. Targets are grouped according to their most recent location/responsible laboratory. Each line contains information about the target (i.e., identifier, experimental status), homologous structure (i.e., PDB identifier, chain identifier, deposition date, release date, and name of the last author of the deposit), and sequence alignment (i.e., percentage of sequence identity and e-value). Members of every research group get information about their pool of targets by an automated email. The results omit targets flagged as ligand studies or functional studies.

In addition to the overviews from the statistics section, the overall advance of the experimental work of the CSGID can be monitored using the progress data dashboard (Figure 20), which is accessible from the header navigation menu of the webpage. The progress page shows the cumulative growth of the number of targets/experiments that reached certain stage in the crystallographic pipeline. Growth of the number of experiments is measured monthly using the end of the month as a cutoff. This chart gives a general idea of how many targets were stuck in one of the four major progress bottlenecks of the pipeline, which include production of expressing clones, expression of a soluble protein, successful purification and production of well-diffracting crystals.

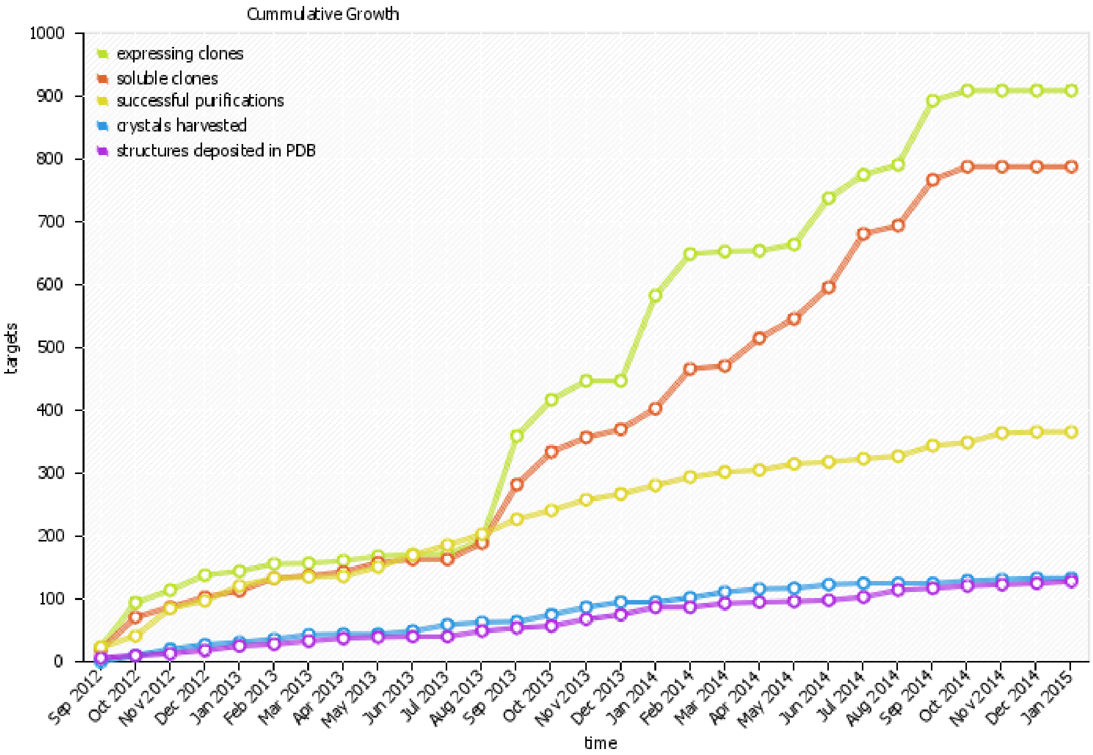
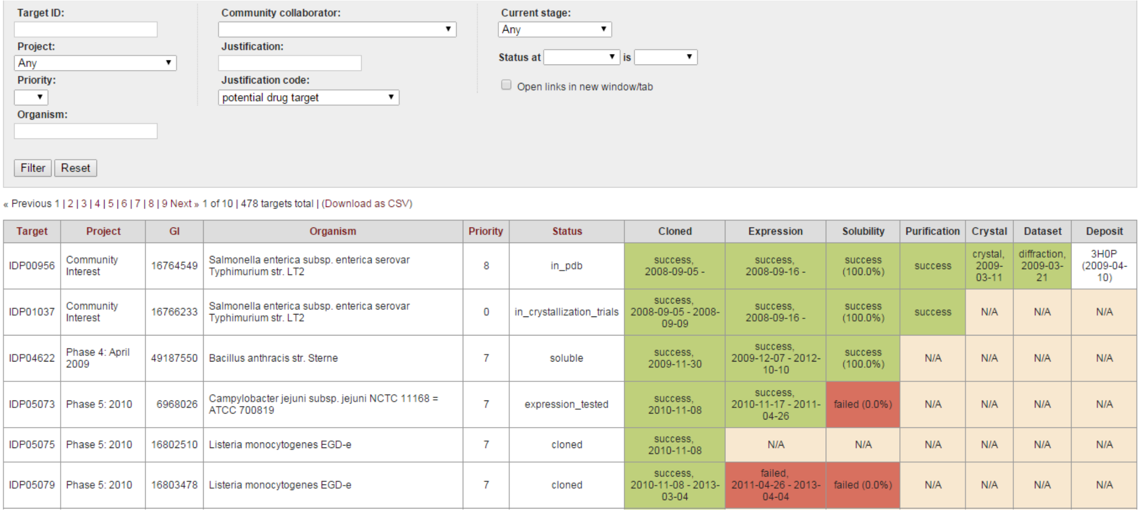


Figure 20 Plot of the cumulative progress for the CSGID center during second phase of the project. Four major bottlenecks of the crystallographic pipeline are clearly visible at this dashboard.

Finally, the latest media coverage and awards are displayed in the 'latest news' box on the main page of the web portal. Users can see all of the previous press releases by clicking 'see all milestones and media coverage' link above the previously mentioned box. The most important structures featured on the NIAID web site in section 'structure of the month' are also presented there.

5.2.8.2 Performance reports

Two of the dashboards serve specifically for generation of performance reports. The 'detailed target report' is the last overview in the statistics section. This report contains short information section for every target, i.e., CSGID target identifier, selection phase (project), gene identifier, organism, priority, current experimental status, and information about successful and failed experimental trials for every step of the crystallographic pipeline (Figure 21). Using a set of filters on top of the page users can limit displayed targets to a specific target identifier, selection phase (project), priority, organism, community collaborator, justification, current experimental stage, and status at selected stage of the pipeline. This data can be downloaded in csv format for an external analysis in spreadsheet application. This tool was made for generation of research performance reports for NIAID.



Target ID:
Project: Any
Priority:
Organism:
Community collaborator:
Justification:
Justification code: potential drug target
Current stage: Any
Status at is
□ Open links in new window/tab

« Previous 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 Next » 1 of 10 | 478 targets total | (Download as CSV)

Target	Project	GI	Organism	Priority	Status	Cloned	Expression	Solubility	Purification	Crystal	Dataset	Deposit
IDP00956	Community Interest	16764549	Salmonella enterica subsp. enterica serovar Typhimurium str. LT2	8	in_pdb	success, 2008-09-05 -	success, 2008-09-16 -	success (100.0%)	success	crystal, 2009-03-11	diffraction, 2009-03-21	3H0P (2009-04-10)
IDP01037	Community Interest	16766233	Salmonella enterica subsp. enterica serovar Typhimurium str. LT2	0	in_crystallization_trials	success, 2008-09-05 - 2008-09-09	success, 2008-09-16 -	success (100.0%)	success	N/A	N/A	N/A
IDP04622	Phase 4: April 2009	49187550	Bacillus anthracis str. Sterne	7	soluble	success, 2009-11-30	success, 2009-12-07 - 2012-10-10	success (100.0%)	N/A	N/A	N/A	N/A
IDP05073	Phase 5: 2010	6968026	Campylobacter jejuni subsp. jejuni NCTC 11168 = ATCC 700819	7	expression_tested	success, 2010-11-08	success, 2010-11-17 - 2011-04-26	failed (0.0%)	N/A	N/A	N/A	N/A
IDP05075	Phase 5: 2010	16802510	Listeria monocytogenes EGD-e	7	cloned	success, 2010-11-08	N/A	N/A	N/A	N/A	N/A	N/A
IDP05079	Phase 5: 2010	16803478	Listeria monocytogenes EGD-e	7	cloned	success, 2010-11-08 - 2013-03-04	failed, 2011-04-26 - 2013-04-04	failed (0.0%)	N/A	N/A	N/A	N/A

Figure 21Fragment of the detailed target report dashboard.

The second performance dashboard is restricted to logged in users, i.e., members of the consortium. This internal report is accessible from the user menu, which appears on top left hand side navigation menu after log in. The internal report page is a set of tables summarizing number of experimental trials that were done in each laboratory belonging to the consortium. Each table corresponds to different time interval, e.g., since beginning of the project, second contract, year by year, or last two months. The numbers can by filtered by project and organism, or counted for the selected time period.

5.3 Structure of isochorismate synthase DhbC from *B. anthracis*

5.3.1 Overall structure of DhbC

DhbC crystallized in a primitive cubic space group (*P*$2_1$3) with two homodimers in the asymmetric unit. In order to obtain unbiased electron density map, the structure was solved using a selenomethionine-substituted protein and the single-wavelength anomalous diffraction method. Data-collection and refinement parameters are shown in Table 2. The single polypeptide chain has 399 residues and molecular mass of 44.6 kDa. Due to protein disorder, several fragments were not modeled, including the N-terminal 8–9 residues, the β6–β7 loop and some residues in the β9–β10 loop (Domagalski et al. 2013). The structure contains 36 ligands resulting from the crystallization conditions: 28 sulphate ions, 6 glycerol, and 2 polyethylene glycol molecules. Structural similarity searches suggest that DhbC belongs to the aminodeoxychorismate (ADC) synthase domain family according to the SCOP classification (SCOP class d.161.1.1) (Murzin et al. 1995), Alpha Beta 4-Layer Sandwich according to CATH (CATH class 3.60.120.10), and the chorismate-binding enzyme family (PF00425) in the Pfam classification (Bateman et al. 2004). DhbC adopts the ADC synthase-like fold containing four repeats of α–β2–β motif arranged in a four-layer core structure, where the layers are α/β/β/α with orthogonally packed β-sheets. The first β-sheet is comprised of β6, β2, β1, β9, β10, β19, β20, β8, and β7. The second β-sheet is comprised of β3, β5, β4, β17, β18, β11, β12, β16, and β14, and the small third sheet contains β13, β16, and β15. The β-sheets are surrounded by nine α-helices (Domagalski et al. 2013). The composition of the asymmetric unit, PISA server assemblies analysis (Krissinel and Henrick 2007) and the gel filtration results, suggest that a homodimer is the biologically functional assembly. The two fold symmetric dimer is made up by joining two β-sheets (β3, β5, β4, β17, β18, β11, β12, β16, and β14) into single intermolecular β-sheet. The overall structure of DhbC monomer is shown in Figure 22.

Table 2 Data-collection, structure-determination, and refinement statistics for the crystal structure of DhbC from *B. anthracis* (PDB entry 3os6), from. Values in parentheses are for the highest-resolution shell. Table reprinted from an original article (Domagalski et al. 2013).

