AD_____

Award Number: DAMD17-03-2-0038

TITLE: Structural Genomics of Bacterial Virulence Factors

PRINCIPAL INVESTIGATOR: Robert C. Liddington, Ph.D.
Adam Godzik, Ph.D.
Maurizio Pellecchia, Ph.D.

CONTRACTING ORGANIZATION: The Burnham Institute
La Jolla, CA 92037

REPORT DATE: May 2005

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*



| 1. REPORT DATE *(DD-MM-YYYY)* | 2. REPORT TYPE | 3. DATES COVERED *(From - To)* |
|---|---|---|
| 01-05-2005 | Annual | 1 May 2004 – 30 Apr 2005 |

**4. TITLE AND SUBTITLE**
Structural Genomics of Bacterial Virulence Factors

**5a. CONTRACT NUMBER**

**5b. GRANT NUMBER**
DAMD17-03-2-0038

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**
Robert C. Liddington, Ph.D.
Adam Godzik, Ph.D.
Maurizio Pellecchia, Ph.D.

E-mail: rlidding@burnham.org

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

The Burnham Institute
La Jolla, CA 92037

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

**10. SPONSOR/MONITOR'S ACRONYM(S)**

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

**12. DISTRIBUTION / AVAILABILITY STATEMENT**
Approved for Public Release; Distribution Unlimited

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT**

We are continuing to apply a comprehensive but focused structural genomics approach to determine the atomic resolution crystal structures of key virulence factors from high priority pathogens. The work in our first year focused on proteins encoded by the B. anthracis virulence plasmid, pX01, and the setting up of a virulence factor computational data base. In the second year we expanded our efforts to include genome-encoded proteins of B. anthracis, structural studies on proteins encoded by Variola virus, thecausative agent of smallpox; initiated work to characterize a SARS virus surface protein in complex with a neutralizing antibody; and inititated work on a close homolog of a Yersinia pestis SUMOylase. We have generated a large library of expression vectors for virulence factors, as well as research quantities of pure proteins, which could readily be adapted for vaccine design. In the broader and longer term, the accumulated structural information will generate important and testable hypotheses that will increase our understanding of the molecular mechanisms of pathogenicity, putting us in a stronger position to anticipate and react to emerging pathogens.

**15. SUBJECT TERMS**
x-ray crystallography, structural genomics, bioinformatics, biodefense, virulence factor, toxin

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | USAMRMC |
| U | U | U | UU | 118 | 19b. TELEPHONE NUMBER *(include area code)* |

# Table of Contents

# INTRODUCTION

We are continuing to apply a comprehensive but focused structural genomics approach to determine the atomic resolution crystal structures of key virulence factors from high priority pathogens. The work in our first year focused on proteins encoded by the B. anthracis virulence plasmid, pX01, and the setting up of a virulence factor computational data base. In the second year we expanded our efforts to include genome-encoded proteins of B. anthracis, structural studies on proteins encoded by Variola virus, the causative agent of smallpox; initiated work to characterize a SARS virus surface protein in complex with a neutralizing antibody; and inititated work on a close homolog of a Yersinia pestis SUMOylase. We have generated a large library of expression vectors for virulence factors, as well as research quantities of pure proteins, which could readily be adapted for vaccine design. In the broader and longer term, the accumulated structural information will generate important and testable hypotheses that will increase our understanding of the molecular mechanisms of pathogenicity, putting us in a stronger position to anticipate and react to emerging pathogens.

# BODY

## Task 1: Atomic resolution crystal structures of virulence factors:

### Cloning and expression of novel B. anthracis proteins

**Expression of selected genes from the *B. anthraces* plasmid PX01 in *Bacillus subtilis* and *Bacillus megaterium* cells.**

We have continued our work on the expression and purification proteins encoded by the pX01 plasmid and selected by our bioinformatics approached (summarized below, see Task 2). Our hit rate on soluble protein expression and crystallization has been disappointing when compared with our general success-rate for other bacterial and eukaryotic proteins. We therefore investigated Bacillus expression systems to see if these would provide a superior system for expressing B. anthracis proteins.

Though *Bacillus* strains are broadly used for industrial expression of heterologous proteins, there is only one company that sells the expression system. Furthermore, their shuttle plasmid is underdeveloped - it does not have purification tags and secretion peptides. There are also numerous *Bacillus subtilis* strains and plasmids, but they have been used mostly for functional studies, where overexpression of a protein is not important. We tested two systems, *Bacillus subtilis* and *Bacillus megaterium*. Derivatives of *Bacillus subtilis* strain 168 (1A436, S53, 1A1) and the plasmid pDG148 were obtained from the Bacillus Genetic Stock Center (Ohio University). *Bacillus megaterium* strain WH320 and the plasmid pWH1520 were obtained from MoBiTec (Germany). *Bacillus subtilis* strain 168 has a natural ability for transformation (intake of plasmid DNA through the cell wall). The protein expression, however, is problematic, because this strain undergoes sporulation when the expressed protein is toxic or the growth conditions are not optimal. The value of this system for secreted expression is also limited, because

*B. subtilis* produces too many proteases. *B. megaterium* strain WH320 does not sporulate, the shuttle plasmid is fairly stable there, and it not secrete many proteases. However, *B. megaterium* does not take plasmids by transformation. The alternative protocol, which requires removal of the cell wall by lysozyme, is unreliable.

We successfully adopted the two Bacillus expression systems and tested expression of following genes px01-97, px01-99, px01-118, px01-119, px01-125 which did not express well in *E. coli.* Gene pX01-118, which expressed well in E. coli, was used as a positive control. We found that the level of protein expression in correlates with the level of expression in *E. coli.* The best results were obtained for px01-118 using *B. megaterium*; nevertheless, the expression level per gram of cell mass was about 0.5-2 mg, which is 5 times lower then the expression from pET plasmid in *E.coli.* The expression of other proteins as soluble proteins was detectable by Western blot against His-tag, but insufficient for crystallization. The expression of pX01-118 in *B. subtilis* strains was unstable. Often cells began to sporulate even before induction of protein expression (the IPTG promoter was very leaky). We tested the plasmid PDG148 with *B. megaterium* and the plasmid pWH1520 with *B. subtilis.* Contrary to the claims of MoBiTec, the plasmids did not perform well in foreign cells.

We conclude that intracellular expression in Bacillus species does not give a clear advantage over *E. coli* system, perhaps because the codon usage is similar and *E. coli* has a more developed chaperoning system. However, *B. megaterium* may be beneficial for expression of secreted proteins.

**Expression and purification of AtxA and its homologs on pXO2, AcpA and AcpB**

Full-length AtxA was expressed with or without a histidine-tag fusion and purified by Ni affinity, heparine-sepharose and/or anion-exchange, and gel filtration chromatography. Yields are around 2 mg/liter of cell culture. The presence of up to five species, partially separable by heparin-sepharose affinity chromatography, was evident. Native PAGE evidence at mM to mM concentration shows that AtxA interacts with DNA, as a band corresponding to DNA cannot be detected as the concentration of AtxA increases, but a stable specific complex could not yet be characterized, possibly due to the relatively high concentration of protein or the lack of a specific site on the DNA sequence used, a 300 bp stretch upstream of the transcriptional start site of the pag gene. Current work includes further separation of the above mentioned AtxA species, determining whether they are stable or in slow exchange with each other, and whether this affects binding to DNA. Near-future plans are the characterization of the binding to DNA sequences from the promoters of other AtxA-regulated genes using radioactively labelled DNA, which will allow work at or near the protein-DNA dissociation constant, which is as yet undetermined but usually expected in the nM range.

AcpB waas expressed as a histidine-tag fusion and purified with similar results. AcpA appears to be toxic to E.coli cells as their growth is significantly slowed down when transformed with a plasmid encoding the histidine-tagged protein, and yields were therefore an order of magnitude lower. Current work focuses on the cloning, expression

and purification of native (untagged) AtxA and AtxB and future plans will include the characterization of their binding to DNA, similar to AtxA.

**Structural Studies of inhibitor binding to Lethal Factor**

Our work to determine LF-inhibitor complexes in collaboration with the Bavari and Gussio groups at USAMRIID and NCI continues. The crystal structure of full-length LF was grown under high salt conditions, and this may have hampered in several cases the determination of high quality inhibitor complexes. To try to overcome these problems we have cloned, expressed and crystallized a fragment of LF that lacks domain 1(the PA-binding domain), but that contains the critical catalytic module (Domains 2-4). This protein expresses readily in E. coli, and crystallizes from low salt (PEG) conditions; it also diffracts X-rays to high resolution. We are now in the process of repeating our inhibitors soaks and co-crystallization experiments under these low salt conditions.

**Crystal structure of an anthrax toxin-host cell receptor complex**

Two closely related host cell receptor molecules, TEM8 and CMG2, bind to PA with high affinity and are required for toxicity. We determined the crystal structure of the PA-CMG2 complex at 2.5 Å resolution (published in *Nature*, see **Appendix 1**). The structure reveals an extensive receptor-pathogen interaction surface that mimics the non-pathogenic recognition of the extracellular matrix by integrins. The binding surface is closely conserved in the two receptors and across species, but quite different in the integrin domains, explaining the specificity of the interaction. CMG2 engages two domains of PA, and modeling of the receptor-bound PA63 heptamer suggests that the receptor acts as a pH-sensitive chaperone to ensure accurate and timely membrane insertion.

**Structural studies on a B. anthracis epimerase involved in lysine biosynthesis**

Lysine biosynthesis in bacteria provides the essential components both for L-lysine for protein synthesis and meso-diaminopimelate for construction of the bacterial peptideglycan cell wall. Since lysine biosynthesis is deficient to mammals and unique to bacteria, the enzymes involved in the pathway may be useful for antibiotic design. Recent genome sequence analysis of *B. anthracis* revealed the complete sequences of enzymes involved in lysine biosynthesis. Diaminopimelate epimerase (E.C.5.1.1.7), an enzyme involved in the pathway, catalyzes the racemization of L,L- to D,L-meso-diaminopimelate, the immediate precursor of L-lysine in *B. anthracis*. Several enzymes involved in racemization require pyridoxal 5'-phosphate (PLP) as cofactor for their activity; however, little is known about the structural basis of PLP dependence for the activity of anthrax diaminopimelate epimerase. The object in this study is therefore to determine the crystal structure of the anthrax diaminopimelate epimerase to investigate the structure/function correlation.

1. Cloning, expression and Purification: The gene encoding diaminopimelate epimerase (EP) from *B. anthracis* (gene code: BA5170; MW=32 kDa; 288 residues) was

cloned from genomic DNA and inserted into pET15b at sites of *Bam* HI/*Nde* I. The recombinant protein was not expressed in standard *E.coli* BL21 (DE3) even under several different growth conditions using LB/TB medium, different IPTG concentration, and high/low temperature. The sequence analysis revealed that many numbers of rare codons are involved in the protein sequence, suggesting incomplete translation of the sequence during the protein synthesis in the BL21 (DE3) strain, consistent with non-production of the recombinant protein in the bacteria. Alternatively, another *E.coli* strain, Rosetta (DE3)pLysS, dramatically increased the expression of soluble recombinant EP protein, using 2X YT medium at 37°C.

Large scale protein expression was carried out of the Rosetta (DE3)pLysS bacterial cultures in 2X YT medium. The recombinant EP protein was purified from the cell-free-extract by Ni-affinity chromatography followed by gel filtration chromatography (Superdex200). The MALDI-MS analysis of the purified protein revealed a strong single peak at the expected molecular mass (33.6 kDa). The gel filtration experiment, however, revealed the recombinant EP protein eluted at around the molecular mass of >60 kDa, suggesting it forms a dimer in solution. The purification protocol presented above typically yielded >70 mg of the purified EP protein per 1 liter of the bacterial cultures. 5 mM DTT was always included in the running buffer used for gel filtration experiment. His-tag has NOT been cleaved.

**Crystallographic characterization.**
Crystals of EP were grown in several reservoir conditions using commercial screening kits (Table ). The best crystals were obtained in the reservoir of 0.1 M Na/K-phosphate, pH 6.6, 20% PEG3,350, and 0.2 M ammonium formate. The protein concentration used was 13 mg/mL, in 20 mM Tris, pH 8, 150 mM NaCl, and 5 mM DTT. The crystals diffracted to ~2.5 Å resolution using Rigaku FR-E X-ray generator. The crystals belong to space group $P2_12_12_1$, with cell dimensions a=64.9, b=85.5, c=113.0 Å. The crystal structure has recently been solved using Molecular Replacement, and refinement and inhibitor design is in progress.

Table. Crystallographic characterization of the native EP crystal (sample: EP02) /data/liddingA/koichi/Anthrax/EP2/scale.log

| Parameters | |
|---|---|
| Space group | $P2_12_12_1$ |
| Cell dimensions (Å) | a=64.9, b=85.5, c=113.0 |
| Resolution range (Å) | 30-2.7 (2.8-2.7) |
| No. of observed reflections | 50105 |
| No. of unique reflections | 16106 (1298) |
| Completeness (%) | 90.0 (74.7) |
| $R_{merge}$ | 0.088 (0.288) |
| $I$ / s$I$ | 13.8 (2.7) |
| $V_m$ (Å$^3$/Da) | 2.4 |
| No. of molecules per asym | 2 |
| Solvent content (%) | 49 |

**_B. anthracis_ endolysins studies (manuscript submitted; see Appendix 2)**

Endolysins are cell wall dissolving enzymes used by phage to lyse its host to release its progeny, and are potential antibacterial agents. The aim of this study is to examine if the integrated copies of prophage endolysins within the *B. anthracis* Stern strain can be used as anti-bacterial agents for the treatment and prophylaxis of anthrax and other Gram positive bacterial infection.

Two targets were selected, one prophage amidase and one prophage glycosidase, from the *B. anthracis* Stern strain. They are two-domain proteins, consisting of a N-terminal catalytic domain and a C-terminal 80 amino acid putative cell-wall binding domain. The amidase cleaves the bond between the N-Acetylmuramic acid and the L-Alanine, while the glycosidase cleaves the bond between N-acetylglucosamine and N-Acetylmuramic acid of the cell wall. The C-terminal cell wall binding domain of the two endolysin has very high sequence homology (68% identity). Although they were not in the same prophage region, it is believed that with similar cell-wall binding domain, the two enzymes could be working together synergistically.

Both proteins can be expressed in *E. coli* system at higher than 20 mg/L culture. They can be purified easily by standard techniques, but full length proteins were less soluble than the catalytic domains. Crystallization trials were set up for all constructs. It was found that the full length proteins precipitated in the majority of the screen conditions even at concentration lower than 5 mg/ml. The catalytic domains, however, crystallized readily. The sequences of the endolysins are relative distant (less than 27% identity) to any of the known structures of the similar enzyme classes. Molecular replacement using standard techniques failed to provide the phase information. Multiple Anomalous Diffraction (using SelenoMethionine labeled protein) and Single Isomorphous Replacement (using Methyl Mercury Nitrate derivatized protein crystal) phasing techniques were used to elucidate the structures of the amidase and glycosidase, respectively. The highest resolutions of the catalytic domains were 1.8 and 1.4 angstroms for the amidase and glycosidase, respectively. The anthrax prophage amidase structure resemble that of the T7 amidase fold. The prophage glycosidase, on the other hand adopt the *Chalaropsis* glycosidase family fold.

They were shown to be able to hydrolyse bacterial cell wall peptidoglycan, and kill several bacillus strains (*B. anthracis* Stern, *B. cereus*, *B. megaterium*, and *B. subtilis*) *in vitro* within 15 minutes at sub-micromolar concentration. It was also found, surprisingly, that the N-terminal catalytic domain is significantly more active than the full length protein in most of the bacilli strains tested. The C-terminal domain was later cloned into a expression vector as fusion with the green fluorescence protein (GFP). The GFP fused with the C-terminal domain of amidase was able to coat the surface of *B. cereus* but not other strains (*B. anthracis* not yet tested). These results suggest that the C-terminal domain of the amidase could be a negative regulator, and also at the same time provide selectivity for cell-wall binding. The cell-wall binding domain of the amidase was also crystallized and its structure determination is underway.

In conclusion, we have determined the minimum endolysin protein constructs as potential candidates to use for anti-bacterial treatment. These constructs are likely to be more active than the full length protein including the full length PlyG, because of the

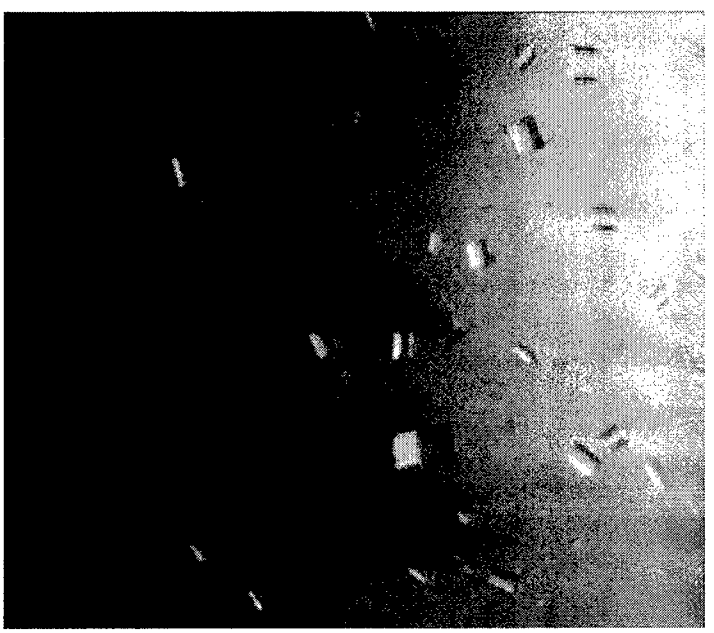
absence of specific cell wall binding domain that may be inhibiting the activity of the catalytic domain when use against non-specific host strains. This minimum catalytic domains will be tested on other Gram positive bacteria strains in the near future, as soon as they become available.

**Collagen binding protein BA5258 of *B. anthracis***

*B. anthracis*, similar to other Gram positive bacteria, attaches to the host via cell-wall-anchoring proteins. Two of such protein from *B. anthracis* were characterized by Xu *et al* (2004), namely BA0871 and BA5258. These two proteins have sequence homology to CNA, a cell wall-anchored collagen adhesin of *S. aureus*. The full length BA5258, excluding the leader sequence, has been cloned into a *E. coli* expression vector. It can be expressed and purified to a final yield of 10 mg/L culture. The protein is extremely soluble and resistant to limited proteolysis with trypsin, elastase, and chymotrypsin. Crystallization trials of the protein by itself and with a collagen peptide are in progress, and small but promising protein crystals have been obtained.

**Structural studies of the SARS S1 (spike protein) and its complex with a high affinity anitbody.**

In collaboration with Dr, Wayne Marasco, Dana Farber Cancer Institute, Boston, we have initiated a structural study of the SARS S1 spike protein with a high affinity antibody ("80R)" (Sui et al., 2004). Both the S1 protein and antibody have been expressed and purified in milligram quanitities, and the S1 protein alone and in complex with the antibody (figure below). X-ray data sets have been collected to 2.3 Å resolution and the structure determination is in progress.



**Structural studies of Variola proteins**

A highly conserved poxvirus protein, N1L (*Vaccinia* gene name), with no significant homology to any non-poxvirus proteins, was recently shown to be a viral inhibitor of the host innate immunity (DiPderna et al., 2004). N1L is a small 14kDa protein, highly conserved among poxviruses, with 94% sequence identity between *Vaccinia* and *Variola* orthologs. N1L is considered one of the most potent virulent factors based on the attenuated phenotype of the recombinant mutant *Vaccinia* virus (Kotwal et al., 1989). N1L associates with several kinases within the multi-subunit IKK complex, N1L interacted most strongly with the TANK-binding kinase 1 (TBK1). The N1L gene, amplified from genomic DNA of *Vaccinia* Western Reserve and Cowpox Brighton Red (a gift from Dr. D.J. Pickup, Duke University). We have successfully produced homogeneous samples of *Vaccinia* N1L, judged by SDSPAGE (Fig. A). We have also obtained small needle crystals of His-tagged N1L (Fig. B). Structural analysis is underway.

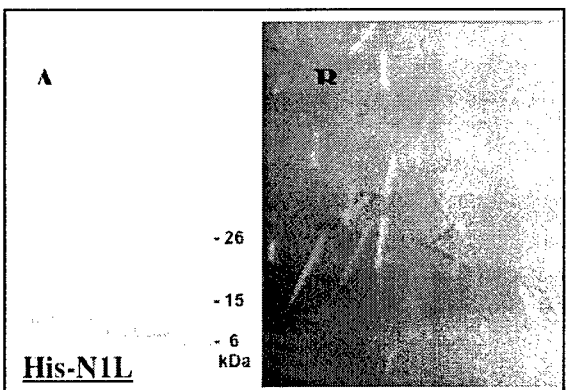



**Figure. Purified His-tagged and non-tagged *Vaccinia* N1L analyzed by SDSPAGE (A) and preliminary small, needle crystals of N1L (B).**

**NMR based structural characterization of virulence factors.**

The overall goals of Dr. Pellecchia's laboratory within this project are to provide support for the determination of the structures of key virulence factors by NMR spectroscopy. In particular, Dr. Pellecchia focused on the identification, expression and purification of novel virulence factors for subsequent NMR analysis. A group of bacterial genes homologous to the human Ubiquitin-like protease (Ulp) or SUMO-specific protease (SUMOylase) have been identified by bioinformatics methods in Dr. Godzik's laboratory. These proteins are also related to the *Yersinia* virulence factor YopP. Dr.

Pellecchia focused his efforts on a particular protein construct from *Salmonella typhimurium* called Virulase ST. In unpublished work, Dr. Reed's laboratory has established that much like YopP, Virulase ST regulates apoptosis and inflammation in infected host cells, presumably via the NK-kB pathway. In order to provide additional insights into the function and role of this protein in the onset and propagation of the infection, Dr. Pellecchia begun to investigate protein constructs from this family of proteins for subsequent structural analysis by NMR. Because the putative catalytic domain of Virulase ST has been shown to induce apoptosis (measured by caspase-3 activation) in transfected human cells, the studies focused on this domain. Recombinant Virulase ST (145-326) was produced from a pET-19b (Novagen) plasmid construct containing the nucleotide sequence for the catalytic domain fused to an *N*-terminal poly-His tag. Unlabeled protein was expressed in *E. coli* BL21 in LB media at 37°C, with an induction period of 3-4 hours with 1 mM IPTG. $^{15}$N-labeled protein was similarly produced, with growth occurring in M9 media supplemented with 0.5 g/L $^{15}NH_4Cl$. Double $^{13}C/^{15}N$-labeled protein as well as triple labeled $^{2}H/^{15}N/^{13}C$ protein were similarly produced in M9 media supplemented with $^{13}$C-glucose (2 g/L) and $^{2}H_2O$ (70%), respectively. Following cell lysis, soluble protein is purified over a Hi-Trap chelating column (Amersham, Pharmacia). Final protein samples were dialyzed into a buffer appropriate for the subsequent experiments. A number of stability tests have subsequently been performed in order to verify that the protein would survive the time needed to collect the NMR experiments for resonance assignments (2-3 weeks). Unfortunately, the protein is not long lived and it tends to aggregate very rapidly (hours) or gets cleaved. In order to increase the stability of the catalytic domain of Virulase ST number of different conditions were tested including temperature, pH, different detergents (TRITON and NP-40, both at 0.1%) and salts. Conditions that lead to samples that are stable for ~ 3-7 days, included a second step purification (ion-exchange purification with a MonoQ (Amersham, Pharmacia) column), pH = 7.2, 100 mM NaCl and 50 mM each of arginine and glutammic acid. Because 3-7 days is still to short lived for a complete set of NMR experiments to be collected, several samples were finally prepared. 2D [$^{1}$H,$^{15}$N] HSQC and TROSY-type experiments were subsequently carried out on a 600 MHz spectrometer at 20 °C and 30°C. A typical 2D [$^{15}$N, $^{1}$H] HSQC spectrum of the resulting protein sample is reported in Figure 1A. The number of peaks and the dispersion are indicative of a folded monomeric protein. Chemical shift dispersion in the $^{13}C^{a}$ (and $^{13}C^{b}$) from initial triple resonance experiments (Figure 1B) suggests a mixed a/b secondary structure, although there is probably a flexible region as well. Therefore, while additional work is needed to complete the acquisition of a minimal data set for structural determination, samples that appear well versed for high resolution studies have been obtained. The isotopically labeled samples and the preliminary NMR data collected lay the foundation for a detailed structure determination project, for which funds are being sought elsewhere. Much in line with the objectives of this project, the work initiated here will shed additional light on the function of this class of essential virulence factors and will represent the starting point for inhibitor design and subsequent target validation studies.

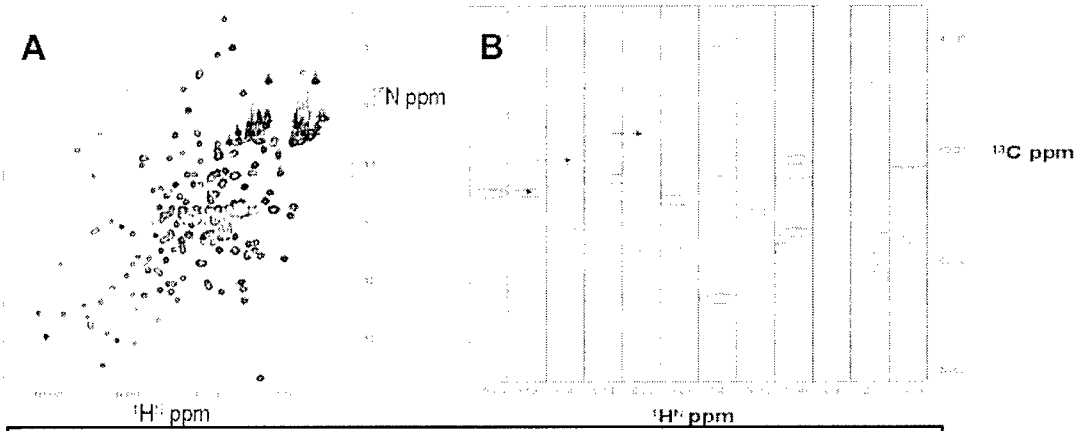



**Figure 1. A)** 2D [$^{1}H$, $^{15}N$] HSQC spectrum of the catalytic domain of Virulase ST acquired on a 1 mM sample in phosphate buffer, pH = 7.2, 50 mM each Arg/Glu, 100 mM NaCl. The spectrum was acquired with ns=16 at 20° C on a 600 MHz Avance Bruker instrument. **B)** Typical $^{13}C^{\alpha}/^{1}H^{N}$ strips taken at different $^{15}N$ chemical shifts from a 3D HNCA experiment measured with a triple $^{2}H/^{13}C/^{15}N$ labeled sample.

**Task 2: Collect expression vectors and purified proteins into a library suitable for use by other interested groups, and post the information on our website.**

This task is ongoing for *B. anthracis* and other Class A pathogens; target selection and experimental updates are done on a monthly basis in the light of new cloning, expression and structural data. The current status is summarized below. We will make this information publicly available if this is deemed appropriate by USAMRMC.

**Summary of cloning, expression and purification of novel pX01 proteins:**

**pXO1-37** (Acetyltransferase) His tagged full-length pXO1-37 (1-193) was solubly overexpressed by E. coli at 30°C. Previous instability problem upon concentrating to higher concentration is solved by adding 100 mM DTT to the protein solution after Ni-column purification. Crystallization setups have begun

**pXO1-47** (Transcription Activator of multidrug-efflux) His tagged full-length pXO1-47 (1-201) was overexpressed in inclusion bodies . Varying expression conditions did not lead to soluble protein. pXO1-47 was purified under denatured condition by Ni-column and refolded as soluble protein. DSC experiment is underway to demonstrate correct folding.

**pXO1-87 and pXO1-99** were expressed, but proved to be difficult to purify. Both proteins were co-purified with a 60 kDa protein, which is suspected to be a heat shock protein or chaperonin. High resolution columns, superdex200HR gel filtration, monoS

and monoQ column could not separate the contaminants. $Mg^{2+}$-ATP has been shown to enhance dissociation of E. coli chaperonin from proteins with large hydrophobic surface area exposed. It will be used in the immediate future for the pXO1-99 and 87 protein purification.

**pXO1-97** was cloned and gave soluble protein, and structural analysis by NMR is in progress.

**pXO1-104** His tagged full-length pXO1-104 (1-61) was overexpressed as inclusion body. Other conditions have been tried to make it expressed solubly without success. Refolding experiments are underway.

**pXO1-109/PagR** Cloning and soluble expression; crystallization trials in progress.

**pXO1-111 (homologous to PA domain 4).** Cloning and soluble expression; crystallization trials in progress.

**pXO1-116** Cloning unsuccessful so far.

**pXO1-117 and 143** cloning successful but no expression in *E. coli.*

**PX01-118 (and pX02-61)** have been crystallized and their structures determined (see **Appendix 3**)

**pXO1-121** His tagged full-length pXO1-121 (1-58) was overexpressed as inclusion body. Other conditions have been tried to express it solubly, without success. Refolding is underway.

**pXO1-125** – cloning and expression successful – protein is insoluble and could not be refolded.

Cloning of all the following target genes as full-length proteins has been completed, and expression trials are in progress. All the genes are now subcloned into the bacterial expression vector, pET28a:

**pXO1-96**, 274 residues, homologue to putative transposase;
**pXO1-103**, 317 residues, homologue to site-specific recombinase;
**pXO1-105**, 67 residues, homologue to regulators of stationary/sporulation gene expression;
**pXO1-126**, 151 residues, homologue to uncharacterized ACR ML0644;
**pXO1-130**, 237 residues, predicted periplasmic or secreted protein.

**pXO1-109 (PagR)** expressed in E. coli and purified.

**pXO1-110 (PA)** expressed in E. coli and purified;
604-735 (domain IV) expressed, partially purified

597-735 (domain IV) expressed, purified
588-735 (domain IV) expressed, partially purified

**pXO1-107 (LF)** expressed in E. coli and purified; catalytic mutants E687C and E786A expressed and purified.
263-776 (domains II-IV) expressed, purified and crystallized

**pXO1-119 (AtxA)** full-length and 1-393 expressed and purified;
1-141 and 1-160 (putative DNA binding domain) expressed, insoluble;
141-475, 162-475, 141-393, 162-393 (putative regulatory domains);
388-475 expressed, soluble, precipitates during purification

**pXO1-138 (PagR homolog)** expressed, soluble

**pXO2-53 (AcpB)** expressed and purified
**pXO2-64 (AcpA)** expressed and purified (low yield << 1 mg/l)

The following gene products of unknown function have been cloned expressed and purified: **pXO1-04**, **pXO1-07**, **pXO1-10**, **pXO1-32**, **pXO1-90**, **pXO1-94**, **pXO1-98**, a truncated form of **pXO1-98**, **pXO1-117**, **pXO1-124**, **pXO1-127**, and **pXO1-132**.

**pXO1-1, pXO1-15, pXO1-125, pXO1-117, pXO1-128 and pXO1-143** were expressed in *E. coli* as insoluble proteins. Refolding with arginine as refolding buffer solubilized the proteins but precipitations occurred during the removal. **pXO1-87 and pXO1-99** could be purified but as soluble aggregates, which precipitate at high concentration.

## Task 3: Develop a computational database of virulence-related genes

**Bioinformatics and Target Selection.** The main focus of the bioinformatics part of the grant is the development of an annotated collection of virulence factors. To this end we developed the VirFact database (http://virfact.burnham.org), which contains information on microbial virulence factors and pathogenicity islands (PAIs) from major pathogens. The database initially contained information manually collected from literature, and then combined this with results obtained by genome context analysis and distant homology recognition. The database can be browsed by virulence factor, PAI or organism name. The annotations, including multiple alignments of proteins homologous to virulence factors, genomic context, models of three dimensional structures (if available) are presented using graphical web interface and standard visualization tools. The VirFact can also be used as a tool to recognize the presence of homologs of known virulence factors in the genome delivered by the user. For instance application of VirFact to *Francisella tularensis* genome allowed us to recognize over 50 known virulence factors in this genome.

We also used several of the annotation tools developed in our group for a detailed analysis of anthrax virulence plasmids. Using a combination of advanced bioinformatics tools, including context analysis, distant homology and fold recognition, we have re-annotated the predicted open reading frames on the pXO1 plasmid, most of which were described as proteins of unknown function in previous analyses. Thanks to improved annotation tools we significantly enhanced the annotation of the pXO1 plasmid, bringing the total number of ORFs with some level of functional annotation from 48 to over 100. The new results also clearly show the mosaic nature of pXO1 and give tantalizing hints about the origin of anthrax virulence. The highlights of the new finding are two type IV secretion system-like clusters present on the pathogenicity island of the pXO1 plasmid, as well as at least three clusters related to DNA processing. Similar annotation of the pX02 plasmid as well as pathogenic islands of several bacteria from the Streptococcus group are now in preparation.

## Key Research Accomplishments

• Development of the VirFact database (http://virfact.burnham.org) of virulence factors

• Successful expression and/or cloning and of more than 50 proteins and domain fragments from the B. anthracis and other Class A pathogens.

• Crystal structures and functional characterization of B. anthracis prophage amidase and lysozyme. The amidase is homologous to a bactericidal phage enzyme that specifically kills B. anthracis.

• Crystallization of virulence factors from Variola and SARS virus.

• Crystal structure of anthrax PA in complex with its host receptor (published in *Nature*).

## Reportable Outcomes

**Published manuscripts:**

Santelli E, Bankston LA, Leppla SH & Liddington RC. Crystal structure of a complex between anthrax toxin and its host cell receptor. ***Nature***. 2004 Aug 19;430(7002):905-8. Epub 2004 Jul 4.

**Manuscripts under review:**

Lieh Yoon Low, Chen Yang, Marta Perego, Andrei Osterman and Robert Liddington "Structure and lytic activity of a Bacillus anthracis prophage endolysin"

Marcin Grynberg, Iddo Friedberg, Marc Robinson-Rechavi, and Adam Godzik "Surprising connections: in-depth analysis of the *Bacillus anthracis* pXO1 Plasmid"

Adrian Tkacz, Leszek Rychlewski and Adam Godzik "VirFact: a relational database of virulence factors and pathogenicity islands (PAIs)"

**Reagents generated:**

• Expression vectors for more than 50 virulence factors.

• Atomic coordinates have been deposited in the Protein Data Bank for the Protective Antigen-host cell receptor complex and are freely available. Atomic coordinates for other crystal structures derived here will be released upon publication.

**Funding arising from these studies:**

Some of the work described here has led to a Program Project grant from NIAID led by Dr. Liddington **(P01 AI 55789-01)**. This proposal was funded, effective 7/04.

Our work on the inhibitors of anthrax Lethal Factor played a large part in out successful application to NIAID to develop a novel class of inhibitors using in silico and NMR-based methods combined with crystallography **(U19 AI56385-01 Dr. Alex Strongin, P.I.)**. Our general approach also led to the successful application for novel therapeutic treatments of Smallpox **(U01 AI061139 - P.I, Dr. Alex Strongin)**

## Conclusions

In this second year of funding we have broadened our approach to (1) carry out focused studies on B. anthracis genome-encoded proteins and (2) strutcural studies of virulence factors from Variola virus, SARS virus, our attention on target selection, protein expression, purification and crystallization of proteins encoded by the Bacillus anthracis pX01 plasmid. We have cloned and expressed a total of 50 new proteins, and structural analysis of several of these is underway. Currently, 6 new crystal structures are essentially complete. We have also determined the first crystal structure of a complex between anthrax protective Antigen and its host cell receptor (published in *Nature*).

**So what section**: Post-exposure therapeutics do not exist for any of the major pathogens likely to be used in biowarfare or bioterrorism. Our work identifies and characterizes structurally and functionally key protein "virulence factors" from these organisms, allowing for the rational structure-based small molecule inhibitor design that can lead to the development of therapeutic drugs to treat anthrax, smallpox, plague and SARS.

## References

DiPerna, G., Stack, J., Bowie, A. G., Boyd, A., Kotwal, G., Zhang, Z., Arvikar, S., Latz, E., Fitzgerald, K. A., and Marshall, W. L. (2004) *J Biol Chem 279*, 36570-8.

Kotwal, G. J., Hugin, A. W., and Moss, B. (1989) *Virology 171*, 579-87.

Sui J, Li W, Murakami A, Tamin A, Matthews LJ, Wong SK, Moore MJ, Tallarico AS, Olurinde M, Choe H, Anderson LJ, Bellini WJ, Farzan M, Marasco WA. Proc Natl Acad Sci U S A. (2004) 101:2536-41

Xu Y, Liang X, Chen Y, Koehler TM, Hook M. J Biol Chem. (2004) 279:51760-8.

## Appendices

**Appendix 1:** Santelli E, Bankston LA, Leppla SH & Liddington RC. Crystal structure of a complex between anthrax toxin and its host cell receptor. ***Nature***. 2004 Aug 19;430(7002):905-8. Epub 2004 Jul 4.

