AD_____

Award Number: DAMD17-03-2-0038

TITLE: Structural Genomics of Bacterial Virulence Factors

PRINCIPAL INVESTIGATOR: Robert C. Liddington, Ph.D.

CONTRACTING ORGANIZATION: Burnham Institute
La Jolla, California 92037-1005

REPORT DATE: May 2004

TYPE OF REPORT: Annual

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

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	May 2004	Annual (1 May 2003 - 30 Apr 2004)

4. TITLE AND SUBTITLE
Structural Genomics of Bacterial Virulence Factors

5. FUNDING NUMBERS
DAMD17-03-2-0038

6. AUTHOR(S)
Robert C. Liddington, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Burnham Institute
La Jolla, California 92037-1005

E-Mail: rlidding@burnham.org

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. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

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

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)

We are applying a comprehensive yet focused structural genomics approach to determine the atomic resolution crystal structures of key bacterial virulence factors from high priority bacterial pathogens. Knowledge of protein structure and inhibitor complexes at atomic resolution is typically a pre-requisite for rational drug design. In this first year of funding we have focused our attention on plasmid annotation, target selection, protein expression, purification and crystallization of proteins encoded by the Bacillus anthracis pX01 plasmid. We have cloned and expressed a total of 35 new proteins, and structural analysis of several of these is underway. Currently, 3 new crystal structures are essentially complete, and 6 crystal structures of anthrax Lethal Factor in complex with small molecule inhibitors provided by our collaborators have been determined, and lodged in the public data base. We have also determined the first crystal structure of a complex between anthrax protective Antigen and its host cell receptor. The work performed under this grant has allowed us to leverage funding in several NIAID-funded research projects to carry out in-depth structure-function studies that will enable the next stages of drug design.

14. SUBJECT TERMS
X-ray crystallography, structural genomics, bioinformatics, biodefense, virulence factor, toxin

15. NUMBER OF PAGES
74

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

INTRODUCTION

We are applying a comprehensive but focused structural genomics approach to determine the atomic resolution crystal structures of key bacterial virulence factors from four high priority bacterial pathogens. These studies will expedite anti-toxin and vaccine design in a number of different ways. Structural data will be made available to all appropriate groups for use in structure based drug design. In addition, we will generate a large library of expression vectors for virulence factors, as well as research quantities of pure proteins, which can readily be adapted for vaccine production; and which are also likely to have applications in detector 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:

Target Selection. We performed a detailed analysis of the Bacillus anthracis virulence plasmids. Using a variety of bioinformatics tools we identified the possible function of about 40 proteins and discovered several likely operons on the pXO1 plasmid. The most interesting discoveries include numerous DNA processing enzymes, several new regulatory proteins and elements of the type IV secretion system. The results of the analysis of pXO1 are now being prepared for publication (a draft manuscript describing this work is provided in **Appendix 1**), and a continuation of the analysis of pXO2 plasmid is now being finalized.

We have identified a new domain in a broad range of bacterial, as well as single archaeal and plant proteins. Its presence in the virulence-related pXO1 plasmid of Bacillus anthracis (pX01-01) as well as in several other pathogens makes it a possible drug target. We term the new domain nuclease-related domain (NERD) because of its distant similarity to endonucleases. This work has been published in Trends in Biochemical Sciences (Grynberg & Godzik 29: 106-110 (2004)) and is included as **Appendix 2.**

Cloning and expression of novel B. anthracis proteins Two target lists were generated from the bioinformatics approaches: proteins with distant homologues in the protein data base of structures, and a second list of proteins with no homologues. Research fellows each chose 5 targets from List 1 and 3 from list 2. The work in progress is summarized below. For the most part, cloning was successful, and expression trials are at various stages, with several undergoing crystallization and NMR trials. Crystal structures of three novel proteins have been determined (described in **Appendices 3 and 4**). The work on pX01-118 and pX02-62 has led to a focus on the structure of the "master regulator" of the toxin genes, AtxA, and we have made a concerted effort to express full-length and domain fragments in different hosts and in a cell-free system (described in detail in **Appendix 5**). Our hit rate on soluble protein expression and crystallization has been somewhat disappointing when compared with our general success-rate for other bacterial

and eukaryotic proteins. The reasons for this are unclear at this stage, although certainly several of the proteins appear to be toxic to the host. We are now trying to find improved general systems for expression, including cell-free, insect cell and Bacillus megaterium expression.

Query	length	range	score	%id	covered by template(s)
gi|10956388|ref|NP_052837.1| pXO1-142 [Bacillus anthracis] (04/03/03)	887	1--634	-112	25	1f7d_A mol:protein length:659 DNA Topoisomerase III
gi|10956265|ref|NP_052714.1| pXO1-18 [Bacillus anthracis] (04/03/03)	316	1--315	-77.1	14	1kbu_A mol:protein length:349 Cre Recombinase
gi|10956379|ref|NP_052828.1| pXO1-132 [Bacillus anthracis] (04/03/03)	361	1--357	-76.4	13	1kbu_A mol:protein length:349 Cre Recombinase
gi|10956351|ref|NP_052799.1| pXO1-103 [Bacillus anthracis] (04/03/03)	317	12--310	-76.3	21	1a0p_ mol:protein length:290 Site-Specific Recombinase Xerd
gi|10956387|ref|NP_052836.1| pXO1-141 [Bacillus anthracis] (04/03/03)	214	57--201	-73.6	34	1ez5_A mol:protein length:149 Staphylococcal Nuclease
gi|10956362|ref|NP_052811.1| pXO1-115 [Bacillus anthracis] (04/03/03)	193	2--186	-64.6	31	1gdt_A mol:protein length:183 Gamma-Delta Resolvase
gi|10956342|ref|NP_052791.1| pXO1-95 [Bacillus anthracis] (04/03/03)	443	1--416	-64	25	1d1i_A mol:protein length:402 Udp-Glucose Dehydrogenase
gi|10956341|ref|NP_052790.1| pXO1-94 [Bacillus anthracis] (04/03/03)	295	1--294	-63.9	29	1am_A mol:protein length:292 Glucose-1-Phosphate Thymidylyltransferase
gi|10956343|ref|NP_052792.1| pXO1-96 [Bacillus anthracis] (04/03/03)	274	1--87	-12.6	10	1mdm_A mol:protein length:149 Paired Box Protein Pax-5
		38--270	-55	16	1k6y_A mol:protein length:288 Pol Polyprotein
		108--267	-62.6	19	1a5v_ mol:protein length:162 Integrase
gi|10956306|ref|NP_052754.1| pXO1-58 [Bacillus anthracis] (04/03/03)	272	4--265	-55.6	16	1ion_A mol:protein-het length:243 Probable Cell Division Inhibitor Mind
gi|10956284|ref|NP_052733.1| pXO1-37 [Bacillus anthracis] (04/03/03)	193	1--162	-55	10	1ghe_A mol:protein-het length:177 Acetyltransferase
gi|10956334|ref|NP_052783.1| pXO1-87 [Bacillus anthracis] (04/03/03)	160	2--160	-27.5	11	1jlu_A mol:protein length:186 Thiol Disulfide Interchange Protein Tlpa
		53--157	-52.2	17	1quw_A mol:protein length:105 Thioredoxin
gi|10956338|ref|NP_052787.1| pXO1-91 [Bacillus anthracis] (04/03/03)	280	1--162	-51.6	14	1d2t_A mol:protein length:231 Acid Phosphatase
gi|10956358|ref|NP_052807.1| pXO1-111 [Bacillus anthracis] (04/03/03)	204	1--154	-51.5	37	1acc_ mol:protein length:735 Anthrax Protective Antigen
gi|10956383|ref|NP_052832.1| pXO1-137 [Bacillus anthracis] (04/03/03)	61	2--60	-51	40	1kq1_A mol:protein length:77 Host Factor For Q Beta
gi|10956292|ref|NP_052741.1| pXO1-46 [Bacillus anthracis] (04/03/03)	435	2--390	-49.5	17	1fsz_ mol:protein length:372 Ftsz
		17--427	-37.3	9	1ffx_A mol:protein length:451 Tubulin
gi|10956340|ref|NP_052789.1| pXO1-93 [Bacillus anthracis] (04/03/03)	366	2--254	-49.2	14	1qg8_A mol:protein length:255 Spore Coat Polysaccharide Biosynthesis Prote
gi|10956306|ref|NP_052755.1| pXO1-59 [Bacillus anthracis] (04/03/03)	477	3--477	-11.8	9	1e32_A mol:protein length:468 P97
		79--406	-47.2	19	1g6o_A mol:protein-het length:330 Cag-Alpha
		197--472	-14.5	12	1jj7_A mol:protein length:280 Peptide Transporter Tap1
gi|10956374|ref|NP_052823.1| pXO1-127 [Bacillus anthracis] (04/03/03)	214	1--87	-12.4	11	1mdm_A mol:protein length:149 Paired Box Protein Pax-5
		37--214	-37.1	14	1k6y_A mol:protein length:288 Pol Polyprotein
		111--214	-46.1	14	1vsd_ mol:protein length:152 Integrase
gi|10956294|ref|NP_052743.1| pXO1-47 [Bacillus anthracis] (04/03/03)	201	11--113	-42.3	13	1jbg_A mol:protein length:109 Transcription Activator Of Multidrug-Efflux
gi|10956257|ref|NP_052706.1| pXO1-10 [Bacillus anthracis] (04/03/03)	383	4--361	-41	14	2adm_A mol:protein length:421 Adenine-N6-DNA-Methyltransferase Taqi
gi|10956376|ref|NP_052825.1| pXO1-120 [Bacillus anthracis] (04/03/03)	137	7--137	-30.5	14	1k6y_A mol:protein length:288 Pol Polyprotein
		55--137	-39	16	1vsd_ mol:protein length:152 Integrase
gi|10956386|ref|NP_052815.1| pXO1-119 [Bacillus anthracis] (04/03/03)	475	8--139	-13.2	10	1j5y_A mol:protein-het length:167 Transcriptional Regulator, Biotin Repres
		161--387	-38.9	14	1h90_A mol:protein length:224 Transcription Antiterminator Licf
gi|10956287|ref|NP_052736.1| pXO1-40 [Bacillus anthracis] (04/03/03)	66	1--65	-33.2	16	1adr_ mol:protein length:76 P22 C2 Repressor (Amino-Terminal DNA-Binding
gi|10956337|ref|NP_052786.1| pXO1-90 [Bacillus anthracis] (04/03/03)	652	224--650	-20.6	11	1cii_ mol:protein length:602 Colicin Ia
		301--593	-32.8	12	2tma_A mol:protein length:284 Tropomyosin - Chain A
gi|10956356|ref|NP_052805.1| pXO1-109 [Bacillus anthracis] (04/03/03)	99	9--97	-29.2	22	1smf_A mol:protein length:122 Transcriptional Repressor Smtb
gi|10956384|ref|NP_052833.1| pXO1-138 [Bacillus anthracis] (04/03/03)	97	10--95	-29.1	18	1smf_A mol:protein length:122 Transcriptional Repressor Smtb
gi|10956326|ref|NP_052775.1| pXO1-79 [Bacillus anthracis] (04/03/03)	1222	3--977	-9.92	9	1iB4_S mol:protein-het length:1184 Smooth Muscle Myosin Heavy Chain
		14--271	-10.2	14	1qle_C mol:protein length:273 Cytochrome C Oxidase Polypeptide III
		936--1169	-28.9	14	1qu7_A mol:protein length:227 Methyl-Accepting Chemotaxis Protein I
		987--1221	-14.3	8	1h8w_A mol:protein length:312 Bacteriophage T4 Short Tail Fibre
gi|10956286|ref|NP_052735.1| pXO1-39 [Bacillus anthracis] (04/03/03)	325	1--323	-28.1	11	1mm8_A mol:protein length:481 Tn5 Transposase
gi|10956283|ref|NP_052732.1| pXO1-36 [Bacillus anthracis] (04/03/03)	484	5--473	-27.8	12	1mm8_A mol:protein length:481 Tn5 Transposase
gi|10956328|ref|NP_052777.1| pXO1-81 [Bacillus anthracis] (04/03/03)	424	1--209	-27.3	13	1qsa_A mol:protein length:618 Soluble Lytic Transglycosylase Slt70
gi|10956282|ref|NP_052731.1| pXO1-35 [Bacillus anthracis] (04/03/03)	478	16--478	-25.3	11	1mm8_A mol:protein length:481 Tn5 Transposase
gi|10956380|ref|NP_052829.1| pXO1-133 [Bacillus anthracis] (04/03/03)	485	6--302	-10.1	12	1e32_A mol:protein length:468 P97
		183--485	-22.5	17	1pjr_ mol:protein length:724 Pcia
gi|10956270|ref|NP_052719.1| pXO1-23 [Bacillus anthracis] (04/03/03)	461	8--337	-20.7	9	1khv_A mol:protein length:516 RNA-Directed RNA Polymerase
		12--443	-16.1	9	1rdr_ mol:protein length:461 Poliovirus 3D Polymerase
gi|10956325|ref|NP_052774.1| pXO1-78 [Bacillus anthracis] (04/03/03)	405	125--398	-20.1	12	1gki_A mol:protein length:437 Conjugal Transfer Protein Trwb
gi|10956368|ref|NP_052817.1| pXO1-121 [Bacillus anthracis] (04/03/03)	57	1--56	-19.8	21	1qb7_A mol:protein length:236 Adenine Phosphoribosyltransferase
gi|10956254|ref|NP_052703.1| pXO1-07 [Bacillus anthracis] (04/03/03)	602	1--566	-12	8	1c2p_A mol:protein-het length:576 RNA-Dependent RNA Polymerase
		7--412	-19.1	11	1khv_A mol:protein length:516 RNA-Directed RNA Polymerase
		513--602	-11.5	17	1bvb_ mol:protein length:211 Cytochrome C-554
gi|10956271|ref|NP_052720.1| pXO1-24 [Bacillus anthracis] (04/03/03)	132	1--113	-18.1	16	1fa0_A mol:protein length:537 Poly(A) Polymerase
gi|10956280|ref|NP_052729.1| pXO1-33 [Bacillus anthracis] (04/03/03)	259	1--257	-15.2	11	1kan_A mol:protein length:253 Kanamycin Nucleotidyltransferase (E.C. 2.7.7
gi|10956352|ref|NP_052801.1| pXO1-105 [Bacillus anthracis] (04/03/03)	67	6--28	-14.5	43	1ek1_A mol:protein length:53 Transcription State Regulatory Protein Abrb
gi|10956260|ref|NP_052709.1| pXO1-13 [Bacillus anthracis] (04/03/03)	1320	358--1310	-13.8	13	1k83_A mol:protein length:1733 DNA-Directed RNA Polymerase II Largest Subu
gi|10956317|ref|NP_052766.1| pXO1-70 [Bacillus anthracis] (04/03/03)	437	80--437	-13.7	15	1dd9_A mol:protein length:336 DNA Primase
gi|10956281|ref|NP_052710.1| pXO1-14 [Bacillus anthracis] (04/03/03)	564	503--556	-11.9	14	1b0n_A mol:protein length:111 Sinr Protein
gi|10956276|ref|NP_052725.1| pXO1-29 [Bacillus anthracis] (04/03/03)	274	1--61	-11.8	30	1i8_A mol:protein length:195 Rad50 Abc-Atpase
gi|10956269|ref|NP_052718.1| pXO1-22 [Bacillus anthracis] (04/03/03)	91	9--91	-11.3	16	1n6_A mol:protein-het length:433 Putative Cell Cycle Protein Mesj