Data collection	
Wavelength (Å)	0.9793
Space group	$P2_13$
Unit-cell parameters	a=b=c=201.4Å, α=β=γ= 90°
Resolution (Å)	50.00-2.40 (2.44-2.40)
No. of unique reflections	105217
Completeness (%)	99.40 (100)
Redundancy	3.7 (3.7)
Mean <I/σ(I)>	20.0 (2.2)
Molecules in asymmetric unit	4
Matthews coefficient ($Å^3$ Da^{-1})	3.78
Solvent content (%)	67.5
$R_{merge}/R_{meas}/R_{pim}$	0.048 (0.532) / 0.054 (0.620) / 0.026 (0.312)
Structure refinement	
R_{work}/R_{free}	0.171 (0.229) / 0.212 (0.262)
No. of residues/protein atoms	1534 / 11630
No. of water atoms	763
Average B factors ($Å^2$)	
Main chain	42.1
Side chains	46.7
Overall	44.3
Waters	38.4
Ramachandran plot (%)	
Most favored	97.9
Allowed	2.1
Disallowed	0.0
R.m.s. deviations from ideal values	
Bond lengths (Å)	0.018
Bond angles (°)	1.71
MolProbity	
Score	1.36
Clashscore	4.39
Poor rotamers	1.03%

Data in the highest resolution shell are given in parentheses

$$^{\#}R_{merge} = \sum_{hkl}\sum_{i}\left|I_i(hkl) - \langle I(hkl)\rangle\right| / \sum_{hkl}\sum_{i}\left|I_i(hkl)\right|$$

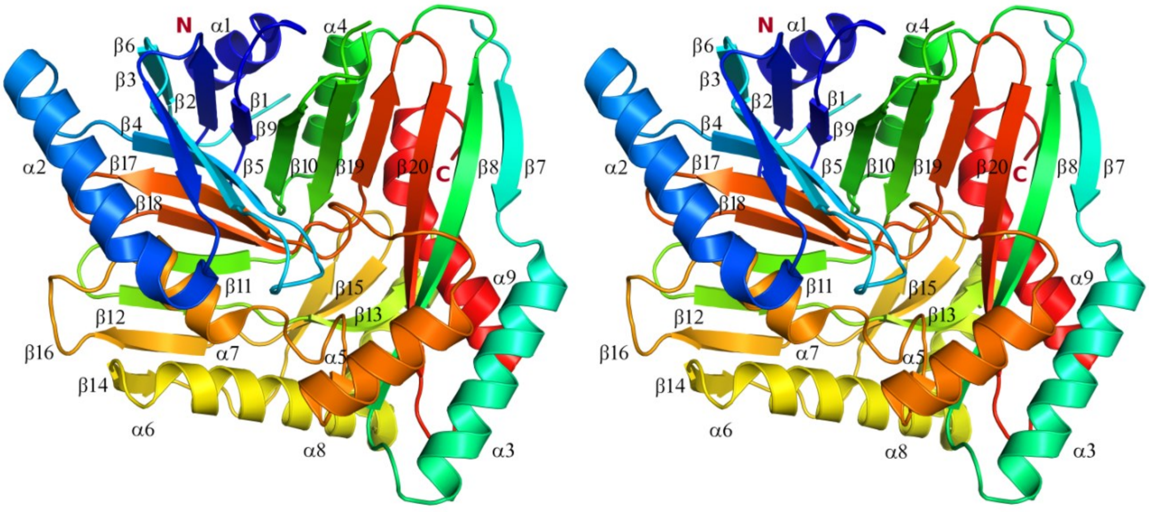


Figure 22 Overall structure of isochorismate synthase DhbC monomer in cross-eyed ribbon representation. The ribbon is colored accordingly to the order of secondary structure elements, i.e., from blue at N-terminus to red at C-terminus. Secondary structure elements are labeled and numbered. Figure reprinted from an original article (Domagalski et al. 2013).

5.3.2 Comparison of DhbC to other chorismate-utilizing proteins with known structures

Structural searches revealed similarity of DhbC to several isochorismate-binding enzymes (Table 3). The closest similarity to DhbC has isochorismate synthase from *E. coli* (Sridharan et al. 2010) as it is shown in Figure 23. EntC is part of biosynthesis pathway of the siderophore enterobactin (Sridharan et al. 2010). Enterobactin is a very similar to bacillibactin catecholate siderophore with three DHB moieties directly attached to a tri-L-serine backbone, whereas in bacillibactin DHB moieties are linked to a tri-threonine via glycine spacers. Similarly to *B. anthracis* and *B. subtilis*, *E. coli* also has the second isochorismate synthase gene, but EntC is not able to fully restore menaquinone deficiency in mutants with a disrupted *menF* gene (Dahm et al. 1998). HHpred analysis showed that salicylate synthetases Irp9 from *Y. enterocolitica* (Kerbarh et al. 2006) and MbtI from *M. tuberculosis* (Manos-Turvey et al. 2010) as well as specific for menaquinone biosynthesis, isochorismate synthases MenF from *E. coli* (Parsons et al. 2008) and *Yersinia pestis*, anthranilate synthase TrpE from *Salmonella typhimurium* (Morollo and Eck 2001), and 2-amino-2-desoxyisochorismate (ADIC) synthase PhzE from *Burkholderia sp.* (Li 2011) are also structurally similar to DhbC. All of the aforementioned proteins belong to ADC synthase structural family according to SCOP and take part in conversion of chorismate to isochorismate (EntC, MenF), salicylate (Irp9 and Mbtl), and anthranilate (TrpE) or ADIC (PhzE) (Domagalski et al. 2013).

Table 3 The closest *B. anthracis* DhbC homologs with known structure. Table reprinted from an original article (Domagalski et al. 2013).

Protein name	Organism	PDB code	Sequence identity (%)	RMSD (Å)	Number of superposed residues
isochorismate synthase EntC	*E. coli*	3HWO	38	1.34	354
salicylate synthetase Irp9	*Y. enterocolitica*	2FN0	25	1.95	323
salicylate synthetase Mbtl	*M. tuberculosis*	3LOG	22	1.95	327
isochorismate synthase MenF	*E. coli*	3BZM	25	1.86	336
isochorismate synthase MenF	*Y. pestis*	3GSE	22	2.25	324
anthranilate synthase TrpE	*S. typhimurium*	1I1Q	14	2.30	322
2-amino-2-desoxyisochorismate (ADIC) synthase PhzE	*Burkholderia sp.*	3R75	16	2.41	324

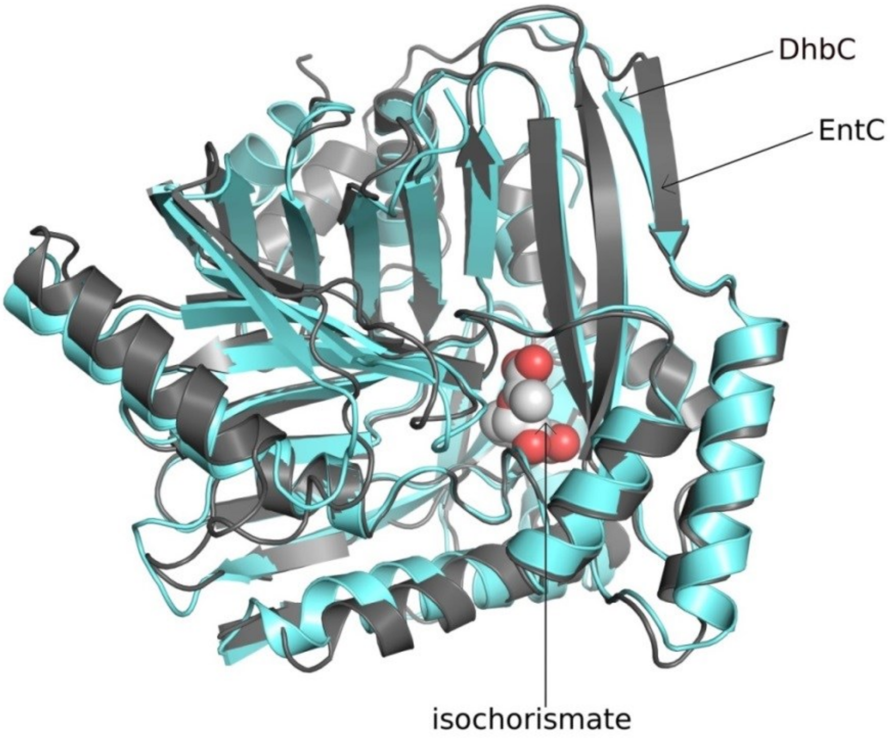


Figure 23 Superposition of the isochorismate synthase DhbC from *B. anthracis* and its closest homolog isochorismate synthase EntC from *E. coli*. Figure reprinted from an original article (Domagalski et al. 2013).

5.3.3 Active site

The active site of *B. anthracis* DhbC is very similar to the active site of *E. coli* EntC, as shown in Figure 24. Only two of the essential active site residues are different between the two proteins: EntC Leu304 is substituted by the chemically similar amino acid Val305 in DhbC, while Phe359 is substituted by Tyr360. The second substitution should not cause a significant change in enzyme activity, as it is also present in *B. subtilis* DhbC and MenF from both *E. coli* and *Y. pestis*. In contrast to EntC, in the DhbC structure neither an isochorismate nor a magnesium ion is bound in the active site. In EntC, a magnesium ion is coordinated between two glutamic acid residues, Glu241 and Glu376, and the C1 carboxylate of isochorismate. In DhbC, the side chain of Glu241 is rotated to the outside of the active site, opening the chorismate-binding pocket. The DhbC structure also contains a sulfate ion located in the position corresponding to the isochorismate C1 carboxylate in EntC and coordinated by the conserved Lys381 and Ser215, and the backbone N atoms of Gly214 and Gly364. The arrangement of the enol-pyruvyl holding residues Lys380, Ile346 and Arg347 (Lys381, Ile347 and Arg348, respectively in DhbC) is not affected by binding of isochorismate in the EntC's active site. The last difference is the orientation of π−stacked pair of aromatic residues. In DhbC, Phe328 and Tyr360 are oriented parallel to potential location of the isochorismate ring, which is consistent with the orientation of corresponding pairs in structures of other chorismate-binding enzymes. In EntC, Phe327 and Phe359 are oriented parallel to pyruvyl group of the isochorismate (Domagalski et al. 2013).

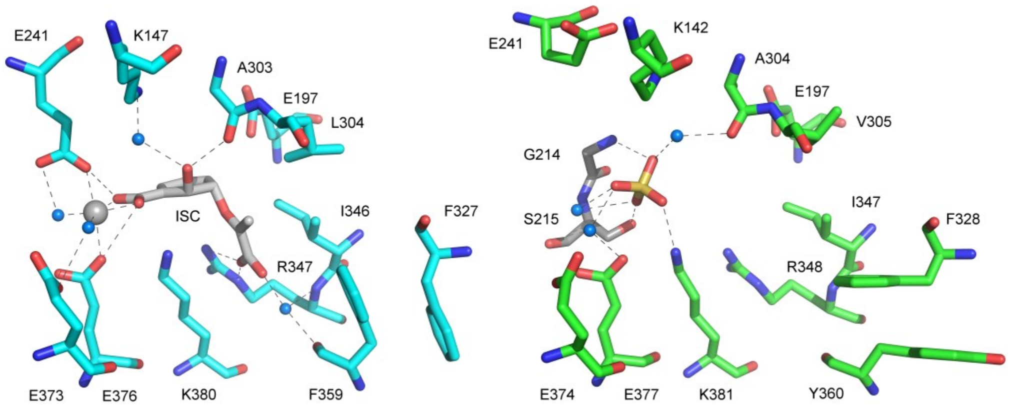


Figure 24 Comparison of the active sites of DhbC from *B. anthracis* and EntC from *E. coli*. Figure reprinted from an original article (Domagalski et al. 2013).

5.3.4 Molecular function assignment and Michaelis-Menten kinetics

Isochorismate synthase enzymatic activity of the DhbC was verified using spectrophotometric assay that measures formation of isochorismate by following the increase in absorbance of monochromatic light at 278 nm. The same type of assay was used for recording reaction rates at different substrate concentrations used to calculate constants of Michaelis-Menten equation. Formation of isochorismate was observed in a reaction mixture containing 10 µg of enzyme, 1 mM of chorismic acid, and 5 mM $MgCl_2$. After 10 minutes reaction time, average conversion rate of 33% (from three repeats of the experiment) of substrate was recorded (Figure 25). Conversion of chorismate to isochorismate was not detected in two control samples that were missing $MgCl_2$ or DhbC. The assay confirmed that DhbC has the enzymatic function of isochorismate synthase and that its activity is dependent on the presence of Mg^{2+} ions (Domagalski et al. 2013).