**Appendix 2:** Lieh Yoon Low, Chen Yang, Marta Perego, Andrei Osterman and Robert Liddington "Structure and lytic activity of a Bacillus anthracis prophage endolysin" (mansucript submitted)

**Appendix 3**: Gudrun Stranzl, Marcin Grynberg, Chandra La Clair, Dorinda Shoemaker, Robert Schwarzenbacher, Eugenio Santelli, Adam Godzik, Marta Perego and Robert C. Liddington "Structural and Functional studies of a *Bacillus anthracis* sensor domain" (manuscript in preparation)

**Appendix 4:** Marcin Grynberg, Iddo Friedberg, Marc Robinson-Rechavi, and Adam Godzik "Surprising connections: in-depth analysis of the *Bacillus anthracis* pXO1 Plasmid" (manuscript submitted)

**Appendix 5**: Adrian Tkacz, Leszek Rychlewski and Adam Godzik "VirFact: a relational database of virulence factors and pathogenicity islands (PAIs)" (mansucript submitted)

# Crystal structure of a complex between anthrax toxin and its host cell receptor

**Eugenio Santelli[1], Laurie A. Bankston[1], Stephen H. Leppla[2] & Robert C. Liddington[1]**

[1]*Program on Cell Adhesion, The Burnham Institute, 10901 North Torrey Pines Road, La Jolla, California 92037, USA*
[2]*Microbial Pathogenesis Section, National Institute of Allergy and Infectious Diseases, NIH, Bethesda, Maryland 20892, USA*

**Anthrax toxin consists of the proteins protective antigen (PA), lethal factor (LF) and oedema factor (EF)[1]. The first step of toxin entry into host cells is the recognition by PA of a receptor on the surface of the target cell. Subsequent cleavage of receptor-bound PA enables EF and LF to bind and form a heptameric $PA_{63}$ prepore, which triggers endocytosis. Upon acidification of the endosome, $PA_{63}$ forms a pore that inserts into the membrane and translocates EF and LF into the cytosol[2]. Two closely related host cell receptors, TEM8 and CMG2, have been identified. Both bind to PA with high affinity and are capable of mediating toxicity[3,4]. Here, we report the crystal structure of the PA–CMG2 complex at 2.5 Å resolution. The structure reveals an extensive receptor–pathogen interaction surface mimicking the non-pathogenic recognition of the extracellular matrix by integrins[5]. The binding surface is closely conserved in the two receptors and across species, but is quite different in the integrin domains, explaining the specificity of the interaction. CMG2 engages two domains of PA, and modelling of the receptor-bound $PA_{63}$ heptamer[6–8] suggests that the receptor acts as a pH-sensitive brace to ensure accurate and timely membrane insertion. The structure provides new leads for the discovery of anthrax antitoxins, and should aid the design of cancer therapeutics[9].**

Both TEM8 and CMG2 contain a domain that is homologous to the I domains of integrins, which comprise a Rossmann-like α/β-fold with a metal-ion-dependent adhesion site (MIDAS) motif on their upper surface[10]. Crystal structures of the CMG2 I domain and full-length PA proteins have previously been determined[6,11]. The PA monomer is a long slender molecule comprising four distinct domains. In the PA–CMG2 I domain complex, two of these four domains (II and IV) pack together at the base of PA and engage the upper surface of the CMG2 I domain surrounding the MIDAS motif (Fig. 1), burying a large protein surface (1,900 Å$^2$), consistent with the very high affinity (sub-nanomolar dissociation constant) of this interaction[12]. The I domain adopts the 'open' conformation, typical of integrin–ligand complexes[5,13]. PA mimics the ligand recognition mechanism of the integrins[5] by contributing an aspartic acid side chain that completes the coordination sphere of the MIDAS magnesium ion, as predicted by mutagenesis[14,15] (Fig. 2a, b). This single interaction contributes substantially to binding, as mutation of the aspartic acid to asparagine completely eliminates toxicity, as does mutation of a metal-coordinating residue on the receptor.

However, the MIDAS bond does not fully explain the specificity of the interaction, as it does not distinguish between CMG2 and integrins. Further specificity arises from two additional interactions. First, PA domain IV docks onto the surface of CMG2 adjacent to the MIDAS motif. Domain IV comprises a β-sandwich with an immunoglobulin-like fold, but the mode of binding is quite different from that of antibody–antigen recognition. One of the receptor loops (α2–α3) emanating from the MIDAS motif forms a hydrophobic ridge that inserts into a groove formed by one edge of the β-sandwich where its hydrophobic core is exposed. Flanking this ridge-in-groove arrangement are two further loops from CMG2, which make a number of specific polar interactions and salt bridges (Figs 3 and 4a). Together with the MIDAS contact, CMG2 and PA domain IV bury 1,300 Å$^2$ of surface area, a value very similar to two integrin–ligand interactions that have affinities in the sub-micromolar range[5,13]. CMG2 and TEM8 share 60% identity in their I

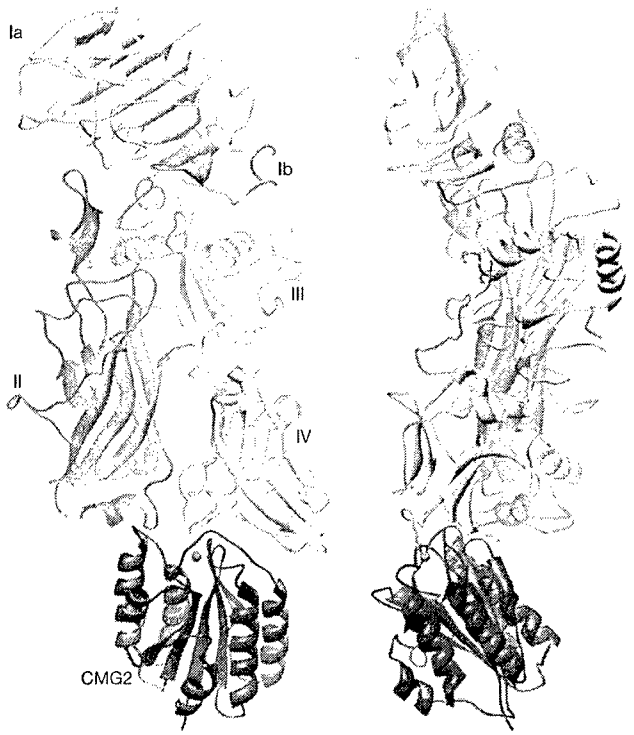



**Figure 1** Structure of the PA–CMG2 complex. Two orthogonal views are shown in ribbon representation. PA is coloured by domain (I–IV). CMG2 is blue; the metal ion is shown as a magenta ball. PA domain I is cleaved after receptor binding, leading to the loss of domain Ia (yellow) and the formation of $PA_{63}$. All molecular graphics images were generated using the UCSF Chimera package[20] (http://www.cgl.ucsf.edu/chimera).

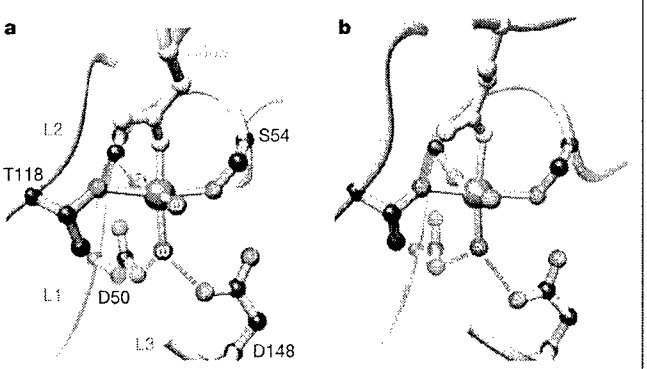



**Figure 2** The MIDAS motifs of the PA–CMG2 complex (**a**) and the collagen–integrin α2β1 complex[5] (**b**). Coordinating side chains and two water molecules (ω) are shown in ball-and-stick representation. The metal is shown in blue. D683 from PA, and a collagen glutamic acid, are in gold. Bond distances to the metal are 2.1 ± 0.2 Å in both cases. The three MIDAS loops (L1–L3) are labelled in **a**.

domains, and homology modelling based on the CMG2 structure shows that this ridge is well conserved in TEM8 and their murine counterparts, implying that they will bind PA in a similar fashion; however, the structure and sequence of the ridge are very different in integrins, explaining their weak binding (Fig. 4b).

The interaction between PA domain II and CMG2 was not anticipated. A β-hairpin from a well-ordered loop (β3–β4) at the bottom of domain II inserts into a pocket on the receptor, burying 600 $Å^2$ of protein surface (Fig. 4b, c). This additional contact may explain the very high affinity of the PA–CMG2 interaction. The pocket is adjacent to the MIDAS motif and is formed by two exposed tyrosine residues (Y119 and Y158) and the β4–α4 loop, which line the sides of the pocket, and by a histidine (H121) at its base. The pocket is conserved in TEM8, but does not exist in the I domains of integrins, thus providing further specificity (Fig. 4b, c). The importance of this loop was shown by systematic mutation of the PA molecule, which revealed three mutations in this loop that reduced toxicity by >100-fold, including G342 at the tip of the β-hairpin that inserts into the pocket[16].

Biophysical studies of channel conductance by $PA_{63}$ pores indicate that the entire region encompassed by residues 275–352 (strands β2 and β3 and flanking loops; see Fig. 3) in domain II rearranges to form a long β-hairpin that lines the channel lumen[7,8]. This requires that the β2 and β3 strands and the β3–β4 loop peel away from the side of domain II. For this to happen, domain IV, which packs against them in the pre-pore, must separate at least transiently from domain II. Thus, by binding to both domains II and IV, CMG2 may restrain the conformational changes that lead to membrane insertion. Indeed, whereas $PA_{63}$ heptamers insert into artificial planar bilayers (in the absence of receptor) when the pH is reduced to 6.5, the pH requirement for receptor-mediated insertion on cells is more stringent, requiring a pH of 5.5 (ref. 17). Thus, we propose that the binding of CMG2 to the β3–β4 loop stabilizes the pre-pore conformation at neutral pH; that is, the receptor may act as a brace to prevent premature membrane insertion on the cell surface before endocytosis. The pH profile of membrane insertion is consistent with the titration of histidine residues, and seven of the nine histidines within $PA_{63}$ cluster at the domain II–IV interface (Fig. 3). In addition, the histidine at the base of the CMG2 pocket

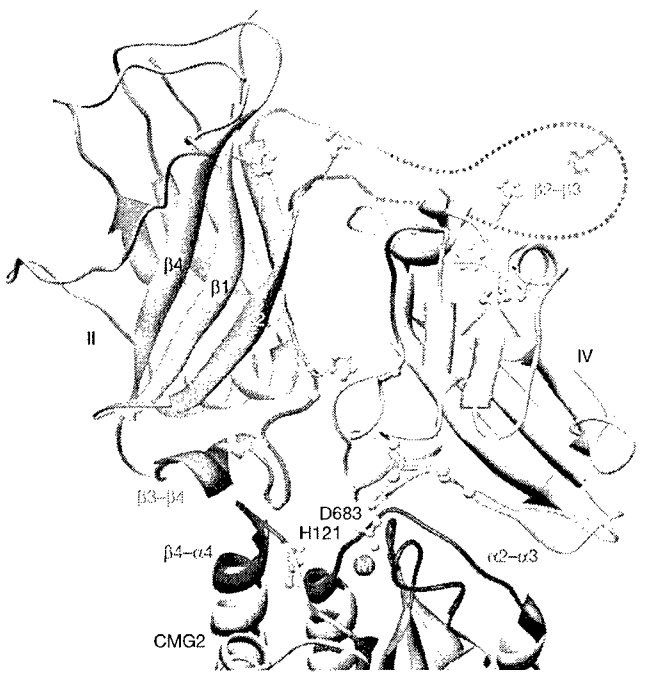



**Figure 3** Intermolecular contacts between PA domains II and IV and CMG2. Contacting regions are coloured blue and green for CMG2 and PA domain IV, respectively. The β2–β3 loop and flanking regions of PA domain II, which are implicated in pore formation, are highlighted in red. The β2–β3 loop is disordered in monomeric PA and is shown schematically as a dashed line. The histidine residues within PA domains II and IV and within the CMG2 I domain are shown coloured cyan and are in ball-and-stick representation. Mutation sites that reduce binding by >100-fold (D683, S337, G342, W346, I656, N657, I665, Y681, N682, P686, L687) are highlighted in gold.

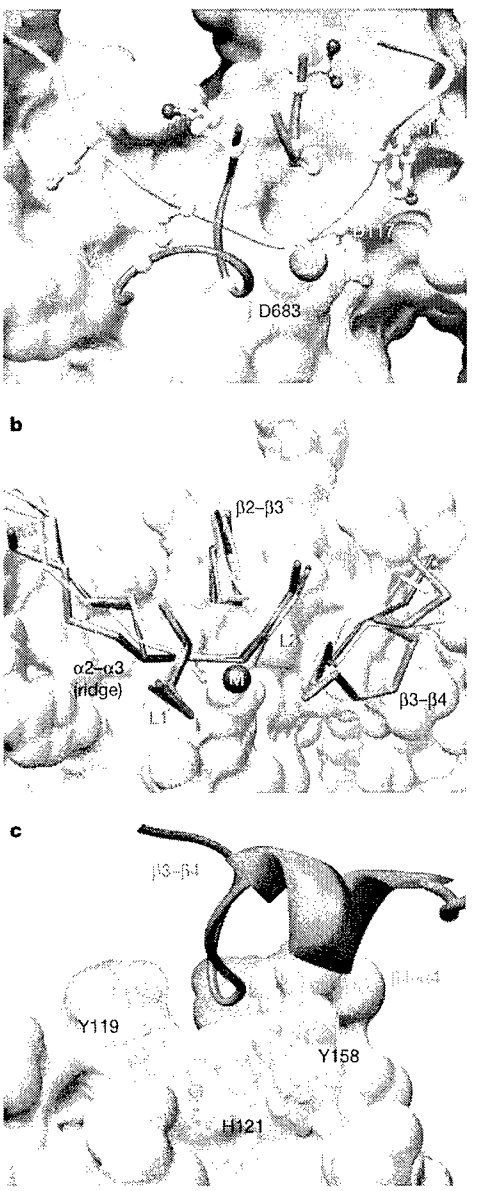



**Figure 4** Key elements of the PA–CMG2 interaction **a**, Solvent-accessible surface of the PA domain IV groove, with key side chains from three CMG2 loops (β1–α1, blue; β2–β3, red; α2–α3, green) shown in ball-and-stick representation. The α2–α3 loop forms the ridge. The MIDAS metal is labelled (M). **b**, Comparison with integrin I domains in the 'open' conformation (CMG2, red; αM, cyan; α2, green; αL, blue) overlaid on the MIDAS motif. **c**, Surface of the CMG2 pocket into which the PA β3–β4 loop (red ribbon) inserts, formed by three CMG2 side chains (shown in ball-and-stick representation) and the β4–α4 loop (cyan).

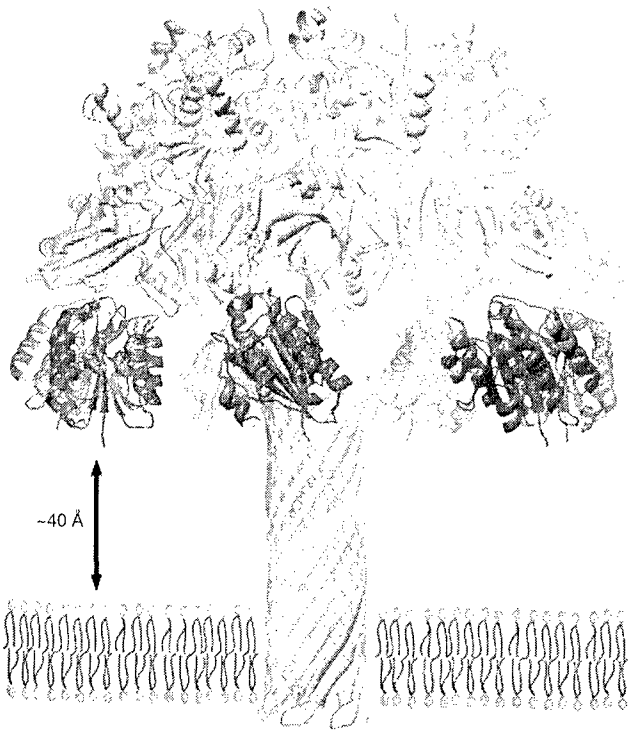



**Figure 5** Hypothetical model of the receptor-bound, membrane-inserted PA pore. The model is based on the pre-pore $PA_{63}$ crystal structure[6], channel conductance studies[8], and the crystal structure of α-haemolysin[19]. The barrel is formed by rearrangement in each monomer of the segment shown in red in Fig. 3. Each $PA_{63}$ monomer is shown in a different colour. Residues 303–324 form the membrane-spanning region of the barrel. Seven copies of the CMG2 I domain bound to the heptamer are in blue. The ~40 Å gap between the CMG2 I domain and the membrane may be occupied by a ~100-residue domain of CMG2, C-terminal to the I domain, which precedes its membrane-spanning sequence.

Table 1 **Data collection and refinement statistics**

| Parameter | Value | |
|---|---|---|
| Space group | $P2_12_12_1$ | |
| Unit cell (Å) | $a = 88.2, b = 94.1, c = 135.6$ | |
| Resolution (Å) | 30–2.5 | |
| Wavelength (Å) | 0.892 | |
| $R_{merge}$ (%) | 17.6 (99.1) | |
| $I/\sigma$ | 11.5 (2.4) | |
| σ-cutoff | None | |
| Average redundancy | 5.3 (5.2) | |
| Completeness (%) | 99.9 | |
| Mosaicity | 0.4 | |
| $R_{work}$ (last shell) | 20.7 (27.5) | |
| $R_{free}$ (last shell) | 26.6 (37.2) | |
| σ-cutoff | None | |
| *B* factors (Å$^2$)* | 32.9, 21.4, 23.3 | |
| r.m.s.d. bond lengths (Å) | 0.17 | |
| r.m.s.d. bond angles (°) | 1.65 | |
| Ramachandran plot (residues, %) | | |
| Most favoured | 655 | 86.3% |
| Additionally allowed | 101 | 13.3% |
| Generously allowed | 3 | 0.4% |
| Disallowed | 0 | 0% |

Values in parentheses refer to the highest resolution shell (2.59–2.50 Å).
*The three values are for Wilson, main chain and side chain, respectively.

(conserved in TEM8) has no H-bonding partners, and is close to an arginine side chain from the β3–β4 loop of PA. Histidine protonation provides a plausible trigger for the release of domain II from CMG2 in the acidified endosome. Indeed, we have shown that the structure of the β3–β4 loop is pH-sensitive, as it becomes disordered when crystals of PA grown at pH 7.5 (in the absence of receptor) are reduced to pH 6.0 (ref. 18).

It is straightforward to model the 7:7 heptameric $PA_{63}$–CMG2 complex, as the crystal structure of the pre-pore is known[6] (Fig. 5). Seven CMG2 I domains lie at the base of the heptameric 'cap', increasing its height by 35 Å. The I domains are well separated, consistent with a 7:7 binding stoichiometry[12], and their amino- and carboxy termini point downwards, towards the membrane. In the transition from pre-pore to pore, the seven hairpin loops, one from each PA monomer[6,8], are predicted to create a 14-stranded, membrane-spanning β-barrel. Assuming an α-haemolysin-like structure[19], the barrel extends ~75 Å below the I domains, with the bottom 30 Å spanning the membrane. This leaves ~40 Å between the bottom of the I domains and the membrane surface, which may be occupied by the second domain of CMG2, which comprises ~100 residues between the I domain and its C-terminal transmembrane sequence. Thus, the receptor may support the heptamer at the correct height above the membrane for accurate membrane insertion, which is stoichiometric on cells but less efficient in the absence of receptor[17].

Soluble versions of the CMG2 and TEM8 I domains protect against anthrax (*Bacillus anthracis*) toxin by acting as decoys[3,15], and our structure will allow for the design of new therapeutic agents that disrupt the PA–receptor interaction. TEM8 is strongly upregulated on the surface of endothelial cells that line the blood vessels of tumours[20,21], allowing for the development of anthrax toxin as an anti-tumour agent[22]; however, toxicity may arise as CMG2 is expressed in most tissues. Although we expect the interactions of TEM8 and CMG2 with PA to be very similar, there are significant differences that may be exploited in the design of PA molecules that would bind better to TEM8 than to CMG2, thus minimizing the side effects from toxin binding to normal tissues. For example, V115 of CMG2, which lies at the heart of the interface with PA domain IV, is a glycine in TEM8, whereas the rim of the pocket that accepts the PA domain II loop has the sequence DGL in CMG2 but is replaced by the sequence HED in TEM8. □

## Methods

### Protein expression and purification

Full-length PA (residues 1–735) was prepared as previously described[14]. The I domain of human CMG2 was cloned as an N-terminal His-tag fusion in pET15b (Novagen) and expressed in *Escherichia coli* strain BL21(DE3). After induction of cell cultures with 0.5 mM IPTG for 2 h at 37 °C, CMG2 was purified from the soluble fraction of the cell lysate by nickel affinity chromatography (HiTrap chelating HP, Pharmacia), followed by removal of the tag with thrombin (Sigma), ion exchange (HiTrap monoQ, Pharmacia) and gel filtration (Superdex S75, Pharmacia), affinity removal of thrombin (HiTrap benzamidine FF, Pharmacia) and incubation in a buffer containing 100 mM EDTA to strip-bound metal. The final product was dialysed and concentrated to 15–20 mg ml$^{-1}$ and flash-frozen in 150 mM NaCl, 20 mM TrisCl pH 7.5, and comprises residues 40–218 of $CMG2^{ATR}$ (GenBank accession number AAK77222) plus an N-terminal extension of sequence GSHMLEDPRG as a result of the cloning strategy. The molecular mass was confirmed by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. To prepare the PA–CMG2 complex, PA was mixed at a final concentration of 4 mg ml$^{-1}$ with a threefold molar excess of CMG2 and a twofold excess of $MnCl_2$, incubated for 20 min at room temperature and purified by gel filtration (Superdex S200, Pharmacia). The complex was extensively dialysed and exchanged, and concentrated to 6 mg ml$^{-1}$ in 20 mM TrisCl pH 7.5, 10 μM $MnCl_2$ for crystallization trials.

### Crystallization and structure solution

Needle-like crystals grew to a size of 10 × 10 × 500 μm in 5–10 days at room temperature in a sitting-drop vapour diffusion set-up using a reservoir buffer containing 50–100 mM CHES pH 9.0–9.2, 25% PEG400. Crystals were flash-frozen at 4 °C in liquid nitrogen using the crystallization buffer with 40% PEG400 as a cryo-protectant before diffraction analysis. The crystals belong to space group $P2_12_12_1$ with unit cell parameters $a = 88.2$ Å, $b = 94.2$ Å, $c = 135.6$ Å. There is one PA–CMG2 complex in the asymmetric unit. A complete native data set to 2.5 Å was collected at beamline 9-1 at SSRL on a ADSC Quantum-315 CCD detector and processed with the HKL package[23] (see Table 1). PA was positioned in the unit cell by Molecular Replacement (Protein Data Bank (PDB) ID code 1acc)[6] using MOLREP, and refined with REFMAC version 5.0 (ref. 24). Density for the MIDAS $Mn^{2+}$ ion and upper loops of the receptor was evident in this map, and a molecule

of CMG2 (PDB ID code 1SHT)[11] was manually placed in the electron density. Model building was performed with O[25] and TURBOFRODO (A. Roussel and C. Cambillau, Silicon Graphics), and the solvent structure was built with ARP/wARP 6.0 (ref. 26). Although the random errors in the diffraction data are high, owing to the small crystal size, the final refinement statistics and maps are excellent (Table 1). Thus, the final *R*-factors are $R_{free} = 26.6\%$ and $R_{work} = 20.7\%$ overall, and $R_{free} = 37.2\%$ and $R_{work} = 27.5\%$ in the outer resolution bin, with root-mean-square deviations (r.m.s.d.) from ideal values of 0.017 Å for bond lengths and 1.65° for angles. Stereochemistry is excellent as assessed with PROCHECK[27], and the model is consistent with composite simulated annealing omit maps (3,000 °C) calculated in CNS[27]. The model comprises residues 16–735 of PA; 41–210 of CMG2, with the exception of three loops (residues 159–174, 276–287 and 304–319) in PA for which no electron density was observed; 139 water molecules; two $Ca^{2+}$ ions in PA domain I; two $Na^{+}$ ions; one PEG molecule; and one $Mn^{2+}$ ion at the MIDAS site. The *B* factors for the $Ca^{2+}$ and $Mn^{2+}$ ions (27–33 $Å^2$) are higher than for the coordinating residues (16–20 $Å^2$). Although the MIDAS metal ion *in vivo* is likely to be $Mg^{2+}$, we have previously shown for integrin I domains that the stereochemistry of the open conformation is not dependent on the nature of the metal ion[5]. The bond lengths to the $Mn^{2+}$ ion are 2.1 ± 0.2 Å, identical to those observed in integrin–ligand complexes[5,13,28]. PA domain I (residues 16–258) undergoes a small rotation as a consequence of crystal constraints when compared with the structure of isolated PA such that the r.m.s.d. values for the superposition of the two molecules are 1.44, 0.58 and 0.79 Å for residues 16–735, 259–735 and 16–258 respectively. CMG2 residues 41–200 superimpose with a r.m.s.d. of 0.60 with the isolated protein[11], while the C-terminal helix (residues 201–210) shifts downwards by one helical turn.

Received 6 May; accepted 18 June 2004; doi:10.1038/nature02763.
Published online 4 July 2004.

1. Mogyeri, M. & Leppla, S. H. The roles of anthrax toxin in pathogenesis. *Curr. Opin. Microbiol.* **7**, 19–24 (2004).
2. Abrami, L., Liu, S., Cosson, P., Leppla, S. H. & Vander Goot, F. G. Anthrax toxin triggers endocytosis of its receptor via a lipid raft-mediated clathrin-dependent process. *J. Cell Biol.* **160**, 321–328 (2003).
3. Bradley, K. A., Mogridge, J., Mourez, M., Collier, R. J. & Young, J. A. Identification of the cellular receptor for anthrax toxin. *Nature* **414**, 225–229 (2001).
4. Scobie, H. M., Rainey, G. J., Bradley, K. A. & Young, J. A. Human capillary morphogenesis protein 2 functions as an anthrax toxin receptor. *Proc. Natl Acad. Sci. USA* **100**, 5170–5174 (2003).
5. Emsley, J., Knight, C. G., Farndale, R. W., Barnes, M. J. & Liddington, R. C. Structural basis of collagen recognition by integrin α2β1. *Cell* **101**, 47–56 (2000).
6. Petosa, C., Collier, R. J., Klimpel, K. R., Leppla, S. H. & Liddington, R. C. Crystal structure of the anthrax toxin protective antigen. *Nature* **385**, 833–838 (1997).
7. Benson, E. L., Huynh, P. D., Finkelstein, A. & Collier, R. J. Identification of residues lining the anthrax protective antigen channel. *Biochemistry* **37**, 3941–3948 (1998).
8. Nassi, S., Collier, R. J. & Finkelstein, A. PA63 channel of anthrax toxin: an extended β-barrel. *Biochemistry* **41**, 1445–1450 (2002).
9. Frankel, A. E., Koo, H.-K., Leppla, S. H., Duesbury, N. S. & Vande Woude, G. F. Novel protein targeted therapy of metastatic melanoma. *Curr. Pharm. Des.* **9**, 2060–2066 (2003).
10. Lee, J.-O., Rieu, P., Arnaout, M. A. & Liddington, R. C. Crystal structure of the A-domain from the α subunit of integrin CR3 (CD11b/CD18). *Cell* **80**, 631–635 (1995).
11. Lacy, D. B., Wigelsworth, D. J., Scobie, H. M., Young, J. A. & Collier, R. J. Crystal structure of the von Willebrand factor A domain of human capillary morphogenesis protein 2: An anthrax toxin receptor. *Proc. Natl Acad. Sci. USA* **101**, 6367–6372 (2004).
12. Wigelsworth, D. J. *et al.* Binding stoichiometry and kinetics of the interaction of a human anthrax toxin receptor, CMG2, with protective antigen. *J. Biol. Chem.* **279**, 23349–23356 (2004).
13. Shimaoka, M. *et al.* Structures of the αL I domain and its complex with ICAM-1 reveal a shape-shifting pathway for integrin regulation. *Cell* **112**, 99–111 (2003).
14. Rosovitz, M. J. *et al.* Alanine scanning mutations in domain 4 of anthrax toxin protective antigen reveal residues important for binding to the cellular receptor and to a neutralizing monoclonal antibody. *J. Biol. Chem.* **278**, 30936–30944 (2003).
15. Bradley, K. A. *et al.* Binding of anthrax toxin to its receptor is similar to α integrin-ligand interactions. *J. Biol. Chem.* **278**, 49342–49347 (2003).
16. Mourez, M. *et al.* Mapping dominant-negative mutations of anthrax protective antigen by scanning mutagenesis. *Proc. Natl Acad. Sci. USA* **100**, 13803–13808 (2003).
17. Miller, C. J., Elliott, J. L. & Collier, R. J. Anthrax protective antigen: prepore-to-pore conversion. *Biochemistry* **38**, 10432–10441 (1999).
18. Petosa, C. in *Crystal Structure of the Anthrax Protective Antigen*. Thesis, Harvard Univ (1995).
19. Song, L. *et al.* Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. *Science* **274**, 1859–1866 (1996).
20. Nanda, A. & St Croix, B. Tumor endothelial markers: new targets for cancer therapy. *Curr. Opin. Oncol.* **16**, 44–49 (2004).
21. Nanda, A. *et al.* TEM8 interacts with the cleaved C5 domain of collagen alpha 3(VI). *Cancer Res.* **64**, 817–820 (2004).
22. Liu, S., Schubert, R. L., Bugge, T. H. & Leppla, S. H. Anthrax toxin: structures, functions and tumour targeting. *Expert Opin. Biol. Ther.* **3**, 843–853 (2003).
23. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
24. Collaborative Computational Project, No. 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
25. Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjelgaard, M. Improved methods for building protein models into electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
26. Morris, R. J., Perrakis, A. & Lamzin, V. S. ARP/wARP and automatic interpretation of protein electron density maps. *Methods Enzymol.* **374**, 229–244 (2003).
27. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
28. Lee, J.-O., Bankston, L. A., Arnaout, M. A. & Liddington, R. C. Two conformations of the integrin A-domain (I-domain): a pathway for activation? *Structure* **3**, 1333–1340 (1995).
29. Sanner, M. F., Olson, A. J. & Spehner, J. C. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* **38**, 305–320 (1996).

**Acknowledgements** We thank the NIH and the DOD for financial support, and the DOE and staff at the SSRL for synchrotron access and support.

**Competing interests statement** The authors declare that they have no competing financial interests.

**Correspondence** and requests for materials should be addressed to R.C.L. (rliddington@burnham.org). The atomic coordinates have been deposited in the Protein Data Bank under accession code 1T6B.

# Cell cycle regulation of central spindle assembly

**Masanori Mishima[1], Visnja Pavicic[1], Ulrike Grüneberg[2], Erich A. Nigg[2] & Michael Glotzer[1]**

[1]*Research Institute of Molecular Pathology, Dr. Bohrgasse 7, A-1030 Vienna, Austria*
[2]*Max-Planck-Institute für Biochemie, Am Klopferspitz 18a, D-82152 Martinsried, Germany*

**The bipolar mitotic spindle is responsible for segregating sister chromatids at anaphase. Microtubule motor proteins generate spindle bipolarity and enable the spindle to perform mechanical work[1]. A major change in spindle architecture occurs at anaphase onset when central spindle assembly begins. This structure regulates the initiation of cytokinesis and is essential for its completion[2]. Central spindle assembly requires the centralspindlin complex composed of the *Caenorhabditis elegans* ZEN-4 (mammalian orthologue MKLP1) kinesin-like protein and the Rho family GAP CYK-4 (MgcRacGAP). Here we describe a regulatory mechanism that controls the timing of central spindle assembly. The mitotic kinase Cdk1/cyclin B phosphorylates the motor domain of ZEN-4 on a conserved site within a basic amino-terminal extension characteristic of the MKLP1 subfamily. Phosphorylation by Cdk1 diminishes the motor activity of ZEN-4 by reducing its affinity for microtubules. Preventing Cdk1 phosphorylation of ZEN-4/MKLP1 causes enhanced metaphase spindle localization and defects in chromosome segregation. Thus, phosphoregulation of the motor domain of MKLP1 kinesin ensures that central spindle assembly occurs at the appropriate time in the cell cycle and maintains genomic stability.**

At the metaphase–anaphase transition, the anaphase-promoting complex triggers proteolysis of cyclin B (an activating subunit of the mitotic kinase Cdk1) and sister chromatid separation. Chromosomes move polewards and non-kinetochore spindle microtubules become bundled, initiating assembly of the central spindle, a structure that has important roles in cytokinesis. In *C. elegans* embryos and other animal cells, central spindle assembly requires centralspindlin[3]. Many proteins that regulate mitosis and cytokinesis re-localize upon anaphase onset. For example, Aurora B and its associated subunits dissociate from centromeres and concentrate on the central spindle[4–6]. Similarly, anaphase onset triggers redistribution of centralspindlin (Fig. 1a, b). In metaphase, centralspindlin is diffuse and in anaphase it localizes to the microtubules positioned between the separating chromosomes, as seen previously[7–10]. ZEN-4 (also known as CeMKLP1) colocalizes with the proline-directed phosphatase CDC-14 (ref. 11) and depletion of CDC-14 prevents ZEN-4 localization[12]. Non-degradable cyclins stabilize Cdk1 activity and prevent central spindle assembly[13,14]. Together these data

# STRUCTURE AND LYTIC ACTIVITY OF A *BACILLUS ANTHRACIS* PROPHAGE ENDOLYSIN

**Lieh Yoon Low[1*], Chen Yang[1*], Marta Perego[2], Andrei Osterman[1] and Robert Liddington[1†].**

**From [1]Infectious and Inflammatory Disease Center, The Burnham Institute, 10901 North Torrey Pines Road, La Jolla, CA92037; and [2]Division of Cellular Biology, Department of Molecular and Experimental Medicine, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037. [*]These authors contributed equally.**

Running title: Bacillus prophage endolysin structure

[†]Address correspondence to: Robert C. Liddington, The Burnham Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037, tel: 858-646-3136; Fax 858-646-3195; E-mail: rlidding@burnham.org

**We report a structural and functional analysis of the λ prophage Ba02 endolysin (PlyL) encoded by the *Bacillus anthracis* genome. We show that PlyL comprises two autonomously folded domains, an N-terminal catalytic domain and a C-terminal cell wall-binding domain (CBD). We determined the crystal structure of the catalytic domain; its three-dimensional fold is related to that of the cell wall amidase, T7 lysozyme, and contains a conserved Zn coordination site and other components of the catalytic machinery. We demonstrate that PlyL is an N-acetylmuramoyl-L-alanine amidase that cleaves the cell wall of several *Bacillus* species when applied exogenously. We show, unexpectedly, that the catalytic domain of PlyL cleaves more efficiently than the full-length protein, except in the case of *B. cereus;* and using GFP-tagged CBD, we detected strong binding of the cell wall-binding domain to *B. cereus* but not to other species tested. To explain these data, and the species specificity of PlyL, we propose that the C-terminal domain inhibits the activity of the catalytic domain through intramolecular interactions that are relieved upon binding of the C-terminal domain to the cell wall. Furthermore, our data show that (when applied exogenously) targeting of the enzyme to the cell wall is not a prerequisite of its lytic activity, which is inherently high. Thus, the catalytic domain of PlyL might be developed as a therapeutic agent with broad efficacy against Gram positive bacteria.**

Endolysins are bacteriophage-encoded enzymes that lyse the host bacterial cell wall during the lytic phase of the phage infectious cycle. They typically consist of an N-terminal catalytic domain and a C-terminal domain that targets the enzyme to the cell wall, providing high species and strain specificity (1,2). For example, the *Listeria monocytogenes* lysins, Ply118 and Ply500, specifically hydrolyse *Listeria* cells, but are inactive in the absence of the cell wall-binding domain (1).

A comparative genome analysis of *Bacillus anthracis* revealed a gene encoding a putative endolysin within the integrated copy of the λ Ba02 prophage, which we will call PlyL. PlyL has a high degree of sequence similarity in its catalytic domain with an endolysin from the bacteriophage γ (PlyG) (3,4), which specifically lyses and kills *B. anthracis* and closely related species when added exogenously to bacterial cultures. For this reason, PlyG is being developed as a diagnostic and therapeutic agent (5).

Here we describe a structural and functional analysis of PlyL. We show that the N-terminal (catalytic) domain is an amidase with high inherent lytic activity against the cell wall of several *Bacillus* species. In contrast to many previously described enzymes, we have found that the presence of the C-terminal domain either reduces or has no effect on the lytic activity. This unexpected finding suggests that the catalytic domain of PlyL could be developed as a relatively broad-spectrum antibacterial agent.

## MATERIALS AND METHODS

**Cloning and expression of full-length endolysin and C-terminal domain** - Full length PlyL was cloned by PCR from the *Bacillus anthracis* Ames

strain total DNA extract prepared by Dr Phil Hanna (University of Michigan Medical School) using the oligonucleotide primers 5'-AAAGGAGATATACATATGGAAATCAGAAAA AAATTAGTT-3' (forward) and 5'-GAATTCGGATCCTCATTATTTATCATCATAC CACCAATC-3' (reverse). We used the forward primer 5'-GGAGATATACATATGGCAAGTGCAACGGTA ACCCCTAAA-3' with the same reverse primer. PCR products were cloned into pET22b (Novagen) via *Nde*I and *Bam*HI restriction sites (without tag). The resulting plasmids were transformed into BL21DE3 (Novagen) for protein expression. Full-length and C-terminal domain proteins were expressed using the same protocol. Transformed cells from overnight plates were used to inoculate 1 L of 2xTY medium (16 g/L Tryptone, 10 g/L yeast extract, and 5 g/ NaCl; with 100 μg/ml ampicillin), and allowed to grow to $OD_{600}$ of 1.0 at 37°C. 1 mM IPTG was added to induce protein expression over three hours at 37°C.