Pfam positive hits	FFAS	pI	GRAVY
gi|10956345|ref|NP_052794.1|	9.45	5.89	-0.533
gi|10956319|ref|NP_052768.1|	10.1	5.02	-0.637
gi|10956335|ref|NP_052784.1|	5.69	9.52	-0.349
gi|10956277|ref|NP_052726.1|	6.15	9.15	-0.565
gi|10956269|ref|NP_052718.1|	11.3	6.72	-0.293
gi|10956378|ref|NP_052827.1|	6.43	6.04	-0.7

Blind' hits	FFAS	pI	GRAVY
gi|10956263|ref|NP_052712.1|	5.47	10.1	-0.51
gi|10956350|ref|NP_052800.1|	6.16	5.1	0.04
gi|10956323|ref|NP_052772.1|	4.99	5.7	-0.62
gi|10956372|ref|NP_052821.1|	5.3	10.1	-0.37
gi|10956312|ref|NP_052761.1|	7.59	9.9	-0.2
gi|10956262|ref|NP_052711.1|	6.2	9.8	-0.44
gi|10956274|ref|NP_052723.1|	6.14	4.4	-0.42
gi|10956302|ref|NP_052751.1|	6.7	6.6	-0.84
gi|10956298|ref|NP_052747.1|	4.85	5	-0.4
gi|10956279|ref|NP_052728.1|	5.88	7.1	-0.39
gi|10956348|ref|NP_052797.1|	7.19	7.5	-0.89
gi|10956272|ref|NP_052721.1|	5.65	10.6	0.23
gi|10956320|ref|NP_052769.1|	6.77	9.1	-0.56
gi|10956297|ref|NP_052746.1|	5.72	4.6	-0.25
gi|10956296|ref|NP_052745.1|	6.39	10.5	-0.6
gi|10956349|ref|NP_052798.1|	5	9.9	-1.01
gi|10956367|ref|NP_052816.1|	5.31	9.6	-0.55
gi|10956289|ref|NP_052738.1|	5.45	4.4	-0.83
gi|10956381|ref|NP_052830.1|	6.47	4.6	-0.25
gi|10956285|ref|NP_052734.1|	5.3	9.5	-0.76
gi|10956329|ref|NP_052778.1|	5.62	5.2	-0.56
gi|10956251|ref|NP_052700.1|	5.76	4.9	-0.37
gi|10956281|ref|NP_052730.1|	5.61	4.9	-0.26
gi|10956331|ref|NP_052780.1|	7.16	10.5	-0.1
gi|10956278|ref|NP_052727.1|	6.18	7.3	-0.17
gi|10956290|ref|NP_052739.1|	6.98	5	-0.37
gi|10956255|ref|NP_052704.1|	7.85	8.5	-0.5
gi|10956377|ref|NP_052826.1|	6.2	6.1	-0.92
gi|10956268|ref|NP_052717.1|	5.25	5.3	-0.68
gi|10956347|ref|NP_052796.1|	4.93	5	-0.35
gi|10956304|ref|NP_052753.1|	6.75	9.8	-0.76
gi|10956365|ref|NP_052814.1|	5.48	8.6	-0.57
gi|10956252|ref|NP_052701.1|	6.12	4.3	-0.63
gi|10956389|ref|NP_052838.1|	6.72	10.3	-0.23
gi|10956318|ref|NP_052767.1|	5.1	4.5	0.05
gi|10956382|ref|NP_052831.1|	6.8	4.5	-0.58
gi|10956346|ref|NP_052795.1|	7.18	3.9	-0.6
gi|10956371|ref|NP_052820.1|	5.12	10.9	-0.62
gi|10956291|ref|NP_052740.1|	6.63	9.4	-0.52
gi|10956359|ref|NP_052808.1|	4.72	10.6	-0.4
gi|10956253|ref|NP_052702.1|	7.14	5	-0.34
gi|10956273|ref|NP_052722.1|	5.87	4.9	-0.22
gi|10956314|ref|NP_052763.1|	5.44	9	-0.67
gi|10956375|ref|NP_052824.1|	5.03	10.7	-1.07
gi|10956364|ref|NP_052813.1|	5.56	10.2	-0.43
gi|10956363|ref|NP_052812.1|	7.02	9.6	-0.43
gi|10956267|ref|NP_052716.1|	4.58	10.8	-0.72
gi|10956275|ref|NP_052724.1|	6.04	4.1	-0.43
gi|10956344|ref|NP_052793.1|	7.14	4.4	-0.65
gi|10956373|ref|NP_052822.1|	5.99	5.8	-0.75
gi|10956258|ref|NP_052707.1|	6.05	8.4	-0.51
gi|10956385|ref|NP_052835.1|	6.87	7.7	-0.39
gi|10956266|ref|NP_052715.1|	3.47	4.4	-0.97
gi|10956293|ref|NP_052742.1|	5.79	6.5	-0.38
NP_052810.1			
NP_052809.2			
NP_052697		9.66	-0.284

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

pXO1-1 has a single transmembrane region and could only be expressed as insoluble protein. Initial trials using high concentration of detergent TritonX-100 extraction failed to produce significant amount of soluble protein. Expression of the fragment excluding the predicted transmembrane also produce insoluble inclusion.

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. **pXO1-04**, **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**.

Structural Studies of inhibitor binding to Lethal Factor

Compounds NSC 12155, NSC 357756, NSC 357777 had been identified as the top 3 hits in the USAMRIID NCI small molecules library high throughput screen for LF inhibition.

We determined the crystal structure of LF-12155-Zn (LF wild-type bound to NSC 12155 in the presence of zinc), and this work in collaboration with Drs. Gussio and Bavari at USAMRIID has been published recently (Panchal et al. Nat. Struct. Mol. Biol. 11: 67-72 (2004) (**Appendix 6**). It showed a compound that is able to bind and inhibit up to 95% of the native catalytic activity of LF. This compound does not require the presence of zinc to bind to the active site of LF, and appears to recognize the substrate-binding site immediately adjacent to the catalytic zinc site through hydrophobic interactions.

Currently, we are working on the structures of LF-357756-Zn and LF-357756-Zn (complex of LF wild-type bound to NSC 357756 or NSC 357777 in the presence of zinc), and the model refinement is continuing, with new data being collected. So far, electron density maps indicate that compound NSC 357756 is bound in the immediate vicinity of the catalytic site, and may be coordinating the zinc atom. NSC 357777 however appears to be relying more on hydrophobic interactions in recognizing the substrate-binding site in LF, while still binding close to the zinc atom. Currently, the focus is on NSC 357756, which has been shown to have better cell permeability abilities than NSC 12155 and better inhibitory abilities than NSC 357777 (unpublished data, sourced from USAMRIID colleagues).

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 (**Appendix 8**). The structure reveals an extensive

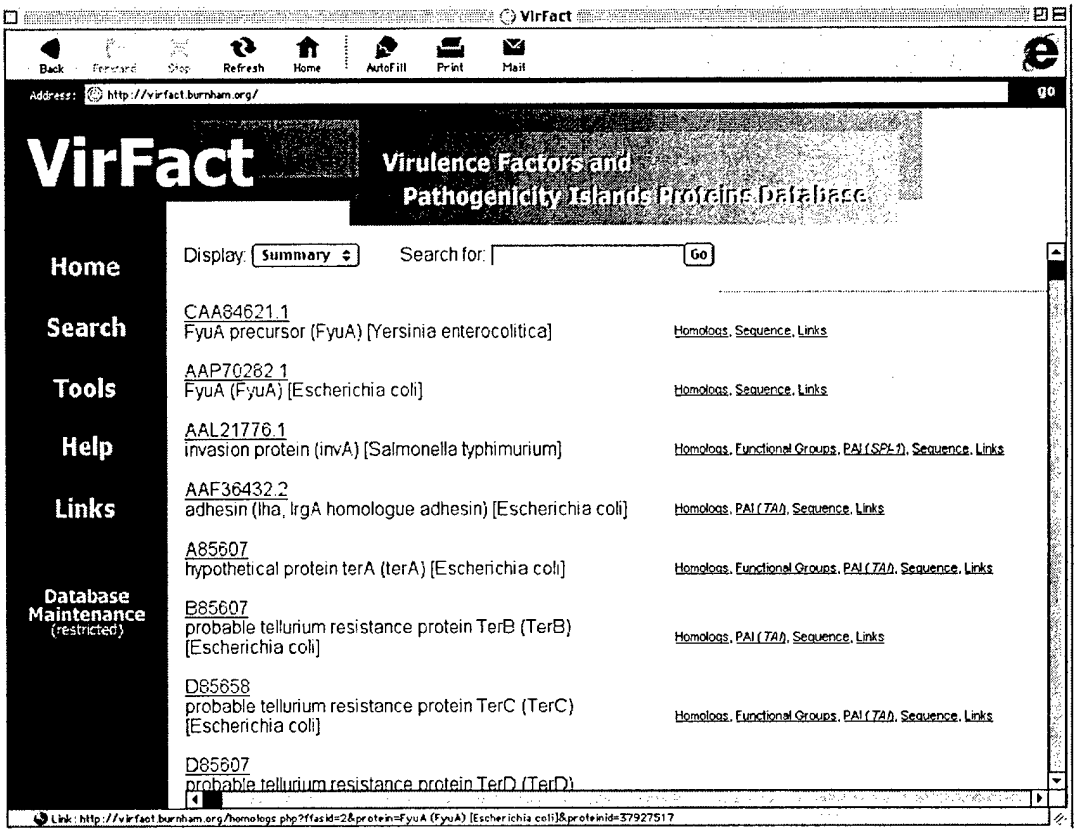
receptor-pathogen interaction surface that mimics the non-pathogenic recognition of the extracellular matrix by integrins. The binding surface is closely conserved in the two receptors and across species, but quite different in the integrin domains, explaining the specificity of the interaction. CMG2 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.

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 has been accomplished for the B. anthracis pX01 proteins, and target selection and experimental updates are done on a monthly basis in the light of new cloning, expression and structural data. We will make this information publicly available if this is deemed appropriate by USAMRMC.

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

We have developed a preliminary version of the virulence factor database (VirFact). It is available at http://virfact.burnham.org/. Currently, this database contains information on about 60 virulence factors and about 10 pathogenic islands, selected mostly based on literature searches (you can see all proteins in the database by entering an empty string is the search window). Each of the proteins in the database was annotated using modeling, distant homology recognition and sequence analysis tools. One of the tools available on-line is the possibility of scanning a new genome for homologues of virulence factors. Several potential virulence factors were identified this way in the Francisella genome.



Task 4: Form a consortium of groups with similar interests who are funded from other sources, developing a common website containing target selections and project status.

We plan to hold an inaugural meeting this Fall. In the first instance we will bring together investigators from the DHHS Region IX - AZ, CA, HI and NV - as this coincides with our attempts to create an NIAID Regional Center of Excellence.

Key research Accomplishments

- In-depth annotation of the anthrax virulence plasmid, and the identification of novel domains.

- Successful expression and/or cloning and of 35 proteins and domain fragments from the B. anthracis virulence plasmid, pX01

- Identification, crystal structure determination and characterization of a putative B. anthracis CO_2 sensor

- Crystal structure of a B. anthracis amidase homologous the bactericidal phage enzyme

• Crystal structure of anthrax PA in complex with its host receptor

• 6 Crystal structures of anthrax Lethal Factor in complex with inhibitors

Reportable Outcomes

Published manuscripts:

1. Panchal RG, Hermone AR, Nguyen TL, Wong TY, Schwarzenbacher R, Schmidt J, Lane D, McGrath C, Turk BE, Burnett J, Aman MJ, Little S, Sausville EA, Zaharevitz DW, Cantley LC, Liddington RC, Gussio R, Bavari S. Identification of small molecule inhibitors of anthrax lethal factor. *Nat Struct Mol Biol.* **11**:67-72 (2004).

2. Turk BE, Wong TY, Schwarzenbacher R, Jarrell ET, Leppla SH, Collier RJ, Liddington RC, Cantley LC. The structural basis for substrate and inhibitor selectivity of the anthrax lethal factor *Nat Struct Mol Biol.* **11**:60-6 (2004)

3. Grynberg M, Godzik A. NERD: a DNA processing-related domain present in the anthrax virulence plasmid, pXO1 *Trends Biochem Sci.* **29**:106-10 (2004)

Manuscript under review:

1. Santelli, E., Bankston, L.A., Leppla, S.H. & Liddington, R.C. "Crystal structure of an anthrax toxin-host cell receptor complex" Submitted to *Nature*.

Reagents generated:

• Expression vectors for 35 proteins from the B. anthracis pX01 plasmid.

• Atomic coordinates have been deposited in the Protein Data Bank for anthrax Lethal Factor-inhibitor complexes, and the Protective Antigen-host cell receptor complex.

Funding applied for:

We developed the initial work on pX01-118, pX02-61 and AtxA funded by this grant into an in-depth structure-function study in an application for a Program Project grant from NIAID led by Dr. Liddington **(P01 AI 55789-01)**. We recently received word that this proposal has been funded, and will start this Summer.

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 a grant to develop novel therapeutic treatments of Smallpox **(U01 AI061139 - P.I, Dr. Alex Strongin)**

On the strength of the work funded by this grant and others, we have been invited to participate in a **Regional Center of Excellence** proposal for Region IX that will be submitted this Fall.

Conclusions

In this first year of funding we have focused 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 35 new proteins, and structural analysis of several of these is underway. Currently, 3 new crystal structures are essentially complete, 6 crystal structures of anthrax Lethal Factor in complex with small molecule inhibitors provided by our collaborators at USAMRIID and elsewhere. We have also determined the first crystal structure of a complex between anthrax protective Antigen and its host cell receptor (under review at Nature). In the next year, in addition to continuing the B. anthracis work, we propose to focus on newly annotated F. tularensis genome and apply a similar set of tools to elucidate virulence in this poorly studied organism.

So what section: Knowledge of protein structure and inhibitor complexes at atomic resolution is typically a pre-requisite for rational drug design. Therapeutics do not exist for any of the major pathogens likely to be used in biowarfare or bioterrorism. Our efforts in the first year were focused towards the design of anthrax therapeutics, for which the need is compelling since antibiotic treatments have limited effectiveness and vaccines are problematic. The work described in this proposal provides the first stages of target identification and characterization, and the structures we have already determined bear directly on inhibitor design. The work has also allowed us to leverage funding in several NIAID-funded research projects to carry out in-depth structure-function studies that will enable the next stages of drug design.

References

References are included in appropriate appendices.