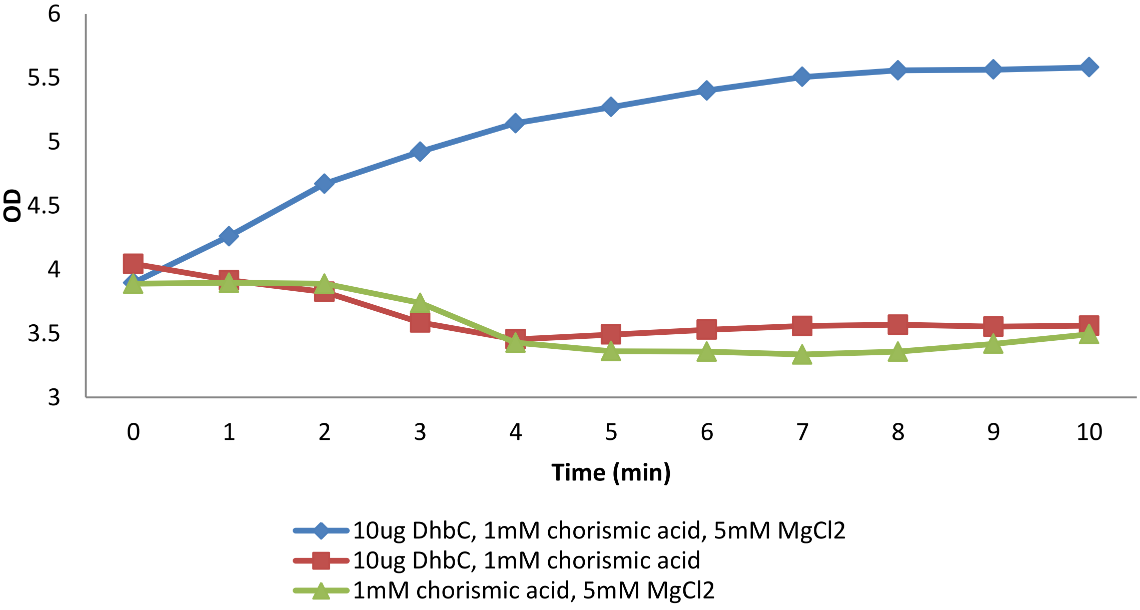


Figure 25 Results of the enzyme activity assay, which monitors formation of isochorismate by measuring increase of absorbance at 278nm ($\Delta\varepsilon_{isochorismate-chorismate}$ = 10211 M^{-1} cm^{-1}). The absorbance curves were made by averaging three repeats of the experiment.

Michaelis-Menten kinetic constants were measured using 1 μM enzyme and nine substrate concentrations ranging from 5 μM to 1 mM. DhbC was found to have K_m= 164.1 ±21.6 μM and K_{cat}= 35.4 ±1.61 min^{-1} within 95% confidence interval (Figure 26). A substrate velocity curve was fit in with nonlinear regression with $R_{squared}$= 0.9918. Kinetic constants of DhbC homologs, *E. coli* isochorismate synthases EntC and MenF were measured previously in several independent studies. MenF was found to have K_m= 195 ±23 μM and K_{cat}= 80 min^{-1} as measured with assay that monitored absorption at 278nm (Daruwala et al. 1997). In study by Dahm et al. MenF was found to have K_m= 166.9 μM and K_{cat}= 144.9 min^{-1} (Dahm et al. 1998). In study that used a coupled enzyme assay with isochorismatase EntB, MenF was found to have K_m= 192 ±7 of and K_{cat}= 213 ±5min^{-1} (Kolappan et al. 2007). EntC was found to have K_m= 14 μM and K_{cat}= 173 min^{-1} in study that used coupled assay of EntC with isochorismatase EntB (Liu et al. 1990). In more recent study of Sridharan et al. that used a coupled enzyme assay with *Pseudomonas aeruginosa* isochorismate-pyruvate lyase (PchB), EntC was found to have Km= 7 ± 0.8 μM and K_{cat}= 37 min^{-1} (Sridharan et al. 2010). The large differences in activity between EntC and DhbC may be caused by two-fold difference in Mg^{2+} concentration, pH, different precision of the enzymatic assay, different composition of buffer, or different enzyme preparation.

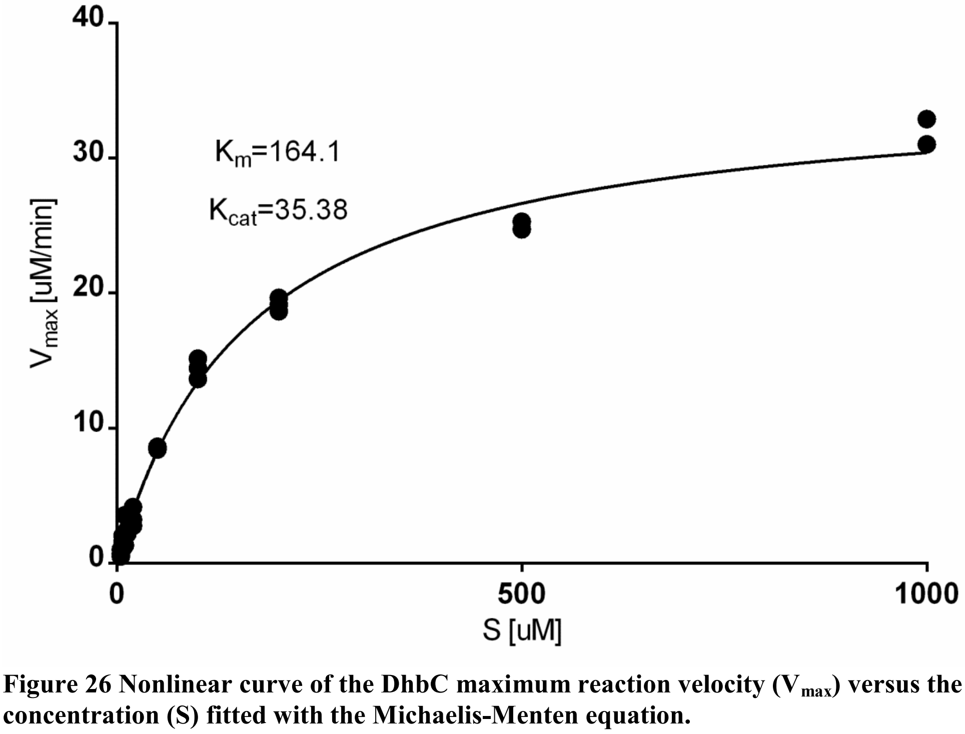


Figure 26 Nonlinear curve of the DhbC maximum reaction velocity (V_{max}) versus the substrate concentration (S) fitted with the Michaelis-Menten equation.

6. Discussion

6.1 Coordination and documentation of DhbC structure determination workflow using UniTrack system

In order to provide more complete information about the *B. anthracis* bacillibactin pathway, the CSGID target selection committee selected four proteins encoded by *bac* biosynthetic operon, i.e., 2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase EntA (IDP04314), isochorismate synthase DhbC (IDP01205), 2,3-dihydroxybenzoate-AMP ligase DhbE (IDP04676), and isochorismatase DhbB (IDP04677) for structure determination. Nomination of the aforementioned targets was justified by their putative drug target roles and involvement in virulence. Apart from members of bacillibactin biosynthetic cluster, multiple other proteins involved in iron acquisition in this pathogenic bacterium were selected. The targets belong to following groups: iron-binding ABC transporters, i.e., FhuC (IDP04427), FhuD (IDP05715), FatB (IDP05050), ferrochelatases, i.e., Hem-H2 (IDP04666) and components of Isd heme scavenging system, i.e., S-layer protein (IDP05156), IsdC (IDP05488), IsdJ (IDP05417), and IsdK (IDP02799).

These targets were incorporated into structure determination pipeline through standard procedure. First, both internally selected and proposed by community targets were reviewed by the CSGID target selection team at University College of London using criteria approved by the NIAID. The target selection team assigned a priority from 0 to 8 to the targets and passed them to a database administrator at University of Virginia. The list of targets contains identifiers in public databases, taxonomy information, protein and DNA sequences, gene name, justification for selection and priority. The database administrator ran the validation tool, which is described in detail in the paragraph 5.1.3. After passing all the validation checks, new targets were added to the CSGID target tracking database with an initial status and priority set to 'selected'. Since that moment, the database update script regularly checked for new experimental records of this target in the XML files produced by LIMSs used in partner laboratories (described in detail in the paragraph 5.1.5). All cloning, expression, purification, and crystallization experiments were transferred automatically through aforementioned mechanism (experimental procedures are described in chapter 4.1). The new records were transferred to the tracking database always within one day from appearing in XML file. Details of the diffraction experiment, structure solution, and PDB deposit were uploaded manually using web forms of the CSGID portal (Figure 14). Finally, the raw diffraction images were compressed and published on CSGID FTP server, after depositing the final structure in PDB. Diffraction images can be downloaded from ftp://danuska.med.virginia.edu/csgid_data/IDP01205_3os6.tar.bz2 (e-mail registration at http://csgid.org/pages/diffraction_images is required to get access to the server). Detailed history of structure determination experiments for *B. anthracis* DhbC is publically available at http://csgid.org/csgid/space_tree/view/IDP01205 and contains hyperlinks to the complete experimental protocols.

6.2 The purpose of research data preservation and public release

Currently in life sciences, most research is ultimately converted into per reviewed articles, which usually contain only a brief description of the experiments. Reproducibility of the published results is a common problem (Prinz et al. 2011; Begley and Ellis 2012) and may arise from insufficient knowledge about the experimental procedure. Additionally, some studies estimate that even 85% of scientific efforts are wasted (Ioannidis 2014). Considering the fact that most of the life sciences research is funded from public funds, it should be expected that scientists would share all of the experimental data and not only the publishable results. In structural biology, publication standards are relatively high, as every three-dimensional structure of a macromolecule must be deposited to public data bank, i.e., PDB, in order to publish the paper that describes it. Moreover, it is a requirement to release not only the structural coordinates of a macromolecule, but also structure factor files that allow other researchers to reprocess diffraction data and verify correctness of the protein model. However, the information about cloning, purification, crystallization, and sample preparation procedure is still limited and their availability depends on the meticulosity of the authors. The completeness of the information about experiment is higher for structures solved by the structural genomics and it is at least partially down to the use of databases (Domagalski et al. 2014). Because of the high-throughput nature of the SG workflow, only small number of structures solved by SG consortia is converted into peer-review papers. Therefore, SG consortia try to make the results available and useful to the scientific community in forms other than publications or PDB deposits (Zimmerman et al. 2014). The UniTrack system helps to reproduce the results by documentation of protocols and detailed specification of the experimental parameters. On every level of the structure determination pipeline, a large volume of metadata, including temperatures, volumes, dates, and name of the scientist that conducted the experiment supplement information. The information is further passed to PSI-SBKB, which coordinates progress of all projects within the Protein Structure Initiative. Public release of data helps to prevent duplication of scientific efforts. It is particularly important that the system release raw experimental data, e.g., unprocessed diffraction images, because it will allow their future analysis with more advanced technology potentially leading to unanticipated discoveries.

6.3 Main features of the UniTrack data management system

UniTrack is a target tracking system, functioning as a top data layer that collects information from LIMSs used in participating laboratories and other data resources. The system consists of the central target tracking database, data dissemination portal, supporting databases (i.e., MetaPDB, and MetaSG), LIMS communication layer, and target validation script. The LabDB LIMS is used by some instances of the system, and its development was not part of this work. The tracking database serves as central physical storage for all of the consortium experimental results. The CSGID data dissemination portal has been developed as a comprehensive web resource to store the data generated by the center as well as to provide both restricted and public access to its content. The web interface of the portal allows users to browse the details of experiments done on individual protein targets. The set of supporting tools consist of auxiliary databases, i.e., annotated database of PDB experimental data – MetaPDB, structural genomics progress tracking database - MetaSG, and applications, i.e., protein target validation system, experimental data import scripts, report generators, and many smaller tools.

The important feature of UniTrack is its distributed design. The system collects all information that is needed for coordination of work on particular targets from laboratories distributed in multiple locations. Research groups are both the senders and recipients of the information from other groups. Information is broadcasted not by one entity, but all research groups simultaneously. The system architecture where one central repository integrates data from many distributed sources is named data warehouse. The main advantage of such an approach is that the data can be fetched from any LIMS without interfering too much with that system. The LIMS systems are designed and maintained to meet needs of particular laboratories and changes in their architecture should be independent from data handling solutions used outside of the laboratory. UniTrack accepts input data in XML files that are relatively easy to prepare and does not require any changes on the LIMS side. Thus, each LIMS does not need to be exposed to outside of the laboratory, which helps maintain the privacy for some of the data. UniTrack has modular design and can be setup up to use different relational database management systems or LIMSs. The application provides customized access for multiple categories of users, which can be grouped into three main categories: researchers involved in project, public community, and advisory committee.

The main mission of CSGID is to support infectious disease scientific community by free of charge determination of important protein structures. Therefore, public dissemination of the experimental results is an important functionality of UniTrack. Information about protein targets, expression constructs, crystallization drops, crystals, and diffraction datasets, and released structures is publically available. Details of all experiments in structure determination pipeline for every target are accessible using an interactive node-link target browser. The target tracking database contains information about thousands of experiments, which may be used in the future for data mining studies, e.g., on protein crystallization conditions (based on approximately 2.5 million of crystallization trials for almost 60,000 of target proteins in four SG centers that use UniTrack). The CSGID portal is also an outreach platform for infectious disease community. The website contains structure annotations, ICM presentations, numerous links to related sites and external resources, moreover, highlights information about workshops, new articles, and the biggest achievements of the consortium.