**Full length PlyL purification** - Cells were harvested by centrifugation at 4°C. 30 ml of lysis buffer (50 mM Na-Mes(morpholinoethanesulfonic acid) pH 6.0, 10 mM β-mercaptoethanol, 0.1 % Triton X-100, and 0.1 mM $ZnSO_4$) was used to resuspend the cell pellet. Resuspended cells were lysed by sonication and clarified by centrifugation for 1 hour at 4°C. Clarified lysate was loaded directly into a HITRAP 5 ml SP column on an Akta FPLC (Amersham Biosciences) equilibrated with 50 ml buffer A (50 mM Na-Mes pH 6.0, 10 mM β-mercaptoethanol, and 0.1 mM $ZnSO_4$). Unbound protein was eluted by washing the column with 50 ml buffer A. A gradient of 0-1 M NaCl in buffer A with a total volume of 50 ml was applied to the column to elute the protein. Fractions containing the full length PlyL, more than 90% pure as verified by SDS-PAGE, were pooled and concentrated to 10-20 mg/ml.

**Purification and crystallization of the catalytic domain**. The N-terminal catalytic domain was generated by limited proteolysis of the full-length PlyL using elastase at a ratio of 1:100 at room temperature for 16 hours. A Superdex S75 16/60 column (Amersham Biosciences) was used as a final column to purify the catalytic domain. The buffer was 20 mM Tris-Cl (pH 7.0), 100 mM NaCl, 10 mM β-mercaptoethanol. The final purified protein was concentrated to 20 mg/ml. Mass spectrometry and amino acid analysis revealed that elastase cleaved after residue Val159. The protein appeared as a single band on SDS-PAGE, and the molecular weight was confirmed by MALDI-MS. Crystals were obtained by hanging-drop vapor-diffusion at 20°C, using a reservoir of 0.6 M $NaH_2PO_4$, 1.0 M $K_2HPO_4$, 0.1 M acetate at pH 6.7. Each drop consisted of 2 μl protein and 1 μl buffer. Crystals grew as hexagonal rods to 0.1 mm x 0.1 mm x 0.3 mm in three days at room temperature. They adopt space group $P6_1$ with cell dimensions a=b=163.2 Å, c=37.3 Å. To prepare for cryo-X-ray data collection, the crystals were soaked in a series of steps with crystallization buffer containing 10% glycerol followed by buffer supplemented with 20% glycerol. All X-ray data sets were collected at 100 K.

**C-terminal domain purification and crystallization** - Bacterial cell extracts were prepared as described above. Supernatant was loaded onto the equilibrated Ni-NTA column, and washed with 10 column-volumes of wash buffer. The elution buffer was similar to the wash buffer, but included 300 mM imidazole. The protein elution was directly linked to an equilibrated gel filtration column (20 mM Tris-Cl and 100 mM NaCl at pH 7.0). The His-tagged C-terminal domain eluted as a monomer. Crystals of the C-terminal 75 amino acid domain were obtained by equilibration against 1.5 M $(NH_4)_2SO_4$ and 10% glycerol in TrisCl pH 7.0 by hanging-drop vapor diffusion. The crystal grew to a size of 0.1 x 0.1 x 0.3 $mm^3$ in seven days at room temperature; they diffract to 2.7 Å resolution using a Rigaku FR-E High Brilliance X-Ray generator and adopt space group $P4_12_12$ with cell dimensions a=b=52.5 Å, c=224.2 Å.

**Structure Determination of the catalytic domain.** MAD data sets were collected at beam-line 9-2 at the Stanford Synchrotron Radiation Laboratory using a MAR345 image plate, and processed using the programs DENZO and SCALEPACK (6). The presence of a zinc ion in the crystal was confirm by a fluorescence scan at the Zn L-I edge. 18 selenomethionine sites were found using SOLVE (7) and used for phase

calculation to a resolution of 2.0 Å. An initial model was generated by RESOLVE (8), further model building was done using O (9) and the model refined with CNS (10) (version 1.1 on Mac OS X). Native crystals were obtained under identical conditions. A data set was collected in-house with a Rigaku FR-E High Brilliance X-Ray generator using the R-axis IV detector. The CNS-refined model of the selenomethionine structure was used as the input template for native refinement to a resolution of 1.86 Å. There are three molecules (A-C) in the asymmetric unit, and a solvent content of 46%. Density for the last two amino acids in the molecules A and C are missing. Molecule B has the most complete density throughout, and its B-factors are lower than for the other two molecules. Refinement statistics are presented in Table 1. The coordinates and structure factors have been deposited with the PDB with accession code 1YB0

**Assay of lytic activity.** The activity of PlyL when applied exogenously to cultures of *B. anthracis* (Sterne 34F2), *B. cereus* ATCC 4342, *B. megaterium* WH320, *B. subtilis* 168 and *Escherichia coli* CFT073 was tested. Cultures were grown to mid-exponential phase, and cells were harvested and resuspended in 10 mM sodium phosphate (pH 7.0). The lysis of cell suspensions upon addition of 2 to 4 μM pure endolysin samples was monitored at 600 nm.

**Determination of the cleavage site in peptidoglycan.** Peptidoglycan suspension (0.5 mg/ml) from *B. subtilis* (Fluka) was incubated at 37°C with purified PlyL (0.4 μM) in 10 ml of Good's buffer (20 mM Na-MES, pH 6.5) containing 100 mM KCl. Boiled PlyL was used as a control. After incubation for 30, 60, and 120 min, samples were boiled and centrifuged at 13000 rpm/min, clear supernatants were analyzed for the release of free amino acids using a modified protocol described in [12]. 100 μl aliquots were mixed with 12 μl of 10% $K_2B_4O_7$ and 10 μl of 1-fluoro-2,4-dinitrobenzene solution (0.1 M in ethanol) was added, and the mixture was heated at 65°C for 45 min in the dark. Following acid hydrolysis in 4 M HCl for 12 h at 95°C, the dinitrophenyl (DNP)-labeled compounds were analyzed by HPLC on a reverse-phase column ($C_{18}$, 4.6x 150 mm, Vydac). The labeled amino acids were eluted with a linear gradient from 90% A + 10% B to 30% A + 70% B (A: 10% acetonitrile in 20 mM acetic acid; B: 90% acetonitrile in 20 mM acetic acid), and detected at 365 nm. The release of free reducing groups during the enzymatic reaction was measured by a modified Morgan-Elson reaction (12) using *N*-acetylglucosamine as the standard.

**C-terminal domain cell binding assay.** A modified Green Fluorescent Protein (GFP) gene (gift of Dr Ruchika Gupta) was PCR-amplified using the following oligonucleotide primers: 5'-CGCGGCAGCCATATGGTGAGCAAGGGCGAGGAGCTGTTC-3' and 5'-GCCCGGATCCTCGAGTTACTTGTACAGCTCGTCCATGCC-3'. The resulting fragment was digested by *Nde*I and *Xho*I (underlined) and ligated with the XhoI-BamHI fragment of the C-terminal domain of PlyL which was amplified using 5'-AGCCATATGCTCGAGATGGCAAGTGCAACGGTAACCCCT-3' (forward) and the same reverse oligonucleotide that was used for the cloning of the full length protein. The GFP-C-terminal domain fusion and a GFP control were cloned into a pET15b vector via NdeI and BamHI or XhoI, respectively. Both proteins were expressed and purified using Ni-NTA affinity chromatography and gel-filtration as described above. Cell samples for the binding assays were obtained by growing Bacilli cultures to late log phase. Cells were harvested, washed with PBS-T (PBS + 0.1% Tween-20), and incubated with 0.4 mM protein samples (GFP-C-domain fusion or GFP control) for 5 min at room temperature, prior to three washes with PBS-T. The washed cells were smeared onto a microscope slide for confocal image analysis with the Biorad Radiance 2100 Multiphoton Laser Scanning Confocal Microscope system equipped with Argon laser (Image Analysis and Histology Facilities, The Burnham Institute). The objective used was 60X LSM with oil immersion, and zoom 5 on the N.A.1.0 (Olympus) microscope. The wavelength of 488 nm was used to excite the GFP.

## RESULTS

**Identification and characterization of PlyL** - A Blast search (http://www.ncbi.nlm.nih.gov/BLAST/) using the γ phage endolysin, PlyG, as the query

sequence identified two genes encoding putative endolysins located within an integrated prophage of *B. anthracis*. The λ Ba01 and λ Ba02 endolysins are annotated as BA3767 and BA4073 ("PlyL"), respectively, in the genome sequence of *B. anthracis* Ames (NCBI accession number NC_003997). Additional endolysins from other *Bacillus* species and their phages were also detected in this search. Those with greater than 30% identity over their catalytic domains are shown in Figure 1. PlyL is most closely related to PlyG in both the enzymatic (93% identity) and C-terminal (60% identity) domains. BA3767 is also very similar but lacks the C-terminal domain.

We cloned and expressed a *B. anthracis* gene encoding BA4073/PlyL. Crystallization trials of the full-length protein were unsuccessful. However, limited proteolysis using elastase allowed us to isolate a stable N-terminal fragment (residues 1-159). Cleavage occurs at the junction between the predicted catalytic and cell-wall binding domains. This fragment was much more soluble than the full-length protein (> 40 mg/ml versus < 3 mg/ml)), and crystallized readily. We also crystallized the C-terminal domain; although we have not yet solved its structure, the existence of crystals that diffract to high resolution indicates that it is an autonomously folded domain.

***N*-acetylmuramoyl-L-alanine amidase activity of PlyL resides in its N-terminal domain** - To assess the enzymatic activity of PlyL, peptidoglycan from *B. subtilis* was treated with full-length PlyL and the elastase-generated N-terminal fragment. No increase in free reducing groups derived from peptidoglycan could be observed, indicating that the enzyme is neither a glucosaminidase nor a muramidase. The free amino groups of the digested (solubilized) products were labeled with 1-fluoro-2,4-dinitrobenzene. After acid hydrolysis, the DNP-labeled compounds were separated by HPLC. Only the amount of DNP-alanine was increased significantly (Supplementary Fig. 1A), which indicates that the enzyme is an *N*-acetylmuramoyl-L-alanine amidase, specifically cleaving the amide bond between *N*-acetylmuramic acid and L-alanine. The same result was observed for the N-terminal proteolytic fragment, showing that it comprises a complete catalytic domain. The N-terminal domain was more active than the full-length protein in this assay (Supplementary Fig. 1B), providing the first indication that the C-terminal domain is autoinhibitory.

**Structure of the PlyL N-terminal domain** - We solved the structure of the PlyL catalytic domain (residues 1-159) at 1.86 Å resolution using MAD phasing from a selenomethionine-substituted protein (Table I). The fold is most similar to those of the T7 lysozyme (13), *Citrobacter* AmpD (14) and the *Drosophila* peptidoglycan recognition protein PGRP-LB (15), with which it shares 10-20% identity. For consistency, we have followed the strand and helix nomenclature of T7 lysozyme. The overall fold consists of a 6-stranded β-sheet flanked by four long α-helices (one at the front (α1) and three at the back (α2 α3 and α4) as well as a number of elaborate loops with short α-helical segments (Fig. 2, 3A). Compared with T7 lysozyme, an N-terminal extension creates an additional β-strand (β0) at one end of the sheet. A zinc ion binds to the front face of the molecule at the center of the active site, coordinated by His29 from strand β1, and by two residues, His129 and Cys137, on either side of strand β5. The fourth ligand is a phosphate (or sulfate) ion from the crystallization buffer.

**Enzyme Active-site** - The active site is solvent-exposed and lies in a shallow groove on the protein surface, consistent with the ability to cleave a highly cross-linked and branched polymer. Helix α1 packs more closely against the β-sheet in PlyL than in T7 lysozyme, so that the pronounced substrate-binding binding groove observed for T7 lysozyme is not seen for PlyL. The active site can be overlaid closely with that of T7 lysozyme (Fig. 3B). The three zinc-coordinating residues (His29, His129 and Cys137) are conserved between PlyL and T7 lysozyme (the third zinc-coordinating residue is an Asp in *Citrobacter* AmpD). PlyL Lys135 is structurally analogous to Lys128 of T7 lysozyme, which has been shown to be important for catalysis (13), perhaps by stabilizing the developing negative charge on the amide carbonyl in the transition state; however, PGRP-LB has a threonine at this position. Tyr46 in T7 lysozyme and Tyr78 in PGRP-LB are important for catalysis, and are thought to act as the general base to activate the

nucleophilic water molecule. On the basis of sequence alignment the analogous residue in PlyL was predicted to be Phe53. However, in the crystal structure the side chain of Phe53 adopts a different orientation and the carboxylate group of Glu90 (from a neighboring strand) occupies the space analogous to the T7 tyrosine. To demonstrate a catalytic role for Glu90 in PlyL, we mutated it to alanine, and indeed this mutation completely abolished the amidase activity (data not shown).

There are only ten amino acid residues different within the N-terminal domains of PlyL and PlyG, so that their 3D structures should be almost identical. These differences are plotted on the three-dimensional model of PlyL (Fig. 3A). Most of the differences are located on the surface of the molecule, and all of them are distant from the active site and a putative substrate binding cleft, suggesting that the two catalytic domains should have similar or identical substrate specificity and catalytic activity.

**Lytic activity of PlyL.** - We next examined the lytic activity of PlyL on whole cells of several bacilli, as measured by light scattering ($OD_{600}$) (Fig. 4) and confirmed by microscopy. We found that the full length PlyL lysed *B. cereus* with an efficiency comparable to that reported for PlyG on *B. anthracis* and some strains of *B. cereus* (5). However, in marked contrast with PlyG, a relatively high lytic activity of PlyL was established on *B. megaterium* and lower but detectable activity on *B. subtilis* and *B. anthracis.*

We found, unexpectedly, that the N-terminal catalytic domain of PlyL is more active than the full-length protein in lysing *B. subtilis*, *B. megaterium* and *B. anthracis* cells. The strongest enhancement was observed on *B. subtilis* (Fig. 4B, C). By contrast, the removal of the C-terminal domain had almost no effect on the lytic activity towards *B. cereus.*

To further assess the role of the C-terminal domain of PlyL, we performed cell-binding studies using a recombinant C-terminal domain fused with GFP. When added to *B. cereus* and viewed under a confocal microscope, a clear green fluorescence can be observed around the cells (Fig 4D). No binding was observed with *B. megaterium* or *B. subtilis* (*B. anthracis* was not tested).

## DISCUSSION

We have shown that the endolysin from the *B. anthracis* λ prophage Ba02, PlyL, is a bona fide cell wall lytic amidase with a modular organization comprising an N-terminal catalytic domain and a C-terminal cell wall-binding domain. We determined the three-dimensional atomic resolution structure of the catalytic domain and showed that the overall fold and active site are similar to but distinct from that of T7 lysozyme and other amidases. The zinc coordinating residues, His29, His129 and Cys137 are invariant among the *Bacillus* endolysins listed in Figure 1, as are the other active-site residues, Glu90 and Lys135. The role of Glu90 was not predicted from sequence alignments with T7 lysozyme, but its side chain occupies a similar spatial location to the general base Tyr in T7 lysozyme, and we demonstrated a critical role for Glu90 in catalysis by mutagenesis. Our results suggest that all of the enzymes listed in Figure 1 should have an *N*-acetylmuramoyl-L-alanine amidase activity and a similar catalytic mechanism (as was already demonstrated for the TP21 endolysin (1)). In particular, the 10 residues that differ between PlyL and PlyG do not lie close to the active site, so that their distinct lytic specificities are presumably endowed by the C-terminal domain, which is less conserved.

We showed that the C-terminal domain is indeed a cell wall-binding domain (CBD) and that it interacts specifically with *B. cereus* cells. We further showed that the presence of the CBD within the full-length PlyL has an inhibitory effect on the lytic activity of the catalytic domain when tested with peptidoglycan or with the whole cells of *B. subtilis* and, to a lesser extent, with *B. megaterium* and *B. anthracis.* By contrast, the presence of the CBD had a negligible effect on the activity of PlyL towards *B. cereus.*

To reconcile these observations we propose that the C-terminal domain of PlyL has a dual function (Fig. 5): (i) in the absence of specific interaction with cognate cell wall, the CBD plays an autoinhibitory role, similar to a propeptide in zymogens, by binding to the catalytic domain and

blocking access to the active site either sterically or allosterically; and (ii) the CBD participates in species-specific cell wall binding (recognition) which disrupts the interaction between the CBD and the catalytic domain, thus relieving the inhibitory effect. For example, the marked difference in the activity of the full-length PlyL and the free N-terminal domain against *B. subtilis* can be explained by very weak binding of the CBD to the *B. subtilis* cell wall, while the cell wall is intrinsically sensitive to the amidase activity. In the case of *B. cereus* where the full-length and truncated enzymes have an almost equally high activity, we propose that strong binding of the CBD to the target cell wall releases the constraints on the catalytic domain. It is surprising, however, that localization of the enzymatic domain to the cell surface does not enhance the rate of lysis via a local concentration effect.

Endolysins are generally observed to be highly specific towards a particular species of bacteria, by virtue of their distinct CBDs that recognize variable cell wall structures (1,2). Our observation that the catalytic domain of PlyL has strong lytic activity against a number of different *Bacillus* species and that this activity does not require (or is inhibited by) the CBD suggests either that the PlyL/PlyG family of endolysins are atypical or that the kinetics of lysis are different when the lysin is applied exogenously rather than endogenously. We note however that there are precedents for such behavior: thus, certain phage hydrolases have been shown to maintain or even increase their exogenous lytic activity when the C-terminus is truncated (16-18). These findings raise the possibility of developing the catalytic domains of certain lysins as broad-spectrum therapeutic agents.

**Table I: Crystallographic statistics**

| **A. Data collection** | Peak | Se-Met Remote | Inflection | **Native** |
|---|---|---|---|---|
| Wavelength (Å) | 0.9792 | 0.8919 | 0.9794 | 1.54 |
| Resolution (Å) | 2.03 | 2.03 | 2.03 | 1.86 |
| Resolution range | 30-2.03 (2.07-2.03) | 30-2.03 (2.07-2.03) | 30-2.03 (2.07-2.03) | 30-1.86 (1.89-1.86) |
| Total observations | 190756 | 182585 | 191392 | 188742 |
| Unique reflections | 39226 | 39277 | 39375 | 53283 |
| Completeness | 98.9(96.8) | 98.4(95.5) | 96.9(98.9) | 100(99.9) |
| Average I/$\sigma$ | 19.1(3.0) | 17.0(2.6) | 18.4(2.7) | 23.6(2.2) |
| $R_{sym}$ | 10.8(44.2) | 9.2(45.3) | 9.2(46.5) | 8.7(52.4) |

Figure of merit after SOLVE = 0.41

| **B. Refinement** | **Native** | | |
|---|---|---|---|
| Refinement range | 30.0-1.86 | | |
| Number of reflections | 48365 | | |
| $R_{work}$ | 20.8 | | |
| $R_{free}$ | 24.3 | | |
| Number of refined residues | 479 | | |
| Number of water molecules | 276 | | |
| rmsd from ideality | | | |
| Bonds lengths (Å) | 0.007 | | |
| Bond Angles (deg) | 1.5 | | |
| Average B-value (Å$^2$) | A | B | C |
| Protein | 27.7 | 25.9 | 40.6 |
| Main-chain | 26.2 | 24.3 | 39.5 |
| Side-chain | 29.3 | 27.4 | 41.7 |
| Solvent | | 34.3 | |
| Ramachandran Plot (%) | | | |
| Most favoured | | 85.8 | |
| Additionally allowed | | 13.9 | |
| Generously allowed | | 0.2 | |
| Disallowed | | 0.0 | |

Figures in parenthesis refer to the highest resolution shell.
$R_{sym} = \Sigma|I_h - \langle I_h \rangle|/\Sigma I_h$, where $\langle I_h \rangle$ is the average intensity over symmetry equivalent reflection.
$R_{work} = \Sigma|F_{obs} - F_{calc}|/\Sigma F_{obs}$, where the summation is over the data used for refinement.
$R_{free}$ was calculated using 5% of data excluded from refinement (Kleywegt, 1996).

## Figure Legend

**Figure 1. Sequence alignment of a family of *Bacillus* endolysins.** Zinc coordinating and active-site residues are colored in cyan and red, respectively. Arrows and cylinders represent the β-strands and α-helices secondary structures of the λ Prophage Lambda Ba02 endolysin (PlyL). XlyB, XlyA, TP21 and ϕ-105 are endolysins from *B. licheniformis*, *B. subtilis*, *B. cereus* and *B. subtilis*, respectively. Alignment was performed using the program CLUSTALX version 1.82 (19).

**Figure 2. Three-dimensional structure of PlyL and related amidases.** Molscript (version 2.1; (20,21)) ribbon representations of the structures of *Bacillus* endolysin, T7 lysozyme (PDB: 1LBA) , PGRP-LB (PDB: 1OHT) and AmpD (PDB: 1J3G). The zinc ion is shown as grey sphere.. The colors represent the secondary structures arrangement. There are three molecules in the asymmetric unit, and these are essentially identical in structure, with backbone (Cα) RMS deviations of 0.29 Å. The backbone RMS differences with T7 lysozyme and PGRP-LB, are 1.8 Å (for 107 atoms) and 2.0 Å (for 106 atoms), respectively.

**Figure 3. Stereo views of PlyL and active site comparisons (A)** Stereo Cα representation of PlyL. Amino acids differences between PlyG and PlyG are indicated. Most of these are surface-exposed except for Val55, which makes hydrophobic contacts with Trp68 in PlyL. In PlyG, the Val55 is replaced by the larger residue Ile, but this is complemented by a change to the smaller Leu in place of Trp68. **(B)** Stereo view of the active site residues of PlyL (light gray), T7 lysozyme (PDB: 1LBA) (Medium gray), and PGRP-LB (PDB: 1OHT) (dark gray).

**Figure 4. Lytic and cell wall binding activity of PlyL.** Lysis of viable cells of 4 different *Bacillus* species by **(A)** full-length PlyL and **(B)** the N-terminal catalytic domain. The protein concentration was 0.4 μM except for *B. anthracis* where 2 μM was used. **(C)** The time required for the full-length and catalytic domain of PlyL to reduce the $OD_{600}$ by half ($t_{1/2}$). Error bars indicate the standard deviation from at least three independent experiments. **(D)** Confocal image of the GFP-CBD fusion protein binding to the cell wall of *B. cereus,* showing the rod-shape cells with green fluorescence. No fluorescence was observed for other *Bacillus* species or for the control with GFP alone (data not shown).

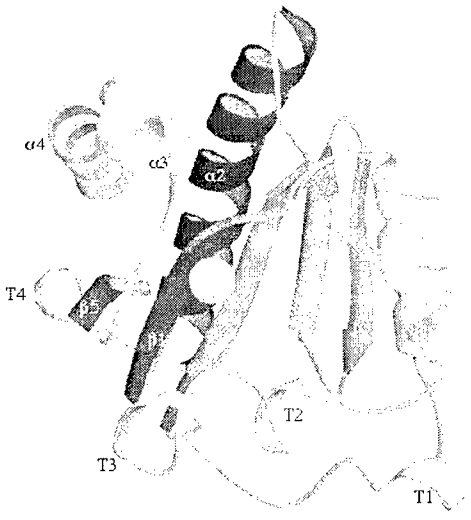
**Figure 5. A proposed model of species-specific activation of PlyL. (A)** In the unbound full-length PlyL, a C-terminal domain (gray oval) suppresses a catalytic activity of the N-terminal domain (blue square) by preventing the access of peptidoglycan substrate (light blue circles) to the active site, either sterically or allosterically. **(B)** An alternative, active, conformation of PlyL is stabilized by specific interactions of the C-terminal domain with a cell-wall component (shown by black cross) characteristic of cognate bacteria (such as *B. cereus*). In the absence of such an interaction partner, as in the case of *B. subtilis*, *B. megaterium* or a free peptidoglycan in vitro, the full-length PlyL would exist mostly in the inactive (closed) conformation. **(C)** A truncation of the C-terminal domain maintains the enzyme in a constitutively active form.

## REFERENCES


1. Loessner, M. J., Kramer, K., Ebel, F., and Scherer, S. (2002) *Mol Microbiol* **44**, 335-349
2. Lopez, R., Garcia, E., Garcia, P., and Garcia, J. L. (1997) *Microb Drug Resist* **3**, 199-211
3. Inglesby, T. V., O'Toole, T., Henderson, D. A., Bartlett, J. G., Ascher, M. S., Eitzen, E., Friedlander, A. M., Gerberding, J., Hauer, J., Hughes, J., McDade, J., Osterholm, M. T., Parker, G., Perl, T. M., Russell, P. K., and Tonat, K. (2002) *Jama* **287**, 2236-2252
4. Brown, E. R., and Cherry, W. B. (1955) *J Infect Dis* **96**, 34-39
5. Schuch, R., Nelson, D., and Fischetti, V. A. (2002) *Nature* **418**, 884-889
6. Otwinowski, Z., and Minor, W. (1997) *Methods in Enzymology* **276**, 307-326
7. Terwilliger, T. C., and Berendzen, J. (1999) *Acta Crystallogr D Biol Crystallogr* **55 ( Pt 4)**, 849-861
8. Terwilliger, T. C. (2000) *Acta Crystallogr D Biol Crystallogr* **56 ( Pt 8)**, 965-972
9. Jones, T. A., Zou, J.-Y., Cowan, S. W., and Kjelgaard, M. (1991) *Acta Crystallogr.* **A47**, 110-119
10. Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. (1998) *Acta Crystallogr D Biol Crystallogr* **54 ( Pt 5)**, 905-921
11. Rygus, T., and Hillen, W. (1991) *Appl Microbiol Biotechnol* **35**, 594-599
12. Ghuysen, J.-M., Tipper, D. J., and Schneewind, O. (1966) *Methods in Enzymology* **8**, 685-699
13. Cheng, X., Zhang, X., Pflugrath, J. W., and Studier, F. W. (1994) *Proc Natl Acad Sci U S A* **91**, 4034-4038
14. Liepinsh, E., Genereux, C., Dehareng, D., Joris, B., and Otting, G. (2003) *J Mol Biol* **327**, 833-842
15. Kim, M. S., Byun, M., and Oh, B. H. (2003) *Nat Immunol* **4**, 787-793
16. Loessner, M. J., Gaeng, S., Wendlinger, G., Maier, S. K., and Scherer, S. (1998) *FEMS Microbiol Lett* **162**, 265-274
17. Loessner, M. J., Gaeng, S., and Scherer, S. (1999) *J Bacteriol* **181**, 4452-4460
18. Baba, T., and Schneewind, O. (1996) *EMBO J* **15**, 4789-4797
19. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997) *Nucleic Acids Res* **25**, 4876-4882
20. Kraulis, P. J. (1991) *J. Appl. Crystallog.* **24**, 946-950
21. Merritt, E. A., and Murphy, M. E. P. (1994) *Acta Crystallogr.* **D50**, 869-873


Figure 1

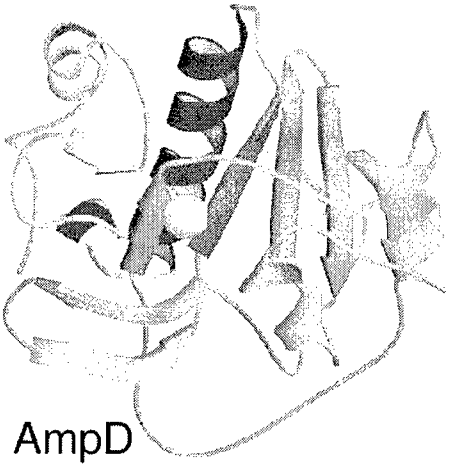
**L. Y. Low *et al.***




PlyL

T7 lysozyme

PGRP-LB

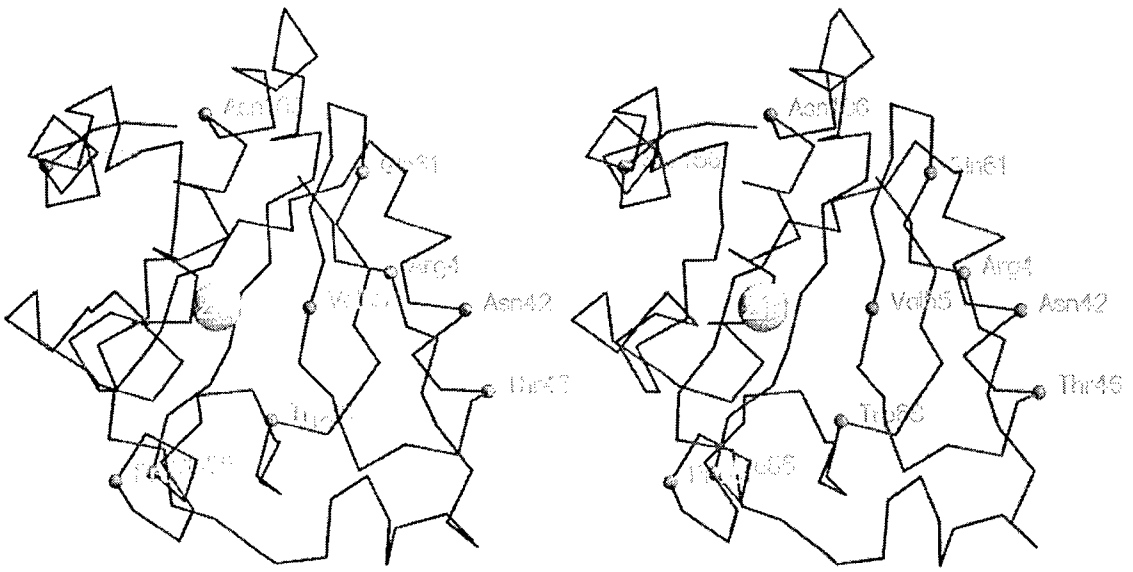


AmpD

Figure 2

**L.Y. Low *et al.***

A


Asn42
Thr45

B

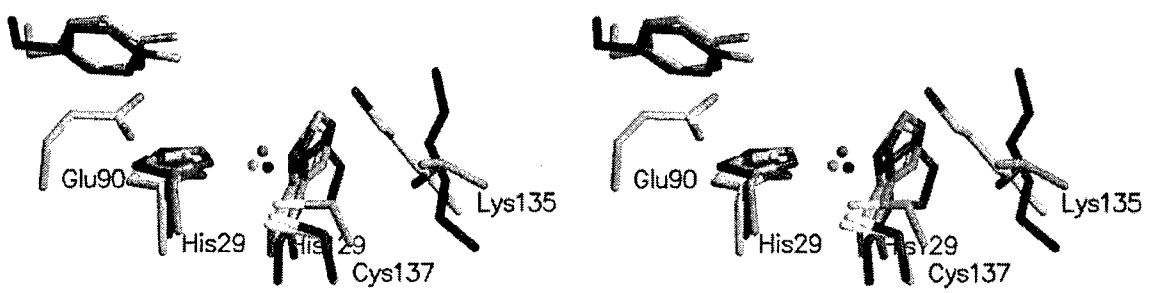

Glu90
His29
His29
Cys137
Lys135
Glu90
His29
Cys137
Lys135

**Figure 3**

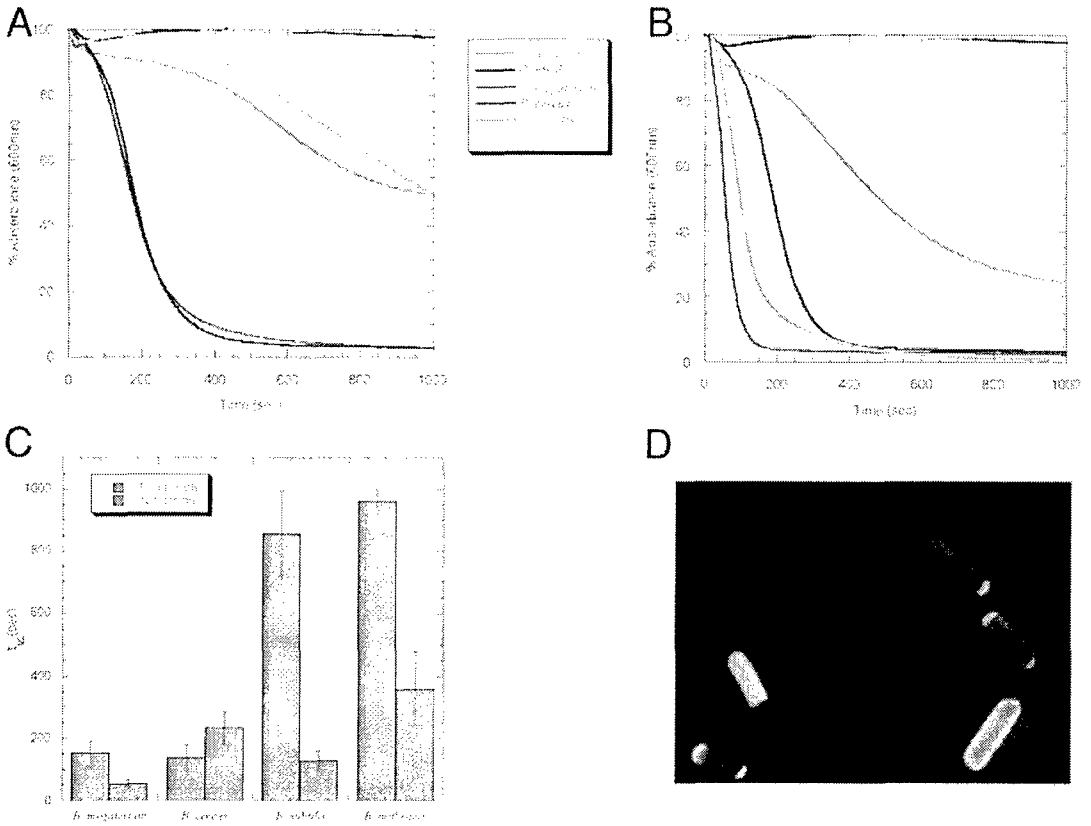



**Figure 4**

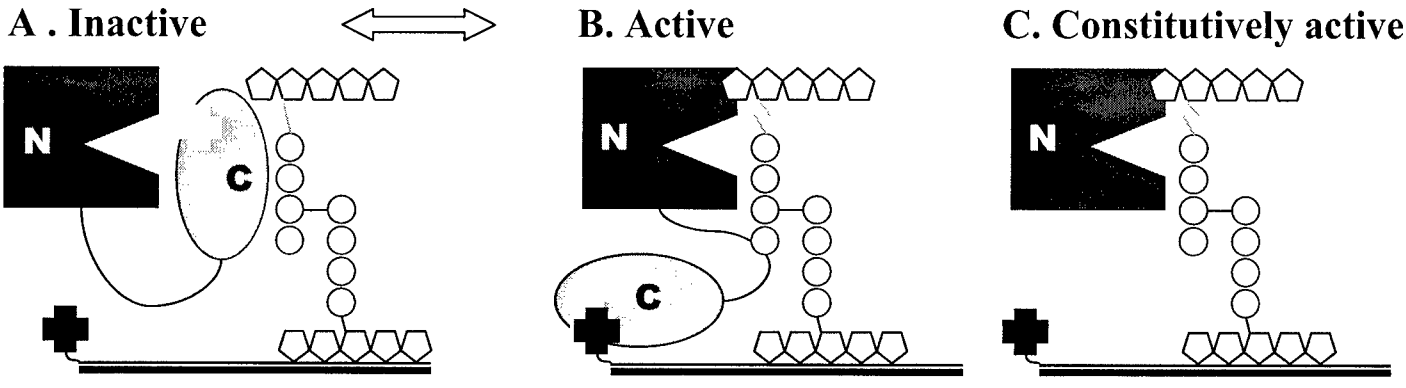
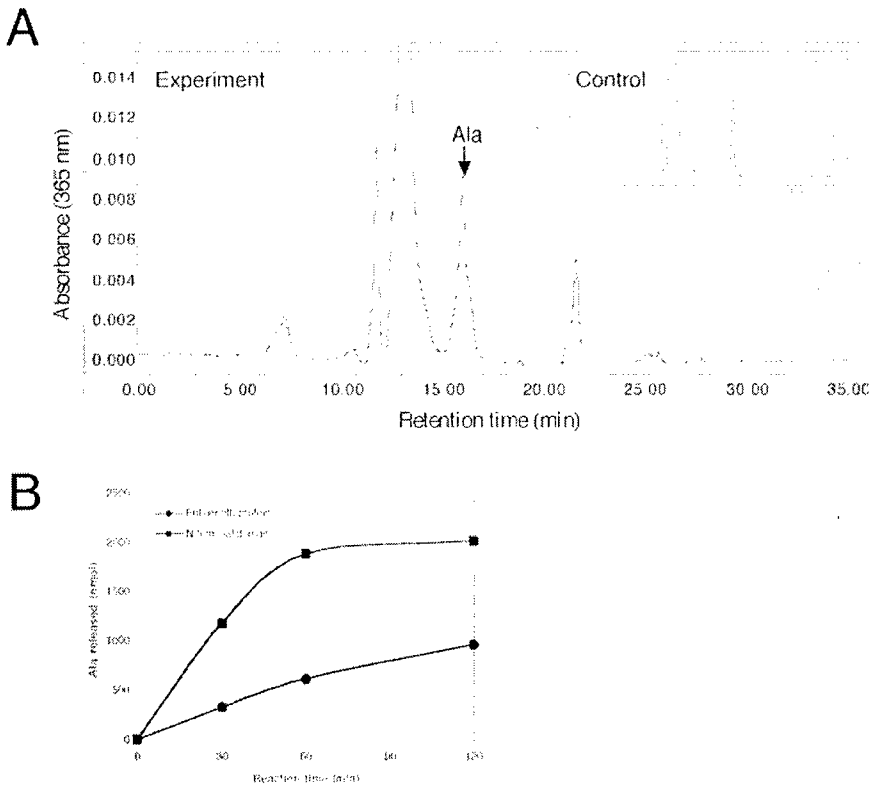



Figure 5

**Supplementary Figure** (**A**) HPLC analysis of DNP-labeled amino groups after digestion of *B. subtilis* peptidoglycan with the purified PlyL. A significant amount of DNP-alanine was observed from the released fraction. (**B**) Time course of the alanine released from peptidoglycan with the N-terminal domain and full-length protein of PlyL.