Appendices

Appendix 1: Surprising connections: in-depth analysis of the *Bacillus anthracis* pXO1 plasmid (manuscript in preparation)

Appendix 2: NERD: a DNA processing-related domain present in the anthrax virulence plasmid, pXO1 (published manuscript)

Appendix 3: Discovery, crystal structures and characterization of a putative CO_2 sensor domain, "BACO"

Appendix 4: Crystal Structure of B. anthracis amidase homologous to a bacteriophage lysin

Appendix 5: Structural studies of AtxA, a member of the PRD family of transcriptional activators.

Appendix 6: "Identification of small molecule inhibitors of anthrax lethal factor" (Published paper)

Appendix 7: The structural basis for substrate and inhibitor selectivity of the anthrax lethal factor (published paper).

Appendix 8: Crystal structure of an anthrax toxin-host cell receptor complex (manuscript submitted to Nature)

Manuscript in Preparation

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

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

[1]Bioinformatics and Systems Biology, The Burnham Institute, La Jolla, CA 92037, USA; [2]Joint Center for Structural Genomics, San Diego Supercomputer Center, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

ABSTRACT

Anthrax disease is caused by *Bacillus anthracis*. Virulence of this bacterium has been associated with two plasmids, pXO1 and pXO2. We used the DNA and proteome sequences of pXO1 to understand where it comes from and what are the functions of, so called, unknown open reading frames. For this purpose, we used context analysis and distant homology tools that allowed us to discover, among others, two type IV secretion system-like operons. We also significantly increased the description of many pXO1 ORFs and showed its mosaic nature.

[Supplemental material available online at bioinformatics.ljcrf.edu/pXO1. The genome sequence data is available at NCBI: http://www.ncbi.nlm.nih.gov/genomes/framik.cgi?db=Genome&gi=16452.]

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

[4]Corresponding author. E-mail: adam@burnham.org; FAX: +1 (858) 713 9930.

INTRODUCTION

Anthrax is an ancient disease primarily affecting herbivores but also attacking other mammals, including humans. Scientific studies of this disease started in the middle of the XVIII century and almost a century later the agent was recognized as *Bacillus anthracis*, a bacterium causing many diverse manifestations of the same infection. From pioneering work of Pasteur (1881) a quest for an effective vaccine started and until the end of sixties of the XX century, the work progressed much, not only in the creation of human anthrax vaccines but also in the elucidation of the three-component nature of the anthrax toxin (for review see [Turnbull, 2002]). A couple of incidents and a threat of bioterrorism, however, started an era of even more intensive work on *B.anthracis*. The sequencing projects [Okinaka, 1999; Pannucci, 2002; Read, 2003] and projects focusing on the mechanisms of toxin functions, regulation and release (for review see [Koehler, 2002]), advanced significantly the understanding of anthrax pathogenesis. The *B.anthracis* genome consists of a 5.23-Mb chromosome and two megaplasmids, pXO1 (181.7 kb) and pXO2 (94.8 kb)[Okinaka, 1999; Read, 2003]. Genetic analysis focused mainly on toxins

(PagA, LEF, CyaA), cell envelope and germination geness (Cap, S-layer and Ger proteins), and the regulatory mechanisms triggering the virulence. From genetic experiments and informatic analyses, we know that both the chromosomally- and plasmid-encoded factors control the toxin genes. The chromosomal copy of the AbrB-encoding gene is a negative controller of toxin genes (*pagA*, *cyaA* and *lef*) as well as of the atxA gene [Saile, 2002], the main pXO1-encoded activator of the pXO1-encoded toxin genes [Dai, 1995]. AtxA was also shown to activate pXO2 genes. It triggers the production of capsule proteins *via* the activation of homologous activators from the pXO2 plasmid, AcpA and AcpB [Drysdale, 2004]. In total, at least 7 pXO1 and 10 pXO2 genes are regulated by the AtxA protein [Bourgogne, 2003]. In addition to that another regulator, the PagR protein, has a weak negative effect on the *pag* operon and regulates the cell envelope genes, the S-layer genes, *sap* and *eag* [Hoffmaster, 1999; Mignot, 2003]. Virulence-related genes are also known to be regulated by temperature and CO_2/bicarbonate levels [Bartkus, 1994; Sirard, 1994].

Recent works have focused on bioinformatic analyses of the anthrax genome and phylogenetically-related plasmids [Ariel, 2002, Berry, 2002; Ariel, 2003; Rasko, 2004]. The studies confirmed that *B.anthracis* is closely related to *B.thuringiensis* and *B.cereus*, and showed many previously unknown features of the deadly plasmids, pXO1, pBtoxis and pBc10987, respectively. These works did not answer our questions: (I) where do the elements of the pXO1 plasmid come from, and (II) what do the unknown genes encoded on pXO1 do? For analysis, we used the most recent sequence of the pXO1 plasmid [Read, 2003], and we did not focus on similarities to other bacilli plasmids. We were interested in operon conservation and twilight zone homologies that can reveal hypothetically important features for the virulence function of that plasmid.

RESULTS

Statistics

DNA level analysis

Previous analyses have been performed to analyze the DNA sequence of the pXO1 plasmid [Okinaka, 19999; Read, 2002; Pannucci, 2002]. 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.

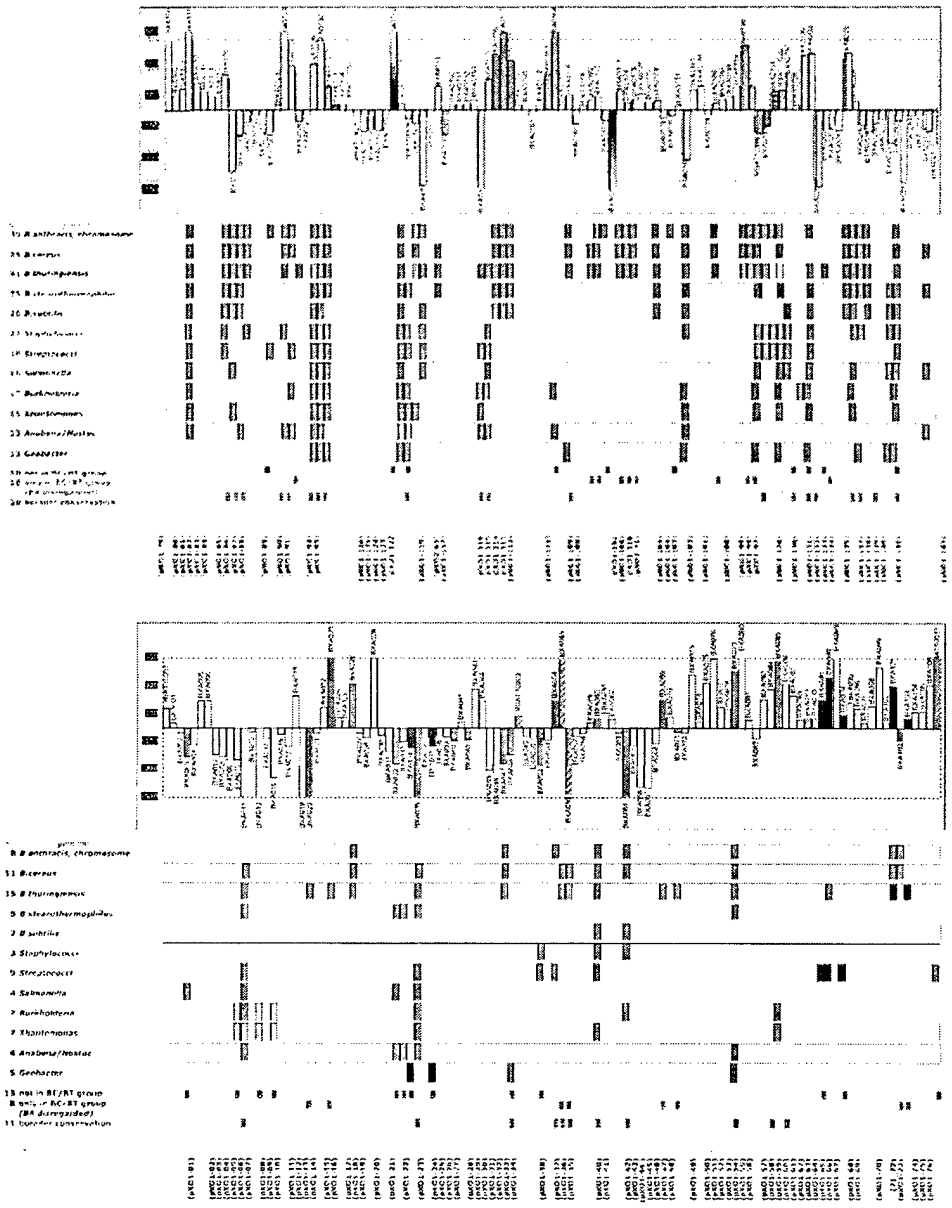


Figure: Summary of the pX01 annotation

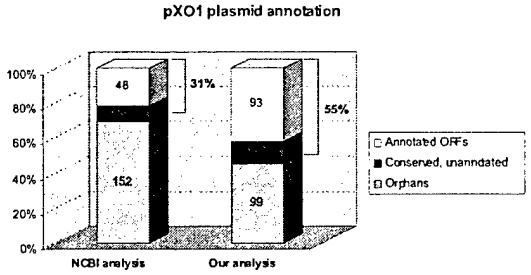


Figure. **Improvement of pXO1 plasmid annotation using FFAS profile-profile algorithm [Rychlewski, 2000 #6] and context analysis (ERGO).**

One goal was to find proteins directly involved in 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 neighborhood of hypothetical proteins, with no recognizable homology to proteins from publicly available databases. It is located between ORFs BXA0076 (previously 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

In our efforts to understand the function of pXO1 we focused on the plasmid proteome, looking for a consistent picture of its function and phylogeny. We used context analysis methods to reveal interesting connections with other regulons. As already mentioned in the introduction, we did not focus on obvious similarities to other bacilli and their plasmids, even if this data is available in Figure 1. From the total analysis of pXO1

ORFs we realized two major features. First, pXO1 has genes collected from many different species (Figure 1). It also contains a lot of truncated and mutated sequences with homology to existing genes (Supplementary data). Second, pXO1 has a significant number (15%) of proteins related to DNA metabolism (Supplementary data).

Gram-positive bacteria connection

BXA0010, BXA0013 and BXA0015 proteins are similar to proteins from two Gram-positive species, *Xanthomonas* and *Burkholderia* (Figure 2A), and from the proteobacterial *Pseudomonas* group. BXA0010 and BXA0013 are homologues of the *Xanthomonas* orf8 (gi number: 21242978), *Burkholderia* protein (gi number: 22985671) and a number of *Pseudomonas* genus proteins (gi numbers: 24461733, 37955709, 40019205). All of these 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). One can hypothesize that this insertion occurred after the duplication and disrupted the BXA0010 gene. BXA0013 forms an operon with BXA0015; this feature is also conserved as an operon in the species mentioned above (*Xanthomonas* 21242977, *Burkholderia* 22985672, and *Pseudomonas* spp. 24461735, 37955708, 40019206). BXA0015 has strong similarity to the N-terminal part of its homologues. This region of the protein encodes the coenzyme-binding domain of various DNA methyltransferases (FFAS score: -42.200 to 1g38a). Except for anthrax, other species preserve numerous proteins in the operon. Unfortunately, even the most advanced tools could not recognize any homology of other ORFs to known proteins; therefore their function remains a mystery. From the function of known members of this operon one can imply a DNA modifying function.

Actinobacteria/Cyanobacteria connection

BXA0032 and BXA0033, if fused, can be a part of the COG0175 family. These proteins belong to 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*. The functions of other proteins from the regulon 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, 2003], and a clear connection with BXA0037, a nucleotidyltransferase domain protein, is obvious. As a complex they may catalyze the addition of a nucleotidyl group to unknown substrates, perhaps to antibiotics or other poisonous substances, as their structural homolog kanamycin nucleotidyltransferase does [Matsumura, 1984]. The function of these operons is unknown.

Bacilli connection

BXA0091 and BXA0094 are homologous to each other and to proteins from several other bacilli; *Enterococcus*, *Listeria*, *Lactococcus*, *Lactobacillus*, or other *Bacillus* species. This protein family is of unknown function and is hypothesized to be an extracellular protein [Nakai, 1999]. 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*, *Listeria innocua* and *E.faecalis* the BXA0091 homologues colocalize with a surface layer domain protein (FFAS score: -11.500 to COG1361 for the BXA0092 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 RTX toxin and related Ca2+-binding proteins family (FFAS: -12.000 to COG2931) and a regulatory protein homologous to transcription positive regulators MGA (FFAS: -85.100). LysM domain binds peptidoglycans and was first identified in bacterial lysins [Ponting, 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, 2000], and MGA are important in streptococci virulence [McIver, 2002]. Other unknown proteins from these operons are predicted to be extracellular. All these facts strongly suggest that these related operons can be involved in pathogenesis.

Type IV secretion system machinery

In two operons we identified proteins strongly resembling known type IV secretion system proteins. The first is composed of the BXA0083 protein involved in type IV pili biogenesis, CpaB/RcpC (COG3745)(CDD score: 1e-18), the VirB11 family protein (BXA0085)(CDD: 1e-41) and two homologous unknown proteins that belong to the TadC family (COG2064), often found in the operon with the VirB11 protein (BXA0086 and BXA0087)(FFAS: -19.400 and -12.300, respectively). The second operon contains the homolog of the VirB4 protein (BXA0107)(FFAS: -64.100) 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)(FFAS: -11.600 to VirB6, -11.700 to myosin tail and -23.200 to the MCP domain).

In the first operon the most studied is the VirB11 [Dang, 1999; Krause, 2000; Yeo, 2000; Christie, 2001; Savvides, 2003]. The model predicts that VirB11 family of ATPases "function as chaperones reminiscent of the GroEL family for translocating unfolded proteins across the cytoplasmic membrane" [Christie, 2001]. Both the BXA0083 and two homologues BXA0086 and BXA0087, as well as BXA0085 VirB11 homologue, are distant homologues of proteins also forming an operon in many gram-negative bacterial species [Kachlany, 2000; Skerker, 2000]. The genes CpaB/RcpB and TadC form large families, widespread in bacteria and archaea [Kachlany, 2000]. The *cpa* operon in *Caulobacter crescentus* was proven to be required for pilus assembly [Skerker, 2000]. Amazingly, we couldn't identify the homolog of pilA gene or any other pilin subunit, which is necessary for pilus formation.

The BXA0107 protein from the second operon belongs to the large VirB4 family. It is one of the elements of the type IV secretion system important in the delivery of effector molecules to the host cell [Christie, 2000; Christie, 2001 and citations therein]. 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, 1999]. It contains the Walker A motif responsible for ATP binding, which is well conserved in BXA0107 (200-207 fragment: GISGSGKS).

BXA0107 forms an operon with the BXA0108 protein which 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 homology at the C-end is to a domain that is thought to transduce the external chemotaxis signal to the two-component histidine kinase CheA (for review see [Stock, 2002]).