6.4 Comparison of UniTrack with other SG data management systems

Data management in high-throughput structural biology has three distinguishable, but partially overlapping levels, i.e., experiment tracking, target tracking, and project tracking. LIMSs fulfill the base role of organizing laboratory resources and monitoring results of the experiments, those systems are prerequisite for every laboratory that does large-scale research. Sesame and HalX are two efficient and highly customized for SG research LIMSs. Traditional laboratories often use simpler LIMS-type tools, which are commonly called electronic laboratory notebooks (ELNs). Xtrack is a lightweight ELN system, which was designed specifically to manage collection of X-ray diffraction data. The second level of data management comprises functionality related to target selection, tracking experimental progress over targets, data dissemination, generation of statistics and reports. Spine, Spex Db, and UniTrack are examples of target tracking systems that were developed specifically for SG. The line between experiment and target tracking is arbitrary and many of the LIMS systems have limited target tracking functionality as well as target tracking systems often have features that are typical for LIMS. Nevertheless, classification into experiment or target tracking categories gives the best overall description of the system characteristics. On top of experiment and target tracking, the PSI-SBKB provides an additional third level of data management that aims to monitor progress and synchronize outcomes of all SG projects founded within Protein Structure Initiative.

Certain system features are universal and applicable at every level of the data management. Each of aforementioned systems, except the simplest Xtrack, has to some extent a modular architecture of its software components. A modular architecture allows for relatively straightforward implementation of additional functionality and assembling with complementary systems, e.g., UniTrack works with many instances of different LIMSs. Likewise, all systems that are used for multi-laboratory collaborative research have distributed architecture. Another feature used by all of the systems, except PSI-SBKB, is role-based access control. Users have different job functions and responsibilities, so some operations are assigned to certain roles, e.g., in UniTrack access to selected reports is restricted to people with lab contact and PI roles. Users have privileges that are not only role specific, but also group specific, e.g., in UniTrack users have authorization for editing experimental records if they are members of the same laboratory as the author of the experiment. Each SG data management system stores some metadata, i.e., information that characterize data. In UniTrack, metadata include temperatures, composition and volumes of media, expression strains, protein concentrations, type of crystallization plates, dates of experiments, names of the experimenters, references to protocols, software, and log files. Finally, all of the systems are based on the relational data model, which provides good data integrity and consistency, but has lower performance for very large databases.

UniTrack is very similar to other target tracking systems, i.e., Spex DB and SPINE. All three systems have analogous target list pages with search engines that allow filtering targets by multiple combined parameters, homology search tools, and statistics/report sections. The UniTrack's "node-link target browser" is the equivalent of the Spex's Db "tree view". However, in Spex Db experimental pipeline is presented in a table form, while in UniTrack analogical pipeline has a form of an interactive graph. SPINE and UniTrack web applications also share some similarities, i.e., statistics sections, progress summaries, and structure galleries. A unique feature of SPINE is a web tools and servers section, i.e., collection of bioinformatics web-tools developed by NESG, e.g., primer design, disorder prediction, homology searches, homology modeling, structure validation, and functional annotation tools. SPINE has some LIMS features, i.e., tracking of target sample tubes and detailed sample histories, however, it cannot be considered as a complete LIMS as NESG uses several distributed LIMS-type tools, e.g., LIMS dedicated only to crystallization. Unfortunately, public access to experimental details of particular targets in SPINE is practically limited to information about status and date of the last experiment. On the contrary, UniTrack makes all results publicly available at a very detailed level. Other features of UniTrack that make it distinctive among target tracking systems are the FTP repository for diffraction images, availability of virtual screening results, and automatically generated ICM presentations.

UniTrack as well as other target tracking systems has workflow-centered design based on strictly defined data model, which is an abstraction of high-throughput structure determination pipeline used by SG projects. The system is meant to serve *ad hoc* solutions for CSGID and therefore does not have some of the functionalities typical for the LIMS system. LIMS-type systems are more flexible and allow defining new types of experiments and modifying workflows. Moreover, LIMSs are often distributed as open-source software. LIMSs are often able to track target status, but does not have data dissemination features, statistics and progress reports, and contain much less annotations than target tracking systems. HalX and Sesame control laboratory equipment, store gel images, elution profiles, and pictures of crystals, while in UniTrack this functionality was passed to LabDB LIMS system. Nevertheless, UniTrack shares some features with experiment tracking systems, e.g., data extraction from log files, which is also one of the features of Xtrack.

Several systems, i.e., UniTrack, Xtrack, Sesame, Spine, and PSI-SBKB allow exporting data in CSV or TSV file formats that can be further analyzed in spreadsheet program or imported to other database. This feature enables generation of reports or statistics, which are not available from the web interface. Comparison of the SG data management systems by offered features is presented in Table 4.

Table 4 Comparison of the SG data management systems features.

	UniTrack	**Xtrack**	**Sesame**	**HalX**	**Spine**	**Spex Db**	**PSI-SBKB**
LIMS-type system	✗	✓	✓	✓	✓/✗	✗	✗
Target tracking system	✓	✗	✗	✗	✓	✓	✓
Distributed architecture	✓	✗	✓	✓	✓	✓	✓
Modular architecture	✓	✗	✓	✓	✓	✓	✓
Flexible workflows/ possibility to define new types of experiments	✗	✗	✓	✓	✗	✗	✗
Automated data import from other resources	✓	✗	✓	✗	✓	✗	✓
Target validation system / data integration tools	✓	✗	✓	✗	✓	✗	✓
Search tools	✓	✗	✓	✓	✓	✓	✓
Open-source	✗	✓	✓	✓	✗	✗	✗
Metadata management	✓	✓	✓	✓	✓	✓	✗
Data dissemination	✓	✗	✗	✗	✓	✓	✓

Data available for download (e.g., in CSV, TSV, XML, or MS EXCEL formats)	✓	✓	✓	✗	✓	✗	✓
Homology search tools	✓	✗	✗	✗	✓	✓	✓
Access control	✓	✓	✓	✓	✓	✓	✗
User roles	✓	✓	✓	✓	✓	✓	✗
Communication with laboratory instruments	✗*	✗	✓	✓	✗	✗	✗
Storage of gel images, elution profiles, etc.	✗*	✓	✓	✓	✓	✗	✗
Extraction of information from diffraction log files	✓	✓	✗	✗	✗	✗	✗
Exchange of data with other systems	✓	✗	✓	✓	✓	✓	✓
Diffraction images repository	✓	✗	✗	✗	✗	✗	✗
Report generation	✓	✗	✓	✗	✓	✓	✓
Structure annotations	✓	✗	✗	✗	✓	✓	✓
Statistics / progress reports	✓	✗	✗	✗	✓	✗	✓

* - functionality is available in LabDB.

6.5 Current use of UniTrack

The UniTrack is currently used by Center for Structural Genomics of Infectious Diseases (http://www.csgid.org), Midwest Center for Structural Genomics (http://www.mcsg.org), New York Structural Genomics Consortium (http://kiemlicz.med.virginia.edu/nysgrc), and Enzyme Function Initiative (http://kiemlicz.med.virginia.edu/efi). Each entity of the system has a common central database architecture and set of tools developed for handling targets and experimental data. Project portals are based on a template developed for CSGID, which contains shared parts like target search browser, tree-view target explorer, statistics section, and data input forms. However, some of the portal features are available in single center, e.g., virtual screening results explorer in CSGID. UniTrack was designed to complement each other with the LabDB LIMS (Zimmerman 2005). The systems serve as a complete data management solution for NYSGRC and EFI projects as well as for selected groups inside CSGID and MCSG. The view layer of the UniTrack is highly customized for the needs of particular center or consortium of research laboratories. The total number of database records stored by the system proves it scalability and usability for data management of large-scale projects. Considering all four instances of the system, UniTrack stores 300 protocols and controls progress of almost 60,000 protein targets, 95,000 cloning, 30,000 expressions, 26,000 purifications, 2,600,000 crystallization trials, and over 4,000 X-ray diffraction experiments, that resulted in more than 1,700 structures deposited to PDB. The total number of database records for selected experimental categories in four instances of UniTrack is presented in Table 5.

Table 5 Number of experimental records stored in four instances of the UniTrack (as on 20/01/2015). For each category, number of targets and number of experiments is displayed.

	CSGID	**MCSG**	**NYSGRC**	**EFI**	**total**
Protein targets	7,448	8,520	33,221	10428	59,617
Cloning trials	6,733/ 16,878	6,116/ 30,000	19,291/ 36,798	6,298/ 11,757	38,438/ 95,433
Soluble expressing clones	3,980/ 6,561	4,024/ 11,072	6,814/ 8,442	3,343/ 4,651	18,161/ 30,726
Successful purifications	2,306/ 6,895	2056/ 11913	2,903/ 4,168	1,779/ 3,459	9,044/ 26,435
Crystallization trials	1,564/ 528,221	1,084/ 205,453	1,887/ 1,084,467	933/ 787,165	5468/ 2,605,306
Crystals harvested	690/4,760	605/2,712	345/3,543	215/698	1,855/ 11,713
X-ray diffraction data sets	516/1,733	329/1,238	235/1,259	201/360	1,281/ 4,590
Structures solved	512/826	325/513	218/284	188/284	1,243/ 1,907
Structures deposited to PDB	508/739	319/441	213/261	188/281	1228/1722
Protocols	72	83	110	34	299
Diffraction images uploaded	548	170	-	-	718

6.6 Future development prospects for UniTrack

The UniTrack data management system has proved its ability to facilitate high-throughput SG projects. Nevertheless, the SG field is constantly evolving and several things can be improved upon in this project. First, there are areas where the target tracking database architecture could be optimized, e.g., implementation of crystallization experiment. Crystallization experiments are often automated and usually done on screening plates containing two or three crystallization drops for every of 96 different crystallization conditions. In its current implementation, each crystallization drop that was setup on the plate is stored in the database as a separate record. The excessive amount of crystallization data is particularly burdensome while the complete experimental history of a target is being displayed using the tree view browser. Crystallization drop records not only slow down the database queries, but also impede finding the successful experimental paths on the target tree. The problem was partially eliminated by hiding from the tree view browser all crystallization drops that did not produce any crystals and by highlighting the most advanced experimental paths with darker color. Since most of the crystallization experiments are setup using a limited number of commercial sparse matrix sampling kits, introduction of crystallization plate template and replacement of crystallization drop entity with a crystal entity would limit number of records for all crystallization experiments that use crystallization screens. The proposed change would force changes in structure of the XML files used for transferring data from LIMSs and therefore it cannot be easily implemented.

While browsing web interface, users may experience long response times to some complex queries, e.g., when accessing real time statistics. In order to limit the problem, database queries were optimized, database tables were indexed and defragmented, and CakePHP cache was turned on for selected pages. Although well-structured SG data naturally fit to relational database model used by UniTrack, depending on the speed of database growth and related loss of its performance, upgrade to NoSQL database (i.e., graph database, key-value store, or columnar database) might be worth considering.

The other possible improvement specific to CSGID would be implementation of infectious disease and protein ontologies to represent the relationship between studied protein targets and the bacteria pathogenicity in humans. Bio-ontologies define relationships among different kinds of entities providing interoperability between databases and supporting the annotation and analysis of large-scale data. In the future, it will be possible to use them to combine heterogeneous data from different resources, and then analyze with systems biology methods towards discovery of relevant biomedical patterns.

One feature that could be introduced to UniTrack that could have the highest impact on efficiency improving is automated structure deposition to PDB. Unfortunately, in UniTrack's workflow the final step of the structure determination pipeline, i.e., structure deposition to PDB is still manual and time-consuming process. In order to deposit a structure in PDB, the responsible crystallographer must solve all problems detected by ADIT validation software, convert and check crystallographic structure factors, extract information from log files, and fill-up multiple detailed web forms. The required information includes contact authors details (i.e., name, e-mail address, postal address, phone and fax numbers), a title for the deposited structure and any relevant keywords, macromolecule names, sequence and chain ID for each macromolecule, including expression tags and residues missing due to disorder, information about source organism, expression systems, citation, ligand names and chemical diagrams. A large part of this information is stored in target tracking database and could be used for generation of annotated PDB file. Additional information stored in log files could be retrieved using PDB deposition software. The RCSB PDB already tried to automate the deposition process for SG during first phase of PSI, but the final goal was never achieved due to many revisions and updates to the PDB format and deposition procedure itself.