Supplementary Figure 1

# Structural and Functional studies of a *Bacillus anthracis* sensor domain

Gudrun R. Stranzl [1], Marcin Grynberg [2], Chandra La Clair [3], Dorinda Shoemaker [3], Robert Schwarzenbacher [1], Eugenio Santelli [1], Adam Godzik [1], Marta Perego [3], Robert C. Liddington [1,4]

[1]Infectious & Inflammatory Disease Research Center (IIDC), The Burnham Institute, La Jolla, CA 92037, USA

[2]Department of Genetics; Institute of Biochemistry and Biophysics; Pawinskiego 5A; 02106 Warsaw; Poland

[3]The Scripps Research Institute, Department of Molecular and Experimental Medicine, Divison of Cellular Biology, La Jolla, CA 92037, USA

[4]Correspondence: rliddington@burnham.org

## Summary

We have determined the crystal structures of two proteins BAS-1 and BAS-2, which are encoded by the genes pXO1-118 and pXO2-61 from the virulence plasmids from *Bacillus anthracis*. The gene pXO1-118 belongs to the pathogenicity island on the pXO1 plasmid. Both structures adopt a globin fold while their amino acid sequence reveals a conserved sensory motif, KIAxER, found in sensor histidine kinases from different *Bacilli* and in a so called trans-acting positive regulator from *Bacillus cereus*. In the BAS-1 structure, density corresponding to a putative ligand was observed. GC-MS identified this ligand as palmitic acid. Isothermal titration calorimetry (ITC) experiments showed reasonable binding of fatty acids to BAS-1 and BAS-2. These studies indicate that BAS-1 and BAS-2 function as fatty-acid binders and may play a role in fatty acid transport or regulation within the cell.

## Running Title

## Introduction

*Bacillus anthracis*, the spore forming Gram-positive *Bacillus* is the causative agent of anthrax, a potentially lethal infectious disease in humans and animals. Fully virulent forms of *Bacillus anthracis* carry two plasmids, pXO1 (182 kb) and pXO2 (96 kb). These plasmids encode major virulence factors such as those responsible for toxin production and capsule formation. The transcription and synthesis of anthrax toxin, capsule and certain chromosomal genes are regulated by atxA (anthrax toxin activator) [Uchida, 1993 #16], which is located on pXO1 and its homologue acpA (anthrax capsule activator) [Vietri, 1995 #52], which is located on pXO2. BAS-1 (Bacillus anthracis sensor domain also known as pXO1-118 gene) has its ORF 358 base pairs close to atxA (Figure 1-A) and is transcribed in a different direction than atxA. BAS-2 (also known as pXO2-61 gene) has its ORF 5658 base pairs from acpA (Figure 1-C) away, and it has been shown that pXO2-61 (BAS-2) is regulated by atxA [Bourgogne, 2003 #45]. BAS-1 and BAS-2 share an amino acid sequence identity of 61 %, however BAS-1 has a higher sequence identity of 81 % to a homologue protein (locus ZP_00236329) [Hoffmaster, 2004 #47] from *Bacillus cereus*, the spore forming Gram-positive *Bacillus* is found in soil and many other sources, it is an opportunistic pathogen that causes food poisoning manifested by diarrhoeal or emetic syndromes. Few genes from the pXO1 pathogenicity island [Okinaka, 1999 #57], pXO1-96 to pXO1-127, appear to be present in the various *B. cereus* group strains, too [Read, 2003 #46]. Interestingly is the fact that the

ORF of the homologue protein ZP_00236329 (gene BCE_G9241_pBC218_0049) from *Bacillus cereus*, which has a high homology to BAS-1, has its ORF 349 base pairs from atxA from *B. cereus* (Figure 1-B) [Hoffmaster, 2004 #47].

By searching GenBank at the NCBI with the amino acid sequence of BAS-1, using the program PSI-BLAST with default values, we identified the amino acid sequence of the N-terminal sensor domains of several bacterial sensor histidine kinases, however their sequence identity to BAS is 27 %. All aligned amino acid sequences do have a KIAxER motif in common. The sensor histidine kinases were found in *Bacillus anthracis*, *Bacillus cereus*, *Bacillus thuringensis*. The two-component regulatory system to which Sensor histidine kinases belong are composed of two domains: an N-terminal signal input domain that often possesses sub-domains with recognized signaling functions and a C-terminal autokinase domain [Stephenson, 2002 #17]. The sensing domain monitors changes in light, redox potential, and small ligands, which can further cause protein-protein interaction, DNA-binding and function to activate and/or repress transcription of specific genes [Stock, 2000 #49].

It has been reported that there is a two-component signal transduction system composed of a sensor kinase, DesK and a response regulator, DesR, which are responsible for cold induction of the *des* gene coding for the Δ5-lipid desaturase from *Bacillus subtilis*. In this case unsaturated fatty acids (UFAs), act as negative signalling molecules of *des* transcription. Further they report that the difference in potency among 16:1 Δ5 and other fatty acids tested strongly suggests that fatty acids with a double bond at the Δ5 position act as specific signals regulating the DesK-DesR signal transduction [Aguilar, 20011 #50].

Further, two-component regulatory proteins are involved in the initiation of sporulation in *Bacillus subtilis*. There, a signal activates the autophosphorylation of histidine kinases, KinA and KinB, which transfer the phosphoryl group to Spo0F, a single domain of the two-component response regulator. Phosphorylated Spo0F passes the phosphate to the final transcriptional regulator, Spo0A, through a phosphotransferase, Spo0B [Tzeng, 1997 #59].

The existence of antibiotic-resistant bacterial strains that arise either naturally or through deliberate engineering emphasizes the need for alternative therapeutic approaches. Vaccines are typically problematic for prophylactic treatment of large civilian groups, because of possible side effects. The two-component systems and phosphorelays have been recognized as targets for antimicrobial intervention [Stephenson, 2002 #58]. Therefore different approaches are necessary to get to know more about structure and function of proteins involved in the molecular mechanism of *Bacillus anthracis* virulence and pathogenicity.

In this study, we have determined the structures of both proteins BAS-1 and BAS-2 by x-ray crystallography. After successfully determining a fatty acid bound to BAS-1 by GC-MS, we show through isothermal calorimetry experiments that BAS-1 and BAS-2 bind saturated as well as unsaturated fatty acids. We also can exclude that atxA does not bind BAS-1 from a gel filtration experiment. However, in order to get to know more about the function of these new homologue proteins of the sensor domain of the sensor histidine kinases more experiments have to be done.

## Results and Discussion

### Structure of *Bacillus anthracis* BAS-1 protein

We crystallized the full-length *Bacillus anthracis* BAS-1 protein and determined its structure to 1.76 Å resolution. Selenium Singlewavelength-anomalous dispersion (SAD) phasing techniques were used to solve the structure. The asymmetric unit contains one molecule, which forms with another molecule an asymmetric crystallographic dimer. The model includes amino acid residues –2 to 150 (of 150 total residues), including 3 residues (Gly, Ser, His) from the N-terminal His-tag. The *B. anthracis* BAS-1 dimer comprises a single structural domain characterized by six helices which form a so-called globin fold. Helices 1 to 4 are 45° twisted against helices 5 to 6. At the C-terminus there is one turn α-helix (figure 2-B).

The putative active site, which is represented by the KIAxR domain is located on helix α4. Figure 5-C shows the KIAxR residues (K-69, I-70, A-71, R-74) which are involved in a complex hydrogen and salt bridge bonding network. The residue R-74 is buried into the cavity through salt bridges with D-33 (OD1/NH2 is 3.53 Å and OD2/NH1 is 2.77 Å) and the fatty acid ligand 11A, further through hydrogen bonding with HOH-95 and HOH-63. The HOH-63 has hydrogen bonding to residues R-74, E-73 and to the ligand 11A. This network stabilizes the R-74 in a hydrophobic environment, which would force the R-74 to point out onto the surface of the protein. Due to salt bridges and the hydrogen bonding of R-74, it stays buried into the cavity. In coordinating distance of the carboxyl oxygen of the fatty acid ligand 11A a small piece of electron density has been observed, which is very likely a potassium ion. Residue K-69 is engaged in salt bridges with E-38, E-73 and N-42 all bonds are within 3 Å distance. While the core of the globin fold is dominated primarily by hydrophobic residues, the two major surfaces of the globin fold show a distinct polarity in the abundance of charged residues. The dimerization interface shows both hydrophobic as well as charged interactions which involves residues mainly from helices α5 and α6 like Ala-82, Ile-85, -113, -124, -128,

Asn-89, -117, Lys-92, -93, -110, -135, -136, Leu-96, Phe-100, Gly-106, Cys-109, Glu-114, -121, Tyr-131, -132 (figure 3-A). There are two Cysteines namely A Cys-109 and B Cys-109, which are in a distance of 4.2 Å. No disulfide bond could be observed in the electron density maps. Each monomer has a salt bridge (A Asn-117 with A Lys-92) which is located within the dimer interface, however has no influence to form an intermolecular salt bridge. Figure 2-B shows the crystallographic BAS-1 dimer which is consistent with gel filtration experiments where a stable BAS-1 dimer was observed in solution (figure2-D).

The BAS-1 exhibit a cavity (figure 5-A) which is lined with mostly hydrophobic residues which are listed in table 2. The residue Phe-19 (figure 5-E) has two conformations and it looks like that it stabilizes the ligand in the cavity. The volume of the mostly hydrophobic cavity is about 123 $Å^3$ (calculated with program VOIDOO [Kleywegt, 1994 #39] ), its length is about 20 Å (distance between Arg-74NE and Phe-19CB) and its width is 8 Å (from Phe-84CD1 to Trp-23CH2). No water molecules are found in this region, apart from water 62, 63, 65 and water 95. Additional electron density close to the ligand is assumed a potassium ion according to B values. In figure 5-A the cavity is shown as a pink grid, with the ligand is shown in ball and sticks together with residues (His-30, His-32, Val-79, Glu-83) which are close to the entrance of the cavity.

While the core of the globin fold is dominated primarily by hydrophobic residues, the two major surfaces of the dimer show a distinct polarity in the abundance of charged residues. The surface of the BAS-1 dimer contains 18 of 30 glutamate residues, 8 of 12 aspartic acid residues, 6 of 8 arginine residues, and 24 of 38 lysines residues. Resulting a slightly negative net charge of the dimer. Figure 3-B shows the electrostatic potential surface maps, whereas charged residues are grouped in E1 to E4 zones for the different charged parts of the dimer. Zone E1 is a small positively charged part which is provoked by residues Lys-5, Arg-6, Tyr-7, and Arg-55. Zone E2 is a bigger negatively charged part which is provoked by residues Glu-17, Glu-38, Glu-57, Asp-62, Glu-64, Asp-65, Glu-73, Asp-83, Glu-90 and Asp-121. Zone E3 are the C-termini of each monomer which is positively charged which is provoked by residues Lys-78, Lys-133, Lys-135, Lys-136, Lys-146, Lys-147 and Lys-148. Zone E4 is a smaller positively charged part which is provoked by residues Lys-18, Lys-24, Lys-25, Arg-26, Lys-41, Arg-55 and Lys-93.

A DALI [Holm, 1993 #29] search showed closest structural homology to an dimeric oxygen sensor from *Bacillus Subtilis* (PDB 1OR4, [Zhang, 2003 #40]) (figure 4-A), a light-harvesting protein (PDB 1QGW, [Wilk, 1999 #41], and a dimeric bacterial hemoglobin from *Vitroscilla*

*sp.* (PDB 2VHB, [Tarricone, 1997 #42]. Based on its three-dimensional structure, BAS-1 belongs to the globin superfamily. These hits were within a RMSD of 3.0 Å, whereas their sequence identity to BAS-1 was around 13% and the length was about 150 residues (figure4-B). This means that through primary sequence homology searches a globin fold would have never been identified. From superposition with the bacterial hemoglobin from *Vitroscilla s.p* we could find a slight difference in the hydrophobic pocket. The pocket of BAS-1 has no haeme as co-factor and it is also slightly narrower than in the oxygen sensor of *Bacillus Subtilis* (figure 4-A). Due to the same fold with haeme binding proteins a reconstitution with haemin has been tried, however failed, because of the differences in the hydrophobic pocket compared with the oxygen sensor in *Bacillus Subtilis*. The helix 3 from BAS-1 is closer to helix 4 than helix F and E in the oxygen sensor of *B. Subtilis* structure (figure 4-A). From a superposition we see that Leu-92, Ile-83, Tyr-70, Phe69, Leu-96, His-123,Thr-95 from the oxygen sensor of *B. Subtilis* is a Ile-39, Lys-36, Lys-24, Trp-23, Gly43, Arg-74, Asn-42 in the BAS-1 structure, respectively. Residues Ile-39 and Asn-42 would clash into the haeme. This space problem may explain why a reconstitution with haemin failed. The His-123 from the oxygen sensor of *B. Subtilis* is is the residue which covalently binds the iron in the heme, through superposition with BAS-1 we identified the residue Arg-74 which is at the same position as the His-123 it seems that the ligand binding pockets pretty well conserved. It could be that BAS-1 and BAS-2 are a new sub-family within the globin superfamily due to similar features in their architectures, however quite different in their amino acid sequence.

**Description of the BAS-1 ligand.**

The BAS-1 structure contains additional electron density indicative of a non covalently bound ligand. Figure 5-E shows the undecanoic acid ligand fit into the remaining electron density, which has a long hydrophobic tail which fits perfectly in the hydrophobic cavity of BAS-1. His-tagged BAS-1 protein was mixed with the crude extract from *Bacillus cereus* and incubated for one hour. After this procedure BAS-1 was purified and crystallized. Still the same ligand could be observed. Gas chromatography mass spectrometry (GC-MS) analysis revealed a mass pattern indicative of a hexadecanoic acid molecule. The hydrophobic tail of the hexadecanoic acid is flexible and can be only partly modeled. Therefore, a shorter fatty acid has been used in the model. The chemical environment in the cavity is also compatible that the ligand being a fatty acid. Its hydrophobic tail contacts a hydrophobic residue (Phe-

19), which is seen in two conformations and its carboxylic group within coordinating distance of an argine (Arg-74). The question arises, where does the ligand comes from? We did not add any hexadecanoic acid to buffer solutions during the protein purification. We think that during cell disruption the BAS-1 protein binds the hexadecanoic acid, which is part of all gram-negative bacteria cell walls [Kaneda, 1967 #43]. We tried different buffer solutions and could still find the ligand, therefore we think that it maybe a physiological ligand.

**Structure of Bacillus anthracis BAS-2 protein**

We crystallized the full-length *Bacillus anthracis* BAS-2 protein and determined its structure to 1.49 Å resolution. The BAS-1 model was used for solving the BAS-2 structure with molecular replacement. The asymmetric unit contains two molecules, which form an asymmetric crystallographic dimer. The dimerization interface shows both hydrophobic as well as charged interactions which involves residues mainly from helices H5 and H6 like Ala-82, Ile-85, -93, -124, Asn-89, -117, Lys-92, -114, -135, Met-96, Thr-100, Leu-106, Gln-110, Val-113, -128, Tyr-109, -131, -132, Asp-121 (figure 3-A). Residue Lys-92 is engaged with Asn-117 to perform a salt bridge within a BAS-2 monomer. The model includes amino acid residues 5-136 (of 136 total residues), excluding 3 residues (Gly, Ser, His) from the N-terminal His-tag and 4 residues (Met-1, Glu-2, Glu-3, Ile-4) from the BAS-2 protein. These residues can not be observed in the electron density. The *B. anthracis* BAS-2 dimer (figure2-C) comprises a single structural domain characterized by six helices which form a so-called globin fold. There are two Cysteines namely A Cys-6 and A Cys-9, which form a disulfide bond within a distance of 3.2 Å, observed in the electron density maps. It looks like that the disulfide bond causes a distortion of the helix at the beginning of the amino acid sequence, which may explain, why this part is more flexible and no density could be observed for residues 1 to 4.

The putative active site, which is represented by the KIAxR domain is located on helix $\alpha$4. Figure 5-D shows the KIAxR residues (K-69, I-70, A-71, R-74) which are involved in a complex hydrogen and salt bridge bonding network. The residue R-74 is buried into the cavity through a salt bridges with D-33 and through hydrogen bonding with HOH-17 and HOH-23. The HOH-23 has hydrogen bonding to residues R-74 and E-73. This network stabilizes the R-74 in a hydrophobic environment, which would force the R-74 to point out onto the surface of the protein. Due to salt bridges and the hydrogen bonding of R-74, it stays buried into the cavity. A small piece of density has been observed, which is very likely a

iodide ion, which is too far away to coordinate the R-74 or waters. Residue K-69 is engaged in salt bridges with E-38, E-73 and N-42 all bonds are within 3 Å distance.

The BAS-2 exhibit a cavity (figure 5-B) which is lined with mostly hydrophobic residues which are listed in table2. The volume of the mostly hydrophobic cavity is about 123 $Å^3$ (calculated with program VOIDOO [Kleywegt, 1994 #39] ), its length is about 26 Å (distance between Arg-32NH1 and Ile-95CD1) and its width is 8 Å (from Trp-23 CH2 to Ile-70 CD1). No water and no ligand molecules were found in the cavity. In figure 5-B the cavity is shown as a pink grid and residues which are at the entrance are shown in ball and stick (Arg-30, Arg-32, Val-79, Glu-83).

A DALI [Holm, 1993 #29] search showed closest structural homology to the same hits as for BAS-1. An oxygen sensor in *Bacillus Subtilis* (PDB 1OR4, [Zhang, 2003 #40]), a light-harvesting protein (PDB 1QGW, [Wilk, 1999 #41], and a dimeric bacterial hemoglobin from *Vitroscilla sp.* (PDB 2VHB, [Tarricone, 1997 #42]. These hits were within a RMSD of 3.0 Å.

The electrostatic surface potential (figure 3-C) shows 4 large charged spots. One negatively charged we have called E1 is provoked by residues (Asp-33, Asp-76, Glu-37, Glu-38, Glu-73). One positively charged one is called E2, which is provoked by residues (Lys-24 and Lys-25 ), another positively charged one is called E3, which is provoked by residues (Arg-30, Lys-78 and Lys-135). There is a huge positively charged spot E4, which is provoked by residues ( Lys-5, Arg-10, Lys-13, Lys-55, Lys-92, Lys-114, Lys-115, Lys-117). Altogether we can observe 7 out of 14 glutamic acid residues, 4 out of 5 aspartic acid residues, 11 out of 18 lysine residues and 2 out of 6 argine residues on the surface. Resulting a slightly positive net charge of the monomer.

**Sequence and Structure homology of BAS-1 and BAS-2**

A BLAST search with the amino acid sequence of BAS-1 revealed a homologue from *Bacillus anthracis* BAS-2 with an e-value of $e^{-35}$ and a homologue from *Bacillus cereus* ZP_00236329 / gi:47565287 with an e-value of $e^{-49}$ and the N-terminal domains of several *Bacilli* sensor histidine kinases. The amino acid sequence identities between BAS-1 and BAS-2 and the homologue protein ZP_00236329 from *Bacillus cereus* are 62% and 81 %, respectively. The amino acid sequence identity between BAS-1 and the N-terminal domains of several *Bacilli* sensor histidine kinases are 27 %. The amino acid sequence identity between BAS-2 and the homologue protein ZP_00236329 from *Bacillus cereus* is 61 %. All

aligned amino acid sequences (figure 3-A) do have a KIAxER motif in common. In the BAS structures these residues belong to Lys-69, Ile-70, Ala-71, Glu-73 and Arg-74 (figure 6). Residue Pro-34 is conserved in both BAS structures as well as Pro-104, which are both responsible for the turn in the secondary structure. Another interesting difference is residue 91 in BAS-1 a glycine and in BAS-2 an alanine, these residues are located on helix α5 and are located in the hydrophobic cavity of both proteins. Therefore is could be possible that BAS-1 and BAS-2 have a different affinity to fatty acids. One significant structural and sequence difference between BAS-1 and BAS-2 is the C-terminal extension. The BAS-2 structure is 11 amino acids shorter and therefore the C-termini do not intersect. The BAS-1 and BAS-2 structures have a RMSD of 0.753. RMS calculations have been carried out using the program LSQMAN [Kleywegt, 2001 #20]. Figure 2-A shows a $c_\alpha$ superposition of BAS-1 and BAS-2, which demonstrates how similar the $c_\alpha$ is, however the Electrostatic Potential surface of the BAS-1 and the BAS-2 dimer looks different. Especially the major positively charged spots E3 from BAS-1 and E4 from BAS-2 (figure 3) are different placed on the their surfaces and may explain their different behavior during protein purification and crystallization.

To demonstrate how the electrostatic surface potentials of the sequence aligned proteins from figure 3-A look alike, residues of all the candidates have been mapped onto the BAS-1 model. BAS-1 has been modified at the N- and C-terminus for better comparison. As orientation helices α1 through α4 are exposed. Apparently, the electrostatic surface potentials of BAS-1 and its homologue protein ZP_00236329 from *Bacillus cereus* are very similar. BAS-2 has some same patches with BAS-1 in common, but the the N-terminal domains of several *Bacilli* sensor histidine kinases are much more negatively charged as the BAS protein group (figure 3-D). However within the histidine kinases the are highly conserved. The structures Ba.c.3 and Ba.a. have the same amino acid sequence, however Ba.c.1, has two different amino acids V45A and N120A, Ba.c.2 has two different amino acids E119D and N125A, Ba.th. has eight different amino acids like V13I, D37E, R42K, F55L, I60T, E63D, N120A, Q121K. In figure 5-C and D the putative active sites of BAS-1 and BAS-2 show that the waters in these sites are very conserved, however the fact that we observe a potassium ion in the BAS-1 structure and an iodide ion in the BAS-2 structure reveals the question that maybe because of the iodides charge and ion size no fatty acid binding in the crystal could be observed. However the ITC experiments binding to fatty acids could be observed, because there is no presence of iodide.

**ITC binding studies with BAS-1 and BAS-2**

We have measured affinity of BAS-1 and BAS-2 to saturated (myristic, palmitic), saturated branched (12-methyltetradecanoic, 13-methyltetradecanoic), and non-saturated (palmitoleic) acids using ITC (Fig. 7). All binding data were best described by a one to one stoichiometry model. We did not detect any selectivity among tested acids. All bound BAS-1 and BAS-2 were exothermically (Fig. 7), with very similar affinities (Table 3).

**AtxA and other stuff**

In order to test BAS-1 binds atxA both proteins have been expressed, mixed and applied to a size exclusion column. No complex of atxA and BAS-1 could be observed, both proteins eluted separately. In addition to that we performed an electrophoretic mobility shift assay (EMSA) with BAS-1, atxA and the pagA promoter DNA in the presence and absence of $Na_2CO_3$. We tried different concentrations of BAS-1, atxA and the pagA promoter DNA, and also different combinations, like atxA without BAS-1 and could also see no binding. Our conclusion is that BAS-1 does not bind atxA.

BAS-1 and BAS-2 has been tested for hydrolase and oxidoreductase activity and it has been tested negativ.

**Gene deletion analysis**

In an attempt to define a physiological function for the product of ORF118, a 34F2 derivative strain carrying a spectinomycin resistance cassette in place of ORF118 was constructed. The strain, named 34F2_118 did not show any growth or sporulation defect when compared to the parental strain 34F2 (Fig.8-A and data not shown). Both strains were transformed with the pTCVlac construct carrying the *atxA* promoter and the transcription of this gene was analyzed by means of _-galactosidase assays. As shown in Fig. 8-A, no difference in transcription was observed between the parental strain and 34F2_118 indicating that ORF118 does not affect AtxA production. As a consequence, the product of ORF118 did not affect the transcription of the *pagA* gene encoding the protective antigen (our unpublished results).

**Yeast Two-hybrid system analysis**

The yeast two-hybrid system (Clontech) was used to test whether ORF118 could interact with AtxA. Both genes were singly cloned in the bait plasmid pGBT9 and in the prey plasmid pGAD424. When the interaction assays were carried out in the yeast strain AH109, we detected interaction in the control strain carrying ORF118 on both pGBT9 and pGAD424 plasmids but we did not detect any interaction with AtxA either as a bait or as a prey. These results confirm that ORF118 can dimerize but do not support the hypothesis that it may interact with AtxA.

**Gene transcription analysis**

The transcription profile of the ORF118 and ORF61 promoters were determined by means of _-galactosidase analysis carried out on the promoter-*lacZ* fusion constructs. The pTCVlac plasmid derivatives carrying either the ORF118 or the ORF61 promoter were transformed in the Sterne strain 34F2 or in its derivative carrying a deletion of the *atxA* gene (34F2_atxA). The results of this analysis are shown in Figure 8-B. The transcription from both promoters was induced in late exponential phase and it increased during the early hours of stationary phase. The absence of AtxA prevented this induction from the ORF61 promoter but not from the ORF118 promoter. A similar pattern of transcription was observed when the cells were grown in Schaeffer's sporulation medium which induces sporulation of *B. anthracis* cells at a faster rate than the LB medium. Thus while transcription of the ORF118 gene is independent of AtxA, the transcription of ORF61 depends on this virulence factor as previously indicated by microarray study [Bourgogne, 2003 #45].

## Experimental Procedures

### Cloning, expression and purification

The plasmid pXO1 from *Bacillus anthracis*, sterne strain (provided by Philip Hanna) served as a template for cloning the hypothetical protein BAS-1. The forward and reverse primers for pXO1-118 (5' – GAGTGGACATATGGAAGCAACAAAACG - 3' , 5' – CTATAGGAT CCAAAAATTTCAAGGTG - 3') were used for amplification and the ORF-encoding BAS-1 was subsequently cloned with restrictions sites NdeI/BamHI into a pET-28a vector (Novagen) and transformed into BL21(DE3) cells, and plated on a selective medium containing kanamycin. Colonies were grown at 37 °C and 500ml of LB media ( Bacto™ Yeast Extract, Bacto™ Tryptone purchased by Difco and NaCl, pH adjusted to 7.5) were inoculated. Cultures were grown at 32 °C overnight to an optical density (600nm) of about 0.6 – 0.7 and protein expression induced by the addition of Isopropyl-β-thiogalactopyranoside (IPTG) to 0.1 mM. Shaking proceeded for a further 4 h at 220 rpm at 32 °C. BAS-1 was expressed in E. coli BL21 (DE3) for selenomethionine (SeMet) incorporation [Harrison, 1994 #56]. Cells were harvested by centrifuging the cells at 6000 rpm. The supernatant was discharged and the cell pellets resolved in lysis buffer (20 mM Tris pH 7.4, 0.5 M NaCl, 5 mM Imidazole, 1% Triton), sonicated and centrifuged at 16000 rpm for 10 min. the supernatant was applied to a nickel column (Amersham, Pharmacia). BAS proteins have been eluted with 250 mM Imidazole, 0.5 M NaCl, 20 mM TRIS HCl pH 7.4 and subsequently dialyzed against 1 M NaCl, 20mM TRIS HCl pH 7.4 over night. After cleavage of the His-tag with thrombin the solution was concentrated and further purified using a Superdex75 gel filtration column connected to an Äkta-FPLC (Amersham, Pharmacia). Protein has been concentrated (AMICON) and dialyzed into 20mM TRIS HCl buffer pH 7.4, 1M NaCl, 50µM KCl, 5mM DTT and flash frozen in liquid nitrogen for long term storage at –80 °C. The protein runs on SDS-PAGE gel as expected with a molecular weight of 18.510 kDa, also confirmed by MALDI-TOF. On a sizing column (Superdex75), the estimated molecular weight is ~32 kDa, suggestive of a dimer in solution. The synthetic pXO2-61 gene was obtained from GenScript Co, NJ, USA and subcloned, expressed and purified as previous described. BAS-2 has been dialyzed into 20mM TRIS HCl buffer pH 7.4, 500 mM NaCl, 50µM KCl, 5mM DTT and concentrated to 16mg/ml and flash-frozen in liquid nitrogen for storage at –80 °C. The protein runs on SDS-PAGE as expected with molecular weight of 16.4 kDa, confirmed by MALDI-TOF. On a sizing column (Superdex75), the estimated M.W. is ~32 kDa, suggestive of a dimer in solution.

**Cloning, expression and purification of atxA**

The ORF-encoding atxA sequence was cloned in a pET-15b vector (Novagen) and transformed into BL21(DE3) cells. The protein runs on SDS-PAGE gel as expected with a molecular weight of 55.561 kDa, which was confirmed by MALDI-TOF. On a sizing column, the estimated M.W. is ~115 kDa, suggestive of a dimer in solution.

**Crystallization, data collection, and structure solution**

Purified native and SeMet-substituted BAS-1 was crystallized by vapor diffusion at room temperature using sitting and hanging drops of 3 µl of precipitant solution (40% (v/v) PEG-300, 100mM Tris-HCl pH 5.4, 5% (w/v) PEG-1000) and 3µl of protein solution (14mg/ml) yielded crystals within 3 days and belonged to space group $P3_221$ with unit cell parameters a=b= 89.86 Å c= 35.25 Å and $\alpha=\beta= 90°$ $\gamma= 120°$ and a Matthews coefficient 2.2 (44.2% solvent). Crystals were grown rod shaped with dimensions 0.1 mm x 0.05 mm x 0.05 mm. Data collection statistics are summarized in Table 1. Crystals were already grown in cryo protectant solution and flash cooled in liquid nitrogen. One native and one Se-SAD (single anomalous dispersion) datasets were collected at SLAC (Stanford Linear Accelarator Center) SSRL beamline 9-2 and BNL (Brookhaven National Lab) NSLS beamline X26C, respectively. Diffraction images were processed and scaled with HKL [Otwinowski, 1997 #23]. The program SOLVE [Terwillinger, 1999 #24] was used to locate four Se positions in the BAS-1 structure, which were used to obtain initial phases (figure of merit [FOM]=0.32). Following phase improvement using the program RESOLVE [Terwillinger, 2001 #25] ([FOM]=0.60) and automatic model building with RESOLVE resulted in model fragments of 9 chains with 83 residues at a model completeness of 77%. Further model building was performed manually in O [Kleywegt, 2001 #20].

Purified BAS-2 was crystallized by microbatch under paraffin oil. One good diffracting crystal could be obtained from 1M NaJ, 20% (v/v) PEG3350 solution, which grow within two days. Crystals grow in Tetragonal and Orthorhombic crystal systems. The Tetragonal crystal form was crystallized by vapour diffusion method at room temperature using 0.1 M TRIS HCl pH 8.5, 30% (v/v) PEG4000, 2M $Li_2SO_4$. Crystals were grown rod shaped with dimensions 0.7 mm × 0.3 mm × 0.3 mm. The crystal used for structure solution belonged to space group $P2_12_12_1$ with unit cell parameters a=44 Å, b=62, Å, c=124 Å, $\alpha = \beta = \gamma = 90°$. Data were processed with DENZO and SCALEPACK [Otwinowski, 1997 #23]. Data collection statistics

are summarized in Table 1. High and low resolution data sets were collected at SLAC (Stanford Linear Accelarator Center) up to 1.49 Å. High and low resolution data sets were combined and the structure was solved with molecular replacement using the BAS-1 structure as a search model. Model building and refinement were carried out in O [Kleywegt, 2001 #20] and REFMAC5 [Murshudov, 1997 #40]. The asymmetric unit contains two molecules and a Matthews coefficient of 2.9 (56.7% solvent).

**Refinement.**

The initial model for BAS-1 refinement contained one chain of residues 2 to 147. One round of rigid-body refinement was carried out against data to 1.76 Å resolution (native), followed by a simulated-annealing step against a maximum-likelihood target with the programs CNS [Brünger, 1998 #21] and REFMAC5 [Murshudov, 1997 #40]. In the active site an electron density of a ligand was located. The initial model for BAS-2 contained two chains of residues 5-136. The fatty acid model undecanoic acid was constructed in the program PRODRG [Schuettelkopf, 2004 #35] and manually positioned with the program O into the remaining density of the active site of BAS-1. Each cycle of refinement was followed by manual model rebuilding with the program O. During the final refinement stages of BAS-1 structure, alternate conformations were modeled and refined. The final BAS-1 model has an $R_{work}$= 18.50 % and $R_{free}$=24.10 % for data between 76.7 and 1.76 Å resolution. Average B factors (in $Å^2$) were 25.00 for main chain and 32.65 for side chain atoms, 34.62 for solvent atoms, 37.42 for ligands and ions. The final BAS-2 model has an $R_{work}$= 17.70 % and $R_{free}$=20.90 % for data between 62.02 and 1.49 Å resolution. Superposition of model BAS-2 chain A with chain B over 132 atoms reveals a RMSD of 0.434 Å. Average B factors (in $Å^2$) were 18.90 for main chain and 23.95 for side chain atoms, 35.90 for solvent atoms and 39.66 for ions.

**Quality and deposition of the Crystallographic Models.**

The BAS-1 structure has been evaluated with the program PROCHECK [Laskowski, 1993 #32] 97.3 % of the residues are in the most favoured region of the Ramachandran plot and 2.7 % are in allowed regions. The residues Ser15 and Phe19 are in two alternate conformations in chain A. The BAS-2 structure has been evaluated with the program PROCHECK [Laskowski, 1993 #32] 98.4 % of the residues are in the most favoured region of the Ramachandran plot and 1.6 % are in allowed regions. The residues Ser21, Ser48, Asp76, Lys114 are in two alternate conformations in chain A. However in chain B could be only Ser48, Ser67 and

Tyr132 in two alternate conformations observed. The coordinates of the BAS-1 and the BAS-2 structures have been deposited in the Protein Data Bank access codes 1Y87 and 1YKU, respectively.

**Gas Chromatography-Mass Spectroscopy (GC-MS) of the ligand.**

To 100 µl of a BAS-1 protein solution (10mg/ml) 200 µl of chloroform has been added. This two phase system has been treated with ultrasonic for 10 minutes. Then the two phase system has been incubated at 70 °C for one hour and centrifuged. The organic phase was separated with a syringe. In order to verify the carboxylic group of the fatty acid it was derivatived with 20 µl of BSTFA (bis trimethylsylil trifluoroacetamide) and 20µl pyridine and incubated for 1.5.hrs. at 65 °C. Samples were evaporated under a stream of $N_2$ gas to dryness and reconstituted with 100µl methylene chloride and analyzed with GC-MS (Scripps center for Mass Spectrometry, CA, USA).

**Isothermal Titration Calorimetry**

Isothermal titration calorimetry (ITC) was performed on a VP-ITC calorimeter from Microcal (Northampton, MA). Eight microliters of fatty acid (myristic acid n-C14:0 and palmitic acid n-C16:0 purchased by Sigma-Aldrich Co, MO, St.Luis, USA, 12-methyltetradecanoic acid anteiso-C15:0 and 13-methyltetradecanoic acid iso-C15:0 purchased by Indofine chemical Co, NJ, USA, palmitoleic acid purchased by Fluka) solution (1.6-2.6mM) were injected into the cell containing 100 µM protein (BAS-1 or BAS-2). In each experiment 37 injections were made. All titrations were performed at 23°C. ITC samples also contained 20mM Tris pH 7.4 and either 500 mM (BAS-2) or 1000 mM (BAS-1) NaCl. Experimental data were analyzed using Microcal Origin software provided by the ITC manufacturer (Microcal, Northampton, MA).

**Bacterial strains and growth conditions**

Functional analysis was carried out in the *B. anthracis* Sterne strain 34F2. Cells were grown in LB medium or Schaeffer's sporulation medium [Schaeffer, 1965 #65]. Transformation by electroporation was carried out according to Koehler et al [Koehler, 1994 #62]. Unmethylated DNA was obtained by passing plasmid constructs into the *dam*⁻ strain SCS110 (Stratagene). *E. coli* DH5_ was used for plasmid construction and propagation. Antibiotics were used at the

following concentrations in *E. coli* or *B. anthracis*, respectively: kanamycin 30_g/ml and 7.5_g/ml; chloramphenicol 10_g/ml and 7.5_g/ml; spectinomycin 100_g/ml and 200_g/ml. Ampicilin was used at 100_g/ml for *E. coli* only. The _-galactosidase assays were carried out as previously described [Brunsing, 2005 #60; Ferrari, 1985 #61; Miller, 1972 #63]. Protein interaction analysis was carried out essentially as described by the Clontech Two-hybrid system manual.