The next protein in this operon resembles the C-terminus of a *Bacillus firmus* integral membrane protein, consisting of 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 [Sparrow, 1985; Icho, 1985]. The function of the C-terminal part of the *B.firmu* 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 is completely unexpected. Unfortunately, we were not able to detect any other elements of this machinery in the plasmids or chromosome. Is the presence of incomplete operons an evolutionary artifact, a minimal complex to deliver molecules to the host, or a part of a larger complex not yet recognized with the use of available tools? These operons are good targets for experimental analysis. The discovery of putative molecules secreted by this system may be crucial for our understanding of diverse roles of pXO1 in virulence.

Particular cases

The statistical analysis of the pXO1 megaplasmid is a convenient way to describe the general physiology. It does not, however, allow one to understand the complexity of each protein's function. In the detailed analysis of *B.anthracis* ORFs we focused on particular cases, especially from the pathogenic region [Okinaka, 1999; Sirard, 2000] that are of special interest to the scientific community.

BXA0139

The BXA0139 protein is located close to the edema factor (CyaA) on the pXO1 sequence. It is 150 amino acids long, located on an operon with two unknown hypothetical proteins, BXA0138 and BXA0140. The only known fact about these proteins is the similarity of BXA0138 to BXA0149 (Table 1). The most interesting finding is the homology of BXA0139 to the C-terminal end of the hemolysin II from *B.cereus* [Miles, 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, 1999; Miles, 2002]. Hemolysins form heptameric rings [Song, 1996; Gouaux, 1997], in which the C-terminal domain would reside in the outside part of each monomer [Miles, 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 may 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 "recreate" a real sequence, not the one that is an automatic translation deposited at NCBI, 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. It may contribute to some attack related function that has nothing to do with the hemolytic activity.

BXA0167

This hypothetical ORF has no known homologs and no distant 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 conducted, however, a BLASTX analysis along its sequence and found an interesting homology coded by the opposite strand. Loaded 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?

pXO1 regulators

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, or even regulators regulate virulence factors as well as other genes not related with pathogenicity [for review see: Hacker, 2000; Hentschel, 2001]. We think then that it is essential to describe these regulators on anthrax pathogenicity vector in order to decipher the physiology of pXO1.

BXA0020

BXA0020 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 (FFAS score: -11.900) is the SinR repressor domain [Gaur, 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, 1992; Mandic-Mulec, 1995]. In *B.anthracis* pXO1 plasmid, BXA0020 does not form an operon with sin genes. Instead, it is located close to a protein (BXA0019) 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, 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 (e-value: 6e-06 to the *A.thaliana* protein, gi: 22327992), 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 [Woehlke, 2000; Mandelkow, 2002].

BXA0048

The N-terminal part of BXA0048 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, 1994], BetI, a repressor of the osmoregulatory choline- glycine betaine pathway [Lamark, 1996], MtrR, a regulator of cell envelope permeability that acts as a repressor of *mtrCDE*-encoded and activator of *farAB*-encoded efflux pumps [Lee, 1999; Lee, 2003]. 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). The presence on the same operon of the nucleotidyltransferase (gi: 5459398) with a superfamily II DNA and RNA helicase family protein (gi: 5459399) in *Streptomyces coelicolor* can be a suggestion that BXA0048 is involved in the DNA metabolism.

BXA0060

The BXA0060 belongs to the 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 and do not preserve the operon structures. It seems then that BXA0060 homologues are involved in very diverse functions/pathways.

BXA0069

BXA0069 belongs to the family of global transcription activators of membrane-bound multidrug transporters, responsible for bacterial multidrug resistance (MDR)[Paulsen, 1996]. The closest homologue is the *B.subtilis* MtnA regulator that belongs to the MerR family (FFAS: -42.300)[Summers, 1992]. It is known to activate two MDR transporters (*bmr* and *blt*), a transmembraneous protein-coding gene *ydfK* and its own gene [Baranova, 1999]. It acts independent from two specific activators, BmrR and BltR, that are encoded by *bmr* and *blt* operons [Ahmed, 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, 2001]. BXA0069 perfectly fits this description, it possesses quite conserved two distal regions and a 90 amino acid region of no homology that has an almost 80% probability of a coiled-coil structure [Lupas, 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.

BXA0122

The FFAS analysis revealed low score similarity (FFAS: -10.500) of BXA0122 to the MarR regulators of the multiple antibiotic resistance locus [Seoane, 1995; Grkovic, 2002]. 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, 1995]. The repressive DNA-binding by MarR can be inhibited by several anionic compounds, e.g. salicylate [Alekshun, 1999].

AtxA

AtxA is a proven regulator of anthrax toxin genes [Uchida, 1993; Koehler, 1994; Dai, 1995]. It is also known to influence the expression of other genes on pXO1, pXO2 plasmids and the anthrax genome [Bourgogne, 2003]. AtxA is a member of the PTS (the phosphoenolpyruvate-dependent, sugar transporting phosphotransferase system) regulatory domain-containing family [Greenberg, 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. The presence of PTS EII homology domains is consistent with its being an activator, since these domains are lacking in antiterminators [Greenberg, 2002]. Knowing the architecture of this family, we searched the whole anthrax genome in order to find all similar regulators. Among the ones we found, apart from the obvious AtxA and AcpA proteins, there is a very recent confirmation of the activity of the BXB0060 (pXO2-53), named AcpB [Drysdale, 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.

BXA0166 + BXA0207

Both BXA0166 and BXA0207 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, 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, 1993], and ArsR, that acts as the arsenic/antimony-responsive repressor of the *ars* operon in *Escherichia coli* [Wu, 1991]. Another, less well studied, group in the ArsR/SmtB family are the transcriptional activators, with *Vibrio cholerae* HlyU as the founding member [Williams,

1993]. HlyU is known to upregulate the expression of hemolysin and of two *hcp* genes, which are coregulated with hemolysin [Williams, 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, 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 PagR is known to act as a repressor, but in a weak manner [Hoffmaster, 1999] and is suspected of having an activation function as well [Mignot, 2003]. A more detailed phylogeny of close homologues of the pXO1 proteins (Fig. xxxB) shows that there has been a wave of gene duplications in the ancestor of *B.antracis* and *B.cereus*. 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*. 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.

BXA0178

BXA0178 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 [Trowsdale, 1978; Philips, 2002]. Recently, Saile and Koehler [Saile, 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.

BXA0180

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 (Structural Classification of Proteins)(FFAS: -12.200)[Andreeva, 2004]. The ORF is truncated after the first half, and experiments are needed to check whether a shortened domain can exert any function.

BXA0206

BXA0206 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 [Wachi, 1999; Masse, 2002], mRNA stability [Vytvytska, 1998], stabilization and degradation of RNAs [Tsui, 1997; Takada, 1999]. Hfq proteins are similar to eukaryotic Sm proteins involved in RNA splicing [Moller, 2002]. The function of 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.

DISCUSSION

We have uncovered several novel features of the pXO1 plasmid. We showed 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 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 in Gram-negative bacteria. It remains to be seen whether we are dealing with a minimal set indispensable for the formation and function of secretion, or if it is a remnant, unfunctional set of proteins, or perhaps if the function(s) of these operons has changed from the original function. The discovery of type IV secretion system can have a significant impact on our understanding of anthrax virulence. If this system is functional, an unknown pathogenic delivery pathway may be important 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, the borrowing of diverse ideas allowed the formation of this killing agent. 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).

The discovery of previously unknown systems would not be important if we did not ask questions about regulation. External signals, cell state or host-pathogen interaction certainly trigger bacterial response(s), and a couple of them are already known (for review see [Koehler, 2002]). All these signals finally activate transcription of virulence-related genes. We were trying to describe all possible regulators that we could find using more sensitive programs than BLAST. Some of the regulatory proteins are known not to influence the toxin function (e.g. the homologue of AbrB), but others may be of interest to researchers studying *B.anthracis*. Further experimental studies are needed to prove whether the newly discovered factors regulate plasmid genes or chromosome genes.

We don't know how reliable 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) questions its reliability. We don't know, however, if all the genes are well predicted and if there are no not-recognized ORFs 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.

Interesting was also the description of the ArsR/SmtB family of regulators. Apparently, *B.anthracis* is armed with all kinds of ArsR homologues, however the majority of them are related to the activators subfamily, with pXO1 homologues among others. 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%). It seems that DNA is a central point of the pXO1's function. Is this function related with the processing of pXO1, chromosomal DNA, transposons, or host DNA? We don't know, however not one 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. 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. It seems to be the subject of constant evolutionary pressure. This plasmid should have a tag: "under construction."

METHODS

DNA level analysis

The *Bacillus anthracis* strain A2012 pXO1 plasmid sequence was used for analysis (accession: NC_003980)[Read, 2003]. For the analysis of common DNA features in promoter regions of AtxA-dependent genes [Bourgogne, 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, BXA0124, BXA0125, BXA0137, BXA0142 (cyaA), BXA0164 (pagA), BXA0172 (lef), BXB0045, BXB0060, BXB0066, BXB0074, BXB0084. We used MEME and MITRA programs to search for common motifs [Bailey, 1994; Eskin, 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, 1990]. To analyze the protein sequences, we used the following programs: BLAST tools [Altschul, 1990; Altschul, 1997], SMART tool [Letunic, 2002], Pfam [Bateman, 2002], CDD [Marchler-Bauer, 2003], TMHMM2.0 [Sonnhammer, 1998], **SEED** [], Radar [Heger, 2000], FFAS03 [Rychlewski, 2000], Metaserver.pl [Elofsson, 2003], Superfamily [Gough, 2001]. To align sequences we used: T-COFFEE [Notredame, 2000], AliBee [Nikolaev, 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 independantly for each repetition.

NERD: a DNA processing-related domain present in the anthrax virulence plasmid, pXO1

Marcin Grynberg and Adam Godzik

Program in Bioinformatics and Systems Biology, The Burnham Institute, 10901 N. Torrey Pines Road, La Jolla, CA 692037, USA

Corresponding author: Adam Godzik (adam@burnham-inst.org).

We have identified a new domain in a broad range of bacterial, as well as single archaeal and plant proteins. Its presence in the virulence-related pXO1 plasmid of *Bacillus anthracis* as well as in several other pathogens makes it a possible drug target. We term the new domain nuclease-related domain (NERD) because of its distant similarity to endonucleases.

Anthrax, a disease of herbivores and primates (including humans), is caused by a gram-positive, spore-forming bacterium, *Bacillus anthracis*. The virulence of this bacterium is dependent on two megaplasmids: pXO1, which is required for the synthesis of the toxin protein [1]; and pXO2, which is required for the synthesis of an anti-phagocytic capsule [2–4]. Strains lacking either of the two megaplasmids are avirulent.

The pXO1 plasmid has been analyzed in several recent genome sequence studies [5–7], by using standard tools such as BLAST. Using sensitive homology-detection algorithms, we have found that a 117-amino acid fragment of the pXO1–01 protein, previously annotated as a hypothetical protein, defines a new domain that is shared by multiple proteins in other eubacteria and is also present in small numbers in archaea and plant proteins. We call it NERD for nuclease-related domain.

The NERD domain

Starting from the amino acid sequence of the *B. anthracis* pXO1–01 protein, a cascade of PSI–BLAST searches [8] identified >40 proteins with a region displaying statistically significant sequence similarity to the seed protein and to each other (Figure 1) and with varied domain combinations (Figure 2). The NERD domain partly overlaps two Pfam-B domains – Pfam-B_22501 and Pfam-B_26882 [9]. However, the Pfam-B families contain only a few sequences (5 and 4, respectively) with single domain context each. An alignment of NERD is presented in Figure 1 and covers 117 amino acids.

The NERD sequence is characterized by three conserved regions interspersed among weakly conserved or very diverse regions (Figure 1). Conserved hydrophobic, mainly aliphatic motifs (consisting of Leu, Ile and Val) and polar, mainly charged positions (e.g. Asp, His, Glu and Lys), alternate in the alignment. The first and most conserved region is formed by the N-terminal Glu followed by the [Gln/Glu]-[Ile/Val/Leu]-Asp motif, then a stretch of hydrophobic residues with two polar (Glu and Lys) and two hydrophobic (Gly and [Ile/Leu/Val]) residues at the end. The next 20 amino acids are not conserved, but the [Ser/Asn]-Pro-[Ile/Leu/Val/Met] motif with a neighboring Gln form a second conserved region. The third is at the C-terminal 25 amino acids, with mainly the hydrophobic amino acids conserved. An interesting feature of NERD is the existence of subgroups that have no conservation in motifs that are conserved in all other members of the family (e.g. two N-terminal glycine residues are missing in the plant domain) or with a charge difference (e.g. Glu instead of Gln in the most conserved [Gln/Glu]-[Ile/Val/Leu]-Asp motif). We can only hypothesize that these differences account for functional diversity within the NERD family.

The predicted α–β–β–β–β–(weak β/long loop)-α–β–β secondary structure of NERD domain helps rationalize the conservation of specific regions of the domain (Figure 1) because all the conserved residues coincide with secondary-structure elements, especially the third and fourth β strands. The only exception is the fifth β strand, which is likely to be a terminal strand or a long loop (Figure 1).

NERD-domain associations

The majority of NERD-containing proteins are single-domain, in several cases with additional (predicted) transmembrane helices. In only a few instances, proteins containing NERD have additional domains that, in 75% of these cases, are involved in DNA processing. In all cases in which NERD is present in multidomain proteins, it is found at the N terminus. There is also no evident operon conservation for NERD-containing proteins and no apparent connection between phyla and domain fusions.

Most NERD-containing proteins, including the group-defining *B. anthracis* pXO1–01 protein, consist entirely of the NERD domain, sometimes with short tails of several amino acids on both C and N termini. All proteins in this group are hypothetical open reading frames (ORFs). In addition, in several proteins the NERD domain is associated with one or two predicted transmembrane motifs, which could be located either at the N or C terminus (Figure 2).

In a hypothetical *Clostridium perfringens* protein (gi: 9118309656), the NERD domain is followed by the helicase and RNaseD C-terminal (HDRC) domain (PF00570; Figure 2). HRDC is an 80-amino acid protein domain usually found at the C terminus of RecQ helicases and RNase D homologs from various organisms, including human [10]. An HRDC domain is present in genes linked

to the human diseases Werner and Bloom syndromes [11,12]. The HRDC domain is involved in the binding of DNA to specific DNA structures (e.g. long-forked duplexes and Holliday junctions) that are formed during replication, recombination or transcription [13]. Interestingly, in the many HRDC-containing proteins, the N-terminal region in the 3′→5′ exonuclease domain (PF01612) that is responsible for the 3′→5′ exonuclease proofreading activity of the DNA polymerase I and other enzymes and catalyzes the hydrolysis of unpaired or mismatched nucleotides [14,15]. One can speculate that NERD, existing in analogous arrangement with the HRDC domain, has a related function.