The determination of the quality of X-ray structures is still very challenging problem. Statistics provided by the UniTrack system are helpful for detecting structures that does not fulfill the quality metrics. The CSGID portal is also hosting CheckMyMetal server, which enables validation of metal binding sites. Incorporation of existing and development of new structure validation tools would be a suitable addition to the project.

A very important aspect of a large-scale project is public outreach. In order to accomplish long-term project objectives we need to build the public awareness of the scientific problem and promote project achievements. The RCSB PDB provides an educational service named PDB 101, which regularly releases short articles about the most interesting structures in PDB (i.e., 'Molecule of the month' authored by David

Goodsell). NIAID runs a similar service featuring the most interesting structures of CSGID and SSGCID. PSI-SBKB publishes short notes about featured structures and articles. The CSGID portal contains a ‘Milestones and Press Coverage’ section that is a list of short notes about CSGID structures featuring in press and other services. This page can be refactored to regularly present some of the most interesting research with more detail, e.g., in the form of extended Molsoft ICM presentations. New updates of the page should be spread to the community through an email newsletter.

6.7 Active site composition and putative catalytic mechanism of DhbC

B. anthracis, similar to closely related *B. subtilis* and *E. coli,* contains two isochorismate synthase genes *dhbC* (*entC* in *E. coli*) and *menF*, which are located in the biosynthetic operons of catecholate siderophore bacillibactin (enterobactin in *E. coli*) and respiratory chain component menaquinone, respectively. The crystal structure of *B. anthracis* DhbC shows a high degree of similarity, including nearly identical active sites, to both *E. coli* EntC and MenF structures, which have been previously solved. The ISC-type active site is also very similar to the active site of the other members of ADC synthase-like fold, i.e., anthranilate synthase (AS), 4-amino-4-deoxychorismate synthase (ADCS), and salicylate synthase (SS), suggesting that the proteins utilize very similar catalytic mechanism to convert chorismate into different products. Analysis of the conservation and spatial arrangement of the active site residues combined with findings from mutational studies done previously on members of ADC synthase superfamily support a putative Mg^{2+}-dependent catalytic mechanism originally formulated by Walsh and refined by He, Kolappan and other researchers (Walsh et al. 1987; He et al. 2004; Kolappan et al. 2007; Ziebart et al. 2010). Roles of the essential active site residues in DhbC were deduced from the aforementioned analyses.

According to mutational studies of *E. coli* EntC (Sridharan et al. 2010) Ala304 (Ala303 of EntC; Ala344 of *E. coli* MenF) plays an important role in positioning the peptide-bond carbonyl, enabling the formation of a proper hydrogen bond to the isochorismate C2 hydroxyl. In EntC, the A303T mutation as well as mutation of the neighboring Leu304 (Val305 of DhbC; Val345 of *E. coli* MenF) to alanine resulted in complete loss of enzyme activity. In other chorismate-utilizing enzymes, the two corresponding residues are Ser338 and Met339 in *Cytophaga hutchinsonii* ADC synthase PabB, Ser366 and Ile367 in *E. coli* ADC synthase PabB, Thr348 and Ala349 in *Y. enterocolitica* salicylate synthase Irp9, and Thr361 and Ala362 in *M. tuberculosis* salicylate synthase Mbtl. In fact, presence of a threonine in the position corresponding to EntC's Ala304 is the only difference between the conserved active site residues of salicylate synthases and isochorismate synthases. Strict conservation of the Ala-Val/Leu residue pair in DhbC, EntC, and MenF and aforementioned mutational studies in *E. coli* indicate its importance in sustaining isochorismate synthase activity.

In the EntC structure, a magnesium ion is coordinated by Glu241, Glu373, and Glu376 sidechains (Glu241, Glu374, and Glu377 of DhbC), and C1 carboxylate of isochorismate. In the structure of DhbC, neither Mg^{2+} ion nor chorismate or isochorismate are present. However, a sulfate ion occupies the position corresponding to location of C1 carboxylate, while the Glu241 sidechain flips to the surface of the protein opening up the active site cavity. Unfortunately, trials to crystallize DhbC with chorismate or isochorismate bound resulted in crystals diffracting to ~2.8 Å, which was not enough for accurate interpretation of the details of the protein-ligand interactions.

The Lys142 of DhbC corresponds to Lys147 of *E. coli* EntC and Lys190 of *E. coli* MenF, which are thought to act as a catalytic base by activating a nucleophilic water molecule that is also hydrogen bonded to the C2 hydroxyl group of isochorismate (He and Toney 2006; Kolappan et al. 2007; Ziebart and Toney 2010). Lysine is conserved in this position in isochorismate and salicylate synthases, while anthranilate and ADC synthases have a glutamine or asparagine residue. In ADC synthases from *E. coli* (PDB entry 1k0e) (Parsons et al. 2002) and *C. hutchinsonii* (PDB entry 3h9m; New York SGX Research Center for Structural Genomics, unpublished work) corresponding positions are occupied by Glu210 and Glu182, respectively.

The role of the general base for the loss of the C4 hydroxyl from chorismate is fulfilled by Glu197 (Glu197 in EntC and Glu240 in MenF) that points towards C4 of the bound isochorismate (Sridharan et al. 2010).

The function of the two aromatic residues Phe328 and Tyr360, which form parallel-displaced π-stacking interaction, oriented parallel to the expected position of chorismate ring, is unknown. The stacking interaction is present in *E. coli* isochorismate synthases, i.e., EntC and MenF. In the structure of MenF, the orientation of the aromatic pair Tyr368–Tyr399 is identical to DhbC. However, in the structure of EntC, aromatic rings of two phenylalanine residues, Phe327 and Phe359, are oriented parallel to the pyruvyl group of chorismate as can be clearly seen in Figure 24. The F327Y EntC mutation results in a 48-fold decrease in enzyme efficiency, while the double mutant F327Y/I346L results in a 750-fold decrease of wild-type activity (Sridharan et al. 2010). Phe327 is conserved in DhbC and EntC, while in MenF and Irp9 it is substituted by tyrosine. The aromatic pair is not preserved in *C. hutchinsonii* ADC synthase (corresponding residues are Phe362 and Glu396; PDB entry 3h9m) and *E. coli* ADC synthase (Trp390 and Ser422; PDB entry 1k0e) and in salicylate synthase from *Y. enterocolitica* (Tyr372 and Gln403; PDB entry 2fn0) (Kerbarh et al. 2006). The presence of the stacking interaction in isochorismate synthases may be one of the functional determinants.

In conclusion, Ala304 positions chorismate for nucleophilic attack at the C2 position of chorismate by forming a hydrogen bond with the C2 hydroxyl. Lys142 is the catalytic base that activates the water molecule for nucleophilic attack at the C2 hydroxyl group of chorismate via an Mg^{2+}-bound transition state. Glu197 is a general acid for subsequent loss of the C4 hydroxyl (Domagalski et al. 2013). The S_N2' attack on C2 hydroxyl is a possible common mechanism for isochorismate synthase, salicylate synthase, anthranilate synthase, and ADC synthase. Subsequent events differentiate the enzymes, isochorismate synthase simply releases the product, while salicylate and anthranilate synthases additionally remove pyruvate, and ADC synthase performs second S_N2' attack replacing C4 hydroxyl with amino group (He et al. 2004).

6.8 DhbC as a potential drug target

The exact role of isochorismate synthase DhbC in *B. anthracis* pathogenicity is not clear. The enzyme catalyzes the first step in the biosynthesis of catecholate siderophore bacillibactin, which is one of two siderophores produced by this organism. Studies on iron acquisition in *B. anthracis* showed that siderophores are essential for mouse virulence and that petrobactin, but not bacillibactin, is necessary for initiating infection in this model organism (Cendrowski et al. 2004). However, the authors of this research indicate that is does not necessarily exclude bacillibactin from playing a role in other host species. They note the example of mutagenesis in the *Brucella abortus* catechol siderophore biosynthetic pathway (brucebactin), which does not influence virulence in mice but results in an avirulent phenotype in cows, its primary host.

The *B. anthracis* infection has two stages: an establishment stage within phagocytes and an extracellular stage that leads to sepsis and death. The aforementioned study showed that petrobactin is key siderophore during the intracellular stage, but did not explain what happens during the extracellular stage of the infection. Other studies on siderophores secretion during *B. anthracis* spore germination and overgrowth in culture also indicated that spore development may need petrobactin early in an infection, while delayed bacillibactin production suggests that it plays a role in the later stages of infection (Wilson et al. 2010). The main reason for preferential expression of petrobactin at early stages of the infection is the fact that it is not recognizable by host protein siderocalin, which recognizes and deactivates catechol siderophores (Abergel et al. 2006). This feature of *B. anthracis* allows the anthrax bacteria to trick a host immune system. When the bacteria start to replicate rapidly and iron concentration drops down, expression of bacillibactin, which has much higher affinity for iron, is possibly triggered. The complex regulation of the bacillibactin operon expression by ferric uptake regulator (Fur), catabolite control protein A (CcpA) (Wunsche et al. 2012), oxygen depletion, and salinity suggests that bacillibactin is important for pathogenicity. It cannot be excluded that *B. anthracis* uses different sources of iron and therefore a different acquisition strategies depending on the route of infection. Recent studies on closely related species *B. cereus* (*B. anthracis* is a member of *B. cereus sensu lato* group) demonstrated that cooperation of bacillibactin and surface protein IlsA is crucial for iron acquisition from host ferritin and therefore effective virulence in insects (Segond et al. 2014). The *B. anthracis* genome encodes two proteins homologous to IslA, i.e., BslL, which is nearly identical to last three fourths of IlsA and BslK, which shares similarity with NEAT and SLH domains of IslA. BslK was shown to bind heme and mediate heme delivery to Isd system. Iron-regulated surface determinant (Isd) system of *B. anthracis* takes part in extraction of heme from hemoglobin during the extracellular phase of infection. It remains unknown if BslK, analogous to IslA, requires cooperation with bacillibactin. The proteins that consist on *B. anthracis* Isd system, i.e., IsdC (IDP05488), IsdJ (IDP05417), and IsdK (IDP02799) were selected for structure determination by CSGID. Unfortunately, thus far all experimental attempts for the proteins failed on expression or purification trials. Additional studies on the regulation of the *B. anthracis asb* and *bac* siderophore operons during the extracellular stage of infection and when heme is used as an iron source are necessary to elucidate if targeting bacillibactin biosynthesis has anti-virulence potential.

Regardless of whether the bacillibactin production is necessary for *B. anthracis* pathogenicity, DhbC is still very promiscuous drug target. The product of its reaction, isochorismate is needed for synthesis of an important component of the electron transport chain, menaquinone. DhbC, but not the opposite can compensate lack of menaquinone biosynthesis-specific isochorismate synthase MenF. Sequence alignment of the two enzymes suggests that they share an active site composition and a catalytic mechanism, raising the chances to developing a single inhibitor compound for both enzymes. Moreover, a similar active site composition is characteristic for anthranilate synthase, salicylate synthase, aminodeoxychorismate synthase, and 2-amino-2-desoxyisochorismate synthase, according to several crystal structures solved to date. The active site similarity suggests that all the aforementioned enzymes use a related mechanisms that includes the addition of either nitrogen or oxygen nucleophiles to C2 of chorismate. The availability of multiple structures from this protein superfamily may allow finding a single inhibitor of chorismate-binding enzymes. Development of such compound requires a better understanding of the mechanistic features that differentiate catalytic activities of chorismate-binding enzymes.

Succeeding studies of DhbC will focus on identifying the inhibitors of isochorismate synthase activity by structure-based virtual screening of large compound libraries. The effectiveness of multienzyme inhibitors identified by Ziebart (Ziebart et al. 2010) should be also tested on DhbC using inhibition assays.