**Plasmid constructions**

The plasmid for *E. coli* over expression and purification of ORF118 was obtained by cloning the PCR amplified coding sequence using oligonucleotides BaORF1185'Nde (5'-GAGTGGACATATGGAAGCAACAAAACG-3') and BaORF1183'Bam (5'-CTATAGGATCCAAAAATTTCAAGGTG-3') into plasmid pET28a (Stratagene) digested with *NdeI* and *BamHI*. Transcriptional fusions to the *E. coli lacZ* gene were constructed in the replicative vector pTCVlac [Poyart, 1997 #64]. The promoter region of ORF118 was amplified using oligonucleotides p1185'Eco2 (5'-CTATTGAATTCATTGATAAAGTGTAG-3') and p1183'Bam2 (5'-TAAATGGATCCTGGCTTTCTTTTAGG-3'). The promoter region of ORF61 was PCR amplified using oligonucleotides pX0261-5'Eco (5'-GTTTAGAATTCTGAAATATTTTAATAGAC-3') and pX0261-3'Bam (5'-CTTTTGGATCCAATCAGATATAAATTTTTC-3'). The fragments were digested with *EcoRI* and *BamHI* and cloned in pTCVlac similarly digested. Plasmid pORICm was used for the construction of the ORF118 deletion strain [Brunsing, 2005 #60]. A 720bp fragment downstream ORF118 was PCR amplified using oligonucleotides Delta 118Kpn (5'-AATAAGGTACCTTAAGTAATAAATAC-3') and Delta 118Bam (5'-ATATTGGATCCTAAAAAAGAAATATAAC-3') and cloned in pORICm at the *KpnI* and *BamHI* sites. A 860bp fragment upstream of ORF118 was also PCR amplified using oligonucleotides Delta118Sal (5'-CATAAGTCGACTCCTTAATTCCTTAAAAATC-3') and Delta118Pst (5'-TATTACTGCAGGGAAACGGCCAATAATC-3') and cloned in the resulting plasmid at the *SalI-PstI* sites. Finally, a blunt-ended spectinomycin cassette was cloned at the *HincII* site positioned in between the two cloned fragments in the vector multiple cloning site. The resulting plasmid was transformed into strain 34F2 and used to generate a deletion-spectinomycin replacement of ORF118 essentially as described [Brunsing,

2005 #60]. The promoter region of *atxA* was PCR amplified using oligonucleotides Delta 118Eco2 (5'-TTCCAGAATTCCACTCCTTAATTCC-3') and AtxA3'Bam (5'-CAAATGGATCCAGGGCATTTATATTATC-3'); the fragment was digested with *EcoRI* and *EcoRV* (the latter is naturally present in the *atxA* gene) and the 360bp fragment was cloned in pTCVlac digested with *EcoRI* and *SmaI*.

Plasmid pORICm was also used for the construction of the *atxA* deletion strain. The *atxA* coding region and upstream sequences were PCR amplified using oligonucleotide Ba118delta (5'-TTAATGAATTCTCGCATATACATTGTGAATAC-3') and AtxA3'Bam (5'-CAAATGGATCCAGGGCATTTATATTATC-3') and cloned in the *EcoRI-BamHI* sites of pORICm. The resulting plasmid was digested with *BclI* and *EcoRV* and the 670bp excised fragment was replaced by the spectinomycin cassette as a *BamHI-HincII* fragment. The resulting plasmid was used to transform strain 34F2 and generate a deletion-replacement of the *atxA* gene essentially as described [Brunsing, 2005 #60]. The gene encoding ORF118 was cloned in the two hybrid system vector pGBT9 and pGAD424 (Clontech) as an *EcoRI-BamHI* fragment obtained by PCR amplification using oligonucleotides THS1185'Eco (5'-AATTAGAATTCGGAGGAATGGAAGCAACAAAACGATAC-3') and BaORF1183'Bam described above. The *atxA* gene was cloned in the pGBT9 and PGAD424 plasmids using oligonucleotides AtxA5'EcoRI (5'-TTATAGAATTCCTAACACCGATATCCATA-3') and AtxA3'Bam (5'-CAAATGGATCCAGGGCATTTATATTATC-3'). An EcoRI linker with the sequence (5'-GAATTCTTGCCGGGACCTCTTCCGGGTCCGGAACTTCCTGGACCGGAGGGAATTC-3') was then inserted in the EcoRI site to provide flexibility to the fusion protein. All PCR reactions were carried out on the full genome of strain 34F2 extracted using the UltraClean Microbial DNA Isolation Kit (Mo Bio, Solana Beach, California) or on purified pXO2 plasmid DNA ( generously provided by Philip Hanna).

**Acknowledgments**

We would like to thank Annie Heroux for measuring crystals taking advantage of the FedEx crystallography access from the National Synchrotron Light Source, Brookhaven National Laboratory, which is supported by the U.S. Department of Energy, Division of Materials Sciences and Division of Chemical Sciences, under Contract No. DE-AC02-98CH10886. Portions of this research were carried out at the Stanford Synchrotron Radiation Laboratory, a national user facility operated by Stanford University on behalf of the U.S. Department of Energy, Office of Basic Energy Sciences. We would like to thank the staff of the SSRL

Structural Molecular Biology Program, which is supported by the Department of Energy, Office of Biological and Environmental Research, and by the National Institutes of Health, National Center for Research Resources, Biomedical Technology Program, and the National Institute of General Medical Sciences. G.R. Stranzl received an Erwin Schrödinger fellowship from the Austrian Science Fund. Work at The Scripps Research Institute was supported in part by grant GM055594 from the National Institute of General Medical Sciences and AI055860 from the National Institute of Allergies and Infectious Diseases, National Institutes of Health. The Stein Beneficial Trust supported in part oligonucleotide synthesis and DNA sequence.

## References

## Figure Legends

**Figure 1**. Predicted pXO1, pBC218 and pXO2 plasmids ORFs and physical map.

**A** shows a part of the pXO1 plasmid of *Bacillus anthracis* and the directions of the arrows indicates the direction of transcription in each ORF relative to all the other ORF. BAS-1 which is encoded by pXO1-118 is shown as well as *atxA* (anthrax toxin activator), *cya* ( edema factor gene), *pagA* (protective antigen gene).

**B** shows a part of the pBC218 plasmid from *Bacillus cereus* strain G9241 [Hoffmaster, 2004 #47] and the directions of the arrows indicates the direction of transcription in each ORF relative to all the other ORF. The homologue protein (locus ZP_00236329) which is encoded by 0049 is shown as well as *atxA* (anthrax toxin activator).

**C** shows a part of the pXO2 plamid of *Bacillus anthracis* and the directions of the arrows indicates the direction of transcription in each ORF relative to all the other ORF. BAS-2 which is encoded by pXO2-61 is shown as well as *acpA* (gene encoding a positive trans activator of capsule synthesis), *capB* (capsule biosynthesis operon B) [Drysdale, 2004 #53].

**Figure 2**. Stereo view, Ribbon representation and Size exclusion runs of BAS-1 and BAS-2.

**A** The BAS-1 Cα backbone is shown in black and the BAS-2 is shown in red. N- and C-termini are labeld and every tenth Cα is numbered.

**B** The BAS-1 dimer in ribbon representation coloured from the N-terminus (blue) to the C-terminus (red). The helices H1-H6 are indicated.

**C** The BAS-2 monomer in ribbon representation coloured from the N-terminus (blue) to the C-terminus (red). The helices H1-H6 are indicated. Figure A and B are produced with PYMOL [DeLano, 2002 #9].

**D** BAS-1 and BAS-2 have been detected at 280nm. BAS-1 is shown in red and elutes at 10.99 ml and BAS-2 is shown in black and elutes at 10.98 ml. In green a molecular weight standard is shown, the molecular weight of the standards are indicated above the arrows.

**Figure 3**. Sequence alignment of BAS-1 with conserved amino acid sequences and Electrostatic Potential surface of the BAS-1 and the BAS-2 dimer.

**A** By searching GenBank at the NCBI with the amino acid sequence of BAS-1, using the program PSI-BLAST with default values, we identified a homologue from *Bacillus anthracis* BAS-2 with an e-value of $e^{-35}$ and a homologue from *Bacillus cereus* ZP_00236329 / gi:47565287 with an e-value of $e^{-49}$. The KIAxER domain, which is highlighted in the above alignment with green, was also found in the N-terminal domain of one *Bacillus anthracis* sensor histidine kinase (Ba.a.: NP_844676 / gi:30262299), three sensor histidine kinases from *Bacillus cereus* (Ba.c.1: NP_978635 / gi42781388, Ba.c.2: ZP_00236689 / gi47565649, Ba.c.3: YP_083662 / gi52143167), and one sensor histidine kinase from *Bacillus thuringensis* (Ba.th.: I_40575 / gi:2127280). The alignment was carried out using CLUSTAL W [Higgins D., 1994 #38]. The secondary structure determined for BAS-1 and BAS-2 is shown above the alignment. The numbering of the α helices is the same as in figure 1A and 1B. Residues which are highlighted yellow are found in the hydrophobic cavity. Residues which are highlighted red are within the dimer interface. Cysteins are highlighted cyan, only the two cystenines of BAS-1 and the homologue from *Bacillus cereus*, which are within the dimer interface are highlighted magenta.

**B** The surface of the BAS-1 dimer is colored by its electrostatic surface potential at ∀12 KBT/e for positive (blue) or negative (red) charge potential. Zones E1 to E4 are shown,

whereas residues 1 to 5 are not shown for better comparision to the electrostatic surface of BAS-2.

**C** The surface of the BAS-2 dimer is colored by its electrostatic surface potential at ∀12 KBT/e for positive (blue) or negative (red) charge potential. Zones E1 to E4 are shown. Both dimers are oriented in the same way and just one side of the dimers has to be shown, because of the symmetry. This figure was prepared by SPOCK [http://mackerel.tamu.edu/spock/, #44].

**D** The BAS-1 model has been used to map the amino acid sequences from *Bacillus cereus* (ZP_00236329 / gi:47565287), the N-terminal domain of one *Bacillus anthracis* sensor histidine kinase (Ba.c.: NP_844676 / gi:30262299), three sensor histidine kinases from *Bacillus cereus* (Ba.c.1: NP_978635 / gi42781388, Ba.c.2: ZP_00236689 / gi47565649, Ba.c.3: YP_083662 / gi52143167), and one sensor histidine kinase from *Bacillus thuringensis* (Ba.th.: I_40575 / gi:2127280) onto it, further the surface of BAS-2 is shown. The surfaces are colored by its electrostatic surface potential at ±12 KBT/e for positive (blue) or negative (red) charge potential. This figure was prepared by SPOCK.

**Figure 4**. Side by side view of the BAS-1 structure and the oxygen sensor from *Bacillus subtilis*.

The BAS-1 structure has been superimposed with the oxygen sensor structure from *Bacillus subtilis*. Both are shown in ribbon presentation and helices has been colored in the same way. Helices 4 and 5 show the space difference in both structures very well. In the BAS-1 structure the undecanoic acid is shown and in the oxygen sensor from *Bacillus subtilis* the haemin is shown.

**Figure 5**. Cavity grid presentation of BAS-1 and BAS-2 and Stereo view of the remaining electron density in the cavity of BAS-1.

**A** shows a ribbon representation of BAS-1 and a pink grid demonstrates the calculated hydrophobic cavity. Residues (His-33, His-35, Val-82, Glu-86) which are at the entrance are shown.

**B** shows a ribbon representation of BAS-2 and a pink grid demonstrates the calculated hydrophobic cavity. Residues (Arg-30, Arg-32, Val-79, Glu-83) which are at the entrance are shown.

**C** Electron density map of the $2F_{obs}$ - $F_{calc}$ electron density map at sigma level 1 is shown. In the remaining electron density map an undecanoic acid has been modeled. Close to the carboxylic group of the ligand two sphere shaped electron density can be observed which is probably a $Na^+$ and a $Cl^-$ ion. As red spheres are W1and W2 presented, as a pink sphere a chloride ion and a blue sphere indicates a sodium ion.

**D** putative active site of BAS-1.

Residues are shown in ball and sticks, the KIAxR domain residues are shown as well residues His-30, Asp-33, Tyr-35, Glu-38, Asn-42, Lys-69, Ile-70, Ala-71, Glu-73, Arg-74, Asp-83, Phe-84, Asn-87, waters 62, 63, 65,95, potassium ion and fatty acid ligand 11A, which are involved in hydrogen bonding and salt bridges.

**Figure 6**. superposition of the KIAXER motif.

The green colored worm represents the BAS-1 structure and the yellow colored worm represents the BAS-2 structure. Especially the residues of the motif are shown as Lys-70, Ile-71, Ala-72, Glu-74 and Arg-75.

**Figure 7**

**A** Representative ITC titration for binding of palmitoleic acid to BAS-1. A 2.3 mM plamitoleic acid solution was titrated into 100μM BAS-1. Binding mode to BAS-1 of other tested fatty acids (data not shown) was similar to that of palmitoleic acid.

**B** Representative ITC titration for binding of 13-methyltetradecanoic acid to BAS-2. A 1.6 mM 13-methyltetradecanoic acid solution was titrated into 100μM BAS-2. Binding mode of other tested fatty acids to BAS-2 (data not shown) was similar to that of 13-methyltetradecanoic acid. Experimental conditions were as described in Experimental Procedures.

**Table 3.** Thermodynamic parameters obtained from ITC titrations of BAS-1 and BAS-2 with selected fatty acids.

**Figure 8**

**A** Transcription analysis of the *atxA* promoter in the ORF118 deletion strain. _-galactosidase assays were carried out on *B. anthracis* cultures grown in LB supplemented with kanamycin at 7.5_g/ml. Open symbols: growth curves; closed symbols: Miller Units

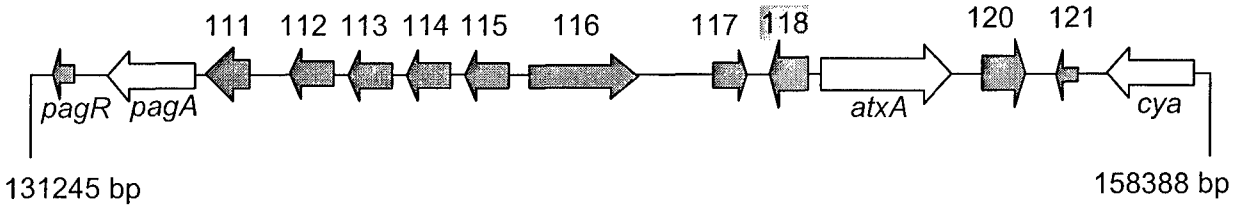
Strains and symbols: 34F2/pTCVlac-atxA: - ○ -; 34F2_118/pTCVlac-atxA: - ▽ -.

**B** Transcription analysis of the ORF118 and ORF62 promoters in the *atxA* deletion strain. _-galactosidase assays were carried out on *B. anthracis* cultures grown in LB medium containing kanamycin at 7.5_g/ml. Open symbols: growth curves; closed symbols: Miller Units. Strains and symbols: 34F2/pTCVlac-118: - ▽ -; 34F2_atxA/pTCVlac-118: - ◇-; 34F2/pTCVlac-62: - ○ -; 34F2_atxA/pTCVlac62: - △ -.

Stranzl_figure1

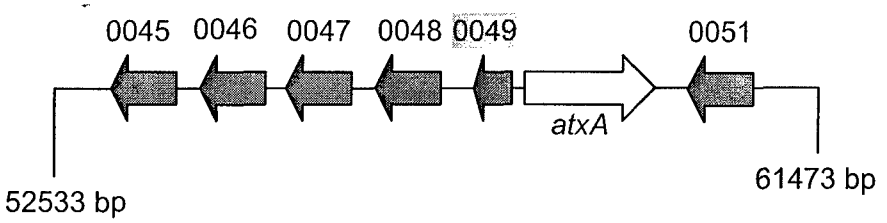
A

**pXO1**



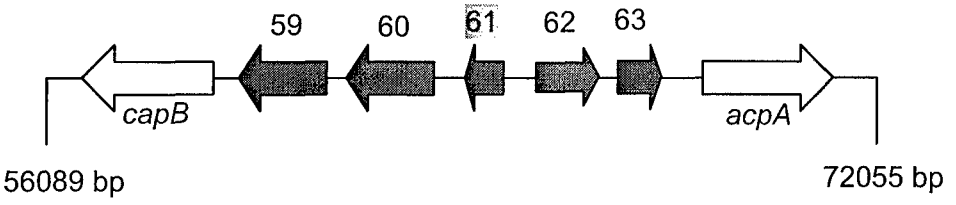

B

**pBC218**




**C**

**pXO2**

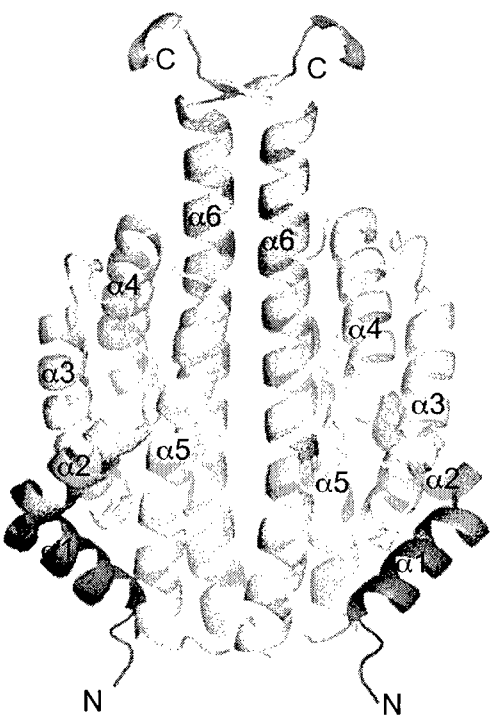



Stranzl_figure2

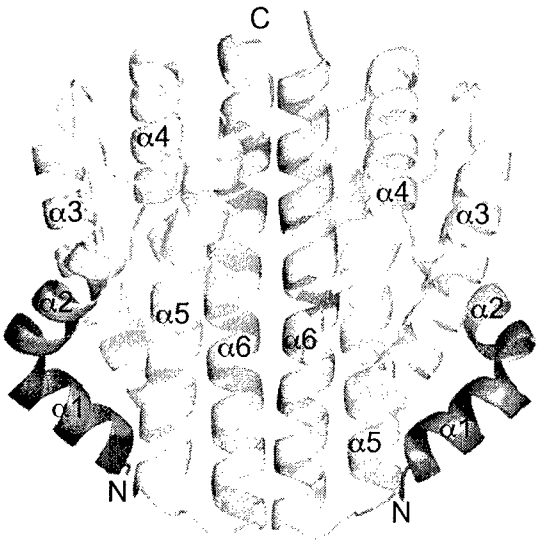
A

Stranzl_figure2

B



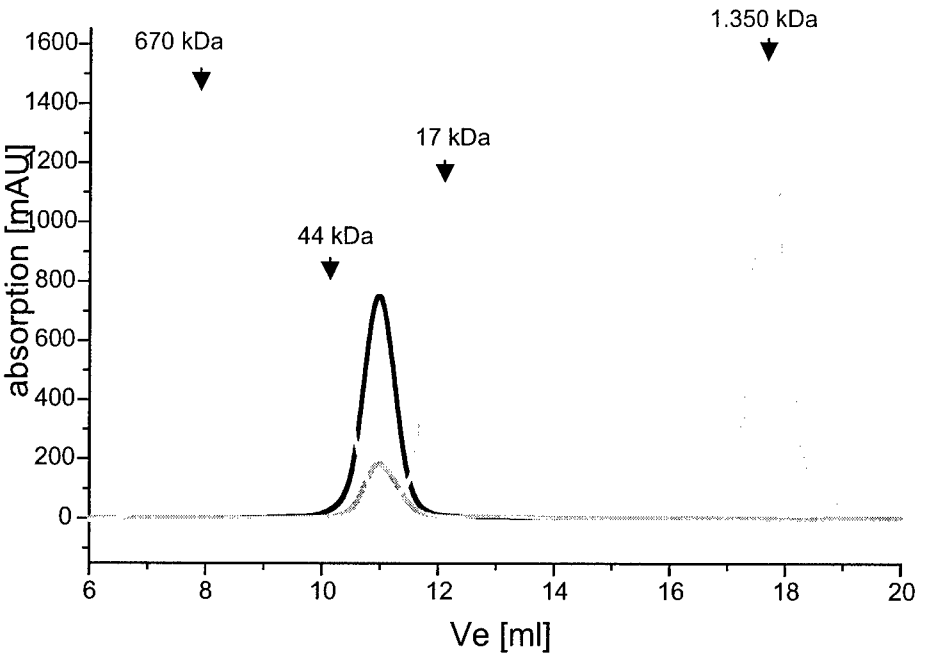

C




C

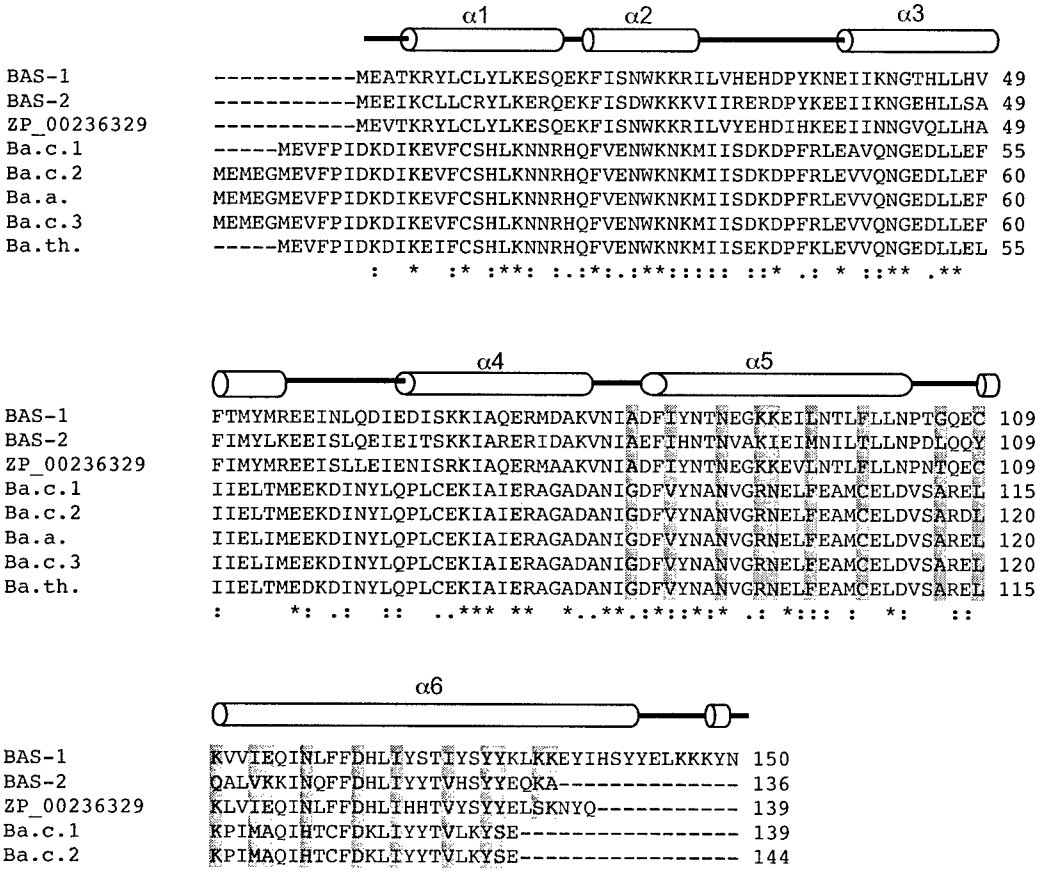
Stranzl_figure2

**D**



stranzl_figure3

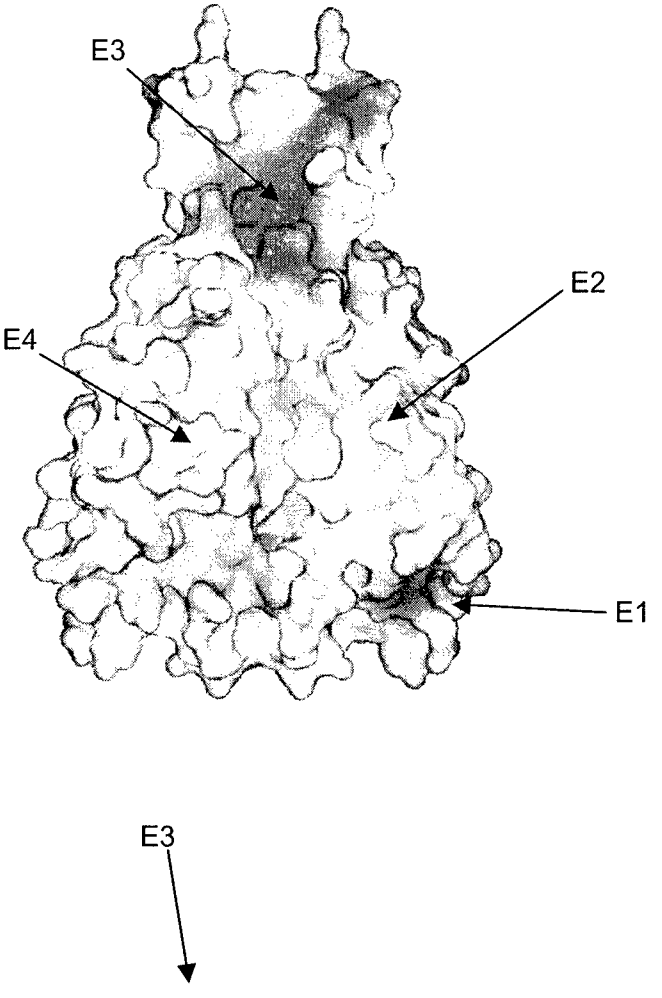
A

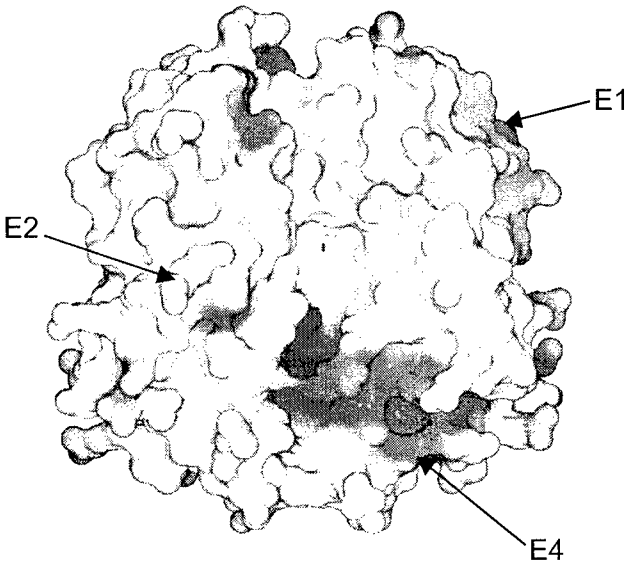


```
Ba.a.          KPIMNQIHTCFDKLIYYTVLKYSE----------------- 144
Ba.c.3         KPIMNQIHTCFDKLIYYTVLKYSE----------------- 144
Ba.th.         KPIMAKIHTCFDKLIYYTVLKYSE----------------- 139
               : :: :*:  **:**: *: .* :
```

Stranzl_figure3

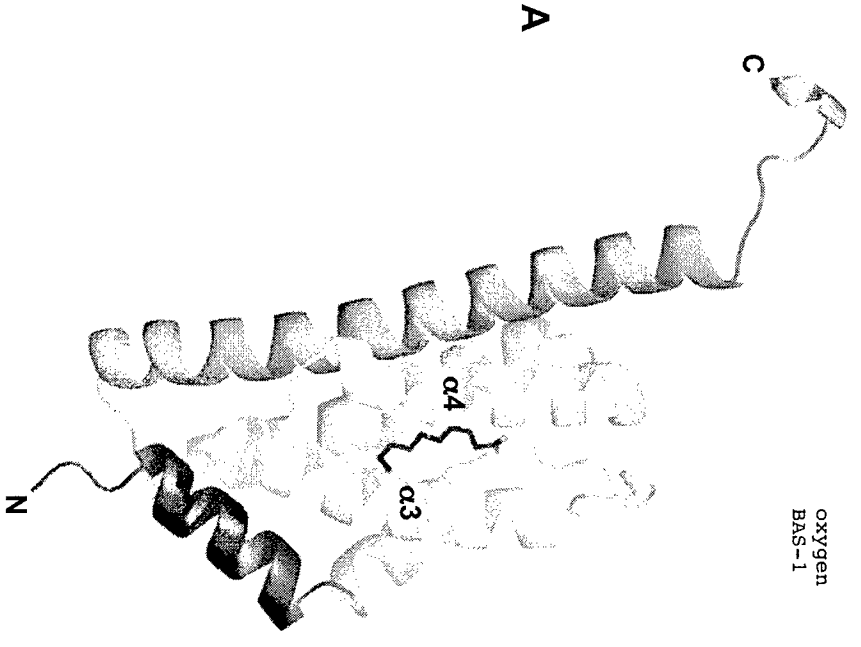
B





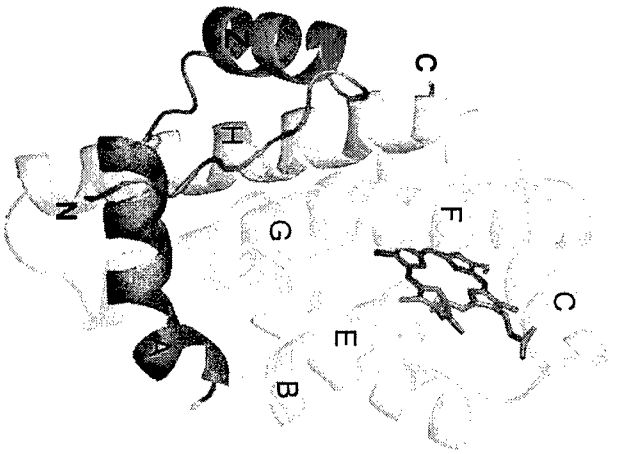

C

Stranzl_figure 4



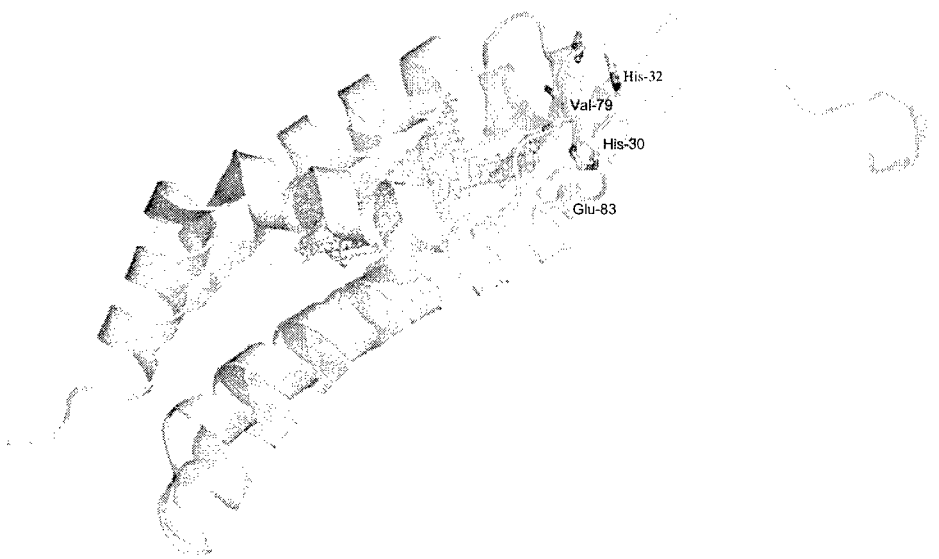

B

```
oxygen      MLFKKDRKQETAYFSDSNGQQKNRIQ
BAS-1       --MEATKRYLCLYLKES---------
                ::  ::   *:.:*

oxygen      NIVNIVDAFYKNLDHESSLMDIINDH
BAS-1       NGTHLLHVFTMYTMREEINLQDIEDIS
            * .:::..*  : .* .* ** :

oxygen      SIHLRIGLLPKWYMGAFQELLLSMID
BAS-1       NTLFLLNPTGQECKVVIEQINLFFDH
            . : :.  :   .:::: * : .
```

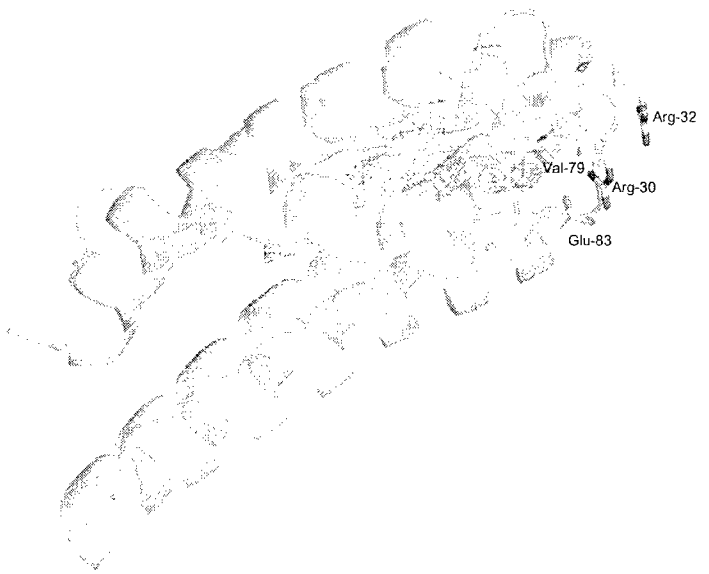



Stranzl_figure5

A



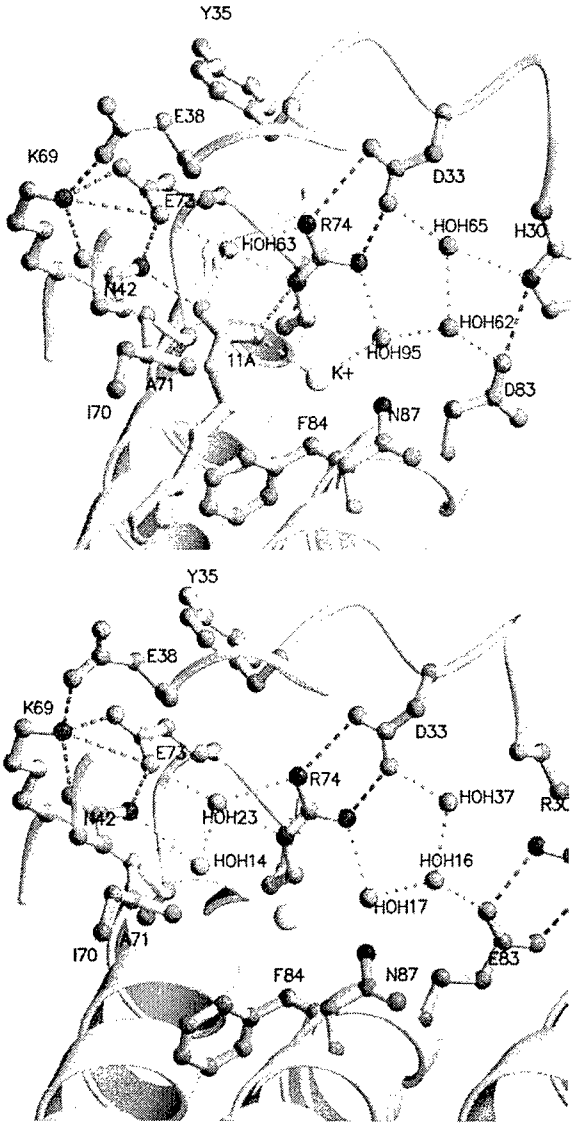
B



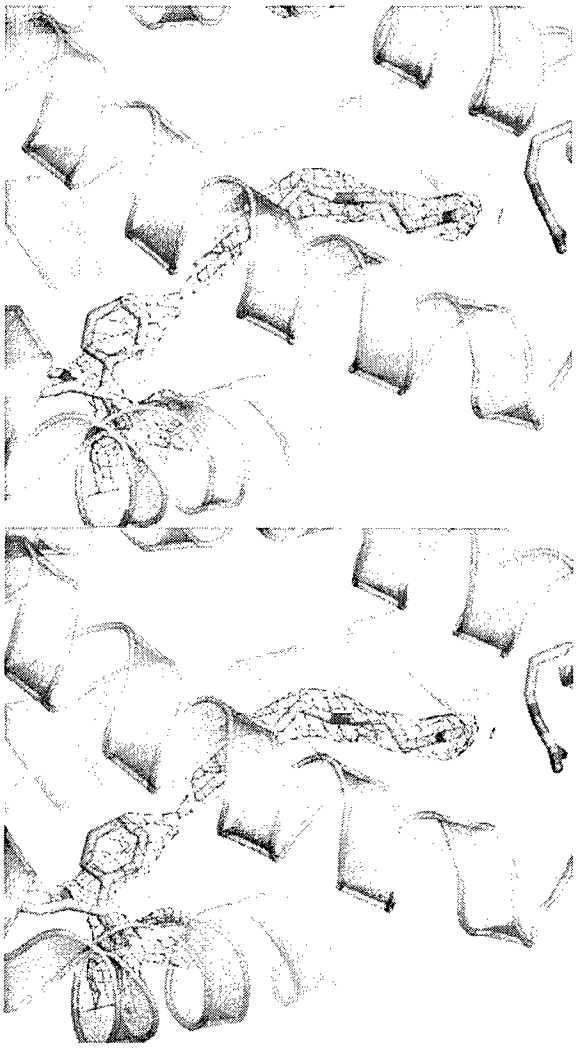

Stranzl_figure5

C

D


Y35
E38
K69
D33
E73
R74
HOH63
HOH65
H30
N42
HOH62
HOH95
11A
K+
A71
D83
I70
F84
N87
Y35
E38
K69
D33
E73
R74
HOH37
R3C
N42
HOH23
HOH14
HOH16
HOH17
A71
I70
E83
F84
N87