In at least three proteins, including the hypothetical protein (gi: 22972752) from *Chloroflexus aurantiacus*, the NERD domain is found at the N terminus of the UvrD/Rep 3′→5′ DNA helicases (PF00580), which catalyze the ATP-dependent unwinding of double-stranded to single-stranded DNA (ssDNA) [16]. DNA helicases are essential for processes such as DNA replication, recombination and repair [17]. This domain co-occurs with the HRDC domain in several bacterial species (i.e. *Streptomyces coelicolor*, *Corynebacterium glutamicum*, *Mycobacterium leprae* and *Mycobacterium tuberculosis*).

In two proteins, in *Pseudomonas aeruginosa* (gi: 26406504) and the *Bacteroides* (gi: 8308027), NERD is followed by the DNA-binding C4 zinc finger (PF01396), which is a short motif present in two NERD proteins (Figure 2), usually a C-terminal region of prokaryotic topoisomerases I [18]. The role of topoisomerase in the bacterial cell is to remove excessive negative supercoils from DNA to maintain the optimal superhelical state [19]. The zinc motifs do not cleave or recognize the topoisomerase substrate, rather, they are believed to interact with ssDNA to relax negatively supercoiled DNA [20]. Apart from topoisomerases, there are a few proteins with proximally located restriction endonucleases (PF04471) or unknown N termini that possess the C4 zinc fingers. However, their role is unknown.

In five proteins, the NERD domain is followed by two STYKc domains (PF00069). STYKcs are protein kinases with possible dual serine, threonine and tyrosine kinase specificity [21]. For example, in the cases of *Thermomonospora fusca* and *Streptomyces coelicolor*, there are genomic associations with DNA polymerase III and transposase, and an adenine-specific methyltransferase, respectively, which can suggest a nucleotide-related function of these large proteins (ERGO database: http://ergo.integratedgenomics.com).

In most cases, only one copy of the NERD domain is present in a given organism. We found that in only three bacteria there are two copies of NERD per genome (in *Burkholderia fungorum*, *Oceanobacillus iheyensis* and *Desulfitobacterium hafniense*).

pXO1-01 function

None of the NERD-containing proteins have been studied by experiment, therefore, its exact function is not known. However, bioinformatics analyses offer some clues.

The closest homolog of pXO1-01 is the orf8 protein from *Bacteroides* spp. It is an ORF from the non-replicating *Bacteroides* unit 1 (NBU1) , a 10.3-kbp integrated element that can be excised and mobilized in *trans* by tetracycline-inducible *Bacteroides* conjugative transposons [22,23]. The elements responsible for integration and excision were recognized [24–26], but *orf8* is probably not involved in these processes. The large G+C content difference between *orf6*, *orf7* and *orf8* (35%), and other *Bacteroides* genes (42%) suggests a possible recent acquisition that is involved in a yet-undiscovered transposition process. The presence of NERD in a unique archaeal and only two plant species supports such a transposon-type transfer of the domain.

A more detailed prediction can be made based on the domain structure similarity between NERD proteins that contain the HRDC domains and the N-terminal region of exonuclease proteins that contain the HRDC domains [Au: correction was not clear, is edit correct?]. This is further supported by distant homology between NERD and the COG0792 family, a predicted endonuclease family distantly related to archaeal Holliday junction resolvase, members of which are involved in DNA replication and/or recombination, and/or repair. This homology is predicted by a profile–profile search algorithm FFAS (fold and function assignment system) [27], albeit with low statistical significance. Several fold-recognition algorithms (e.g. Superfamily and BASIC) [27–29] identify matches to the Holliday junction resolvase structure (PDB codes: 1gefA and 1hh1A) with statistically significant scores [30,31]. The alignment between the NERD and COG0792 families and the sequence of the Holiday junction resolvase (PDB code: 1gefA) is shown in Figure 1 (both alignments were obtained by the FFAS [27] algorithm). The alignment covers only the N-terminal half of NERD, and the 3D model of this is shown in Figure 3. Interestingly, all active-site residues of resolvase (black arrows in the alignment and residues shown in atomic detail in the Figure 3) are conserved in most NERD family members, which strongly supports the functional prediction. The common denominator of all these predictions suggests a nuclease function for NERD.

Concluding remarks

We have discovered a novel domain, NERD, with predicted connection to DNA processing. Genomic context analysis and distant homology analysis suggest a nuclease function.

The finding of this domain is important for the understanding of anthrax virulence. The location of pXO1-01 in the vicinity of other DNA processing-related ORFs, on the anthrax virulence plasmid, suggests an orchestrated function of the products of these genes. Is this machinery an anthrax DNA-remodeling system or is it involved in the eukaryotic cell attack? Maybe further advances in the studies of the NBU1 element will reveal its function.

The presence of NERD in only few non-bacterial species not only suggests that this domain might be involved in some mobility processes, but also that the species transfer must have happened quite recently.

Acknowledgements

We thank Marc Robinson-Rechavi for his critical reading of the article. This work was supported by the National Institutes of Health, grant GM60049.

References

1 Vodkin, M.H. and Leppla, S.H. (1983) Cloning of the protective antigen gene of *Bacillus anthracis*. *Cell* 34, 693–697

2 Uchida, I. *et al.* (1985) Association of the encapsulation of *Bacillus anthracis* with a 60 megadalton plasmid. *J. Gen. Microbiol.* 131, 363–367

3 Mikesell, P. *et al.* (1983) Evidence for plasmid-mediated toxin production in *Bacillus anthracis*. *Infect. Immun.* 39, 371–376

4 Green, B.D. *et al.* (1985) Demonstration of a capsule plasmid in *Bacillus anthracis*. *Infect. Immun.* 49, 291–297

5 Pannucci, J. *et al.* (2002) *Bacillus anthracis* pXO1 plasmid sequence conservation among closely related bacterial species. *J. Bacteriol.* 184, 134–141

6 Okinaka, R.T. *et al.* (1999) Sequence and organization of pXO1, the large *Bacillus anthracis* plasmid harboring the anthrax toxin genes. *J. Bacteriol.* 181, 6509–6515

7 Read, T.D. *et al.* (2003) The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. *Nature* 423, 81–86

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

9 Bateman, A. *et al.* (2002) The Pfam protein families database. *Nucleic Acids Res.* 30, 276–280

10 Morozov, V. *et al.* (1997) A putative nucleic acid-binding domain in Bloom's and Werner's syndrome helicases. *Trends Biochem. Sci.* 22, 417–418

11 Yu, C.E. *et al.* (1996) Positional cloning of the Werner's syndrome gene. *Science* 272, 258–262

12 Ellis, N.A. *et al.* (1995) The Bloom's syndrome gene product is homologous to RecQ helicases. *Cell* 83, 655–666

13 von Kobbe, C. *et al.* (2003) Werner syndrome protein contains three structure specific DNA binding domains. *J. Biol. Chem.* 278, 52997–53006

14 Moser, M.J. *et al.* (1997) The proofreading domain of *Escherichia coli* DNA polymerase I and other DNA and/or RNA exonuclease domains. *Nucleic Acids Res.* 25, 5110–5118

15 Joyce, C.M. and Steitz, T.A. (1994) Function and structure relationships in DNA polymerases. *Annu. Rev. Biochem.* 63, 777–822

16 Hickson, I.D. *et al.* (1983) The *E. coli uvrD* gene product is DNA helicase II. *Mol. Gen. Genet.* 190, 265–270

17 Matson, S.W. *et al.* (1994) DNA helicases: enzymes with essential roles in all aspects of DNA metabolism. *Bioessays* 16, 13–22

18 Tse-Dinh, Y.C. and Beran-Steed, R.K. (1988) *Escherichia coli* DNA topoisomerase I is a zinc metalloprotein with three repetitive zinc-binding domains. *J. Biol. Chem.* 263, 15857–15859

19 Drlica, K. (1990) Bacterial topoisomerases and the control of DNA supercoiling. *Trends Genet.* 6, 433–437

20 Ahumada, A. and Tse-Dinh, Y.C. (2002) The role of the Zn(II) binding domain in the mechanism of *E. coli* DNA topoisomerase I. *BMC Biochem.* 3, 13

21 Hanks, S.K. and Hunter, T. (1995) Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB J.* 9, 576–596

22 Li, L.Y. *et al.* (1993) Characterization of the mobilization region of a *Bacteroides* insertion element (NBU1) that is excised and transferred by *Bacteroides* conjugative transposons. *J. Bacteriol.* 175, 6588–6598

23 Shoemaker, N.B. and Salyers, A.A. (1988) Tetracycline-dependent appearance of plasmidlike forms in *Bacteroides uniformis* 0061 mediated by conjugal *Bacteroides* tetracycline resistance elements. *J. Bacteriol.* 170, 1651–1657

24 Shoemaker, N.B. *et al.* (1996) NBU1, a mobilizable site-specific integrated element from *Bacteroides spp.*, can integrate nonspecifically in *Escherichia coli*. *J. Bacteriol.* 178, 3601–3607

25 Shoemaker, N.B. *et al.* (1996) The *Bacteroides* mobilizable insertion element, NBU1, integrates into the 3' end of a Leu-tRNA gene and has an integrase that is a member of the λ integrase family. *J. Bacteriol.* 178, 3594–3600

26 Shoemaker, N.B. *et al.* (2000) Multiple gene products and sequences required for excision of the mobilizable integrated *Bacteroides* element NBU1. *J. Bacteriol.* 182, 928–936

27 Rychlewski, L. *et al.* (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.* 9, 232–241

28 Ginalski, K. *et al.* (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* 19, 1015–1018

29 Gough, J. *et al.* (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

30 Nishino, T. *et al.* (2001) Crystal structure of the archaeal Holliday junction resolvase Hjc and implications for DNA recognition. *Structure (Camb)* 9, 197–204

31 Bond, C.S. *et al.* (2001) Structure of Hjc, a Holliday junction resolvase, from *Sulfolobus solfataricus*. *Proc. Natl. Acad. Sci. U. S. A.* 98, 5509–5514

32 Nikolaev, V.K. *et al.* (1997) Building multiple alignment using iterative analyzing biopolymers structure dynamic improvement of the initial motif alignment. *Biochemistry* 62, 578–582

33 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

34 McGuffin, L.J. *et al.* (2000) The PSIPRED protein structure prediction server. *Bioinformatics* 16, 404–405

35 Karplus, K. *et al.* (1999) Predicting protein structure using only sequence information. *Proteins* 37 (Suppl. 3), 121–125

36 Rost, B. and Eyrich, V.A. (2001) EVA: large-scale analysis of secondary structure prediction. *Proteins* 45 (Suppl. 5), 192–199

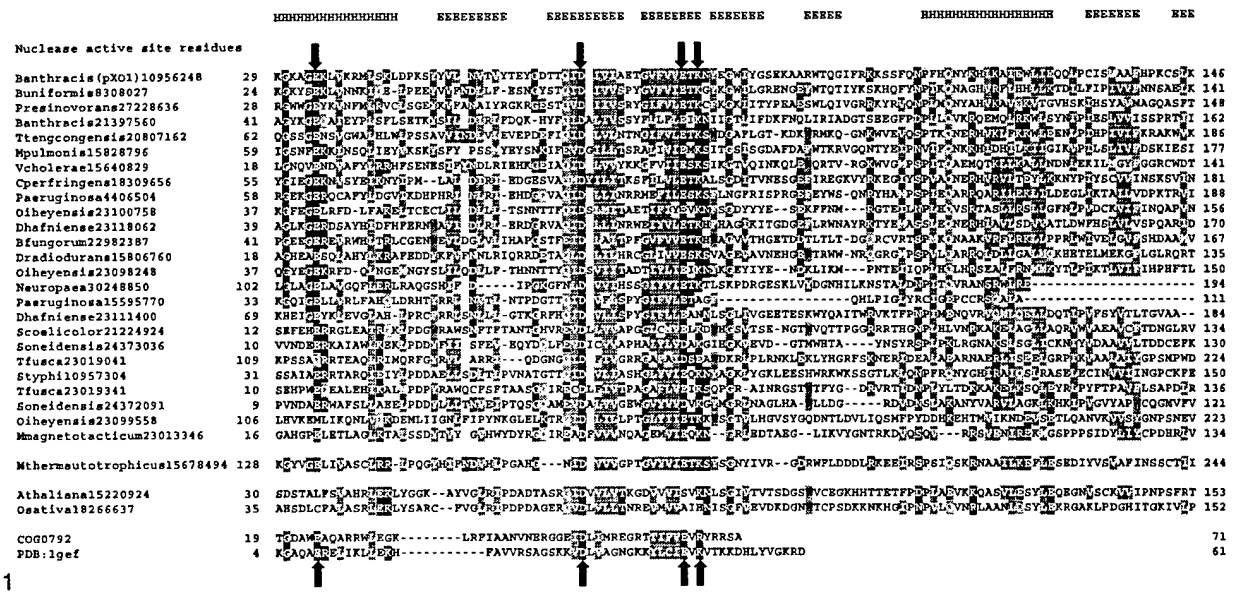
37 States, D.J. and Gish, W. (1994) Combined use of sequence similarity and codon bias for coding region identification. *J. Comput. Biol.* 1, 39–50

38 Gish, W. and States, D.J. (1993) Identification of protein coding regions by database similarity search. *Nat. Genet.* 3, 266–272

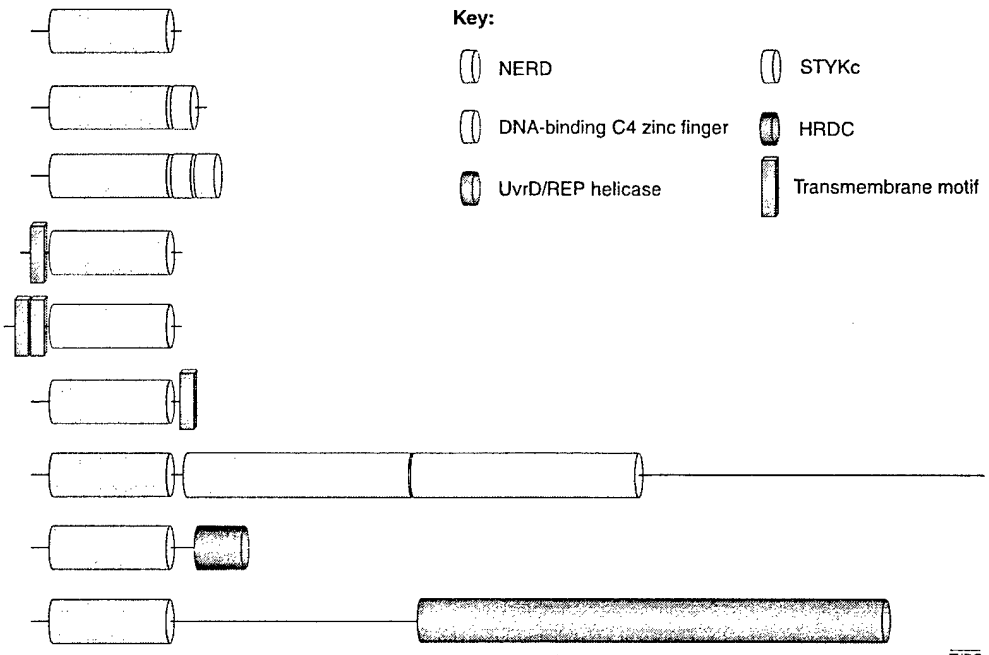
39 Letunic, I. *et al.* (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.* 30, 242–244