7. Conclusions

The result of this work is the three-dimensional structure of isochorismate synthase DhbC from *B. anthracis,* which was solved using the CSGID high-throughput gene-to-structure pipeline under control of the UniTrack data management system. The DhbC's putative molecular function was confirmed and the enzyme was kinetically characterized using spectrophotometric assays. Because chorismate is a branch point metabolite in multiple solely bacterial pathways, DhbC and other chorismate-utilizing enzymes are promising targets in research for new generation of anti-pathogenic drugs. The structure of DhbC will guide the search for potent inhibitors of bacillibactin formation, and hence potentially of bacterial iron uptake. Development of key components of UniTrack was an essential part of this work. The system was designed to monitor, share, document, and publically release experimental details of the structure determination process for every CSGID protein target. The UniTrack-derived system consists of a central database of experimental data and a set of auxiliary databases and applications, which collect and integrate experimental data provided by distributed LIMSs in participating laboratories. Customized variants of the UniTrack system are deployed in the Center for Structural Genomics of Infectious Diseases, the Midwest Center for Structural Genomics, the New York Structural Genomics Consortium, and the Enzyme Function Initiative. The common components of the UniTrack-based data management system are a target tracking database, knowledge dissemination portal, target validation tool, communication scripts, and supporting databases. Additionally, instances used by NYSGRC and EFI are integrated with experiment the tracking system LabDB LIMS. The target tracking database stores data for all structure determination experiments of the CSGID, MCSG, NYSGRC, and EFI protein targets. The knowledge dissemination portal provides an access to experimental information, numerous statistics concerning general progress of the consortium and manual annotations for protein structures solved within the project. UniTrack makes the results and protocols for all steps of experimental pipeline starting from target selection to structure solution and deposition publicly available. Failed trails are also accessible, giving an overview on bottlenecks for a particular target and saving precious time for researchers that would like to continue the studies.

Bibliography

Abagyan, R., Totrov, M., Kuznetsov, D. (1994). "ICM—A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation." Journal of Computational Chemistry **15**(5): 488-506.

Abergel, R. J., M. K. Wilson, et al. (2006). "Anthrax pathogen evades the mammalian immune system through stealth siderophore production." Proc Natl Acad Sci U S A **103**(49): 18499-18503.

Abergel, R. J., A. M. Zawadzka, et al. (2008). "Petrobactin-mediated iron transport in pathogenic bacteria: coordination chemistry of an unusual 3,4-catecholate/citrate siderophore." J Am Chem Soc **130**(7): 2124-2125.

Albeck, S., P. Alzari, et al. (2006). "SPINE bioinformatics and data-management aspects of high-throughput structural biology." Acta Crystallogr D Biol Crystallogr **62**(Pt 10): 1184-1195.

Almo, S. C., J. B. Bonanno, et al. (2007). "Structural genomics of protein phosphatases." J Struct Funct Genomics **8**(2-3): 121-140.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-3402.

Anderson, W. F. (2009). "Structural genomics and drug discovery for infectious diseases." Infect Disord Drug Targets **9**(5): 507-517.

Andrews, S. C., A. K. Robinson, et al. (2003). "Bacterial iron homeostasis." FEMS Microbiol Rev **27**(2-3): 215-237.

Apweiler, R., A. Bairoch, et al. (2004). "Protein sequence databases." Curr Opin Chem Biol **8**(1): 76-80.

Arakawa, T., K. Tsumoto, et al. (2007). "The effects of arginine on protein binding and elution in hydrophobic interaction and ion-exchange chromatography." Protein Expr Purif **54**(1): 110-116.

Aslanidis, C. and P. J. Dejong (1990). "Ligation-Independent Cloning of Pcr Products (Lic-Pcr)." Nucleic Acids Res **18**(20): 6069-6074.

Athamna, A., M. Athamna, et al. (2004). "Selection of Bacillus anthracis isolates resistant to antibiotics." J Antimicrob Chemother **54**(2): 424-428.

Bagg, A. and J. B. Neilands (1987). "Ferric uptake regulation protein acts as a repressor, employing iron (II) as a cofactor to bind the operator of an iron transport operon in Escherichia coli." Biochemistry **26**(17): 5471-5477.

Baichoo, N., T. Wang, et al. (2002). "Global analysis of the Bacillus subtilis Fur regulon and the iron starvation stimulon." Mol Microbiol **45**(6): 1613-1629.

Barbosa, T. M. and S. B. Levy (2000). "The impact of antibiotic use on resistance development and persistence." Drug Resist Updat **3**(5): 303-311.

Bateman, A., L. Coin, et al. (2004). "The Pfam protein families database." Nucleic Acids Res **32**(Database issue): D138-141.

Begley, C. G. and L. M. Ellis (2012). "Drug development: Raise standards for preclinical cancer research." Nature **483**(7391): 531-533.

Beierlein, J. M. and A. C. Anderson (2011). "New developments in vaccines, inhibitors of anthrax toxins, and antibiotic therapeutics for Bacillus anthracis." Curr Med Chem **18**(33): 5083-5094.

Benson, D. A., M. Cavanaugh, et al. (2013). "GenBank." Nucleic Acids Res **41**(Database issue): D36-42.

Bergman, N. H. (2011). Bacillus anthracis and anthrax. Hoboken, N.J., John Wiley & Sons.

Berman, H., K. Henrick, et al. (2003). "Announcing the worldwide Protein Data Bank." Nat Struct Biol **10**(12): 980.

Berman, H., K. Henrick, et al. (2007). "The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data." Nucleic Acids Res **35**(Database issue): D301-303.

Berman, H. M., J. Westbrook, et al. (2000). "The Protein Data Bank." Nucleic Acids Res **28**(1): 235-242.

Berman, H. M., J. D. Westbrook, et al. (2009). "The protein structure initiative structural genomics knowledgebase." Nucleic Acids Res **37**(Database issue): D365-368.

Bernal, J. D., Fankuchen, I., Perutz, M.F. (1938). "An X-ray study of chymotrypsin and haemoglobin." Nature **141**: 523-524.

Bernstein, F. C., T. F. Koetzle, et al. (1977). "The Protein Data Bank: a computer-based archival file for macromolecular structures." J Mol Biol **112**(3): 535-542.

Bertone, P., Y. Kluger, et al. (2001). "SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics." Nucleic Acids Res **29**(13): 2884-2898.

Binkowski, T. A., W. Jiang, et al. (2014). "Virtual high-throughput ligand screening." Methods Mol Biol **1140**: 251-261.

Bio-Rad (2012). Image Lab^{TM}.

Bonanno, J. B., S. C. Almo, et al. (2005). "New York-Structural GenomiX Research Consortium (NYSGXRC): a large scale center for the protein structure initiative." J Struct Funct Genomics **6**(2-3): 225-232.

Bradaric, N. and V. Punda-Polic (1992). "Cutaneous anthrax due to penicillin-resistant Bacillus anthracis transmitted by an insect bite." Lancet **340**(8814): 306-307.

Bragg, W. L. (1913). "The Diffraction of Short Electromagnetic Waves by a Crystal." Proceedings of the Cambridge Philosophical Society **17**: 43-57.

Brenner, S. E. (2001). "A tour of structural genomics." Nat Rev Genet **2**(10): 801-809.

Brenner, S. E. and M. Levitt (2000). "Expectations from structural genomics." Protein Sci **9**(1): 197-200.

Brook, I., T. B. Elliott, et al. (2001). "In vitro resistance of Bacillus anthracis Sterne to doxycycline, macrolides and quinolones." Int J Antimicrob Agents **18**(6): 559-562.

Brunger, A. T., P. D. Adams, et al. (1998). "Crystallography & NMR system: A new software suite for macromolecular structure determination." Acta Crystallogr D Biol Crystallogr **54**(Pt 5): 905-921.

Bullen, J. J. (1972). "Iron-binding proteins in milk and resistance to Escherichia coli infection in infants." Proc R Soc Med **65**(12): 1086.

Bullen, J. J., L. C. Leigh, et al. (1968). "The effect of iron compounds on the virulence of Escherichia coli for guinea-pigs." Immunology **15**(4): 581-588.

Bullen, J. J., C. G. Ward, et al. (1991). "The critical role of iron in some clinical infections." Eur J Clin Microbiol Infect Dis **10**(8): 613-617.

Cairo, G., F. Bernuzzi, et al. (2006). "A precious metal: Iron, an essential nutrient for all cells." Genes Nutr **1**(1): 25-39.

Carlson, P. E., Jr., K. A. Carr, et al. (2009). "Transcriptional profiling of Bacillus anthracis Sterne (34F2) during iron starvation." PLoS One **4**(9): e6988.

Carrano, C. J. and K. N. Raymond (1978). "Coordination chemistry of microbial iron transport compounds: rhodotorulic acid and iron uptake in Rhodotorula pilimanae." J Bacteriol **136**(1): 69-74.

Caza, M. and J. W. Kronstad (2013). "Shared and distinct mechanisms of iron acquisition by bacterial and fungal pathogens of humans." Front Cell Infect Microbiol **3**: 80.

CDC (2013). Antibiotic Resistance Threats in the United States, 2013, U.S. Centers for Disease Control and Prevention.

Cendrowski, S., W. MacArthur, et al. (2004). "Bacillus anthracis requires siderophore biosynthesis for growth in macrophages and mouse virulence." Mol Microbiol **51**(2): 407-417.

Chen, L., R. Oughtred, et al. (2004). "TargetDB: a target registration database for structural genomics projects." Bioinformatics **20**(16): 2860-2862.

Chen, V. B., Arendall III, W.B., Headd, J.J., Keedy, D.A., Immormino, R. M., Kapral, G.J., Murray, L.W., Richardson, J.S. and Richardson, D.C. (2010). "MolProbity: all-atom structure validation for macromolecular crystallography." Acta Crystallographica Section D **66**: 12-21.

Choffnes ER, R. D., Mack A (2010). Antibiotic Resistance: Implications for Global Health and Novel Intervention Strategies: Workshop Summary. Washington (DC).

Chruszcz, M., M. Domagalski, et al. (2010). "Unmet challenges of structural genomics." Curr Opin Struct Biol **20**(5): 587-597.

Cock, P. J., T. Antao, et al. (2009). "Biopython: freely available Python tools for computational molecular biology and bioinformatics." Bioinformatics **25**(11): 1422-1423.

Cowtan, K. D. and P. Main (1993). "Improvement of Macromolecular Electron-Density Maps by the Simultaneous Application of Real and Reciprocal Space Constraints." Acta Crystallographica Section D-Biological Crystallography **49**: 148-157.

Cowtan, K. D. and K. Y. Zhang (1999). "Density modification for macromolecular phase improvement." Prog Biophys Mol Biol **72**(3): 245-270.

Coy, M. and J. B. Neilands (1991). "Structural dynamics and functional domains of the fur protein." Biochemistry **30**(33): 8201-8210.

Dahm, C., R. Muller, et al. (1998). "The role of isochorismate hydroxymutase genes entC and menF in enterobactin and menaquinone biosynthesis in Escherichia coli." Biochim Biophys Acta **1425**(2): 377-386.

Daruwala, R., D. K. Bhattacharyya, et al. (1997). "Menaquinone (vitamin K2) biosynthesis: overexpression, purification, and characterization of a new isochorismate synthase from Escherichia coli." J Bacteriol **179**(10): 3133-3138.

Dasgupta, S., G. H. Iyer, et al. (1997). "Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers." Proteins-Structure Function and Genetics **28**(4): 494-514.

Derewenda, Z. S. and P. G. Vekilov (2006). "Entropy and surface engineering in protein crystallization." Acta Crystallogr D Biol Crystallogr **62**(Pt 1): 116-124.

Dertz, E. A., J. Xu, et al. (2006). "Bacillibactin-mediated iron transport in Bacillus subtilis." J Am Chem Soc **128**(1): 22-23.

Dixon, T. C., M. Meselson, et al. (1999). "Anthrax." N Engl J Med **341**(11): 815-826.

Domagalski, M. J., K. L. Tkaczuk, et al. (2013). "Structure of isochorismate synthase DhbC from Bacillus anthracis." Acta Crystallogr Sect F Struct Biol Cryst Commun **69**(Pt 9): 956-961.

Domagalski, M. J., H. Zheng, et al. (2014). "The quality and validation of structures from structural genomics." Methods Mol Biol **1091**: 297-314.

Dong, A., X. Xu, et al. (2007). "In situ proteolysis for protein crystallization and structure determination." Nat Methods **4**(12): 1019-1021.

Dosselaere, F. and J. Vanderleyden (2001). "A metabolic node in action: chorismate-utilizing enzymes in microorganisms." Crit Rev Microbiol **27**(2): 75-131.

Drenth, J. (1999). Principles of protein x-ray crystallography. New York, Springer.

Emsley, P. and K. Cowtan (2004). "Coot: model-building tools for molecular graphics." Acta Crystallographica Section D-Biological Crystallography **60**: 2126-2132.

Eschenfeldt, W. H., S. Lucy, et al. (2009). "A family of LIC vectors for high-throughput cloning and purification of proteins." Methods Mol Biol **498**: 105-115.

Evans, G. (2002). "Bioterrorism watch. Anthrax aftermath: adverse drug reactions, vaccine controversy undercut CDC extended treatment offer." Hosp Peer Rev **27**(3): suppl 1-4.