E



stranzl_figure 6

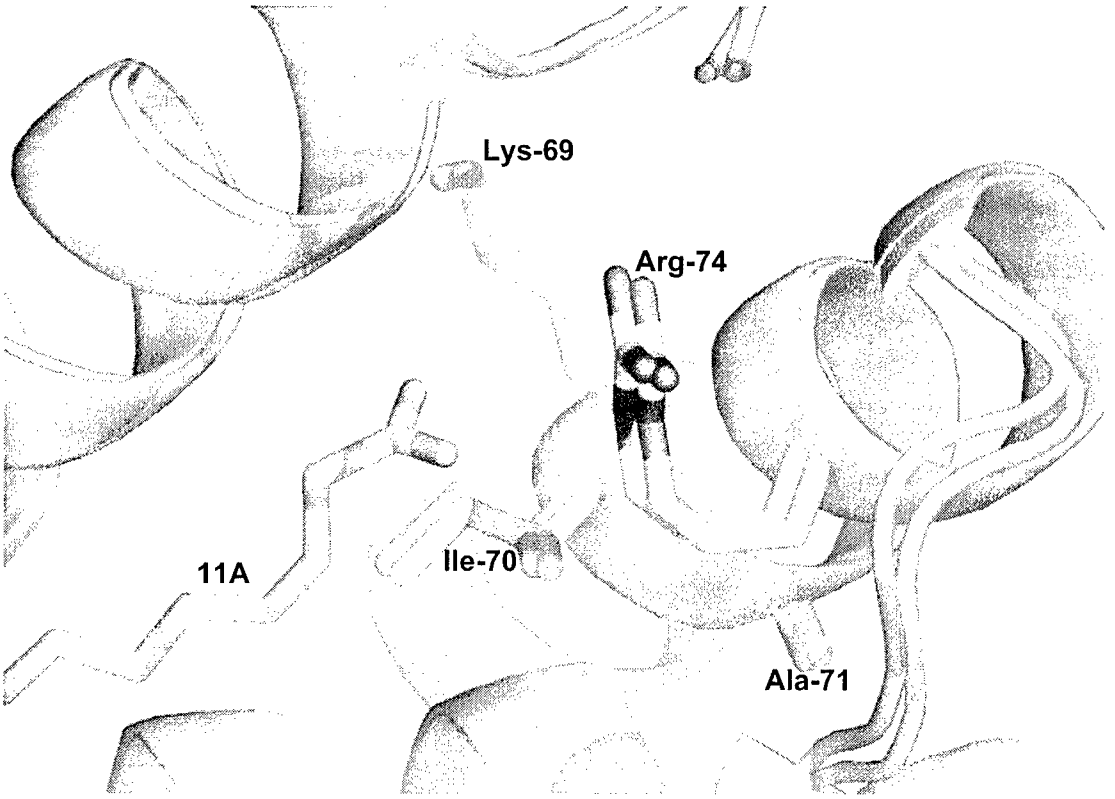



Table 1 Statistics of SAD data collection, phasing and refinement.

| | SAD phasing | Model refinement |
|---|---|---|

| | $\lambda_{peak}$ BAS-1 | BAS-1 | BAS-2 |
|---|---|---|---|
| Wavelength (Å) | 0.9781 | 0.97923 | 0.97923 |
| Resolution range (Å) | 50-2.5 | 76.7-1.76 | 62.02-1.49 |
| Observations | 70890 | 173637 | 408633 |
| Unique reflections | 5888 | 16461 | 56717 |
| Completeness[1] (%) | 99.8(100.0) | 99.5(95.0) | 98.9(94.5) |
| $R_{sym}$[1,2] (%) | 6.9(24.7) | 5.7(46.3) | 6.8(31.5) |
| $R_{cryst}$[3]/$R_{free}$[4] (%) | | 18.50/24.10 | 17.70/20.90 |
| Protein atoms | | 1408 | 2628 |
| Water molecules | | 95 | 364 |
| Ligand atoms | | 1 | 34 |
| Ligand molecules | | 1 | |
| R.M.S. deviations [Engh, 1991 #36] | | | |
| Bonds (Å) | | 0.027 | 0.012 |
| Angles (°) | | 1.697 | 1.367 |
| Average B-factor | | | |
| Protein (Å$^2$) | | | |
| Main chain | | 25.00 | 18.90 |
| Side chain | | 32.65 | 23.95 |
| Water (Å$^2$) | | 34.62 | 35.90 |
| Ligands (Å$^2$) | | 37.417 | 39.66 |

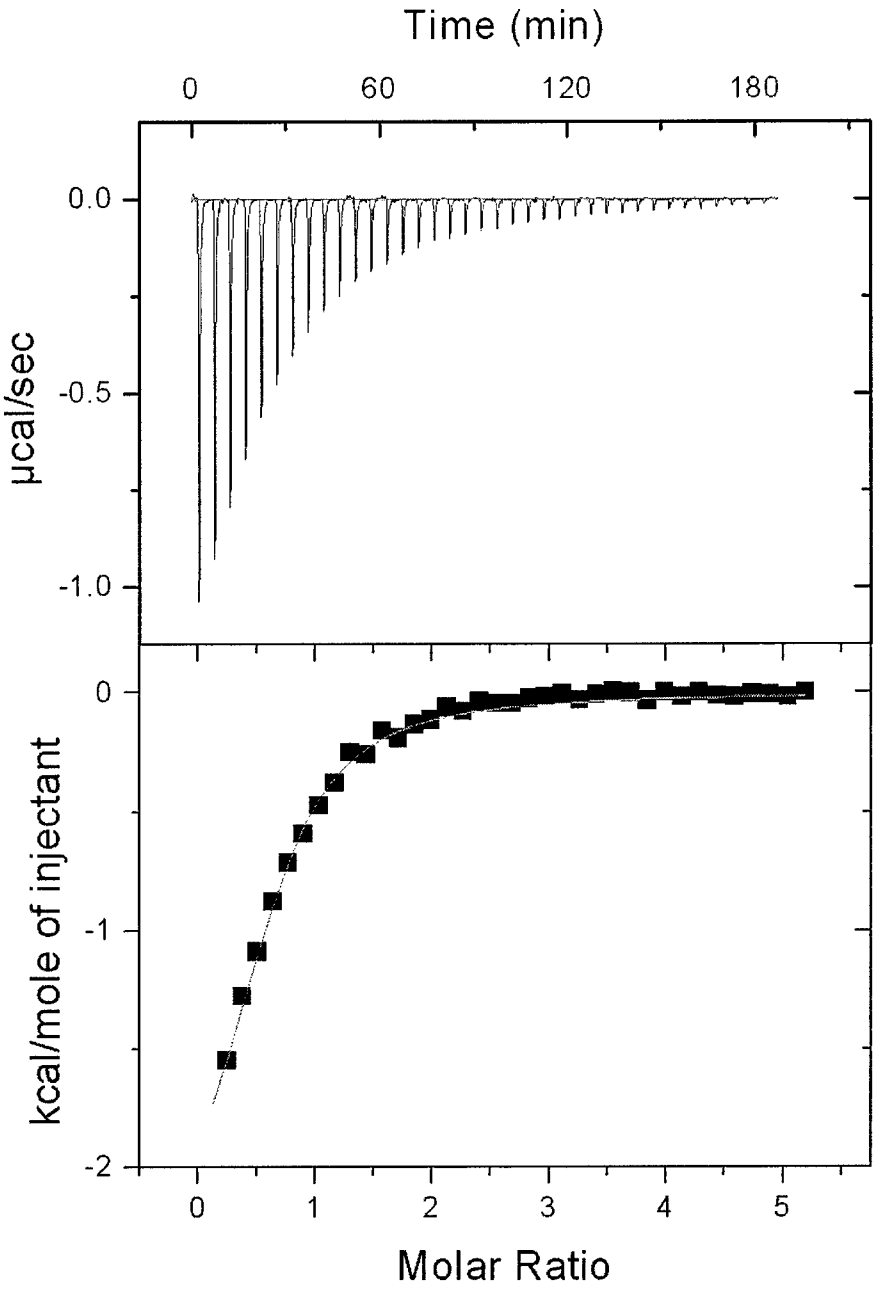
[1]Number in parentheses is for highest resolution shell. [2]$R_{sym}=\Sigma|I_h-\langle I_h\rangle|/\Sigma I_h$ , where $\langle I_h\rangle$ is the average intensity over symmetry equivalent reflection. [3]R-factor $=\Sigma|F_{obs}-F_{calc}|/\Sigma F_{obs}$ , where summation is over the data used for refinement. [4]$R_{free}$ was calculated using 5% of data excluded from refinement.

Table2 residues which line out the hydrophobic cavity of BAS-1 and BAS-2.

| **Hydrophobic cavity residues from BAS-1** | **Hydrophobic cavity residues from BAS-2** |
|---|---|
| Phe-19, Phe-50, Phe-84, Phe-119, Phe-120 | Phe-19, Phe-50, Phe-84, Phe-119, Phe-120 |
| Ile-20, Ile-39, Ile-70, Ile-81, Ile-95 | Ile-20, Ile-28, Ile-29, Ile-39, Ile-70, Ile-81, Ile-95 |
| Trp-23 | Trp-23 |
| Leu-28, Leu-46, Leu-47, Leu-123 | Leu-47, Leu-123 |
| Val-29, Val-79 | Val-79 |
| Pro-34 | Pro-34 |
| Asp-33, Asp-83 | Asp-33 |
| Ala-77 | Ala-77, Ala-91 |
| Glu-38, Glu-73 | Glu-38, Glu-73, Glu-83 |
| Arg-74 | Arg-30, Arg-32, Arg-74 |
| Asn-42, Asn-87 | Asn-42, Asn-87 |
| Gly-43, Gly-91 | Gly-43 |
| Thr-88 | Thr-88 |
| His-30, His-32 | |
| Tyr-35 | Tyr-35 |

Stranzl_figure7

A


Time (min)
0
60
120
180
0.0
-0.5
-1.0
µcal/sec
0
-1
-2
kcal/mole of injectant
0
1
2
3
4
5
Molar Ratio

**B**

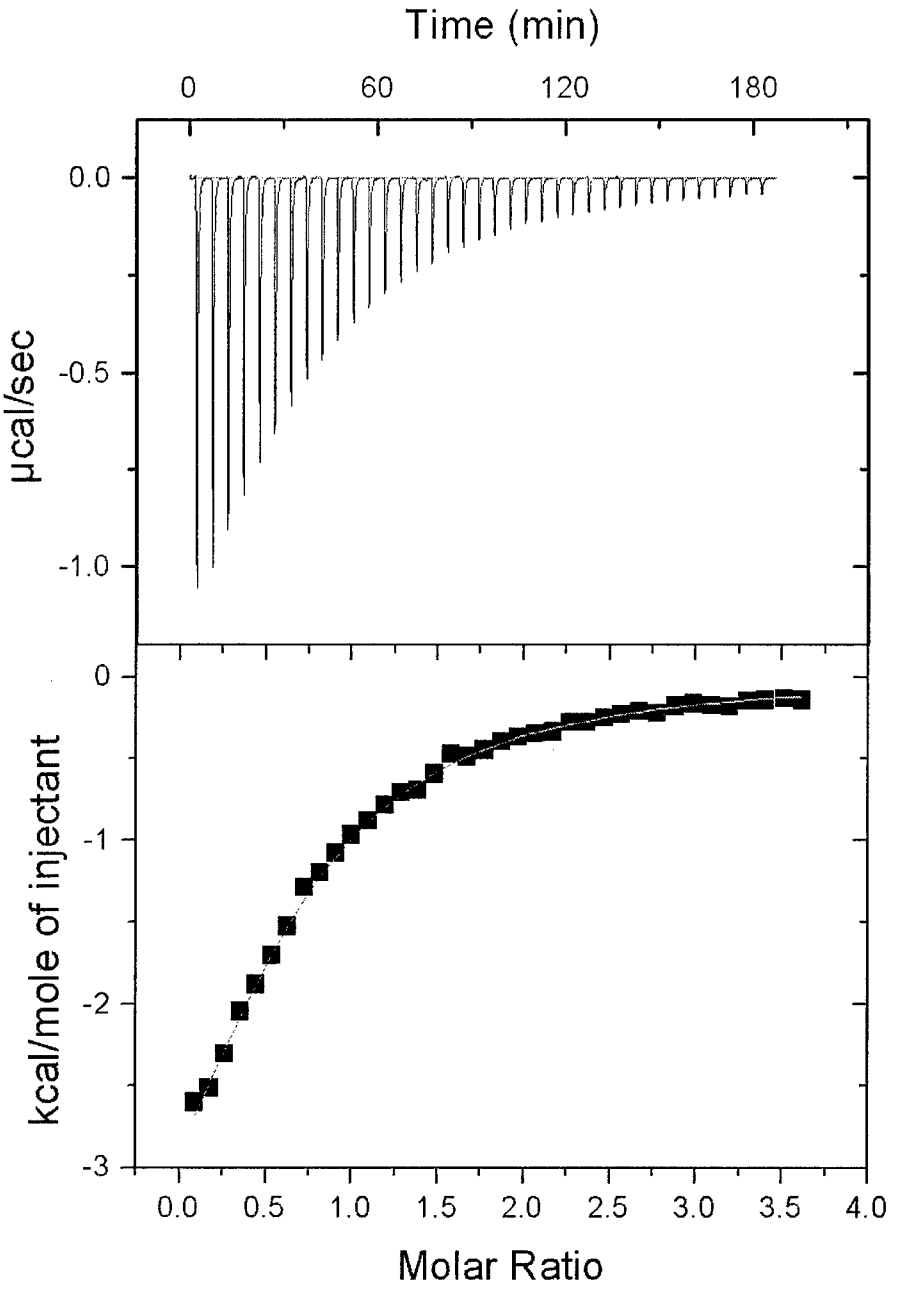
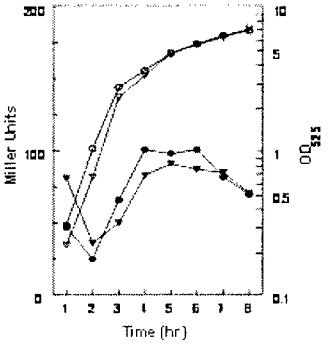

Time (min)
0
60
120
180
0.0
-0.5
-1.0
µcal/sec
0
-1
-2
-3
kcal/mole of injectant
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Molar Ratio

**Table 3.** Affinities of pOX2-61 and pOX1-118 to selected fatty acids.

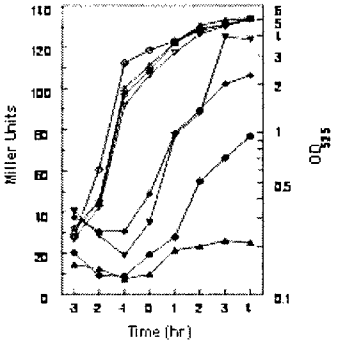
| Protein | Ligand | Kd (μM) |
|---|---|---|
| pOX2-61 | Myristic acid | 40 ± 15 |
| pOX2-61 | Palmitic acid | 41 ± 17 |
| pOX2-61 | 12-Methyltetradecanoic acid | 41 ± 19 |
| pOX2-61 | 13-Methyltetradecanoic acid | 46 ± 6 |
| pOX2-61 | Palmitoleic acid | 20 ± 1 |
| pOX1-118 | Myristic acid | 25 ± 9 |
| pOX1-118 | Palmitic acid | 24 ± 7 |
| pOX1-118 | 12-Methyltetradecanoic acid | 14 ± 3 |
| pOX1-118 | 13-Methyltetradecanoic acid | 13 ± 3 |
| pOX1-118 | Palmitoleic acid | 30 ± 8 |

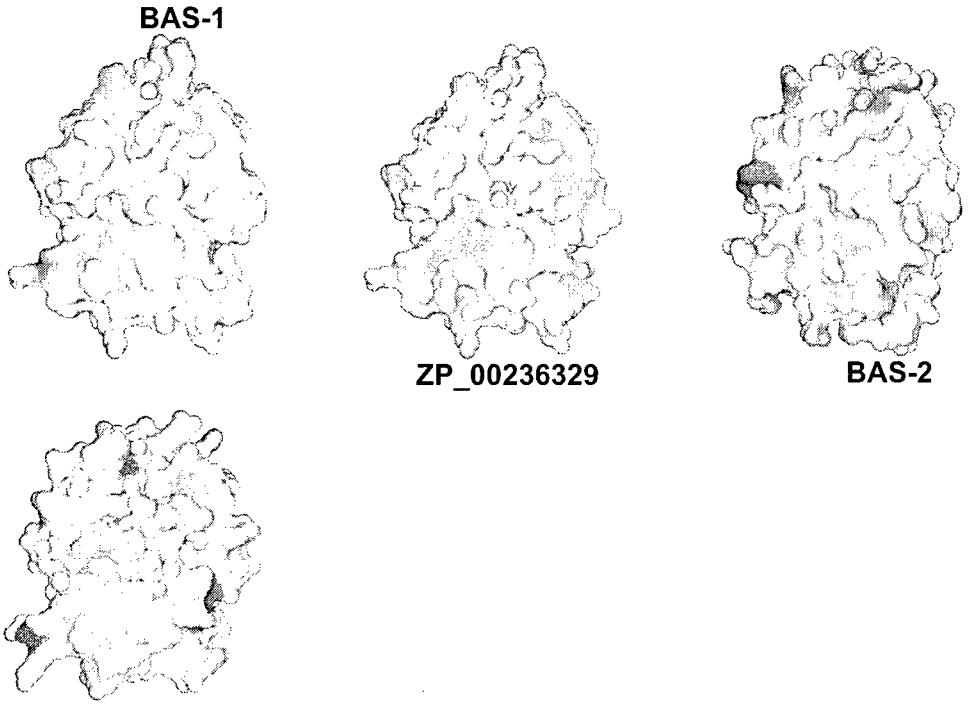
Stranzl_figure8

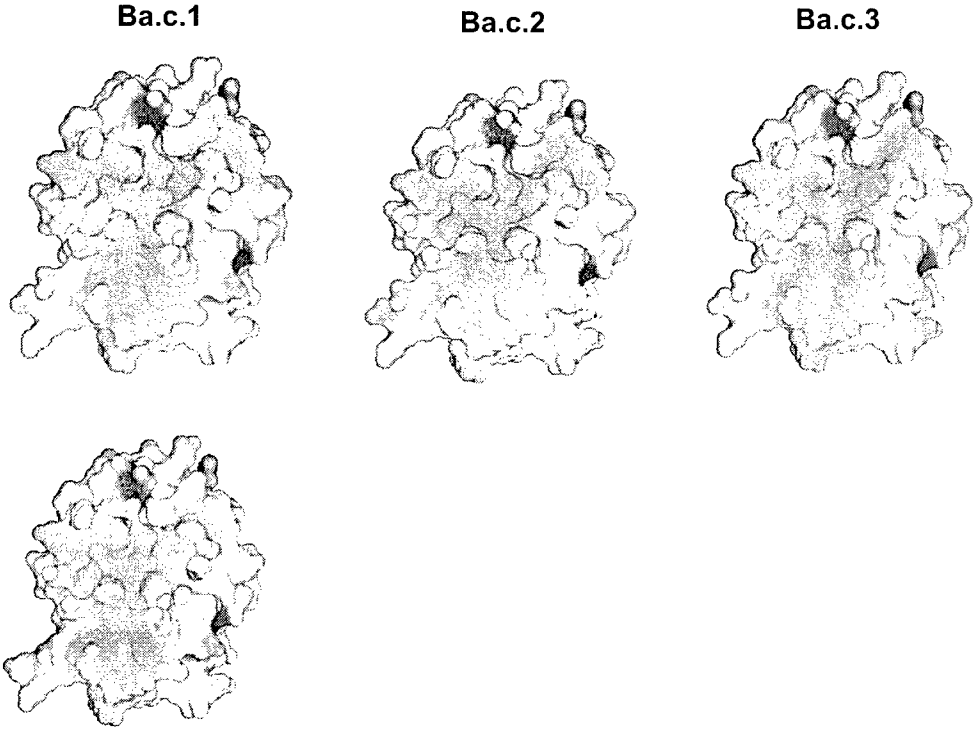
A




B




BAS-1
ZP_00236329
BAS-2


Ba.c.1
Ba.c.2
Ba.c.3

# Surprising connections: in-depth analysis of the *Bacillus anthracis* pXO1 plasmid

**Marcin Grynberg[1,2], Iddo Friedberg[1], Marc Robinson-Rechavi[3], and Adam Godzik[1,3,4]**

*[1]Bioinformatics and Systems Biology, The Burnham Institute, La Jolla, CA 92037, USA; [2]Institute of Biochemistry and Biophysics, PAS, Pawinskiego 5A, 02-106 Warsaw, Poland; [3]Joint Center for Structural Genomics, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA*

Keywords: anthrax, *Bacillus anthracis*, pathogenicity, virulence, pXO1, type IV secretion system, type IV pilus assembly system, context analysis, distant homology, ArsR, SmtB, regulators.

**[3]Corresponding author.**

E-mail: adam@burnham.org;

**PHONE +1 (858) 646 3168**

**FAX: +1 (858) 713 9930.**

## ABSTRACT

Anthrax disease is caused by a bacterium *Bacillus anthracis*. Its virulence has been associated with two plasmids, pXO1 and pXO2. Using a combination of advanced bioinformatics tools, including context analysis, distant homology and fold recognition, we have re-annotated the predicted open reading frames on the pXO1 plasmid, most of which were described as proteins of unknown function in previous analyses. Thanks to improved annotation tools we significantly enhanced the annotation of the pXO1 plasmid, bringing the total number of ORFs with some level of functional annotation from 48 to over 100. The new results also clearly show the mosaic nature of pXO1 and give tantalizing hints about the origin of anthrax virulence. The highlights of the new finding are two type IV secretion system-like clusters present on the pathogenicity island of the pXO1 plasmid, as well as at least three clusters related to DNA processing.

Supplemental material available online at http://bioinformatics.burnham.org/pXO1.

## *INTRODUCTION*

Anthrax is a disease primarily affecting herbivores but also sporadically attacking other mammals, including humans. Anthrax is known since antiquity and the quest for an effective treatment of anthrax is closely related to the birth of modern microbiology (Pasteur 1881). More recent work concentrated mostly on the anthrax toxin, leading to extensive structural and functional analysis of its components (for a review see (Turnbull 2002)). However, until recently the general level of interest in anthrax was limited, since it is not a major threat to human health. An era of more intensive work on *B.anthracis* has started since anthrax was adopted by military as a biological weapon, resulting in a threat of large scale anthrax outbreaks. These threats were kept alive by several large scale incidents, and more recently the threat of anthrax as a bioterror weapon.

At the same time, the origin and mechanism of *B.anthracis* virulence are very interesting on their own. Only very few *B.anthracis* virulence related proteins were studied in detail, among them the toxins (PagA, LEF, CyaA), cell envelope and germination genes (Cap, S-layer and Ger proteins), and the regulatory mechanisms triggering the virulence (Fouet and Mesnage 2002; Lacy and Collier 2002) and citations therein). The sequencing of the *B.anthracis* genome (Okinaka et al. 1999; Pannucci et al. 2002; Read et al. 2003) especially in the context of other *Bacilli* genome projects, highlighted the complex and little understood mechanism of its virulence (Koehler 2002). The *B.anthracis* genome consists of a single chromosome and two virulence associated megaplasmids, pXO1 and pXO2 (Okinaka et al. 1999; Read et al. 2003). The two plasmids together convey the pathogenic phenotype and are responsible for most of the difference between *B.anthracis* and its relatives with different pathogenicity profiles, such as *B.cereus* or *B.thuringiensis*. However, little is known about most proteins encoded by the two plasmids and only a few have been studied by experiment and shown to be directly involved in virulence. Most pXO1 and pXO2 proteins have no obvious sequence similarity to any other known genes. Therefore, the interest in *B.anthracis* pathogenicity transcends its immediate applications in bioterrorism and human health, and bears on fundamental questions of how novel and complex lifestyles, such as pathogenicity, can evolve.

Several earlier works focused on bioinformatic analysis of the anthrax genome and plasmids, often in the context of related organisms (Ariel et al. 2002 2002; Ariel et al. 2003; Rasko et al. 2004). These studies confirmed close relations between *B.anthracis*, *B.thuringiensis* and *B.cereus*, and identified previously unknown features of the virulence related plasmids, pXO1, pBtoxis and pBc10987, respectively. However, a vast majority of pXO1 genes remain uncharacterized, both in terms of their function and origin. A possible reason for this apparent novelty of pXO1 genes is that pathogenic plasmid encoded genes evolve rapidly and often bear little sequence similarity to their homologs from other species, hampering the detection of homology with most tools of sequence analysis. In this study we take advantage of recent improvements in super-sensitive tools for distant homology recognition. These include a profile based variant of the BLAST algorithm (Altschul et al. 1997), algorithms based on Hidden Markov Models (Bateman et al. 2002), and profile-profile based methods (Rychlewski et al. 2000). These algorithms are most often tested in the context of structural and fold predictions (Kinch et al. 2003), where predictions can be easily validated by comparing three dimensional structures. They are gaining acceptance also in function prediction and evolutionary analysis (Altschul and Koonin 1998; Sadreyev et al. 2003). In addition, context analysis, which takes advantage of the operon structure, has emerged as a powerful tool of annotation in prokaryotes (Overbeek, et al. 1999; Huynen, et al. 2000; Wolf, et al. 2001), and we have combined these results with those of distant homology to improve annotation of the pXO1 plasmid.

The origin of pathogenicity plasmids has often proved elusive, all the more that most of their ORFs were not annotated. Our annotation also allows us to put forward hypotheses on the evolutionary origin of the ORFs encoded in pXO1, which represent an interesting mix of vertical and horizontal transfer. Thus we are able to shed new light on the evolution of pathogenicity in the *Bacillus* genus.

## RESULTS

### *Overview of the results*

The results of our annotation effort are summarized in figure 1. All details are available as supplementary material tables on http://bioinformatics.burnham.org/pXO1. Despite previous reports, these results show that many pXO1 proteins do have recognizable homologues in other species. Overall, over 60 ORFs, previously described as unique, could be reliably identified as members of known protein families. Still, for many of them we are not able to confidently assign a molecular function. First, the full functional groups (operons, pathways) of many of the newly characterized proteins seem to be missing in pXO1. These groups may be completed by other proteins from anthrax plasmids or genome which are as yet uncharacterized, or the protein may have acquired a different functional context in anthrax. Second, many ORFs appear truncated and mutated to the point that it is unclear whether they have conserved the same function, or, in fact, whether they have any function at all (Supplementary data). This in turn might be related to the continuing evolution of the plasmid, where some genes are only partly degraded and still recognizable, like the region homologous to a part of the lethal factor (see: Particular cases section in Results) or a fragment of the NADH dehydrogenase (see: Supplementary data).

Despite these reservations, interesting tendencies emerge from our functional annotations: pXO1 contains many regulatory proteins, such as SinR (BXA0020, pXO1-14), AtxA (BXA0146, pXO1-119) or the MerR homologue (BXA0069, pXO1-47), with predicted DNA binding domains. Another interesting trend is that pXO1 has a significant number (15% of the whole plasmid) of proteins related to DNA metabolism (Supplementary data). We have also identified several probable operons, conserved among different groups of bacteria.

**DNA level analysis**

Several analyses were performed in order to analyze the DNA sequence of the pXO1 plasmid [Okinaka, 19999; Read, 2002; Pannucci, 2002]. The ORF prediction programs were used, the DNA motifs were discovered and a connection between promoter elements and ORFs was already done. Our analysis of the DNA sequence focused on two aspects. First, we were interested in the discovery of the origin of replication since no genes obviously involved in this process could be detected. Second, we searched for specific DNA regions related to pathogenicity.

Our goal was to find proteins directly involved in the plasmid replication. Unfortunately, we could not detect those. Therefore, we used the Oriloc program to predict the bacterial origin of replication [Frank, 2000]. In bacteria, the leading strands for replication are enriched in keto (G, T) basis while the lagging strand is enriched in amino bases (A, C) [Rocha, 1999]. This compositional assymetry allows the identification of probable origin and termination sites of replication. Oriloc analysis indicated a potential origin of replication between bases 66538 to 66558 which is quite close to the origin predicted earlier by Berry and colleagues (60955-62192 region)[Berry, 2002]. The origin is predicted in the neighbourhood of hypothetical proteins, with no recognizable homology to proteins from publicly available databases. It is located in between ORFs BXA0076 (pXO1-51) and BXA0077 (pXO1-52). The termination of replication may lie around the position 173914 on the pXO1 plasmid, between genes BXA0206 (pXO1-137) and BXA0207 (pXO1-138) which encode an RNA-binding Hfq (Host Factor I) protein and the transcription regulator from the ArsR family, respectively.

At the DNA level, we were interested in finding regions connected to the regulation of virulence. We focused on genes regulated by AtxA [Bourgogne, 2003]. Our goal was to characterize DNA regions involved in AtxA binding. For this purpose, we collected intergenic sequences preceding the AtxA-dependent genes (see Table 1 in Bourgogne, 2003] and analyzed it using the MEME [Bailey, 1994] and the MITRA [Eskin, 2002] programs. The only common motif that we could find was ANGGAG which was located in diversified distances (5-600 bp) from the putative ATG translation start codon. Large differences

in the location of the ANGGAG motif can be attributed to unrecognized ORFs located upstream from some of the analyzed genes, in the same operon. Another possibility is that this signal is false. Deletion experiments of these *cis* elements should be performed to check our hypothesis.

**Protein level analysis**

**Proposed operons: function and evolutionary conservation**

**A pathogenicity operon conserved in Bacilli**

BXA0091 (pXO1-65) and BXA0094 are homologous to each other and to proteins from several other bacilli; *Enterococcus*, *Listeria*, *Lactococcus*, *Lactobacillus*, or other *Bacillus* species. Function of proteins from this family is unknown, but the proteins are hypothesized to be extracellular (Nakai and Horton 1999). Many members of this family have additional domains on the C-terminus, often repeats such as WD or LRR repeats, associated with protein-protein and receptor-like activities. Not only in anthrax, but also in *E.faecalis* and *B.thuringiensis*, this gene is represented by at least two copies in each operon. In *B. anthracis*, *B.thuringiensis*, *L.innocua* and *E.faecalis* the BXA0091 homologues colocalize with a surface layer domain protein. Interestingly, in species other than anthrax, these two proteins often colocalize with three proteins: a protein homologous (FFAS score: -10.100) to a protein containing the LysM domain (homology is not in the LysM region), a protein homologous to the RTX toxin and related $Ca^{2+}$-binding proteins family and a regulatory protein homologous to positive transcription regulators MGA. The LysM domain binds peptidoglycans and was first identified in bacterial lysins (Ponting et al. 1999). Several proteins, such as staphylococcal IgG binding proteins and *E.coli* intimins, contain LysM domains. RTX toxins are pore-forming, calcium-dependent cytotoxins encoded by various bacterial genomes (Braun and Cossart 2000), and MGA are important in streptococci virulence (McIver and Myles 2002). Other proteins from these operons in other organisms are also predicted to be extracellular and involved in pathogenesis, in *B. anthracis* this appears to be a minimal variant of this virulence related operon.

**A DNA-modifying operon shared with Gram-positive bacteria**

BXA0010 (pXO1-06), BXA0013 (pXO1-08) and BXA0015 (pXO1-10) form an operon that can also be found in two Gram-positive species, *Xanthomonas* and *Burkholderia* (Figure 1), and in the proteobacterial *Pseudomonas* group. BXA0010 and BXA0013 are homologues of the *Xanthomonas* orf8, of a *Burkholderia* protein and of a number of *Pseudomonas* proteins. Both BXA0010 and BXA0013 anthrax proteins belong to the superfamily II of DNA/RNA helicases, and BXA0010 seems to be a duplication of the middle part of the BXA0013 protein. In between these two proteins, in *B.anthracis*, there is an inserted reverse transcriptase (BXA0011, pXO1-07). One can hypothesize that this insertion occurred after the duplication and disrupted BXA0010. BXA0013 forms an operon with BXA0015, a protein with strong similarity to the N-terminal part of its homologues that encodes the coenzyme-binding domain of various DNA methyltransferases. The co-occurence of the DNA/RNA helicase and DNA methyltransferase is also conserved as an operon in other species mentioned above. *Xanthomonas*, *Burkholderia* and *Pseudomonas*, but not anthrax, preserve numerous other proteins in BXA0013-BXA0015 analogous operons. The function of these additional proteins is however unclear. From the functions of known members of this operon one can imply its DNA modifying function.

**A nucleotide metabolism operon shared with Actinobacteria and Cyanobacteria**

BXA0032 and BXA0033 (pXO1-22), if fused, would belong to the COG0175 family, members of the 3'-phosphoadenosine 5'-phosphosulfate sulfotransferase (PAPS reductase)/FAD synthetase group of enzymes which are linked to ATPase involved in DNA repair/chromosome segregation from *Anabaena* spp., *Nostoc* spp., *Bacillus stearothermophilus* and *Streptomyces avermitilis*. Functions of other proteins from this cluster are unknown. In *B.anthracis* however, it is located close to BXA0034. We described the members of this family as a new HEPN nucleotide-binding domain (Grynberg et al. 2003), and a connection with BXA0037 (pXO1-24), a nucleotidyltransferase domain protein, is obvious. As a complex they may catalyze the addition of a nucleotidyl group to unknown substrates, maybe to antibiotics or other poisonous substances,

as their structural homolog kanamycin nucleotidyltransferase does (Matsumura et al. 1984). The specific function of the HEPN-nucleotidyltransferase operon in pXO1 is unknown.

**Type IV secretion system machinery: two operons and missing links**

Two operons in *B.anthracis* contain proteins strongly resembling elements of type IV secretion system proteins (Fig. 2). This specific secretion system is important in the delivery of effector molecules to the host cell (Christie 2001; Christie and Vogel 2000).

The first operon consists of four proteins (BXA0083/pXO1-57, BXA0085/pXO1-59, BXA0086/pXO1-60 and BXA0087/pXO1-61), of which the first is homologous to a protein involved in type IV pili biogenesis, CpaB/RcpC (COG3745). The next protein, BXA0085, belongs to the VirB11 family, and the remaining two are two paralogs belonging to the TadC family (COG2064), whose members are often found in the same operons with the VirB11. VirB11 family is well studied, (Christie 2001; Dang et al. 1999; Krause et al. 2000; Savvides et al. 2003; Yeo et al. 2000) and members of this family are ATPases that function as chaperones reminiscent of the GroEL family for translocating unfolded proteins across the cytoplasmic membrane (Christie 2001). Homologues of all four proteins from the pili biogenesis-like operon form operons in many Gram-negative bacterial species (Kachlany et al. 2000; Skerker and Shapiro 2000). To date, only in *Caulobacter crescentus* this operon was experimentally proven to be required for pilus assembly (Skerker and Shapiro 2000). Distant homologs of pilA and other pilin subunits necessary for pilus formation can be found scattered on pXO1 (for instance BXA0092) and on pXO2 (work in preparation).

The second operon contains the homologue of the VirB4 protein (BXA0107) and a fusion of the VirB6 homology region with a surface-located repetitive sequence, similar to coiled-coil proteins, with a methyl-accepting chemotaxis protein (MCP) signaling domain at the C terminus (BXA0108, pXO1-79). VirB4 family is one of the elements of the type IV secretion system. This system, ancestrally related to the conjugation machinery, is able to deliver DNA molecules as well as proteins. VirB4 is an ATPase that "might transduce information, possibly in the form of ATP-induced conformational changes, across the

cytoplasmic membrane to extracytoplasmic subunits," according to Christie (Christie 2001) and Dang (Dang et al. 1999). It contains the Walker A motif responsible for ATP binding, which is well conserved in BXA0107 (200-207 fragment: GISGSGKS). The BXA0108 protein has at least 7 predicted N-terminal (55-281 aa) transmembrane motifs, similar to the central part of the VirB6 protein, and a surface-located repetitive sequence, most probably forming a coiled-coil structure. The C-terminal of this protein is homologous to a domain that is thought to transduce the external chemotaxis signal to the two-component histidine kinase CheA (for review see (Stock et al. 2002)). The next protein in this operon resembles the C-terminus of a *Bacillus firmus* integral membrane protein, which includes transmembrane domains in the N-terminal part. This region is homologous to the phosphatidate cytidylyltransferase (EC 2.7.7.41), an enzyme that catalyzes the synthesis of CDP-diglyceride, the source of phospholipids in all organisms (Icho et al. 1985; Sparrow and Raetz 1985). The function of the C-terminal part of the *B.firmus* protein is unknown.

The presence of three proteins with features characteristic of type IV secretion system and other ORFs related to type IV pilus formation strongly suggests that such a system may be active on the virulence plasmids in anthrax and may play a role in its virulence. It seems logical then to search for other elements of type IV secretion system in the anthrax plasmids or genome. We are able to detect some other distantly related elements of this machinery, but the system appears incomplete. Is it a fully functional, minimal type IV secretion system? Or are other parts of this system present in anthrax, but impossible to identify with available tools? The operons discussed here are good targets for experimental analysis, since they contain many as yet uncharacterized proteins. It is also not clear what molecules are secreted by this system, the anthrax toxin or other proteins. In any case, understanding of the function of this secretion system would be crucial for our understanding of diverse roles of pXO1 in virulence.