40 Sali, A. and Blundell, T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234, 779–815

41 The PyMOL Molecular Graphics System (2002) http://www.pymol.org



2**Figure 1.** An alignment of a sample set of NERD (nuclease-related domain) sequences. The alignment was generated using AliBee 3(http://www.genebee.msu.su/services/malign_full.html) [32] and colored in BioEdit [33]. PSI-BLAST [8] searches of the nonredundant protein database using the 4*Bacillus anthracis* pXO1-01 protein (gi: 10956248) as query were performed using the default parameters. After five rounds of searching, representatives of all 5subgroups of NERD were found. The highest E value was 2e-06. The uppermost group is composed of prokaryotic proteins, the middle protein is the sole example of 6an archaeal NERD-containing protein and the lowermost are two plant proteins. The shading threshold is 40%. The alignment is colored according to identity and 7similarity according to the default BioEdit amino acid similarity scoring matrix. The secondary-structure prediction is given for pXO1-01 as a combined result of 8PSIPRED [34], Sam-T99-2d [35] and Profsec [36] at the MetaServer (http://bioinfo.pl/Meta/) [28]. The results for other members of the NERD family are almost 9identical. Arrows indicate the conserved residues that are important for the endonuclease activity of resolvases. Two shorter sequences (*Nitrosomonas europaea* 10Q82W50 and *Pseudomonas aeruginosa* Q9I5W3) are coded by genomic sequences with a stop codon and no sequence homology beyond the stop codon, when 11checked using the BLASTX program [37,38] at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/BLAST/). It seems that the N-terminal 12catalytic domain is sufficient for their function. Sequences shown are (species name, gi number, in brackets are the first and last positions in the sequences aligned): 13Banthracis (pXO1)10956248, *Bacillus anthracis* Q8KYT4 (29–146); Buniformis8308027, *Bacteroides uniformis* Q9KIA1 (24–141); Presinovorans27228636, *Pseudomonas* 14*resinovorans* Q8GHQ8; Banthracis21397560, *Bacillus anthracis* Q81XB5 (41–162); Ttengcongensis20807162, *Thermoanaerobacter tengcongensis* Q8RBY3 (62–186); 15Mpulmonis15828796, *Mycoplasma pulmonis* Q98QN6 (59–177); Vcholerae15640829, *Vibrio cholerae* Q9KTS7 (18–141); Cperfringens18309656, *Clostridium* 16*perfringens* Q8XML4 (55–181); Paeruginosa4406504, *Pseudomonas aeruginosa* AAD20003 (58–188); Oiheyensis23100758, *Oceanobacillus iheyensis* Q8ELC6 (37–156); 17Dhafniense23118062, *Desulfitobacterium hafniense* ZP_00101791 (39–170); Bfungorum22982387, *Burkholderia fungorum* ZP_00027654 (41–167); 18Dradiodurans15806760, *Deinococcus radiodurans* Q9RTK3 (18–135); Oiheyensis23098248, *Oceanobacillus iheyensis* Q8ES50 (37–150); Neuropaea30248850, 19*Nitrosomonas europaea* Q82W50 (102–194); Paeruginosa15595770, *Pseudomonas aeruginosa* Q9I5W3 (33–111); Dhafniense23111400, *Desulfitobacterium hafniense* 20ZP_00097061 (69–184); Scoelicolor21224924, *Streptomyces coelicolor* O86560 (12–134); Soneidensis24373036, *Shewanella oneidensis* Q8EGX7 (10–130); 21Tfusca23019041, *Thermobifida fusca* ZP_00058754 (109–224); Styphi10957304, *Salmonella typhi* Q9L5M7 (31–150); Tfusca23019341, *Thermobifida fusca* ZP_00059052 22(109–224); Soneidensis24372091, *Shewanella oneidensis* Q8EJH0 (9–121); Oiheyensis23099558, *Oceanobacillus iheyensis* Q8EPJ8 (106–223); 23Mmagnetotacticum23013346, *Magnetospirillum magnetotacticum* (16–134); Mthermautotrophicus15678494, *Methanothermobacter thermautotrophicus* O26566 (128–24244); Athaliana15220924, *Arabidopsis thaliana* Q9SS58 (30–153); Osativa18266637, *Oryza sativa* Q8W3G9 (35–152). This multiple sequence alignment (alignment 25number ALIGN_000650) has been deposited with the European Bioinformatics Institute (ftp://ftp.ebi.ac.uk/pub/databases/embl/align/ALIGN_000650).



TiBS

Figure 2. The domain architecture of NERD (nuclease-related domain)-containing proteins. In all cases of multidomain proteins, NERD is located in the N terminus. All domains were recognized using the simple modular architecture research tool (SMART) server (http://smart.embl-heidelberg.de/ or http://smart.ox.ac.uk/) [39]. In case of long proteins, the size of domains is not proportional to protein length.

Figure 3. The predicted structure of NERD (nuclease-related domain). The pXO1-01 model was obtained with the Modeller comparative modelling suite [40], on the basis of the FFAS (fold and function assignment system) [27] alignment. The ribbon diagram was prepared using Pymol [41].

Appendix 3

Discovery, crystal structures and characterization of a putative CO_2 sensor domain, "BACO": In computational searches of the sequences of the pX01 and pX02 plasmids, we noticed an ORF (number 118 in pX01) contained within the pX01 pathogenicity island that was previously poorly characterized. By searching GenBank with the amino acid sequence of the pXO1-118 protein sequence, using the BLAST and PSI-BLAST programs with default values, we identified a homologue from the *B. anthracis* pXO2 plasmid (protein pXO2-61) with e-value = 4e-35 in the first iteration. The arrangement of these genes in both pX01 and pX02 is striking similar. In both cases, the gene is next to the activator (AtxA or AcpA) and transcribed in the opposite direction. Recently, the Koehler group (Bourgogne et al., Infect. Immun. 71: 2736-2743 (2003)) have shown that AtxA upregulates gene expression on both plasmids; the most upregulated gene pn the pX02 plasmid is pX02-61, suggesting that it plays an imprtant role in virulence.

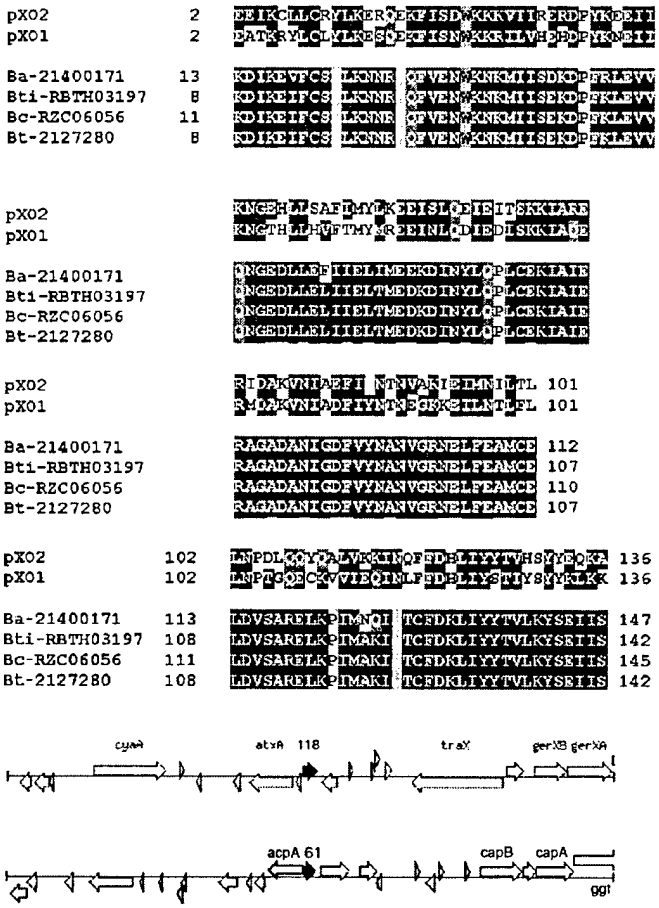


Figure: *A new sensory domain, BACO, found in several Bacillus species. In all cases except pX01 and pX02, it is a part of a sensor histidine kinase protein implicated in sporulation. The presence of this domain on both virulence pXO1 and capsule pX02 plasmids of Bacillus anthracis suggests its involvement in virulence. In the lower panel, the gene organization in the*

neighborhoods of pX01-118 and pX02-61 is seen to be strikingly similar, adjacent to their putatuve response regulators.

We found a second homolog in the *B. anthracis* chromosome (RBAT07138/A2012). In this case, the protein has the organization typical of a sensor histidine kinase. The sequence of the kinase domain suggests that it phosphorylates Spo0F in the phosphorelay system that triggers sporulation. Three other homologs were found in Bacillus species (one each in *B. stearothermophilus*, *B. thuringensis* and *B. cereus*). The homology with pXO1-118 protein reached an e-value = 2e-18 after the first round of searching. The last similar protein was discovered in the genome of *B. stearothermophilus*. It is also a histidine kinase; but is more distantly related (e-value = 2e-11).

B. subtilis	
Bs_KinA	LAAGIAHEIRNPLT
Bs_KinB	LAASVAHEVRNPLT
Bs_KinC	LAAGIAHEVRNPLT
Bs_KinD	LAASTAHEIRNPLT
Bs_KinE	LAAGIAHEIRNPMT
B. anthracis	
Ba_4792	IAAGIAHEVRNPLT
Ba_4659	LTAGIAHEIRNPLT
Ba_4810	LAASVAHEVRNPLT
Ba_4713	MAASISHEIRNPLT
Ba_4660	MAATVGHEIKNPLA
BAKIN	MSASFVHEFRNPLT

Figure: *Sequences of B. Subtilis sporulation sensor histidine kinases around the phosphohistidine site, and B. anthracis kinases implicated in sporulation by sequence similarity. In red, the new member of this family, containing a BACO sensor domain.*

C.2b Crystal Structure of pX01-118/BACO-1: We expressed BACO-1 in *E. coli* as a His-tag protein, cleaved with thrombin and purified using a Ni affinity column and a Superdex75 gel filtration column. The protein runs on SDS-PAGE as expected with a M.W. of ~17 kDa and runs as a dimer on a sizing column. Crystals with typical dimensions 0.1 mm x 0.05 mm x 0.05 mm were grown in space group $P3_221$. Using synchrotron radiation native and SeMet-MAD data sets to 2.5 Å were collected. With phase information from the SeMet-MAD data improved by solvent flattening and phase extension to 1.85 Å an interpretable electron density map was obtained. The asymmetric unit contains one molecule; the molecular two-fold axis coincides with a crystallographic dyad. Model building and refinement were carried out in programs O and CNS. The final BACO-1 model consists of residues 2-147, with Rfree = 25 % and appropriate stereochemistry. The BACO-1 structure reveals a helix bundle with 5 helices (see Figure).

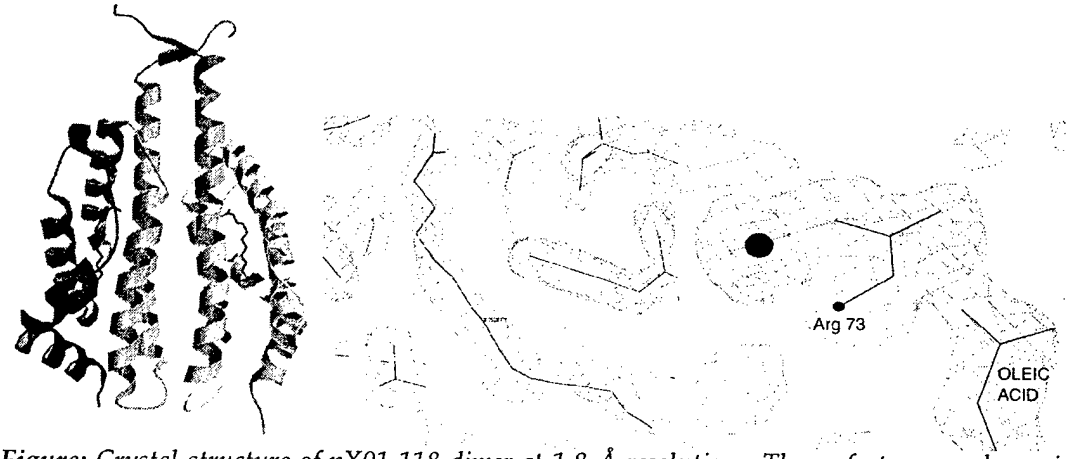


Figure: Crystal structure of pX01-118 dimer at 1.8 Å resolution. The co-factors are shown in red. ***Figure 16:*** *Electron density for the fatty acid (red stick)), perhaps myristic acid, bound in the hydrophobic core of each monomer. (Right panel) Close-up of the electron density near Arg 73, showing strong density connected to a fatty acid at right, and an unknown additional density (red circle).*

The fold is most closely related to the globin fold, which defines a family of proteins that typically bind co-factors in a hydrophobic cavity at the center of the bundle. There is strong electron density at this position that appears to be a fatty acid (perhaps myristic or oleic acid) with at least 13 C-C units. One end of this additional density lies next to a buried arginine (R73), which is part of a motif (KIAxER) that is invariant within this small family of sensor domains. On the other side of the arginine is a strong electron density feature that may be covalently bonded, which may be a bound anion (it is of course a possibility that this represents bound CO_2, since CO_2 was not excluded during crystal growth). We are currently trying to identify the co-factor using mass spectrometry and NMR.

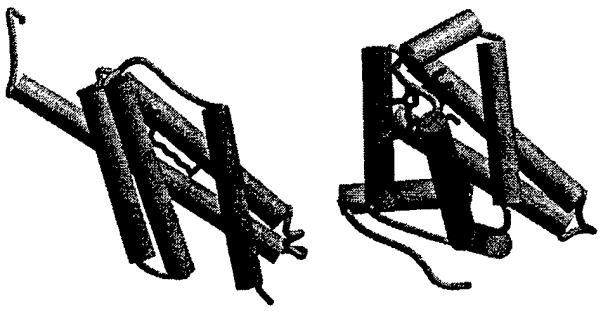


Figure: *Comparison of BACO-1 (left) and the oxygen-sensing domain from Bacillus subtilis (right).*

Very recently, the structure of an oxygen sensor from Bacillus subtilis has been determined. It has a related fold to BACO-1, and binds heme in a location similar to that of the BACO-1 co-factor (BACO-1 has no equivalent of the heme-linked histidine, so

cannot be a heme-binding protein). The oxygen sensor and BACO-1 can be superimposed with an RMSD of 1.8 Å for 70 Cα atoms), although there are different helical extensions at the N- and C-termini, and the helices pack distinctly around the larger heme co-factor. Thus the BACO fold is a subfamily within a larger family of globin-like sensor domains.