Ferreras, J. A., J. S. Ryu, et al. (2005). "Small-molecule inhibition of siderophore biosynthesis in Mycobacterium tuberculosis and Yersinia pestis." Nat Chem Biol **1**(1): 29-32.

Frey, M. (1994). "Water structure associated with proteins and its role in crystallization." Acta Crystallogr D Biol Crystallogr **50**(Pt 4): 663-666.

Frickey, T. and A. Lupas (2004). "CLANS: a Java application for visualizing protein families based on pairwise similarity." Bioinformatics **20**(18): 3702-3704.

Gabanyi, M. J., P. D. Adams, et al. (2011). "The Structural Biology Knowledgebase: a portal to protein structures, sequences, functions, and methods." J Struct Funct Genomics **12**(2): 45-54.

Garfin, D. E. (2003). Gel Electrophoresis of Proteins. Cell Structure, A Practical Approach. J. D. a. J. M. Lord. Oxford UK, University Oxford Press. **1:** 197-268.

Gat, O., G. Zaide, et al. (2008). "Characterization of Bacillus anthracis iron-regulated surface determinant (Isd) proteins containing NEAT domains." Mol Microbiol **70**(4): 983-999.

GE Healthcare, L. S. (2011). PrimeView™.

Gerdts, C. J., M. Elliott, et al. (2008). "The plug-based nanovolume Microcapillary Protein Crystallization System (MPCS)." Acta Crystallogr D Biol Crystallogr **64**(Pt 11): 1116-1122.

Gerlt, J. (2014) "EFI inside: the newsletter for the enzyme function initiative."

Gerlt, J. A., K. N. Allen, et al. (2011). "The Enzyme Function Initiative." Biochemistry **50**(46): 9950-9962.

Goh, C. S., N. Lan, et al. (2003). "SPINE 2: a system for collaborative structural proteomics within a federated database framework." Nucleic Acids Res **31**(11): 2833-2838.

Griffiths, E. (1991). "Iron and bacterial virulence--a brief overview." Biol Met **4**(1): 7-13.

Hampton Research, C. (2013). Crystal Screen HT^{TM} User Guide.

Harris, M. and T. A. Jones (2002). "Xtrack - a web-based crystallographic notebook." Acta Crystallogr D Biol Crystallogr **58**(Pt 10 Pt 2): 1889-1891.

Haun, R. S. and J. Moss (1992). "Ligation-Independent Cloning of Glutathione-S-Transferase Fusion Genes for Expression in Escherichia-Coli." Gene **112**(1): 37-43.

He, Z., K. D. Stigers Lavoie, et al. (2004). "Conservation of mechanism in three chorismate-utilizing enzymes." J Am Chem Soc **126**(8): 2378-2385.

He, Z. and M. D. Toney (2006). "Direct detection and kinetic analysis of covalent intermediate formation in the 4-amino-4-deoxychorismate synthase catalyzed reaction." Biochemistry **45**(15): 5019-5028.

Hoffmann, T., A. Schutz, et al. (2002). "High-salinity-induced iron limitation in Bacillus subtilis." J Bacteriol **184**(3): 718-727.

Holm, L. and P. Rosenstrom (2010). "Dali server: conservation mapping in 3D." Nucleic Acids Res **38**(Web Server issue): W545-549.

Hotta, K., C. Y. Kim, et al. (2010). "Siderophore-mediated iron acquisition in Bacillus anthracis and related strains." Microbiology **156**(Pt 7): 1918-1925.

Hünefeld, F. L. (1840). Chemismus in der Thierischen Organization: 158-163.

Hurd, H. S., S. Doores, et al. (2004). "Public health consequences of macrolide use in food animals: a deterministic risk assessment." J Food Prot **67**(5): 980-992.

Ioannidis, J. P. (2014). "How to make more published research true." PLoS Med **11**(10): e1001747.

Kendrew, J. C., G. Bodo, et al. (1958). "A three-dimensional model of the myoglobin molecule obtained by x-ray analysis." Nature **181**(4610): 662-666.

Kerbarh, O., E. M. Bulloch, et al. (2005). "Mechanistic and inhibition studies of chorismate-utilizing enzymes." Biochem Soc Trans **33**(Pt 4): 763-766.

Kerbarh, O., D. Y. Chirgadze, et al. (2006). "Crystal structures of Yersinia enterocolitica salicylate synthase and its complex with the reaction products salicylate and pyruvate." J Mol Biol **357**(2): 524-534.

Khafizov, K., C. Madrid-Aliste, et al. (2014). "Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative." Proc Natl Acad Sci U S A **111**(10): 3733-3738.

Kierzek, A. M. and P. Zielenkiewicz (2001). "Models of protein crystal growth." Biophys Chem **91**(1): 1-20.

Kim, Y., P. Quartey, et al. (2008). "Large-scale evaluation of protein reductive methylation for improving protein crystallization." Nat Methods **5**(10): 853-854.

Kolappan, S., J. Zwahlen, et al. (2007). "Lysine 190 is the catalytic base in MenF, the menaquinone-specific isochorismate synthase from Escherichia coli: implications for an enzyme family." Biochemistry **46**(4): 946-953.

Kouranov, A., L. Xie, et al. (2006). "The RCSB PDB information portal for structural genomics." Nucleic Acids Res **34**(Database issue): D302-305.

Krewulak, K. D. and H. J. Vogel (2008). "Structural biology of bacterial iron uptake." Biochim Biophys Acta **1778**(9): 1781-1804.

Krissinel, E. and K. Henrick (2004). "Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions." Acta Crystallogr D Biol Crystallogr **60**(Pt 12 Pt 1): 2256-2268.

Krissinel, E. and K. Henrick (2007). "Inference of macromolecular assemblies from crystalline state." J Mol Biol **372**(3): 774-797.

LABTECH, B. (2011). MARS Data Analysis Software.

Laemmli, U. K. (1970). "Cleavage of structural proteins during the assembly of the head of bacteriophage T4." Nature **227**(5259): 680-685.

Lalitha, M. K. and M. K. Thomas (1997). "Penicillin resistance in Bacillus anthracis." Lancet **349**(9064): 1522.

Lathe, G. H. and C. R. Ruthven (1956). "The separation of substances and estimation of their relative molecular sizes by the use of colums of starch in water." Biochem J **62**(4): 665-674.

Lee, D., T. A. de Beer, et al. (2011). "1,000 structures and more from the MCSG." BMC Struct Biol **11**: 2.

Lee, J. Y., K. D. Passalacqua, et al. (2011). "Regulation of petrobactin and bacillibactin biosynthesis in Bacillus anthracis under iron and oxygen variation." PLoS One **6**(6): e20777.

Li, M., F. Dimaio, et al. (2011). "Crystal structure of XMRV protease differs from the structures of other retropepsins." Nat Struct Mol Biol **18**(2): 227-229.

Li, Q. A., Mavrodi, D. V., Thomashow, L. S., Roessle, M., Blankenfeldt, W. (2011). "Ligand binding induces an ammonia channel in 2-amino-2-desoxyisochorismate (ADIC) synthase PhzE." J Biol Chem **286**(20): 18213-18221.

Liu, J., N. Quinn, et al. (1990). "Overexpression, purification, and characterization of isochorismate synthase (EntC), the first enzyme involved in the biosynthesis of enterobactin from chorismate." Biochemistry **29**(6): 1417-1425.

Luft, J. R., J. Newman, et al. (2014). "Crystallization screening: the influence of history on current practice." Acta Crystallogr F Struct Biol Commun **70**(Pt 7): 835-853.

Manos-Turvey, A., E. M. Bulloch, et al. (2010). "Inhibition studies of Mycobacterium tuberculosis salicylate synthase (MbtI)." ChemMedChem **5**(7): 1067-1079.

Martinez, J. L., A. Delgado-Iribarren, et al. (1990). "Mechanisms of iron acquisition and bacterial virulence." FEMS Microbiol Rev **6**(1): 45-56.

May, J. J., T. M. Wendrich, et al. (2001). "The dhb operon of Bacillus subtilis encodes the biosynthetic template for the catecholic siderophore 2,3-dihydroxybenzoate-glycine-threonine trimeric ester bacillibactin." J Biol Chem **276**(10): 7209-7217.

McMahon, B. and R. M. Hanson (2008). "A toolkit for publishing enhanced figures." Journal of Applied Crystallography **41**(Pt 4): 811-814.

MCSG (2014). "MCSG project overview."

MCSG. (2014). "MCSG technologies." from http://kiemlicz.med.virginia.edu/mcsg/pages/technology.

MCSG. (2014). "Structural Genomics Timeline." from http://kiemlicz.med.virginia.edu/mcsg/pages/sgtimeline.

Minor, W., M. Cymborowski, et al. (2006). "HKL-3000: the integration of data reduction and structure solution - from diffraction images to an initial model in minutes." Acta Crystallographica Section D-Biological Crystallography **62**: 859-866.

Mock, M. and A. Fouet (2001). "Anthrax." Annu Rev Microbiol **55**: 647-671.

Montelione, G. T. (2012). "The Protein Structure Initiative: achievements and visions for the future." F1000 Biol Rep **4**: 7.

Morollo, A. A. and M. J. Eck (2001). "Structure of the cooperative allosteric anthranilate synthase from Salmonella typhimurium." Nat Struct Biol **8**(3): 243-247.

Moult, J. (2005). "A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction." Curr Opin Struct Biol **15**(3): 285-289.

Murshudov, G. N., P. Skubak, et al. (2011). "REFMAC5 for the refinement of macromolecular crystal structures." Acta Crystallogr D Biol Crystallogr **67**(Pt 4): 355-367.

Murzin, A. G., S. E. Brenner, et al. (1995). "SCOP: a structural classification of proteins database for the investigation of sequences and structures." J Mol Biol **247**(4): 536-540.

Myler, P. J., R. Stacy, et al. (2009). "The Seattle Structural Genomics Center for Infectious Disease (SSGCID)." Infect Disord Drug Targets **9**(5): 493-506.

Nikaido, H. (2009). "Multidrug resistance in bacteria." Annu Rev Biochem **78**: 119-146.

O'Brien, I. G., G. B. Cox, et al. (1970). "Biologically active compounds containing 2,3-dihydroxybenzoic acid and serine formed by Escherichia coli." Biochim Biophys Acta **201**(3): 453-460.

OECD (2000). Workshop on Structural Genomics: Final Report to the Global Science Forum. Florence.

Ollinger, J., K. B. Song, et al. (2006). "Role of the Fur regulon in iron transport in Bacillus subtilis." J Bacteriol **188**(10): 3664-3673.

Otwinowski, Z. (1991). Isomorphous Replacement and Anomalous Scattering. Daresbury Study Weekend Proceedings, SERC Daresbury Laboratory, Warrington, U.K.

Otwinowski, Z. and W. Minor (1997). Processing of X-ray diffraction data collected in oscillation mode. Macromolecular Crystallography, Pt A. **276:** 307-326.

Parsons, J. F., P. Y. Jensen, et al. (2002). "Structure of Escherichia coli aminodeoxychorismate synthase: architectural conservation and diversity in chorismate-utilizing enzymes." Biochemistry **41**(7): 2198-2208.

Parsons, J. F., K. M. Shi, et al. (2008). "Structure of isochorismate synthase in complex with magnesium." Acta Crystallogr D Biol Crystallogr **64**(Pt 5): 607-610.

Payne, S. M. (1989). "Iron and virulence in Shigella." Mol Microbiol **3**(9): 1301-1306.

Pearson, W. R. and D. J. Lipman (1988). "Improved tools for biological sequence comparison." Proc Natl Acad Sci U S A **85**(8): 2444-2448.

Perrakis, A., R. Morris, et al. (1999). "Automated protein model building combined with iterative structure refinement." Nat Struct Biol **6**(5): 458-463.

Pieper, U., R. Chiang, et al. (2009). "Target selection and annotation for the structural genomics of the amidohydrolase and enolase superfamilies." J Struct Funct Genomics **10**(2): 107-125.

Pishchany, G., Skaar, E.P. (2011). Iron Acquisition by Bacillus anthracis. Bacillus anthracis and Antrax. N. H. Bergman, John Wiley & Sons.

Pomerantsev, A. P., N. A. Shishkova, et al. (1992). "[Comparison of therapeutic effects of antibiotics of the tetracycline group in the treatment of anthrax caused by a strain inheriting tet-gene of plasmid pBC16]." Antibiot Khimioter **37**(4): 31-34.