### *Putative pXO1 regulator proteins*

The most important elements in the description of unknown biological systems are the regulatory proteins. They decide when, who and how is expressed in the cell. In pathogenic systems, frequently

regulators of virulence genes are located in pathogenic regions. However, various permutations are known, where regulators regulate genes outside of the pathogenicity island, or regulators encoded outside of the pathogenicity island regulate genes located in the virulence regions (Hacker and Kaper 2000; Hentschel and Hacker 2001). Anthrax pX01 plasmid contains many uncharacterized regulatory proteins. We think that it is essential to describe the regulators on the anthrax pathogenicity vector in order to decipher the physiology of pXO1.

**Specific duplications in the ArsR/SmtB family: BXA0166 and BXA0207**

Both BXA0166 (pXO1-109) and BXA0207 (pXO1-138) are members of the ArsR/SmtB family of metalloregulatory transcriptional regulators. The vast majority of known family members are repressors. Indeed, BXA0166 has been characterized as the gene for repressor PagR (Hoffmaster and Koehler 1999). They act on operons linked to stress-inducing concentrations of diverse heavy metal ions. Derepression results from direct binding of metal ions by ArsR/SmtB transcription regulators. The founding members of the family are SmtB, the Zn(II)-responsive repressor from *Synecchococcus* PCC 7942 (Morby et al. 1993), and ArsR, that acts as the arsenic/antimony-responsive repressor of the *ars* operon in *Escherichia coli* (Wu and Rosen 1991). Another, less well studied, group in the ArsR/SmtB family are the transcriptional activators, with *Vibrio cholerae* HlyU as the founding member (Williams et al. 1993). HlyU is known to upregulate the expression of hemolysin and of two *hcp* genes, which are coregulated with hemolysin (Williams et al. 1996). We have conducted a phylogenetic analysis of this vast family, with a focus on the evolutionary history of ArsR/SmtB proteins in bacilli, notably in anthrax, and on the relation between phylogeny and function (i.e. repressor or activator).

In a phylogeny of representative members of the ArsR/SmtB family (Fig. 5A), the two pXO1 proteins are closely grouped with other *Bacillus* proteins. This group has very long branches in the tree, indicative of rapid evolution of the proteins. The only two known activators (HlyU and NolR) of the family appear closely related, in a clade with proteins of unknown function. These latter include clear orthologs of

HlyU or of NolR. It is thus reasonable to predict that these proteins form a clade of transcriptional activators. Interestingly, this "activator" clade appears closely related to the clade including both pXO1 proteins (clades boxed in Fig. 5A). PagR is known to act as a repressor, but in a weak manner (Hoffmaster and Koehler 1999) and is suspected of having an activation function as well (Mignot et al. 2003). A more detailed phylogeny of close homologues of the pXO1 proteins (Fig. 5B) shows that there has been a wave of gene duplications in the ancestor of *B.antracis* and *B.cereus* (full circles in Fig. 5B). All seven of the resulting paralogues were retained in *B.antracis*, including the two which were transferred to pXO1, while four were secondarily lost in *B.cereus*. There was an independent duplication in *B.thuringiensis* (open circle in Fig. 5B). Interestingly, these are the only bacilli represented in this clade of close homologues, all three have duplications of the gene, and all three are pathogens.

Overall, the phylogenetic analysis shows that both pXO1 ArsR/SmtB proteins are closely related members of a clade of fast evolving proteins, which have duplicated several times in pathogenic bacilli, and which are related to the only clade of transcriptional activators of the family.

**Other putative regulators**

BXA0020 (pXO1-14) is 564 amino acids long. The C-terminal 60-70 aa are homologous to DNA-binding domains of several repressor families (SCOP: a.35.1 superfamily of lambda repressor-like DNA-binding domains). The one that is the most similar is the SinR repressor domain (Gaur et al. 1986). In *Bacillus subtilis* the proteins of the *sin* (sporulation inhibition) region form a component of an elaborate molecular circuitry that regulates the commitment to sporulation. SinR is a tetrameric repressor protein that binds to the promoters of genes essential for entry into sporulation and prevents their transcription (Mandic-Mulec et al. 1995; Mandic-Mulec et al. 1992). In pXO1, BXA0020 does not form an operon with sin genes. Instead, it is located close to a protein (BXA0019, pXO1-13) that is characterized as similar to the middle fragment (417-1236 aa) of the 236 kDa rhoptry protein from *Plasmodium yoelii yoelii*, involved directly in the parasite attack of red blood cells (Khan et al. 2001). It is not certain whether they form one operon since

both genes have putative independent ribosome binding sites. The N-terminal region of BXA0020 is not well described and has the strongest similarity to the α-helical part of the chromosome-associated kinesin, or the kinesin-like domain (KOG0244). Kinesins are microtubule-dependent molecular motors that play important roles in intracellular transport of organelles and in cell division (Mandelkow and Mandelkow 2002; Woehlke and Schliwa 2000).

The N-terminal part of BXA0048 (pXO1-34) is the DNA-binding helix-turn-helix motif that belongs to the TetR family (PF00440). Members of this family take part in the regulation of numerous pathways/operons, e.g. TetR is a tetracycline inducible repressor (Hillen and Berens 1994), BetI, a repressor of the osmoregulatory choline-glycine betaine pathway (Lamark et al. 1996), MtrR, a regulator of cell envelope permeability that acts as a repressor of *mtrCDE*-encoded and activator of *farAB*-encoded efflux pumps (Lee et al. 2003; Lee and Shafer 1999). We were unable to determine any reasonable homology to the distal part of BXA0048, therefore no functional hypothesis can be drawn. The only indication for the function of that regulator is the probable placement on one operon with a nucleotidyltransferase (BXA0047, pXO1-33). The presence on the same operon of the nucleotidyltransferase with a superfamily II DNA and RNA helicase family protein in *Streptomyces coelicolor* can be a suggestion that BXA0048 is involved in DNA metabolism.

BXA0060 (pXO1-40) belongs to a large superfamily of repressors (SCOP: a.35.1). It is composed of the DNA-binding domain only. Homologues of BXA0060 are present in numerous archaeal and eubacterial genomes, with no preservation of operon structure. It seems then that BXA0060 homologues are involved in very diverse functions/pathways.

BXA0069 (pXO1-47) belongs to the family of global transcription activators of membrane-bound multidrug transporters, responsible for bacterial multidrug resistance (MDR)(Paulsen et al. 1996). The closest homologue is the *B.subtilis* MtnA regulator that belongs to the MerR family (Summers 1992). It is known to activate two MDR transporters (*bmr* and *blt*), a transmembraneous protein-coding gene *ydfK* and its own gene (Baranova et al. 1999). It acts independently from two specific activators, BmrR and BltR, that

are encoded by the *bmr* and *blt* operons (Ahmed et al. 1995). MtnA and other members of the MerR family are composed of three regions; N-terminal DNA-binding domain (winged helix-turn-helix motif), middle all-helical dimerization region and the C-terminal part specific for each protein that is probably involved in specific ligand binding (Godsey et al. 2001). BXA0069 perfectly fits this description, it possesses two quite conserved distal regions, and a 90 amino acid region of no homology that has an almost 80% probability of a coiled-coil structure (Lupas et al. 1991). Because of lack of resemblance of the C-terminus to any known regulatory domain, it is difficult to propose in what metabolism/gene(s) activation is the BXA0069 protein involved.

The FFAS analysis revealed low score similarity of BXA0122 (pXO1-89) to the MarR regulators of the multiple antibiotic resistance locus (Grkovic et al. 2002; Seoane and Levy 1995). This regulon consists of the *marRAB* operon and the *marC* gene. MarR acts as a repressor by binding as a dimer to promoter regions of the *mar* regulon (Martin and Rosner 1995). The repressive DNA-binding by MarR can be inhibited by several anionic compounds, e.g. salicylate (Alekshun and Levy 1999).

AtxA is a proven regulator of anthrax toxin genes (Dai et al. 1995; Koehler et al. 1994; Uchida et al. 1993). It is also known to influence the expression of other genes on pXO1, pXO2 plasmids and the anthrax genome (Bourgogne et al. 2003). AtxA is a member of a large, PTS (the phosphoenolpyruvate-dependent, sugar transporting phosphotransferase system) regulatory domain-containing family (Greenberg et al. 2002). Members of this family usually have a duplicated DNA/RNA binding domain and also duplicated PTS regulatory domain. Different variants of this structure are known, and additional domains are often present. Most probably, the presence of PTS EII homology domains is the necessity to act as an activator, since these domains are lacking in antiterminators (Greenberg et al. 2002). Because of its structure (Fig. 4), AtxA is believed to be a transcriptional activator. Knowing the architecture of this family, we searched the whole anthrax genome in order to find all similar regulators. Among the ones we found (Fig. 4), apart from the obvious AtxA and AcpA proteins, there is a very recent confirmation of the regulatory activity of the BXB0060 (pXO2-53), named AcpB (Drysdale et al. 2004). Diversity of domain composition and subtle

structural differences in the group of evolutionary related anthrax regulators are certainly elements of a very fine regulation of stages of infection.

BXA0178 (pXO1-105) belongs to the AbrB family of "transition state regulators." AbrB was first described in *Bacillus subtilis* as an activator and repressor of numerous genes during transitions in growth phase (Phillips and Strauch 2002). Recently, Saile and Koehler (Saile and Koehler 2002) showed that the genomic copy of AbrB in *B.anthracis* regulates the expression of three toxin genes, whereas the truncated pXO1 version (BXA0178) of AbrB does not affect toxin gene expression. We can speculate then that the truncation could be crucial for BXA0178 function, or its influence on pXO1 function is not yet understood.

According to FFAS analysis, BXA0180 is an N-terminal part of the lambda repressor-like DNA-binding domain superfamily (a.35.1), as classified by the SCOP database (Andreeva et al. 2004). The ORF is truncated after the first half, and experiments are needed to check whether a shortened domain can exert any function.

BXA0206 (pXO1-137) belongs to a large family of Hfq proteins. Members of this family are known to be involved in various metabolic processes, like the regulation of iron metabolism (Masse and Gottesman 2002; Wachi et al. 1999), mRNA stability (Vytvytska et al. 1998), stabilization and degradation of RNAs (Takada et al. 1999; Tsui et al. 1997). Hfq proteins are similar to eukaryotic Sm proteins involved in RNA splicing (Moller et al. 2002). The function of the pXO1 version is not known and the RNA targeted by BXA0206 is not recognized. The question remains whether BXA0206 acts on an RNA encoded by the plasmid itself or has another function, e.g. acts on a chromosomal small RNA or disguises as the human Sm protein.

### *Interesting ORFs from the "pathogenic" region*

The "pathogenic" region is defined as extending from BXA0057 to BXA0191 (Okinaka et al. 1999; Sirard et al. 2000), and is obviously of special interest.

**BXA0139: an ORF implicated in Hemolysis?**

The BXA0139 (pXO1-124) protein is located close to the oedema factor (CyaA) on the pXO1 sequence. It is 150 amino acids long, located on an operon with two unknown hypothetical proteins, BXA0138 (pXO1-125) and BXA0140 (pXO1-123). The only known fact about these proteins is the similarity of BXA0138 to BXA0149 (pXO1-117) (Supplementary data).

The most interesting finding is the homology of BXA0139 to the C-terminal end of the hemolysin II from *B.cereus* (Miles et al. 2002). This homology has already been described by Miles *et al.* (2002), but only as a similarity to a 46-amino acid segment of BXA0139. In reality, however, BXA0139 is a duplication of the same fragment, and C-end of hemolysin II is similar to both the N- and C-terminal parts of BXA0139 (Fig. 3). The significance of the C-terminus of the hemolysin II in *B.cereus* is unknown, and the functional studies suggest it has no influence on the hemolytic activity of the enzyme (Baida et al. 1999; Miles et al. 2002). Hemolysins form heptameric rings (Gouaux et al. 1997; Song et al. 1996), in which the C-terminal domain would reside in the outside part of each monomer (Miles et al. 2002). Miles and colleagues (2002) suggest three possible functions for this domain, however they do not exclude other possibilities. Either it is needed to form lattices or bind to surfaces, or has some catalytic activity. We also hypothesize an auxiliary function for the main monomer domain, maybe a regulatory function. Quite peculiar is the presence of a tandem tail-to-head repeat coded by the pXO1 plasmid. It is not fused to any catalytic domain and no overall function for the whole operon is known. The most attractive hypothesis would be the binding to surfaces. Maybe it serves as an anchor to the host cell membrane during the attack?

An interesting finding can maybe give a clue to a real function of BXA0139. We found a hemolysin II homolog in *B.anthracis* genome (gi: 21400399) that is almost identical to the *B.cereus* enzyme. However, in all anthrax strains sequenced, there is a nonsense mutation (TGG to TGA), instead of tryptophan 372 in *B.cereus*. In order to improve on the prediction of the encoded peptide, we ran the BLASTX program using the genomic sequence with large overhangs on both sides of the recognized ORF. The resulting sequence is given in the alignment in Figure 3. So, if the anthrax mutation is real (and its existence in all anthrax strains

seems to reinforce this notion), we can hypothesize that BXA0139 is auxiliary to the hemolysin's function of the genomic copy of hemolysin.

**Reverse homology of BXA0167**

This hypothetical ORF (pXO1-108) has no identifiable homologs. Its function is also not known. It is a product of automatic translation. We could assume then that it is not an interesting target for analysis.

We performed a BLASTX analysis along its sequence and found an interesting homology coded by the opposite strand. Interspersed with nonsense mutations, we found a strong homology to the N-terminus of the lethal factor (corresponding to 9-176 amino acids of LEF)(data not shown). Noticeably, this homology region is encoded by the opposite strand from the LEF gene. Is it an example of a duplication event covered up by other events that happened later in the course of evolution? Was the part of the N-terminal LEF domain functional in the past?

## DISCUSSION

In our work we described many novel features of the pXO1 plasmid that were not noticed previously. For instance, we show that parts of pXO1 are not only related to other bacilli plasmids, but also to proteins from more distant species. One of the most unexpected findings was the realization that pXO1 possesses two operons with homology to type IV secretion and pilus assembly systems. It is surprising because the type IV system is found mainly in Gram-negative bacteria (Burns 2003). Only some elements of the pilus are present in some Gram-positive bacteria (Grohmann et al. 2003; Wall and Kaiser 1999). It is even more surprising that the operons are not complete. A tempting hypothesis, which should be tested experimentally, is that the proteins present in pX01 constitute a minimal set indispensable for the formation and function of the secretion. Alternatively, these operons may have drifted from the original function. Cases both of minimal functional units, and of drift from original function, are known in pathogens and symbionts. The discovery

of type IV secretion system has the potential for a significant impact on our understanding of anthrax virulence: a new pathogenic delivery pathway can be of major importance in the invasion process.

The similarity to other various bacteria and copying of parts of operons shows the phylogenetic kaleidoscope nature of this megaplasmid. Apparently, this killing agent has developed by collecting genomic pieces from a very broad range of bacteria, including pathogenicity agents as well as other organisms. Some of these pieces may be non-functional (at least in their original way) or not related to anthrax pathogenicity. It is worth noting that pXO1 shares similarity with other pathogenic bacteria also in regions *not* previously recognized as a part of the pXO1 pathogenicity island (see the operon preservation with *Burkholderia* and *Xanthomonas* in Results), whose status may have to be revised.

A detailed analysis of the pXO1 sequence by Okinaka *et al.* (Okinaka et al. 1999) focused mostly on the analysis of mobile elements, their number and possible implication for the evolution of the plasmid. Our findings not only suggest a thorough history of transposition but also allow us to hypothesize on the probable entities that were used to build pXO1. Interestingly, even if the type IV clusters are located inside the putative PAI, one can guess it was an indispensable part of the plasmid sequence, however the presence of the IS DD-E transposases suggest it is a new, independent insertion. Another option would be that we deal with a conjugative transposon, unusually equipped with a set of DD-E transposases instead of Tyr or Ser recombinases. The important question to understand pXO1 as a mobile entity is to localize the replication machinery. We were unable to find it, which makes this even more intriguing, however we identified the putative replication start and termination sites. The nature of replication should be informative on the nature and provenience of the pXO1 plasmid.

The discovery of previously unknown systems on pXO1 plasmid of course begs questions about their regulation. External signals, cell state or host-pathogen interaction certainly trigger bacterial response(s), and several of them are already known (for review see (Koehler 2002)). All these signals finally activate transcription of virulence-related genes. We have attempted to describe all possible regulators that we could find, using sensitive profile-profile alignment programs. Some of the regulatory proteins are known not to

influence the toxin function (e.g. the homologue of AbrB), but others form priority targets for experimental studies of pathogenicity and *B.anthracis* biology. Notably, do the newly discovered factors regulate plasmid genes or chromosome genes?

We don't know how important is the presence of a common motif for AtxA-regulated genes. Its variable location throughout the putative promoter regions (closer or further to the ATG) poses questions. However, there may be ORFs not yet recognized 5' from the ones that are AtxA-dependent. In this case, the recognized ANGGAG sequence would directly precede the operon. Deletion experiments are needed to test whether these *cis* elements have any impact on the function of AtxA-regulated genes.

Another interesting finding is the diversity of ArsR homologs in *B.anthracis*. The majority of these, including those on pXO1, are related to the activator subfamily. The functions of MarR and TetR regulators are also intriguing.

There are two striking features of the whole plasmid that brought our special attention. First, the presence of so many DNA metabolism-related proteins (15%)(Supplementary data). It seems that DNA is a central point of the function of pXO1. Is this function related with the processing of pXO1, chromosomal DNA, transposons, or host DNA? None of these hypotheses can be excluded at the moment. The type IV delivery system could be an indication that some of them could have an external function. Second, when analyzing the DNA and proteome of pXO1 we realized how messy it is. pXO1 is full of incomplete and mutated ORFs (see Results and Supplementary data). There are many traces of ancient duplications, some still fresh (strong homology), but some almost completely faded away (homology barely recognizable), and often disrupted. It also consists of ORFs "borrowed" from other species. pXO1 seems to be the subject of constant evolutionary flux. The pXO1 plasmid should have a tag: "under construction."

## METHODS

### Gene names

The pXO1 plasmid was sequenced at least twice, by two independent research groups. Interestingly, the two sequences differ significantly, both on the DNA and on the (predicted) protein level. The second more recent sequencing identified almost 100 additional genes on pXO1. Several alternative naming conventions for *B.anthracis* plasmid proteins are used in literature. We use the names used by the pXO1 sequencing team (Read et al. 2003) (e.g. BXA007) as our primary names, but where appropriate we also provide the names used by the previous sequencing team (e.g. pXO1-04) or common gene names used in the literature (e.g. AtxA) when available.

### DNA level analysis

The *Bacillus anthracis* strain A2012 pXO1 plasmid sequence was used for analysis (accession: NC_003980)(Read et al. 2003).

We used the Oriloc (Frank and Lobry 2000) program to detect pXO1 origin of replication, using the gene coordinates provided in pXO1 Genbank file.

For the analysis of common DNA features in promoter regions of AtxA-dependent genes (Bourgogne et al. 2003), we used the total DNA sequences between the end of a previous gene and the ATG neighbourhood of the AtxA-regulated gene. We used the 5' regions of the following genes from pXO1 and pXO2 plasmids: BXA0019 (pXO1-13), BXA0124 (pXO1-90), BXA0125 (pXO1-91), BXA0137 (pXO1-126), BXA0142 (cyaA), BXA0164 (pagA), BXA0172 (lef), BXB0045 (pXO1-31), BXB0060 (pXO1-40), BXB0066 (pXO1-58), BXB0074, BXB0084 (pXO1-124). We used MEME and MITRA programs to search for common motifs (Bailey and Elkan 1994; Eskin and Pevzner 2002).

### Protein level analysis

For the analysis of the pXO1 proteome, we used proteins accessible with the BXAxxxx NCBI numbers, enforced with the BLASTX analysis (Altschul et al. 1990).

To analyze the protein sequences, we used the following programs: BLAST tools (Altschul et al. 1990; Altschul et al. 1997), SMART tool (Letunic et al. 2002), Pfam (Bateman et al. 2002), CDD (Marchler-Bauer et al. 2003), TMHMM2.0 (Sonnhammer et al. 1998), SEED (Read et al. 2003), Radar (Heger and Holm 2000), FFAS03 (Rychlewski et al. 2000), Metaserver.pl (Ginalski et al. 2003), Superfamily (Gough et al. 2001).

To align sequences we used: T-COFFEE (Notredame et al. 2000), AliBee (Nikolaev et al. 1997), MultAlin (Corpet 1988), BioEdit (Hall 1999).

Phylogenetic trees were estimated from amino acid alignments using PHYML (Guindon and Gascuel 2003), a fast and accurate Maximum Likelihood heuristic, under the JTT substitution model (Jones, Taylor et al. 1992), with a gamma distribution of rates between sites (eight categories, parameter alpha estimated by PHYML). Bootstrap support of branches was estimated using the programs SEQBOOT and CONSENSE of the PHYLIP package (Felsenstein 2002) with 1000 replicates; the parameter alpha was estimated independently for each repetition.

## ACKNOWLEDGMENTS

We thank Dr. Andrei Osterman for help in the context analysis and Stephane Guindon for help with PHYML. This work was supported by a United States Army Medical Research and Materiel Command Grant DAMD17-03-2-0038.

## REFERENCES

Ahmed, M., L. Lyass, P.N. Markham, S.S. Taylor, N. Vazquez-Laslop, and A.A. Neyfakh. 1995. Two highly similar multidrug transporters of Bacillus subtilis whose expression is differentially regulated. *J Bacteriol* **177:** 3904-3910.

Alekshun, M.N. and S.B. Levy. 1999. Alteration of the repressor activity of MarR, the negative regulator of the Escherichia coli marRAB locus, by multiple chemicals in vitro. *J Bacteriol* **181:** 4669-4672.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Altschul, S.F. and E.V. Koonin. 1998. Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. *Trends Biochem Sci* **23:** 444-447.

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-3402.

Andreeva, A., D. Howorth, S.E. Brenner, T.J. Hubbard, C. Chothia, and A.G. Murzin. 2004. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* **32 Database issue:** D226-229.

Ariel, N., A. Zvi, H. Grosfeld, O. Gat, Y. Inbar, B. Velan, S. Cohen, and A. Shafferman. 2002. Search for potential vaccine candidate open reading frames in the Bacillus anthracis virulence plasmid pXO1: in silico and in vitro screening. *Infect Immun* **70:** 6817-6827.

Ariel, N., A. Zvi, K.S. Makarova, T. Chitlaru, E. Elhanany, B. Velan, S. Cohen, A.M. Friedlander, and A. Shafferman. 2003. Genome-based bioinformatic selection of chromosomal Bacillus anthracis putative vaccine candidates coupled with proteomic identification of surface-associated antigens. *Infect Immun* **71:** 4563-4579.

Baida, G., Z.I. Budarina, N.P. Kuzmin, and A.S. Solonin. 1999. Complete nucleotide sequence and molecular characterization of hemolysin II gene from Bacillus cereus. *FEMS Microbiol Lett* **180:** 7-14.

Bailey, T.L. and C. Elkan. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2:** 28-36.

Baranova, N.N., A. Danchin, and A.A. Neyfakh. 1999. Mta, a global MerR-type regulator of the Bacillus subtilis multidrug-efflux transporters. *Mol Microbiol* **31:** 1549-1559.

Bateman, A., E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S.R. Eddy, S. Griffiths-Jones, K.L. Howe, M. Marshall, and E.L. Sonnhammer. 2002. The Pfam protein families database. *Nucleic Acids Res* **30:** 276-280.

Bourgogne, A., M. Drysdale, S.G. Hilsenbeck, S.N. Peterson, and T.M. Koehler. 2003. Global effects of virulence gene regulators in a Bacillus anthracis strain with both virulence plasmids. *Infect Immun* **71:** 2736-2743.

Braun, L. and P. Cossart. 2000. Interactions between Listeria monocytogenes and host mammalian cells. *Microbes Infect* **2:** 803-811.

Burns, D.L. 2003. Type IV transporters of pathogenic bacteria. *Curr. Opin. Microbiol.* **6:** 29-34.

Christie, P.J. 2001. Type IV secretion: intercellular transfer of macromolecules by systems ancestrally related to conjugation machines. *Mol Microbiol* **40:** 294-305.

Christie, P.J. and J.P. Vogel. 2000. Bacterial type IV secretion: conjugation systems adapted to deliver effector molecules to host cells. *Trends Microbiol* **8:** 354-360.

Corpet, F. 1988. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res* **16:** 10881-10890.

Dai, Z., J.C. Sirard, M. Mock, and T.M. Koehler. 1995. The atxA gene product activates transcription of the anthrax toxin genes and is essential for virulence. *Mol Microbiol* **16:** 1171-1181.

Dang, T.A., X.R. Zhou, B. Graf, and P.J. Christie. 1999. Dimerization of the Agrobacterium tumefaciens VirB4 ATPase and the effect of ATP-binding cassette mutations on the assembly and function of the T-DNA transporter. *Mol Microbiol* **32:** 1239-1253.

Drysdale, M., A. Bourgogne, S.G. Hilsenbeck, and T.M. Koehler. 2004. atxA Controls Bacillus anthracis Capsule Synthesis via acpA and a Newly Discovered Regulator, acpB. *J Bacteriol* **186:** 307-315.

Eskin, E. and P.A. Pevzner. 2002. Finding composite regulatory patterns in DNA sequences. *Bioinformatics* **18 Suppl 1:** S354-363.

Fouet, A. and S. Mesnage. 2002. Bacillus anthracis cell envelope components. *Curr Top Microbiol Immunol* **271:** 87-113.

Frank, A.C. and J.R. Lobry. 2000. Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. *Bioinformatics* **16:** 560-561.

Gaur, N.K., E. Dubnau, and I. Smith. 1986. Characterization of a cloned Bacillus subtilis gene that inhibits sporulation in multiple copies. *J Bacteriol* **168:** 860-869.

Ginalski, K., A. Elofsson, D. Fischer, and L. Rychlewski. 2003. 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* **19:** 1015-1018.

Godsey, M.H., N.N. Baranova, A.A. Neyfakh, and R.G. Brennan. 2001. Crystal Structure of MtaN, a Global Multidrug Transporter Gene Activator. *J. Biol. Chem.* **276:** 47178-47184.

Gouaux, E., M. Hobaugh, and L. Song. 1997. alpha-Hemolysin, gamma-hemolysin, and leukocidin from Staphylococcus aureus: distant in sequence but similar in structure. *Protein Sci* **6:** 2631-2635.

Gough, J., K. Karplus, R. Hughey, and C. Chothia. 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* **313:** 903-919.

Greenberg, D.B., J. Stulke, and M.H. Saier, Jr. 2002. Domain analysis of transcriptional regulators bearing PTS regulatory domains. *Res Microbiol* **153:** 519-526.

Grkovic, S., M.H. Brown, and R.A. Skurray. 2002. Regulation of bacterial drug export systems. *Microbiol Mol Biol Rev* **66:** 671-701, table of contents.

Grohmann, E., G. Muth, and M. Espinosa. 2003. Conjugative plasmid transfer in gram-positive bacteria. *Microbiol Mol Biol Rev.* **67:** 277-301.

Grynberg, M., H. Erlandsen, and A. Godzik. 2003. HEPN: a common domain in bacterial drug resistance and human neurodegenerative proteins. *Trends Biochem Sci* **28:** 224-226.

Guindon, S. and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52:** 696-704.

Hacker, J. and J.B. Kaper. 2000. Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol* **54:** 641-679.

Hall, T.A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl Acids Symp Ser* **41:** 95-98.

Heger, A. and L. Holm. 2000. Rapid automatic detection and alignment of repeats in protein sequences. *Proteins* **41:** 224-237.

Hentschel, U. and J. Hacker. 2001. Pathogenicity islands: the tip of the iceberg. *Microbes Infect* **3:** 545-548.

Hillen, W. and C. Berens. 1994. Mechanisms underlying expression of Tn10 encoded tetracycline resistance. *Annu Rev Microbiol* **48:** 345-369.

Hoffmaster, A.R. and T.M. Koehler. 1999. Autogenous regulation of the Bacillus anthracis pag operon. *J Bacteriol* **181:** 4485-4492.

Icho, T., C.P. Sparrow, and C.R. Raetz. 1985. Molecular cloning and sequencing of the gene for CDP-diglyceride synthetase of Escherichia coli. *J Biol Chem* **260:** 12078-12083.

Jeanmougin, F., J.D. Thompson, M. Gouy, D.G. Higgins, and T.J. Gibson. 1998. Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.* **23:** 403-405.

Kachlany, S.C., P.J. Planet, M.K. Bhattacharjee, E. Kollia, R. DeSalle, D.H. Fine, and D.H. Figurski. 2000. Nonspecific adherence by Actinobacillus actinomycetemcomitans requires genes widespread in bacteria and archaea. *J Bacteriol* **182:** 6169-6176.

Kachlany, S.C., P.J. Planet, R. DeSalle, D.H. Fine, and D.H. Figurski. 2001. Genes for tight adherence of Actinobacillus actinomycetemcomitans: from plaque to plague to pond scum. *Trends Microbiol* **9:** 429-437.

Khan, S., W. Jarra, and P. Preiser. 2001. The 235 kDa rhoptry protein of Plasmodium (yoelii) yoelii: function at the junction. *Mol Biochem Parasitol.* **117.**

Kinch, L.N., J.O. Wrabl, S.S. Krishna, I. Majumdar, R.I. Sadreyev, Y. Qi, J. Pei, H. Cheng, and N.V. Grishin. 2003. CASP5 assessment of fold recognition target predictions. *Proteins* **53 Suppl 6:** 395-409.

Koehler, T.M. 2002. Bacillus anthracis genetics and virulence gene regulation. *Curr Top Microbiol Immunol* **271:** 143-164.

Koehler, T.M., Z. Dai, and M. Kaufman-Yarbray. 1994. Regulation of the Bacillus anthracis protective antigen gene: CO2 and a trans-acting element activate transcription from one of two promoters. *J Bacteriol* **176:** 586-595.

Krause, S., W. Pansegrau, R. Lurz, F. de la Cruz, and E. Lanka. 2000. Enzymology of type IV macromolecule secretion systems: the conjugative transfer regions of plasmids RP4 and R388 and the cag pathogenicity island of Helicobacter pylori encode structurally and functionally related nucleoside triphosphate hydrolases. *J Bacteriol* **182:** 2761-2770.

Lacy, D.B. and R.J. Collier. 2002. Structure and function of anthrax toxin. *Curr Top Microbiol Immunol* **271:** 61-85.

Lamark, T., T.P. Rokenes, J. McDougall, and A.R. Strom. 1996. The complex bet promoters of Escherichia coli: regulation by oxygen (ArcA), choline (BetI), and osmotic stress. *J Bacteriol* **178:** 1655-1662.

Lee, E.H., C. Rouquette-Loughlin, J.P. Folster, and W.M. Shafer. 2003. FarR Regulates the farAB-Encoded Efflux Pump of Neisseria gonorrhoeae via an MtrR Regulatory Mechanism. *J Bacteriol* **185:** 7145-7152.

Lee, E.H. and W.M. Shafer. 1999. The farAB-encoded efflux pump mediates resistance of gonococci to long-chained antibacterial fatty acids. *Mol Microbiol* **33:** 839-845.

Letunic, I., L. Goodstadt, N.J. Dickens, T. Doerks, J. Schultz, R. Mott, F. Ciccarelli, R.R. Copley, C.P. Ponting, and P. Bork. 2002. Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res* **30:** 242-244.

Lupas, A., M. Van Dyke, and J. Stock. 1991. Predicting coiled coils from protein sequences. *Science* **252:** 1162-1164.

Mandelkow, E. and E.M. Mandelkow. 2002. Kinesin motors and disease. *Trends Cell Biol* **12:** 585-591.

Mandic-Mulec, I., L. Doukhan, and I. Smith. 1995. The Bacillus subtilis SinR protein is a repressor of the key sporulation gene spo0A. *J Bacteriol* **177:** 4619-4627.

Mandic-Mulec, I., N. Gaur, U. Bai, and I. Smith. 1992. Sin, a stage-specific repressor of cellular differentiation. *J Bacteriol* **174:** 3561-3569.

Marchler-Bauer, A., J.B. Anderson, C. DeWeese-Scott, N.D. Fedorova, L.Y. Geer, S. He, D.I. Hurwitz, J.D. Jackson, A.R. Jacobs, C.J. Lanczycki, C.A. Liebert, C. Liu, T. Madej, G.H. Marchler, R. Mazumder, A.N. Nikolskaya, A.R. Panchenko, B.S. Rao, B.A. Shoemaker, V. Simonyan, J.S. Song, P.A. Thiessen, S. Vasudevan, Y. Wang, R.A. Yamashita, J.J. Yin, and S.H. Bryant. 2003. CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res* **31:** 383-387.

Martin, R.G. and J.L. Rosner. 1995. Binding of purified multiple antibiotic-resistance repressor protein (MarR) to mar operator sequences. *Proc Natl Acad Sci U S A* **92:** 5456-5460.

Masse, E. and S. Gottesman. 2002. A small RNA regulates the expression of genes involved in iron metabolism in Escherichia coli. *Proc Natl Acad Sci U S A* **99:** 4620-4625.

Matsumura, M., Y. Katakura, T. Imanaka, and S. Aiba. 1984. Enzymatic and nucleotide sequence studies of a kanamycin-inactivating enzyme encoded by a plasmid from thermophilic bacilli in comparison with that encoded by plasmid pUB110. *J Bacteriol* **160:** 413-420.

McIver, K.S. and R.L. Myles. 2002. Two DNA-binding domains of Mga are required for virulence gene activation in the group A streptococcus. *Mol Microbiol* **43:** 1591-1601.

Mignot, T., M. Mock, and A. Fouet. 2003. A plasmid-encoded regulator couples the synthesis of toxins and surface structures in Bacillus anthracis. *Mol Microbiol* **47:** 917-927.

Miles, G., H. Bayley, and S. Cheley. 2002. Properties of Bacillus cereus hemolysin II: a heptameric transmembrane pore. *Protein Sci* **11:** 1813-1824.

Moller, T., T. Franch, P. Hojrup, D.R. Keene, H.P. Bachinger, R.G. Brennan, and P. Valentin-Hansen. 2002. Hfq: a bacterial Sm-like protein that mediates RNA-RNA interaction. *Mol Cell* **9:** 23-30.

Morby, A.P., J.S. Turner, J.W. Huckle, and N.J. Robinson. 1993. SmtB is a metal-dependent repressor of the cyanobacterial metallothionein gene smtA: identification of a Zn inhibited DNA-protein complex. *Nucleic Acids Res* **21:** 921-925.

Nakai, K. and P. Horton. 1999. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci* **24:** 34-36.

Nikolaev, V.K., A.M. Leontovich, V.A. Drachev, and B.L. I. 1997. Building multiple alignment using iterative analyzing biopolymers structure dynamic improvement of the initial motif alignment. *Biochemistry* **62:** 578-582.

Notredame, C., D.G. Higgins, and J. Heringa. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302:** 205-217.

Okinaka, R., K. Cloud, O. Hampton, A. Hoffmaster, K. Hill, P. Keim, T. Koehler, G. Lamke, S. Kumano, D. Manter, Y. Martinez, D. Ricke, R. Svensson, and P. Jackson. 1999. Sequence, assembly and analysis of pXO1 and pXO2. *J Appl Microbiol* **87:** 261-262.

Okinaka, R.T., K. Cloud, O. Hampton, A.R. Hoffmaster, K.K. Hill, P. Keim, T.M. Koehler, G. Lamke, S. Kumano, J. Mahillon, D. Manter, Y. Martinez, D. Ricke, R. Svensson, and P.J. Jackson. 1999. Sequence and organization of pXO1, the large Bacillus anthracis plasmid harboring the anthrax toxin genes. *J Bacteriol* **181:** 6509-6515.

Page, R.D. 1996. TreeView: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci* **12:** 357-358.

Pannucci, J., R.T. Okinaka, R. Sabin, and C.R. Kuske. 2002. Bacillus anthracis pXO1 plasmid sequence conservation among closely related bacterial species. *J Bacteriol* **184:** 134-141.

Pasteur, L. 1881. Compte rendu sommaire des experiences faites a Pouilly-Le-Fort, pres de Meun, sur la vaccination charnonneuse (avec la collaboration de MM. Chamberland et Roux). *Compte Rendus Acad Sci* **XCII:** 1378-1383.

Paulsen, I.T., M.H. Brown, and R.A. Skurray. 1996. Proton-dependent multidrug efflux systems. *Microbiol Rev* **60:** 575-608.

Phillips, Z.E. and M.A. Strauch. 2002. Bacillus subtilis sporulation and stationary phase gene expression. *Cell Mol Life Sci* **59:** 392-402.

Ponting, C.P., L. Aravind, J. Schultz, P. Bork, and E.V. Koonin. 1999. Eukaryotic signalling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer. *J Mol Biol* **289:** 729-745.

Rasko, D.A., J. Ravel, O.K. OA, E. Helgason, R.Z. Cer, L. Jiang, K.A. Shores, D.E. Fouts, N.J. Tourasse, S.V. Angiuoli, J. Kolonay, W.C. Nelson, A.B. Kolsto, C.M. Fraser, and T.D. Read. 2004. The genome sequence of Bacillus cereus ATCC 10987 reveals metabolic adaptations and a large plasmid related to Bacillus anthracis pXO1. *Nucleic Acids Res* **32:** 977-988.