C.2.d NMR evaluation of BACO-1 as a CO_2 sensor. We studied the effect on CO_2 on pX01-118 using 1D ^{1}H NMR. The aliphatic region of the spectra are shown in Figure 19 (upper panel). The protein resonances appear quite broad, suggestive of a large system. Nevertheless, the spectrum changes significantly upon binding CO_2, and a close-up of the amide region (Figure: lower panel) spectra clearly shows new peaks appearing and old disappearing when CO_2 is added to the system. Although these data are preliminary, they suggest a striking effect on CO_2 binding to the protein, and are consistent with a conformational change in the protein.

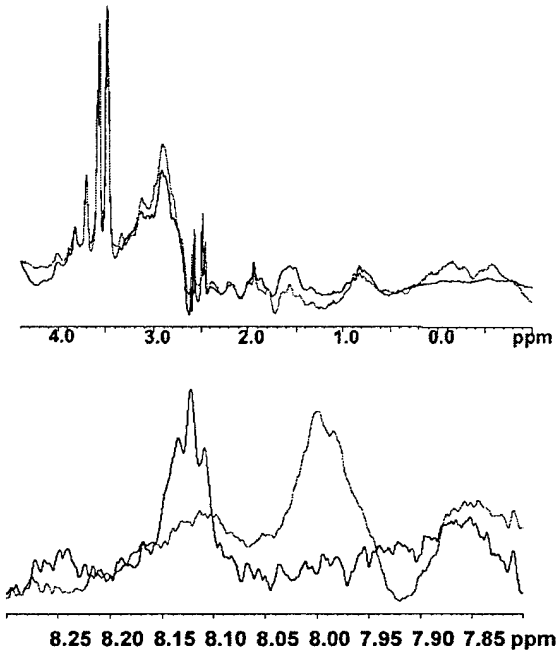


Figure: *1D ^{1}H NMR spectra of recombinant 118 in presence (red) and absence (blue) of CO_2. The aliphatic region of the spectra are shown in upper panel. Zoom-in of the amide region (lower panel) shows new peaks appearing and old peaks disappearing (black, apo; red is + 5 mM Na_2CO_3; the pH did not change (6.0, buffered by 50 mM KPi).*

C.2.e Cloning, Expression and crystallization of BACO-2/pX02-61: The gene encoding pXO2-61/BACO-2 was synthesized by GenScript Corporation and cloned into a pET-28a vector (Novagen). Following cleavage of the His-tag with thrombin, the solution was further purified using a Superdex75 gel filtration column. The protein runs on SDS-PAGE as expected with a M.W. of ~16.4 kDa. On a sizing column, the

estimated M.W. is ~17 kDa, suggestive of a monomer in solution. This behavior contrasts with that of BACO-1, which runs as a dimer.

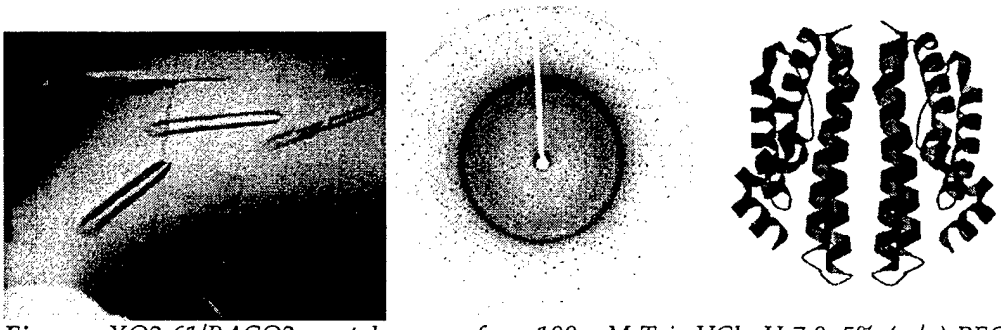


Figure: *pXO2-61/BACO2 crystals grown from 100 mM Tris-HCl pH 7.0, 5% (w/v) PEG-1000. (Right panel) Diffraction pattern from a capillary-mounted crystal*

Purified BACO-2 was crystallized by microbatch under paraffin oil. Orthorhombic crystals were grown from 1M NaI solution within 2 days. The space group is $P2_12_12_1$ with unit cell parameters a=44 Å b=62 Å c=124 Å. One high and one low resolution data set were measured. They have been scaled together and a molecular replacement run with pXO1-118 as model has been done. Using synchrotron radiation, data sets up to 1.5 Å have been collected. The asymmetric unit contains two molecules. Model building and refinement was carried out in programs O and CNS. The current BACO-2 model consists of residues 3-136, with R_{FREE}=27 % and appropriate stereochemistry. As expected, the BACO-2 structure is very simialr to that of BACO-1, although it lacks continuous density for a co-factor.

Appendix 4

·ystal structure of B. anthracis amidase homologous to a bacteriophage lysin

ıcteriophage lysin is a class of protein enzyme that is used by phage to break open its cterial host in order to release its progeny particles. Lysin is an amidase that targets d breaks down peptidoglycan, an important cell wall cross-linking component in cteria. Recently, the lysin (plyG) from the gamma phage of the Bacillus anthracis has en isolated and proved to be lethal to the hosts when applied to bacteria culture as rified protein (Schuch R, Nelson D and Fischetti VA (2002). A bacteriolytic agent that tects and kills Bacillus anthracis. *Nature* **418** 884-889). This discovery opened a new ıy in which anthrax could be treated and/or detected.

:nomic sequence analysis of the *B. anthracis* revealed that there is a gene (N-ɛtylmuramoyl-L-alanine amidase; EC 3.5.1.28) with high sequence homology with the ɣG (82%, see the above sequence alignment). There are only a few amino acid ferences in the N-terminal region 160 amino acids (93% identities) consisting of the talytic amino acids (catalytic domain). The catalytic Tyr and Lys are absolutely nserved, which might suggest that the two proteins could have very similar catalytic ɛchanism. The amino acid sequence differences appear mainly in the C-terminal region, ıich is thought to be a bacterial cell-wall carbohydrates binding domain. It is not known ıether the differences in the binding domain would suggest a different binding site on ɛ bacterial cell wall. But the highly homologous catalytic domain may imply that they ıginated from a single source in a recent time.

quence comparison with the gamma phage plyG

ɔre = 400 bits (1028), Expect = e-111 Identities = 194/234 (82%), Positives = 213/234 (91%), Gaps = 34 (0%)

```
yG    1   MEIQKKLVDPSKYGTKCPYTMKPKYITVHNTYNDAPAENEVSYMISNNNEVSFHIAVDDK 60
          MEI+KKLV PSKYGTKCPYTMKPKYITVHNTYNDAPAENEV+YMI+NNNEVSFH+AVDDK
_ami 1    MEIRKKLVVPSKYGTKCPYTMKPKYITVHNTYNDAPAENEVNYMITNNNEVSFHVAVDDK 60

yG    61  KAIQGIPLERNAWACGDGNGSGNRQSISVEICYSKSGGDRYYKAEDNAVDVVRQLMSMYN 120
          +AIQGIP ERNAWACGDGNG GNR+SISVEICYSKSGGDRYYKAE+NAVDVVRQLMSMYN
_ami 61   QAIQGIPWERNAWACGDGNGPGNRESISVEICYSKSGGDRYYKAENNAVDVVRQLMSMYN 120

yG    121 IPIENVRTHQSWSGKYCPHRMLAEGRWGAFIQKVKNGNVATTSPT-KQNIIQSGAFSPYE 179
          IPIENVRTHQSWSGKYCPHRMLAEGRWGAFIQKVK+GNVA+ + T KQNIIQ+GAFSPYE
_ami 121  IPIENVRTHQSWSGKYCPHRMLAEGRWGAFIQKVKSGNVASATVTPKQNIIQTGAFSPYE 180

yG    180 TPDVMGALTSLKMTADFILQSDGLTYFISKPTSDAQLKAMKEYLDRKGWWYEVK 233
           PD +GAL SL MT   I+  +GLTY ++ PTSD QL+A KEYL+RK WWY+ K
_ami 181  LPDAVGALKSLNMTGKAIINPEGLTYIVTDPTSDVQLQAFKEYLERKDWWYDDK 234
```

order to explore the possibility of using the amidase as a defense or treatment against thrax attack, the mechanism of the enzyme must be studied in detail. We are currently llaborating with Dr. Philip Hanna to test the bacteriocidal effects of the amidase. The romosomal copy of a class II amidase, consisting of 234 amino acids (NP_657904),

ıs cloned into the pET22 expression vector. Expression of protein was performed using ⌐21DE3 using standard protocol. The protein was purified using HITRAP SP- ɔharose (Amersham Pharmacia) and Sephacryl 200 HR column (Amersham armacia) with zinc containing buffers. The DNA sequence and molecular weight were rified by automated DNA sequencing and MALDI mass spectrometry.

ıe full length protein was only soluble up to concentration of 5 mg/ml, but did not ystallize under any of 576 conditions tested. Limited proteolysis using elastase was ɔn used to produce a smaller fragment. Mass spectrometry and peptide sequencing ɔwed that the final product consisted of the first 159 amino acids. This fragment ntains the conserved catalytic domain. The proteolytic reaction was scaled-up, and ɔduct used for crystallization. The catalytic domain was crystallized in 1.2 M $ıH_2PO_4$/0.8 M K_2HPO_4 0.1 M TrisCl pH 7.0. The crystals belong to spacegroup $P6_1$, th cell dimensions a= 164.1 Å, c= 37.6 Å. Phases were generated using a lenomethionine (SeMet) derivative, which was produce in M9 minimal medium pplemented with vitamins, glucose, ammonium sulfate and seleno-L-methionine. The ıstase-digested product was then crystallized under conditions identical to the native ɔtein.

ultiple anomalous diffraction experiments of the SeMet protein were carried out at the anford synchrotron radiation laboratory (SSRL). Three wavelengths were used: peak , gh energy remote, and inflection at 0.9792 A, 0.8919 A and 0.9794 A, respectively. ıe SeMet-amidase diffracted to a resolution of 2.0 A (94% completeness, R-factor of %). Three protein molecules were found in the asymmetric unit. Rebuildingand inement were done using O and CNS. The current R-value is 0.3067 with R_{free} of 3407, B-factor= 33.6399 A^2.

scription of the structure of B. anthracis amidase

uctural comparison between T7 lysozyme and amidase from B. anthracis

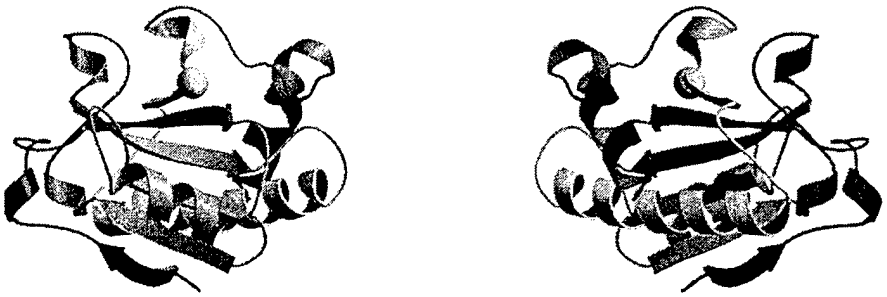
```
                0        1         2         3         4         5         6
                12345678901234567890123456789012345678901234567890123456789012345678901234567890
                 SSSSS  HHHH          SSSSSSSS      HHHHHHHHHH         SSSSS
_ami            EIRKKLVVPSKYGTKCPYTMKPKYITVHNT--YNDAPAENEVNYMITNNNEVSFHVAVDD
BA              --AKQRESTDAIFVHCSATKPSQNVGVREIRQWHKEQGWLDVGYHFIIKRDGTVEAGRDE
                          SSSSSS          HHHHHHHHHH         SSS     SSS
                   *:   ..   .:*. *  .: : *::   ::.  .  :*.* :  :.: :.... *:

                6        7         8         9         10        11        12
                12345678901234567890123456789012345678901234567890123456789012345678901234567890
                 SSSSSSHHHH             SSSSSSSS     HHHHHHHHHHHHHHHHHHHHHHHH
_ami            KQAIQGIPWERNAWACGDGNGPGNRESISVEICYSKSGGDRYYKAENNAVDVVRQLMSMY
BA              MAGVGSHAKGYNHNSIGVCLVGGIDDKGKFDANFTP-------AQMQSLRSLLVTLLAKY
                        HHHSSSSSS                       HHHHHHHHHHHHHHHH
                  .: . .   *  : *     *  :. ..:  ::          :.  .::  *:: *

                12        13        14        15        16
                123456789012345678901234567890123456789012
                     SSSSSS SSS   HHHHHHH HHHHHHHHHHH
_ami            NIPIENVRTHQSWSGKYCPHRMLAEGRWGAFIQKVKSGNVA
```

```
BA              EGAVLRAHHEGGVAPKACPS--FDLKRWWEKNELVTSDRG-
                   SSSSHHH               HHHHHHH
                : .: ..: . . : * **   :   **    : *.*..
```

ıe structure consists of a six-stranded β-sheet surrounded by six helices. The overall ld and topology of resembles that of the T7 enterobacteriophage lysozyme (1LBA; ıeng et al., Proc. Natl. Acad. Sci. USA **91** 4034-4038 (1994)), although the sequence ɛntity with the T7 lysozyme is only 12%. Amino acids coordinating the zinc ion in the o structures are also similar, consisting of 2 histidines, 1 cysteine (shown as bold type ıt in the above sequence alignment) and a water molecule. The active site of the zyme is in the cleft near to the zinc ion. T7 lysozyme uses a Tyr-46 and a Lys-128 for catalytic activity, whereas the B. anthracis enzyme probably uses Tyr-42 and Lys-4. Further analysis fo the structure is in progress.



180° ->

vo views of the B. anthracis amidase. The Zn ion is shown as a gray ball.

Appendix 5

Structural studies of AtxA, a member of the PRD family of transcriptional activators.

Modeling of AtxA: AtxA is the "master regulator" of virulence genes. However, its mechanism of action is unknown. Although recent literature reports have suggested that AtxA bears only limited sequence similarity with proteins of known structure, our analysis using our FFAS tools reveals the domain organization with high confidence, as described below. AtxA is a member of a family of multidomain transcriptional activators and antiterminators that contain a "PRD" (for Phospho-transferase system (PTS) Regulatory Domain). They all contain an N-terminal nucleic acid binding domain, two PRD domains, and in the case of the activators, a C-terminal PTS IIA or IIB domain. Our modeling studies suggest that AtxA contains:

1) A DNA–binding domain at its N-terminus (residues 1-135) that is homologous to the Diphtheria Toxin repressor (>97% confidence). We built a 3D model of this region based on the crystal structure of the DT repressor in complex with DNA. This model reveals the conservation of basic residues on a helix-turn-helix scaffold that would engage the DNA phosphate backbone; it also reveals a conserved hydrophobic dimerization interface in domain 2. Thus, the modeling studies clearly suggest that the N-terminal part of AtxA is a dimeric DNA-binding module.