Price, L. B., A. Vogler, et al. (2003). "In vitro selection and characterization of Bacillus anthracis mutants with high-level resistance to ciprofloxacin." Antimicrob Agents Chemother **47**(7): 2362-2365.

Prilusky, J., E. Oueillet, et al. (2005). "HalX: an open-source LIMS (Laboratory Information Management System) for small- to large-scale laboratories." Acta Crystallogr D Biol Crystallogr **61**(Pt 6): 671-678.

Prinz, F., T. Schlange, et al. (2011). "Believe it or not: how much can we rely on published data on potential drug targets?" Nat Rev Drug Discov **10**(9): 712.

Qiu, L. and S. Dhe-Paganon (2011). "Oligomeric structure of the MALT1 tandem Ig-like domains." PLoS One **6**(9): e23220.

Ratledge, C. and L. G. Dover (2000). "Iron metabolism in pathogenic bacteria." Annu Rev Microbiol **54**: 881-941.

Raush, E., M. Totrov, et al. (2009). "A new method for publishing three-dimensional content." PLoS One **4**(10): e7394.

Raymond, K. M., Denz, E. (2004). Biochemical and Physical Properties of Siderophores. Iron Transport in Bacteria. J. H. Crosa, Mey, A.R., Payne, S.M. Washington, ASM Press**:** 3-17.

Raymond, S., N. O'Toole, et al. (2004). "A data management system for structural genomics." Proteome Sci **2**(1): 4.

RCSB. (2014). "PDB Statistics: Yearly Growth of Total Structures." from http://www.pdb.org/pdb/statistics.

RCSB. (2015). "Advanced Search Interface." from http://www.rcsb.org/pdb/search/advSearch.do?search=new.

Read, T. D., S. N. Peterson, et al. (2003). "The genome sequence of Bacillus anthracis Ames and comparison to closely related bacteria." Nature **423**(6935): 81-86.

Roosenberg, J. M., 2nd, Y. M. Lin, et al. (2000). "Studies and syntheses of siderophores, microbial iron chelators, and analogs as potential drug delivery agents." Curr Med Chem **7**(2): 159-197.

Rosato, A., A. Bagaria, et al. (2009). "CASD-NMR: critical assessment of automated structure determination by NMR." Nat Methods **6**(9): 625-626.

Rosenbaum, G., R. W. Alkire, et al. (2006). "The Structural Biology Center 19ID undulator beamline: facility specifications and protein crystallographic results." J Synchrotron Radiat **13**(Pt 1): 30-45.

Rossman, M. G., Blow, D. M. (1962). "The Detection of Sub-Units within the Crystallographic Asymmetric Unit." Acta Cryst **15**: 24-32.

Rowland, B. M. and H. W. Taber (1996). "Duplicate isochorismate synthase genes of Bacillus subtilis: regulation and involvement in the biosyntheses of menaquinone and 2,3-dihydroxybenzoate." J Bacteriol **178**(3): 854-861.

Rupp, B. (2010). Model building and refinement. Biomolecular Crystallography: Principles, Practice and Application to Structural Biology. S. Scholl. New York, USA, Garland Science.

Russo Krauss, I., A. Merlino, et al. (2013). "An overview of biological macromolecule crystallization." Int J Mol Sci **14**(6): 11643-11691.

Salemme, F. R., Genieser, L., Finzel, B.C., Hilmer, R.M., Wendoloski, J.J. (1988). "Molecular factors stabilizing protein crystals." Journal of Crystal Growth **90**(1-3): 273-282.

Sampathkumar, P., S. A. Ozyurt, et al. (2010). "Structures of the autoproteolytic domain from the Saccharomyces cerevisiae nuclear pore complex component, Nup145." Proteins-Structure Function and Genetics **78**(8): 1992-1998.

SBKB. (2011). "Metrics Describing Progress of the Protein Structure Initiative." from http://sbkb.org/metrics/milestonestables.html.

SBKB. (2015). "About PSI." from http://sbkb.org/about/about-psi.

Schrödinger, L. (2010). The PyMOL Molecular Graphics System.

Segond, D., E. Abi Khalil, et al. (2014). "Iron acquisition in Bacillus cereus: the roles of IlsA and bacillibactin in exogenous ferritin iron mobilization." PLoS Pathog **10**(2): e1003935.

Severn, M. (1976). "A fatal case of pulmonary anthrax." Br Med J **1**(6012): 748.

Shapiro, A. L., E. Vinuela, et al. (1967). "Molecular weight estimation of polypeptide chains by electrophoresis in SDS-polyacrylamide gels." Biochem Biophys Res Commun **28**(5): 815-820.

Sheldrick, G. M. (2008). "A short history of SHELX." Acta Crystallographica Section A **64**: 112-122.

Shimadzu, C. (1998). UVProbe.

Skaar, E. P., A. H. Gaspar, et al. (2006). "Bacillus anthracis IsdG, a heme-degrading monooxygenase." J Bacteriol **188**(3): 1071-1080.

Slabinski, L., L. Jaroszewski, et al. (2007). "The challenge of protein structure determination--lessons from structural genomics." Protein Sci **16**(11): 2472-2482.

Soding, J. (2005). "Protein homology detection by HMM-HMM comparison." Bioinformatics **21**(7): 951-960.

Soding, J., A. Biegert, et al. (2005). "The HHpred interactive server for protein homology detection and structure prediction." Nucleic Acids Res **33**(Web Server issue): W244-248.

Splino, M., J. Patocka, et al. (2005). "Anthrax vaccines." Ann Saudi Med **25**(2): 143-149.

Sridharan, S., N. Howard, et al. (2010). "Crystal structure of Escherichia coli enterobactin-specific isochorismate synthase (EntC) bound to its reaction product isochorismate: implications for the enzyme mechanism and differential activity of chorismate-utilizing enzymes." J Mol Biol **397**(1): 290-300.

Stols, L., M. Gu, et al. (2002). "A new vector for high-throughput, ligation-independent cloning encoding a tobacco etch virus protease cleavage site." Protein Expr Purif **25**(1): 8-15.

Studier, F. W. (2005). "Protein production by auto-induction in high density shaking cultures." Protein Expr Purif **41**(1): 207-234.

Tarlovsky, Y., M. Fabian, et al. (2010). "A Bacillus anthracis S-layer homology protein that binds heme and mediates heme delivery to IsdC." J Bacteriol **192**(13): 3503-3511.

Taylor, G. (2003). "The phase problem." Acta Crystallogr D Biol Crystallogr **59**(Pt 11): 1881-1890.

Terwilliger, T. (2004). "SOLVE and RESOLVE: automated structure solution, density modification, and model building." J Synchrotron Radiat **11**: 49-52.

Touati, D. (2000). "Iron and oxidative stress in bacteria." Arch Biochem Biophys **373**(1): 1-6.

Trillo-Muyo, S., A. Jasilionis, et al. (2013). "Ultratight crystal packing of a 10 kDa protein." Acta Crystallogr D Biol Crystallogr **69**(Pt 3): 464-470.

Tsumoto, K., M. Umetsu, et al. (2004). "Role of arginine in protein refolding, solubilization, and purification." Biotechnol Prog **20**(5): 1301-1308.

Turnbull, P. C. B. a. S., S.V. (2010). Anthrax from 5000 BC to AD 2010. Bacillus anthracis and Anthrax. N. H. Bergman. Hoboken, NJ, USA, John Wiley & Sons, Inc.**:** 1-15.

Vinarov, D. A., C. L. Newman, et al. (2006). "Wheat germ cell-free expression system for protein production." Curr Protoc Protein Sci **Chapter 5**: Unit 5 18.

Vitkup, D., E. Melamud, et al. (2001). "Completeness in structural genomics." Nat Struct Biol **8**(6): 559-566.

Walsh, C. T., M. D. Erion, et al. (1987). "Chorismate aminations: partial purification of Escherichia coli PABA synthase and mechanistic comparison with anthranilate synthase." Biochemistry **26**(15): 4734-4745.

Ward, C. G., J. S. Hammond, et al. (1986). "Effect of iron compounds on antibacterial function of human polymorphs and plasma." Infect Immun **51**(3): 723-730.

Weiss, M. M., P. D. Weiss, et al. (2007). "Anthrax vaccine and public health policy." Am J Public Health **97**(11): 1945-1951.

Wencewicz, T. A., U. Mollmann, et al. (2009). "Is drug release necessary for antimicrobial activity of siderophore-drug conjugates? Syntheses and biological studies of the naturally occurring salmycin "Trojan Horse" antibiotics and synthetic desferridanoxamine-antibiotic conjugates." Biometals **22**(4): 633-648.

WHO (2014). Antimicrobial resistance: global report on surveillance.

Wilson, M. K., R. J. Abergel, et al. (2010). "Temporal production of the two Bacillus anthracis siderophores, petrobactin and bacillibactin." Biometals **23**(1): 129-134.

Winn, M. D., C. C. Ballard, et al. (2011). "Overview of the CCP4 suite and current developments." Acta Crystallographica Section D-Biological Crystallography **67**: 235-242.

Wunsche, A., E. Hammer, et al. (2012). "CcpA forms complexes with CodY and RpoA in Bacillus subtilis." FEBS J **279**(12): 2201-2214.

Yang, H. W., V. Guranovic, et al. (2004). "Automated and accurate deposition of structures solved by X-ray diffraction to the Protein Data Bank." Acta Crystallographica Section D-Biological Crystallography **60**: 1833-1839.

Zheng, H., M. D. Chordia, et al. (2014). "Validation of metal-binding sites in macromolecular structures with the CheckMyMetal web server." Nat Protoc **9**(1): 156-170.

Ziebart, K. T., S. M. Dixon, et al. (2010). "Targeting multiple chorismate-utilizing enzymes with a single inhibitor: validation of a three-stage design." J Med Chem **53**(9): 3718-3729.

Ziebart, K. T. and M. D. Toney (2010). "Nucleophile specificity in anthranilate synthase, aminodeoxychorismate synthase, isochorismate synthase, and salicylate synthase." Biochemistry **49**(13): 2851-2859.

Zimmerman, M., Chruszcz, M., Koclega, K., Otwinowski, Z., Minor, W. (2005). "The Xtaldb system for project salvaging in high-throughput crystallization." Acta Cryst **A61**: 178-179.

Zimmerman, M. D., M. Grabowski, et al. (2014). "Data management in the modern structural biology and biomedical research environment." Methods Mol Biol **1140**: 1-25.

Zolnai, Z. (2014). "Sesame users guide." from http://www.sesame.wisc.edu/.

Zolnai, Z., P. T. Lee, et al. (2003). "Project management system for structural and functional proteomics: Sesame." J Struct Funct Genomics **4**(1): 11-23.

List of figures

List of tables

List of author's publications

Peer-reviewed articles

1. Rashin, A.A., **Domagalski, M.J.**, Zimmermann, M.T., Minor, W., Chruszcz, M., Jernigan, R.L. (2014) Factors correlating with significant differences between X-ray structures of myoglobin. Acta Cryst D70:481-91
2. **Domagalski, M.J.**, Tkaczuk, K.L., Chruszcz, M., Skarina, T., Onopriyenko, O., Cymborowski, M., Grabowski, M., Savchenko, A., Minor, W. (2013) Structure of isochorismate synthase DhbC from Bacillus anthracis. Acta Cryst F69:956-61
3. Trillo-Muyo, S., Jasilionis, A., **Domagalski, M.J.**, Chruszcz, M., Minor, W., Kuisiene, N., Arolas, J.L., Solà, M., Gomis-Rüth, F.X. (2013) Ultratight crystal packing of a 10-kDa protein. Acta Cryst D69:464-70
4. Chruszcz, M., **Domagalski, M.**, Osinski, T., Wlodawer, A., Minor, W. (2010) Unmet challenges of structural genomics. Curr Opin Struct Biol 20:587-97

Book chapters

1. **Domagalski, M.J.**, Zheng, H., Zimmerman, M.D., Dauter, Z., Wlodawer, A., Minor, W. (2014) The Quality and Validation of Structures from Structural Genomics. Methods Mol Biol (Clifton, N.J.) 1091:297-314
2. Zimmerman, M.D., Grabowski, M., **Domagalski, M.J.**, MacLean, E.M., Chruszcz, M., Minor, W. (2014) Data Management in the Modern Structural Biology and Biomedical Research Environment. Methods Mol Biol (Clifton, N.J.) 1140:1-25
3. Chruszcz, M., Borek, D., **Domagalski, M.**, Otwinowski, Z., Minor, W. (2009) X-ray diffraction experiment - the last experiment in the structure elucidation process. Adv Prot Chem and Struct Biol 77:23-40