Read, T.D., S.N. Peterson, N. Tourasse, L.W. Baillie, I.T. Paulsen, K.E. Nelson, H. Tettelin, D.E. Fouts, J.A. Eisen, S.R. Gill, E.K. Holtzapple, O.A. Okstad, E. Helgason, J. Rilstone, M. Wu, J.F. Kolonay, M.J. Beanan, R.J. Dodson, L.M. Brinkac, M. Gwinn, R.T. DeBoy, R. Madpu, S.C. Daugherty, A.S. Durkin, D.H. Haft, W.C. Nelson, J.D. Peterson, M. Pop, H.M. Khouri, D. Radune, J.L. Benton, Y. Mahamoud, L. Jiang, I.R. Hance, J.F. Weidman, K.J. Berry, R.D. Plaut, A.M. Wolf, K.L. Watkins, W.C. Nierman, A. Hazen, R. Cline, C. Redmond, J.E. Thwaite, O. White, S.L. Salzberg, B. Thomason, A.M. Friedlander, T.M. Koehler, P.C. Hanna, A.B. Kolsto, and C.M. Fraser. 2003. The genome sequence of Bacillus anthracis Ames and comparison to closely related bacteria. *Nature* **423:** 81-86.

Rychlewski, L., L. Jaroszewski, W. Li, and A. Godzik. 2000. Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci* **9:** 232-241.

Sadreyev, R.I., D. Baker, and N.V. Grishin. 2003. Profile-profile comparisons by COMPASS predict intricate homologies between protein families. *Protein Sci* **12:** 2262-2272.

Saile, E. and T.M. Koehler. 2002. Control of anthrax toxin gene expression by the transition state regulator abrB. *J Bacteriol* **184:** 370-380.

Savvides, S.N., H.J. Yeo, M.R. Beck, F. Blaesing, R. Lurz, E. Lanka, R. Buhrdorf, W. Fischer, R. Haas, and G. Waksman. 2003. VirB11 ATPases are dynamic hexameric assemblies: new insights into bacterial type IV secretion. *Embo J* **22:** 1969-1980.

Seoane, A.S. and S.B. Levy. 1995. Characterization of MarR, the repressor of the multiple antibiotic resistance (mar) operon in Escherichia coli. *J Bacteriol* **177:** 3414-3419.

Sirard, J.C., C. Guidi-Rontani, A. Fouet, and M. Mock. 2000. Characterization of a plasmid region involved in Bacillus anthracis toxin production and pathogenesis. *Int J Med Microbiol* **290:** 313-316.

Skerker, J.M. and L. Shapiro. 2000. Identification and cell cycle control of a novel pilus system in Caulobacter crescentus. *Embo J* **19:** 3223-3234.

Song, L., M.R. Hobaugh, C. Shustak, S. Cheley, H. Bayley, and J.E. Gouaux. 1996. Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. *Science* **274:** 1859-1866.

Sonnhammer, E.L., G. von Heijne, and A. Krogh. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol* **6:** 175-182.

Sparrow, C.P. and C.R. Raetz. 1985. Purification and properties of the membrane-bound CDP-diglyceride synthetase from Escherichia coli. *J Biol Chem* **260:** 12084-12091.

Stock, J.B., M.N. Levit, and P.M. Wolanin. 2002. Information processing in bacterial chemotaxis. *Sci STKE* **2002:** PE25.

Summers, A.O. 1992. Untwist and shout: a heavy metal-responsive transcriptional regulator. *J Bacteriol* **174:** 3097-3101.

Takada, A., M. Wachi, and K. Nagai. 1999. Negative regulatory role of the Escherichia coli hfq gene in cell division. *Biochem Biophys Res Commun* **266:** 579-583.

Tsui, H.C., G. Feng, and M.E. Winkler. 1997. Negative regulation of mutS and mutH repair gene expression by the Hfq and RpoS global regulators of Escherichia coli K-12. *J Bacteriol* **179:** 7476-7487.

Turnbull, P.C. 2002. Introduction: anthrax history, disease and ecology. *Curr Top Microbiol Immunol* **271:** 1-19.

Uchida, I., J.M. Hornung, C.B. Thorne, K.R. Klimpel, and S.H. Leppla. 1993. Cloning and characterization of a gene whose product is a trans-activator of anthrax toxin synthesis. *J Bacteriol* **175:** 5329-5338.

Vytvytska, O., J.S. Jakobsen, G. Balcunaite, J.S. Andersen, M. Baccarini, and A. von Gabain. 1998. Host factor I, Hfq, binds to Escherichia coli ompA mRNA in a growth rate-dependent fashion and regulates its stability. *Proc Natl Acad Sci U S A* **95:** 14118-14123.

Wachi, M., A. Takada, and K. Nagai. 1999. Overproduction of the outer-membrane proteins FepA and FhuE responsible for iron transport in Escherichia coli hfq:cat mutant. *Biochem Biophys Res Commun* **264:** 525-529.

Wall, D. and D. Kaiser. 1999. Type IV pili and cell motility. *Mol Microbiol.* **32:** 1-10.

Williams, S.G., S.R. Attridge, and P.A. Manning. 1993. The transcriptional activator HlyU of Vibrio cholerae: nucleotide sequence and role in virulence gene expression. *Mol Microbiol* **9:** 751-760.

Williams, S.G., L.T. Varcoe, S.R. Attridge, and P.A. Manning. 1996. Vibrio cholerae Hcp, a secreted protein coregulated with HlyA. *Infect Immun* **64:** 283-289.

Woehlke, G. and M. Schliwa. 2000. Walking on two heads: the many talents of kinesin. *Nat Rev Mol Cell Biol* **1:** 50-58.

Wu, J. and B.P. Rosen. 1991. The ArsR protein is a trans-acting regulatory protein. *Mol Microbiol* **5:** 1331-1336.

Yeo, H.J., S.N. Savvides, A.B. Herr, E. Lanka, and G. Waksman. 2000. Crystal structure of the hexameric traffic ATPase of the Helicobacter pylori type IV secretion system. *Mol Cell* **6:** 1461-1472.

## FIGURE LEGENDS

**Figure 1**. A summary of the distribution of homologs of the predicted proteins (ORFs) encoded in pXO1 plasmid in a set of >100 diverse microbial genomes. Only relatively close homologues (with FASTA P-score above 10-3) were taken into account at this stage of the analysis. Relative size and polarity of ORFs (using the predictions and the nomenclature by TIGR) on the linearized map of pXO1 are illustrated by the heights (cutoff at 500 amino acids) and orientation of the bars along the X-axis (panel A, continued on panel

B). Open bars correspond to proteins for which no homologues have been detected in this analysis. Bars with matching colored borders correspond to "repeats" present in pXO1. Black and colored bars in correspond to proteins for which at least one homolog was detected in this analysis.

Panel C (and its continuation in panel D) mark the presence of respective homologues in at least one of the representative genomes in several groups (as indicated in respective boxes):

Group 1: *B. anthracis* (chromosome or pXO2), *B. thuringiensis* or *B. cereus*.

Group 2: *B. subtilis*, *B. halodurans* or *B. stearothermophilus*.

Group 3: *Staphylococci*, *Streptococci* or *Eneterococci* species.

Group 4: *Salmonella*, *Xanthomonas* or *Burkholderia* species.

Group 5: *Geobacter*, *Anabaena* or *Nostoc* species.

These genomes contain the largest number of homologues of pXO1-borne proteins, and jointly they provide a nearly complete coverage of the phylogenetic space of pXO1 homologues.

**Figure 2**. Type IV secretion and pilus systems representations with homologous genes in *B.anthracis* shown in red. It is worth noting that in the secretion operon representation, the anthrax VirB6 gene is fused to an adhesin-like long sequence, whereas in the pilus assembly operon the last homologue, TadC, has two representations in the anthrax operon. For more detailed comparison to known type IV secretion and pilus assembly systems, see (Christie 2001; Christie and Vogel 2000; Kachlany et al. 2000; Kachlany et al. 2001; Skerker and Shapiro 2000).

**Figure 3**. The multiple alignment of the *Bacillus cereus* terminal hemolysin II domain, two parts of the BXA0139/pXO1-124 protein, the *Streptococcus* phage Cp-1 orf16 and the *B.anthracis* hemolysin II copy with a truncated C terminus. The star represents the stop codon in the anthrax DNA sequence.

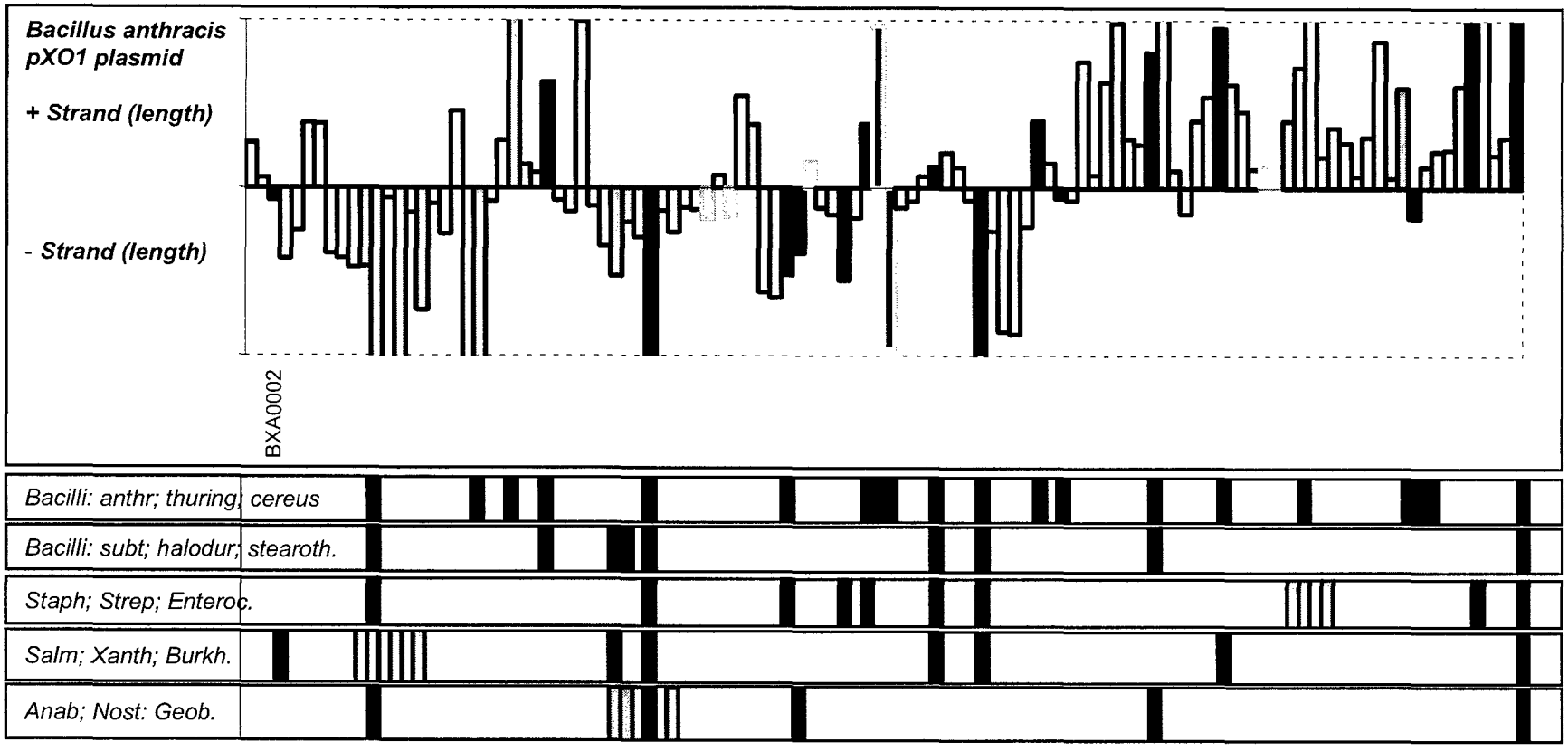
**Figure 4**. The domain structure of the AtxA family of protein from *B.anthracis*. Each colour depicts a family of most homologous sequences. Similar colours describe duplicated sequences.

**Figure 5**. Phylogenetic trees of ArsR/SmtB proteins.

Phylogenies estimated using PHYML (Guindon and Gascuel 2003). Figures at nodes are bootstrap support in % of 1000 replicates; bootstrap proportions under 50% are not reported. Branch length is proportional to the estimated number of substitutions per site. Proteins from pXO1 are boxed.

(A) Phylogeny of representative proteins sampling the diversity of the ArsR/SmtB family. Two *B.anthracis* proteins with short sequences are not included (Q81NE6 and Q81QQ6). Unrooted tree drawn using TreeView (Page 1996); the measure bar represents 0.1 substitutions/site. The boxes indicate clades (monophyletic groups) discussed in the text.

(B) Phylogeny of pXO1 ArsR/SmtB proteins and close homologues. This corresponds to the box "close homologs of pXO1 proteins" in (A), plus all closely related homologs as determined from a phylogeny of all available ArsR/SmtB sequences (487 sequences; tree not shown). Tree rooted according to the phylogeny of all ArsR/SmtB proteins, and drawn using NJplot (Jeanmougin et al. 1998); the measure bar represents 0.5 substitutions/site. Full circles indicate gene duplications in the common ancestor of *B.antracis* and *B.cereus*; the empty circle indicates a gene duplication in *B.thuringiensis*.


Bacillus anthracis
pXO1 plasmid
+ Strand (length)
- Strand (length)
BXA0002
Bacilli: anthr; thuring; cereus
Bacilli: subt; halodur; stearoth.
Staph; Strep; Enteroc.
Salm; Xanth; Burkh.
Anab; Nost: Geob.

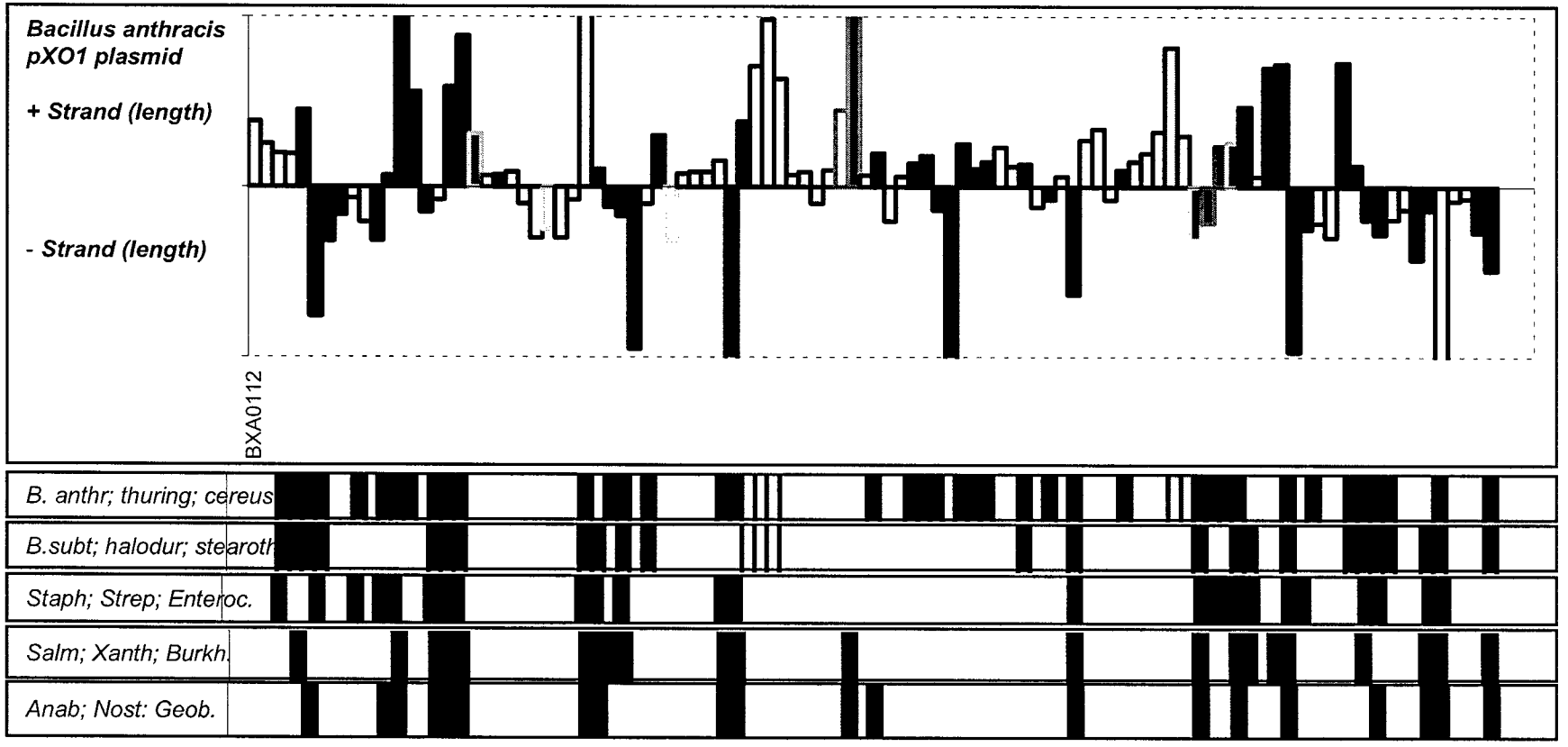

Bacillus anthracis
pXO1 plasmid
+ Strand (length)
- Strand (length)
BXA0112
B. anthr; thuring; cereus
B.subt; halodur; stearoth
Staph; Strep; Enteroc.
Salm; Xanth; Burkh.
Anab; Nost: Geob.

Figure 2

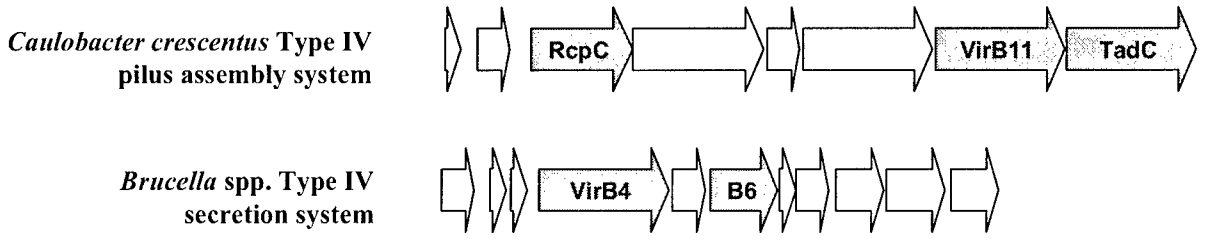

*Caulobacter crescentus* Type IV pilus assembly system
RcpC
VirB11
TadC
*Brucella* spp. Type IV secretion system
VirB4
B6

Figure 3

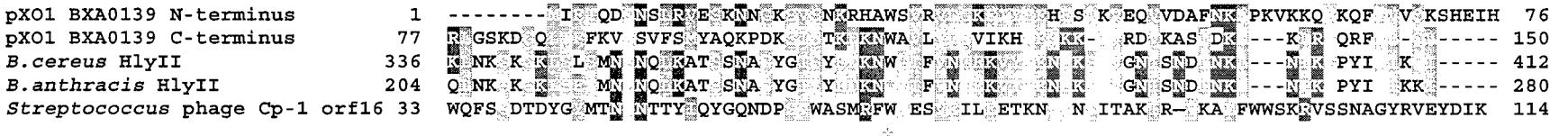
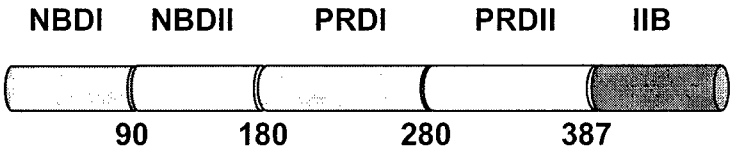
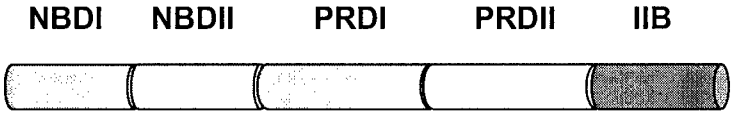

pXO1 BXA0139 N-terminus 1 76
pXO1 BXA0139 C-terminus 77 150
B.cereus HlyII 336 412
B.anthracis HlyII 204 280
Streptococcus phage Cp-1 orf16 33 114

Figure 4

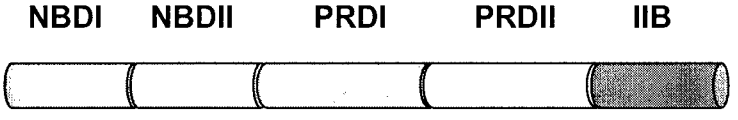
**AtxA**



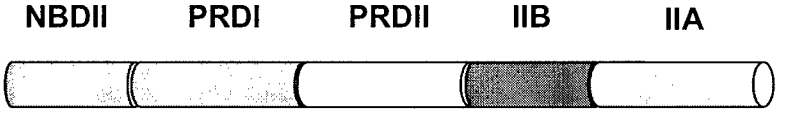

**AcpB-pXO2-53**



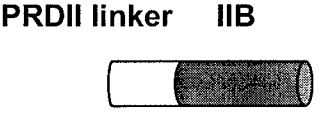

**AcpA-pXO2-63**




**PTS_EIIA_2 (gi 21398751)**



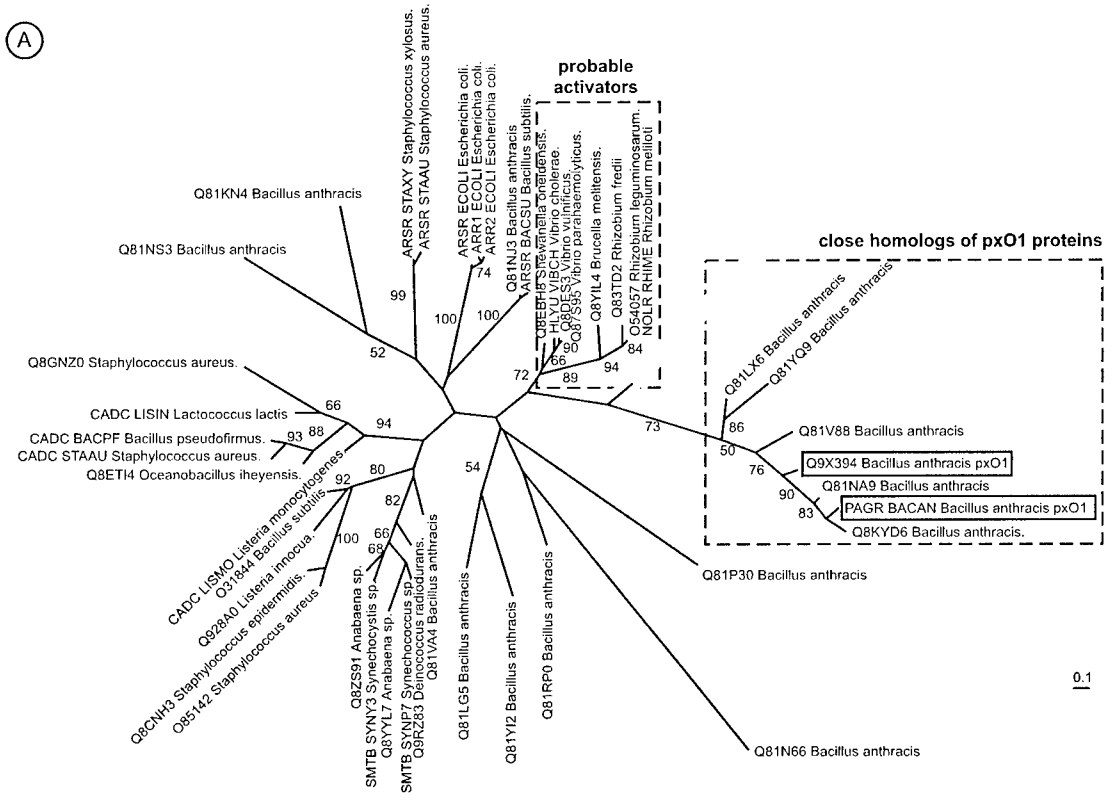

**pXO2-62**




**BglG family (gi 21402105)**


| | |
|---|---|
| **NBDI** | - nucleic acid binding domain I |
| **NBDII** | - nucleic acid binding domain II |
| **PRDI** | - PTS regulatory domain I |
| **PRDII** | - PTS regulatory domain II |
| **IIB** | - PTS EIIB homology domain |
| **IIA** | - PTS EIIA homology domain |
| **CAT RBD** | - Co-AntiTerminator RNA binding domain |

Figure 5


A
probable activators
close homologs of pxO1 proteins
ARSR STAXY Staphylococcus xylosus.
ARSR STAAU Staphylococcus aureus.
ARSR ECOLI Escherichia coli.
ARR1 ECOLI Escherichia coli.
ARR2 ECOLI Escherichia coli.
Q81NJ3 Bacillus anthracis
ARSR BACSU Bacillus subtilis.
Q8EFT5 Shewanella oneidensis.
HLYU VIBCH Vibrio cholerae.
Q8DES3 Vibrio vulnificus.
Q87S95 Vibrio parahaemolyticus.
Q8YIL4 Brucella melitensis.
Q83TD2 Rhizobium fredii
Q54057 Rhizobium leguminosarum.
NOLR RHIME Rhizobium meliloti
Q81KN4 Bacillus anthracis
Q81NS3 Bacillus anthracis
Q8GNZ0 Staphylococcus aureus.
CADC LISIN Lactococcus lactis
CADC BACPF Bacillus pseudofirmus.
CADC STAAU Staphylococcus aureus.
Q8ETI4 Oceanobacillus iheyensis.
CADC LISMO Listeria monocytogenes
O31844 Bacillus subtilis
Q928A0 Listeria innocua
Q8CNH3 Staphylococcus epidermidis.
O85142 Staphylococcus aureus
Q8ZS91 Anabaena sp.
SMTB SYNY3 Synechocystis sp.
Q8YYL7 Anabaena sp.
SMTB SYNP7 Synechococcus sp.
Q9RZ83 Deinococcus radiodurans.
Q81VA4 Bacillus anthracis
Q81LG5 Bacillus anthracis
Q81YI2 Bacillus anthracis
Q81RP0 Bacillus anthracis
Q81N66 Bacillus anthracis
Q81P30 Bacillus anthracis
Q81LX6 Bacillus anthracis
Q81YQ9 Bacillus anthracis
Q81V88 Bacillus anthracis
Q9X394 Bacillus anthracis pxO1
Q81NA9 Bacillus anthracis
PAGR BACAN Bacillus anthracis pxO1
Q8KYD6 Bacillus anthracis.
0.1
B
Q8NNC5 Corynebacterium glutamicum
Q8FNG5 Corynebacterium efficiens.
Q9CCG5 Mycobacterium leprae.
O05840 Mycobacterium tuberculosis.
Q7TYR7 Mycobacterium bovis.
Q8TJ36 Methanosarcina acetivorans.
Q8PUU2 Methanosarcina mazei
O57801 Pyrococcus horikoshii.
Q9V2K4 Pyrococcus abyssi.
Q8TZE4 Pyrococcus furiosus.
Q7NPQ3 Gloeobacter violaceus.
Q7UU43 Rhodopirellula baltica.
Q8KBQ6 Chlorobium tepidum.
Q8ABI9 Bacteroides thetaiotaomicron.
Q8A9L8 Bacteroides thetaiotaomicron.
Q87IV9 Vibrio parahaemolyticus
Q984W4 Rhizobium loti
Q8KNT2 Bacillus thuringiensis
Q8KNP7 Bacillus thuringiensis
Q81V88 Bacillus anthracis
Q81I12 Bacillus cereus
Q81NA9 Bacillus anthracis
PAGR BACAN Bacillus anthracis pxO1
Q8KYD6 Bacillus anthracis.
Q9X394 Bacillus anthracis pxO1
Q81YQ9 Bacillus anthracis
Q81AX4 Bacillus cereus
Q81LX6 Bacillus anthracis
Q818J4 Bacillus cereus
duplications in Bacillus
Q9I5Y9 Pseudomonas aeruginosa.
Q88D61 Pseudomonas putida
Q88AK6 Pseudomonas syringae
Q8VML1 Pseudomonas putida.
Q7V3Y8 Prochlorococcus marinus
Q8YQS6 Anabaena sp.
Q8YQD0 Anabaena sp.
Q8DKT6 Synechococcus elongatus
Q8KEP0 Chlorobium tepidum.
Q9AIS2 Anabaena sp.
Q9KI49 Oscillatoria limnetica.
Q9KI48 Aphanothece halophytica.
0.5

# VirFact: a relational database of virulence factors and pathogenicity islands (PAIs).

**Adrian Tkacz, Leszek Rychlewski and Adam Godzik**

## ABSTRACT

The VirFact database (http://virfact.burnham.org) contains information on microbial virulence factors and pathogenicity islands (PAIs) from major pathogens. The database collects information from literature and combines them with results obtained by genome context analysis and distant homology recognition. The database can be browsed by virulence factor, PAI or organism name. The annotations, including multiple alignments of proteins homologous to virulence factors, genomic context, models of three dimensional structures (if available) are presented using graphical web interface and standard visualization tools. The VirFact can also be used as a tool to recognize the presence of homologs of known virulence factors in the genome delivered by the user.

## INTRODUCTION

Recent development of comparative genomic analysis and experimental molecular biological techniques made it possible to identify specific genes responsible for virulence of pathogenic microbes. Despite some discussions (1), it is widely accepted that virulence of a pathogenic microbe is imparted by a specific set of genes, often localized together on a plasmid (virulence plasmids) or on a genome (pathogenicity islands). Virulence factors are typically identified by comparing genomic sequences of pathogenic and non-pathogenic strains or by studying virulence of deletion mutants. While building VirFact we adhered to a

broad definition of a virulence factor that includes genes specifically involved in interactions between a pathogen and its host, but also genes supporting pathogenic lifestyle and many genes of unknown function if they are part of the genomic structure related to pathogenicity. Virulence factors of many organisms are well studied, but the information about them is usually available only is specialized literature and then usually only in the context of a specific organism. We believe that this scattering of information makes it difficult to study general questions involving pathogenicity, such as for instance similarity between virulence apparatus of unrelated pathogens. At the same time, sequence analysis and annotations of many virulence related genes is very uneven and tools such as distant homology analysis, fold recognition or modeling are seldom used. The goal of the VirFact project is the development of a well annotated database containing information about pathogenicity systems from different organisms and providing a uniform level of annotation, including annotations with most sensitive algorithms.

## THE DATABASE

The VirFact database (http://virfact.burnham.org) is implemented as a relational database containing a collection of virulence factors and pathogenicity islands from major microbial pathogens. The current release of VirFact is divided into five main areas (discussed below) providing different approaches and views to data analysis:

- a collection of individual virulence factors
- a collection of pathogenicity islands
- source genomes
- annotations and prediction results
- links

The first section contains basic information about individual virulence factors, such as their amino acid sequences, annotations collected from literature and links to other fields in database. This area is de facto the core of the system.

Individual virulence factors from a given organisms often form operon like structures called pathogenicity islands (PAIs) – information about them forms the next area of the VirFact database. Additional data, such as a PAI position at the genome, its short characterization and lists of genes it contains is provided here. Since PAIs usually evolve by

lateral transfer, they differ by many features from the host genome. To aid in identifying novel PAIs, the user can view a chart (deposited in database) showing genomic regions that deviate most from the rest of the genome. This diversity is based on three compositional criteria: G+C content, dinucleotide frequency and codon usage (2).

For individual virulence factors, the annotations and results of analysis and prediction tools provide information about homologs and genomic context and other information about a chosen virulence factor, as discussed in detail below.

Finally, the links to sections described above and various addresses that are useful for the user or necessary for the service are listed in a separate area of the website. The current (July 20, 2004) release of VirFact contains about 400 proteins, 12 pathogenicity islands (PAIs) and 7 completely sequenced genomes and it is increasing constantly.

## THE WEB SITE

VirFact is publicly available on the web at http://virfact.burnham.org. The database can be browsed by virulence factor name, PAI or genome using links on the top of the main web page.

- the "Virulence Factors" link: lets the user to see all virulence factors deposited in the database
- the "PAIs" link: allows to display all PAIs that are contained in VirFact. After selection of a specific PAI, the composition of PAI proteins is shown.
- the "Genome" link: leads user to an interface, which allows to check all VirFact proteins that are encoded in selected genome. An additional feature is a chart showing genomic regions that deviate most from the rest of the genome, which could form new, as yet unrecognized PAIs.

For each displayed virulence factor, on the right side of a webpage, there are links to annotation and prediction results, to sequence in FASTA format or to other links that could be potentially useful, like to NCBI PubMed. The link called "Homologs", allows user to view PSI-BLAST (3), FFAS03 (4) or T-Coffee (5) results. PSI-BLAST is used to compare a query sequence with those contained in non redundant protein database at NCBI by performing the iterative BLAST search. It is the most sensitive widely used program for recognizing homologs, making it useful for finding very distantly related proteins. The "FFAS" link shows the results of FFAS03 server, a profile-profile alignment algorithm used for super-

sensitive recognition of distant homologs and fold assignments. Finally, links called "Alignment" and "Tree" leads to T-Coffee results, where a multiple alignment was built using proteins found by the PSI-BLAST search. The T-Coffee results can be visualized with the "JalView" (multiple sequence alignment viewer, 6) and the "A Tree Viewer (ATV)" (phylogenetic tree viewer, 7) applications (Java Virtual Machine is required by both programs).

The "Genomic Context" interface was designed to perform the analysis of the genomic context using The SEED system (http://theseed.uchicago.edu/FIG/index.cgi) for genome annotations. As described by Overbeek et al. SEED is designed to help a researcher study a specific subsystem (set of genes), supporting community-wide annotation of genomes and searching for specific missing genes. SEED focuses on conservation of a genomic context between homologs of the specific gene. In VirFact, we compared genomic context of close homologs of the virulence factor being studied. It is important to note that SEED uses its own definition of a homolog, typically much more conservative then would result from a PSI-BLAST search.

The VirFact can also be queried using the Web-based interface called "Scan" for a presence of homologs of virulence factors covered by VirFact in the genome provided by the user. The search takes some time, up to several minutes, depending on a genome size. The output page shows potential virulence factors in the user genome, with information about the similarity score to known virulence factors, the position on a genome and the sequence alignment to the "parent" virulence factor in the FASTA format. For example, we show here a short analysis of *Francisella tularensis* genome. In the example presented here we focus on the information on how to use VirFact website, the full analysis of the potential virulence factors in *F. tularensis* genome will be presented elsewhere. As is showed in the chart (Fig. 1), there is a peak around 45 kb indicating high diversity of this region from the rest of the genome. In the same region VirFact found a protein similar to "Z0262 gene product" of Escherichia coli. Further analysis indicates that this hypothetical protein of E. coli has a homolog described only in the case of *Francisella tularensis*, called IglB. The last protein is acknowledged as associated in intracellular growth (8). Moreover, a neighborhood of "Z062 gene product" shows the functional coupling with other unknown proteins often present in other pathogens.

## UPDATES

Parsing, annotation and data updates have been automated to minimize human intervention. The VirFact database will be updated at least once per two months to ensure current report of data. The information about PAIs is manually curated.

## FUTURE PERSPECTIVES

VirFact was developed as a relational database of PAIs and virulence factors for the comprehensive representation of pathogenicity in various prokaryotic organisms. A web interface was designed to easy access the various features. To our knowledge, this is the only database devoted exclusively to pathogenicity island and virulence factors that provides a variety of tools for data analysis. We plan to expand the VirFact database to incorporate all annotated PAIs from all completely sequenced genomes and all virulence-related genes/proteins described in the literature. In near future we would like to broad VirFact of new tools predicting surface regions of the proteins and trans-membrane regions. We believe the VirFact will be useful tool for the investigation of the bacterial virulence and for the detection of virulence factors in newly sequenced genomes.

## ACKNOWLEDGEMENTS

We would like to thank Dr Ross Overbeek for The SEED: an Annotation/Analysis Tool that makes possible a development a genome context part of VirFact service. The authors also thank Zhanwen Li for her help in FFAS calculations. The work was partially funded by the 6FP grant MicrobeArray (to LR) and United States Army Medical Research and Materiel Command Grant DAMD17-03-2-0038 (to AG).

## REFERENCES

1. Wassenaar,T.M. and Gaastra,W. (2001) Bacterial virulence: can we draw the line? FEMS Microbiol Lett., 201, 1-7.

2. Tu,Q. and Ding,D. (2003) Detecting pathogenicity islands and anomalous gene clusters by iterative discriminant analysis. FEMS Microbiol. Lett., 221, 269-275.
3. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25, 3389-3402.
4. Rychlewski,L., Jaroszewski,L., Li,W. and Godzik,A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Science, 9, 232-241.
5. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol., 302, 205-217.
6. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The Jalview Java alignment editor. Bioinformatics, 20, 426-427.
7. Zmasek,C.M. and Eddy,S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. Bioinformatics, 17, 383-384.
8. Gray,C.G., Cowley,S.C., Cheung,K.K. and Nano,F.E. (2002) The identification of five genetic loci of Francisella novicida associated with intracellular growth. FEMS Microbiol Lett., 215, 53-56.

Figure 1. Graphic illustration of the using the VirFact for a search of virulence homologs in the genome delivered by the user. The chart of discriminant scores shows a region that deviates most from the rest of the genome. The VirFact has found in this place a homolog similar to "Z062 gene product" of *Escherichia coli*. The PSI-BLAST result show that "Z062 gene product" has a similar sequence: IglB [*Francisella tularensis*]. Moreover, the "Genomic Context" interface shows a significant neighborhood of Z062 with other proteins (in table, the "Z062 gene product" is no. 1, called as "hypothetical protein").