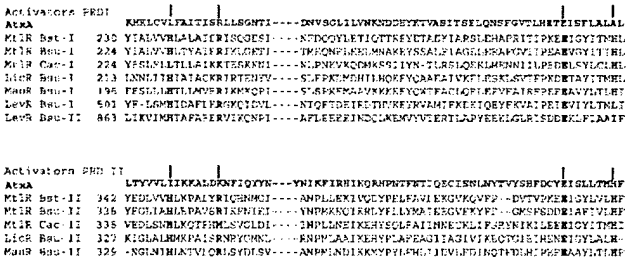


Figure: *Sequence alignments of the PRD domains of transcriptional activators. The vertical bar indicates conserved residues, of which the four histidines are phosphorylated during signal transduction. Note that AtxA contains only one of these histidines.*

2) Two PRD domains (domain 3 and 4; Residues 160-390.) This is predicted with even higher confidence (>99%), and a reliable 3-dimensional model can be built based on the crystal structure of the Lict transcriptional antiterminator [61]. In other members of this family, the duplicated PRD module is phosphorylated on 4 conserved histidines by a phosphotransferase system (PTS) in response to an environmental cue. The phosphorylations are thought to modify the stability of the dimeric proteins and thereby the RNA- or DNA-binding activity of the effector domain. However, sequence alignment of AtxA shows that only one of the 4 histidines is conserved.

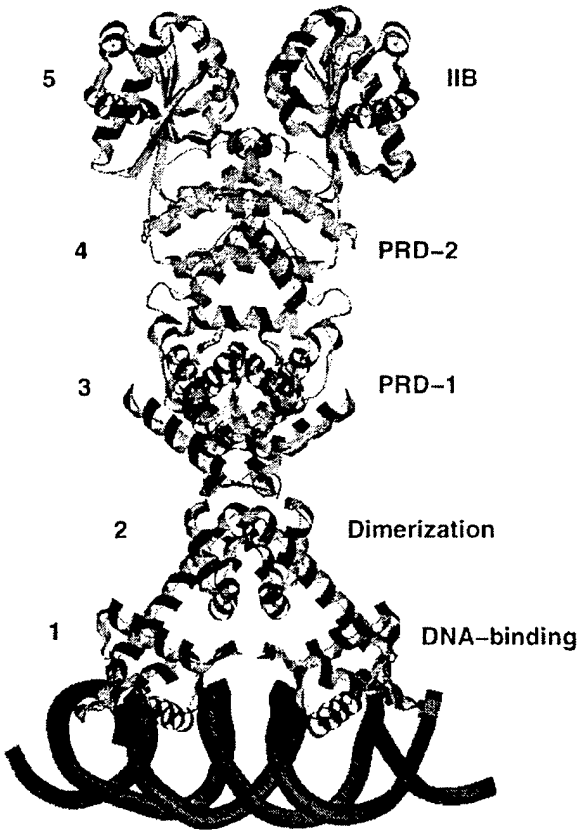


Figure: *Hypothetical model of AtxA bound to DNA, showing domain type at right and domain numbering at left.*

3) At the C-terminus is a PTS IIB domain (>97% confidence), residues 385-475. Previously characterized members of the family have an invariant cysteine residue that is phosphorylated during signal transduction. Atxa does not share this phosphorylatable cysteine.

Taken together, this modeling exercise suggests that AtxA is regulated in a non-canonical fashion, most likely not by phosphorylation of histidines and cysteines.

Expression constructs

Constructs for the expression as His-tag fusions of the following fragments of AtxA in E.coli, strain BL21(DE3), were produced:

Full-length, alone and co-expressed with pXO1-118
Putative DNA-bonding domain: 1-141 and 1-160
Putative regulatory domains including the PRD homology domain:141-393, 141-475, 162-393, 162-475
PTS IIB homology domain: 388-475

Expression of the putative DNA-binding domain

His-tagged AtxA 1-141 and 1-160 were expressed for 5 hours at 37°C and were found in the inclusion body insoluble fraction. After purification over a Superdex 200 10/30 gel filtration column (Amersham) under denaturing condition, fragment 1-141 was refolded with the surfactant/cycloamylose method. A smaller amount of the same fragment was obtained in soluble form using traditional dialysis refolding methods.

In order to characterize the possible DNA binding of this polypeptide by electrophoretic mobility shift assay (EMSA) we have created DNA fragments from the promoter region of the pagA gene, one of the targets of AtxA, by PCR amplification.

Expression of full-length and the PRD homology domain of AtxA

His-tagged, full-length AtxA was expressed at 15 and 37°C and the product was only detectable by SDS PAGE with western blotting using monoclonal antibodies direted against a His_5 tag (Novagen). The product is found in both the insoluble and the soluble fractions. When pXO1-118 and AtxA were co-expressed at 37°C, an increase in total, but not soluble, AtxA was observed, while pXO1-118 expression levels were unaltered, when compared to the expression of each protein alone. Similar results were obtained for the expression of AtxA 141-393, 162-475 and 162-475 at 37°C. We are currently cloning shorter fragments within the regulatory region of atxA in order to identify possible short sequences responsible for the low level of expression observed.

Expression of AtxA (388-475)

AtxA (388-475) could be expressed solubly to satisfactory levels as judged by SDS PAGE.

Other expression systems

We are currently in the process of cloning the protein fragments described above for expression in B. megaterium (MoBiTec), Sf9 cells and in a cell-free system (RTS, Roche Applied Sciences). Full-length AtxA (1-475), as well as the putative regulatory domains (1-141, 141-475) could be expressed in sf9 cells at low level (western blot detectable) with considerable degradation problem (ladders shown in Western blot). To avoid the degradation, secretion expression of full length AtxA is underway. At this time, we have the baculovirus and are doing amplification.

AtxA

Full-length AtxA (1-475), as well as the putative regulatory domain (141-393, 162-393, 141-475, 162-475) could only be expressed at very low levels (detectable only by western blotting) in E.coli despite considerable effort. Expression at lower temperatures, as low as 15 C, improved solubility and increase the amount of full-length product, but not to

satisfactory yields. Co-expression of the full-length with pXO1-118 did not bring any meaningful improvement. The putative DNA binding domain was expressed in the insoluble fraction and could not be solubilized in suitable amounts for DNA-binding studies. Cloning of AtxA homologs AcpA and AcpB (about 25% identical, 50% similar to AtxA) is underway. An attempt to identify possible short fragments of AtxA that are responsible for low level expression has also been started.

nature
structural &
molecular biology

Identification of small molecule inhibitors of anthrax lethal factor

Rekha G Panchal[1], Ann R Hermone[1,5], Tam Luong Nguyen[1,5], Thiang Yian Wong[2], Robert Schwarzenbacher[2], James Schmidt[3], Douglas Lane[1], Connor McGrath[1], Benjamin E Turk[4], James Burnett[1], M Javad Aman[3], Stephen Little[3], Edward A Sausville[1], Daniel W Zaharevitz[1], Lewis C Cantley[4], Robert C Liddington[2], Rick Gussio[1] & Sina Bavari[3]

The virulent spore-forming bacterium *Bacillus anthracis* secretes anthrax toxin composed of protective antigen (PA), lethal factor (LF) and edema factor (EF). LF is a Zn-dependent metalloprotease that inactivates key signaling molecules, such as mitogen-activated protein kinase kinases (MAPKK), to ultimately cause cell death. We report here the identification of small molecule (nonpeptidic) inhibitors of LF. Using a two-stage screening assay, we determined the LF inhibitory properties of 19 compounds. Here, we describe six inhibitors on the basis of a pharmacophoric relationship determined using X-ray crystallographic data, molecular docking studies and three-dimensional (3D) database mining from the US National Cancer Institute (NCI) chemical repository. Three of these compounds have K_i values in the 0.5–5 μM range and show competitive inhibition. These molecular scaffolds may be used to develop therapeutically viable inhibitors of LF.

Anthrax, a disease caused by *Bacillus anthracis*, has recently been the subject of intense interest because of its use as a biological weapon against human populations. The inhalation of *B. anthracis* spores is often fatal if the condition is not properly diagnosed and treated with antibiotics during the early stages of infection. In many cases antibiotic regimes may not be effective, especially if there is bacterium overload, which causes large amounts of lethal toxin to be released. Hence, a new level of adjunct treatment is needed to inactivate the toxins released by *B. anthracis*.

Anthrax toxin (AT) consists of three proteins: lethal factor, protective antigen and edema factor, all of which work in concert to kill host cells. Initially, PA binds to an AT receptor[1,2] on the host cell surface, where it is subsequently cleaved by furin (or furin-like proteases) to produce a 20-kDa N-terminal fragment (PA_{20}) and a 63-kDa C-terminal fragment (PA_{63})[3,4]. After cleavage, seven PA_{63} monomers assemble to form a heptameric prepore capable of binding both LF and EF. Upon binding of LF or EF, the entire complex undergoes receptor-mediated endocytosis. It is hypothesized that the acidic endosomal environment causes a conformational change in the PA_{63} heptamer to produce a functional pore that traverses the membrane and translocates the two enzymatic moieties LF and EF into the cell cytosol. EF is a calmodulin-dependent adenylate cyclase[5]; LF is a Zn-dependent metalloprotease that cleaves several members of the MAPKK family near the N terminus[6,7]. This cleavage prevents interaction with, and phosphorylation of, downstream MAPK[8], thereby inhibiting one or more signaling pathways. Through a mechanism that is not yet well understood, this results in the death of the host. Recent studies suggest that the inactivation of p38 MAPK induces apoptosis in LF-exposed macrophages, thereby preventing the release of chemokines and cytokines, and preventing the immune system from responding to the pathogen[9].

Based on the current understanding of the mechanism of anthrax toxin, methods may be developed to inhibit various steps in toxin assembly and/or function. In one antitoxin therapy approach, dominant-negative PA mutants have been generated that coassemble with the wild-type PA protein, blocking the translocation of LF and EF across the cell membrane. Such PA mutants are potent inhibitors of anthrax toxin in both cell-based assays and *in vivo* animal models[10,11]. In a second approach, a peptide inhibitor that binds to the heptameric PA and prevents the interaction of PA with LF and EF has shown efficacy in animals[12].

The lethal action of anthrax toxin may also be inactivated by molecules that inhibit the protease activity of LF. So far, the only known small molecule inhibitors of LF are nonspecific hydroxymates that are effective at >100 μM concentration[13] and more recently reported hydroxymate derivatives of peptide substrate that inhibit LF at nanomolar concentrations[14]. In this study, we identified several small (nonpeptidic) compounds that inhibit anthrax LF protease activity with K_i values in the 0.5–5 μM range. We approached anthrax therapeutic development (in parallel with the peptidomimetic approach used by Turk *et al.*[15]; this issue) using structure-based discovery to

[1]Developmental Therapeutics Program, NCI Frederick, Frederick, Maryland 21702-1201, USA. [2]The Burnham Institute, 10901 North Torrey Pines Road, La Jolla, California 92037, USA. [3]United States Army Medical Research Institute of Infectious Diseases, 1425 Porter Street, Frederick, Maryland 21702, USA. [4]Division of Signal Transduction, Beth Israel Deaconess Medical Center, Harvard Institutes of Medicine, Room 1007, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA. [5]These authors contributed equally to this work. Correspondence should be addressed to R.G.P. (panchal@dtpax2.ncifcrf.gov) or S.B. (bavaris@ncifcrf.gov).

Published online 29 December 2003; doi:10.1038/nsmb711

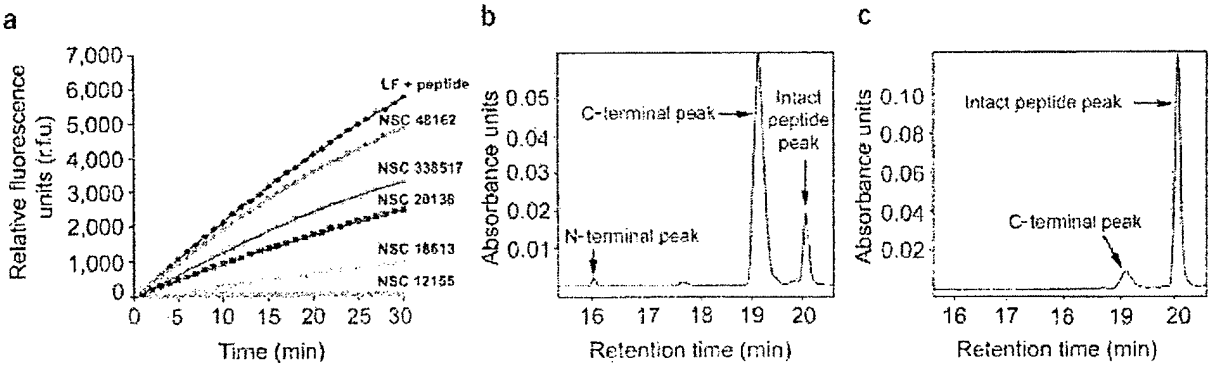


Figure 1 A two-stage assay for screening and validating small molecule inhibitors of anthrax lethal factor. (a) Representative data from a fluorescent plate reader assay showing different degrees of inhibition by compounds from the NCI Diversity Set. (b) HPLC-based assay without inhibitor, showing the N- and C-terminal cleavage products after incubation of the substrate with LF for 30 min. (c) HPLC-based assay with inhibitor NSC 12155 showing a reduced C-terminal peak area at 365 nm, indicating strong inhibition of LF activity.

identify small organic molecules as lead candidates. Specifically, we used molecular diversity screening combined with 3D database searching and molecular modeling. The LF X-ray crystal structure reported by Pannifer *et al.*[16] was useful during the structure-based drug discovery portion of these studies.

The first phase of this study involved a high-throughput screen (HTS) of small molecules from the NCI Diversity Set to identify LF inhibitors. Hits identified from the HTS were verified with an HPLC-based assay. Afterwards, we used X-ray crystallography and molecular modeling (conformational sampling, database mining and molecular docking) to identify additional lead therapeutics. Based on an iterative process of compound selection and biological testing, a pharmacophore for LF inhibitors was developed.

RESULTS

High-throughput screening and hit validation

To screen and identify compounds that inhibit LF activity, we developed a high-throughput fluorescence-based assay. An optimized peptide (KKVYPYPME; B.E.T. *et al.*, unpublished data) with a fluorogenic coumarin group at the N terminus and a 2,4-dinitrophenyl (dnp) quenching group at the C terminus was used as LF substrate for *in vitro* assays. After cleavage by LF, fluorescence increased (excitation and emission wavelengths, 325 and 394 nm, respectively). After standardization of the high-throughput assay, the 1,990 compounds in the NCI Diversity Set were tested (Fig. 1a). Compounds that showed >75% inhibition were selected for validation using an HPLC-based assay. This eliminated false positives due to fluorescence quenching by some of the test compounds. Using the HPLC-based assay (Fig. 1b,c), compounds that showed >50% inhibition were selected for further study. The HPLC assay, in addition to eliminating false positives, was a more rigorous test of LF inhibition, as a lower inhibitor concentration (20 μM) was used (compared with 100-μM concentration used in the fluorescence-based assays). Furthermore, the identified LF inhibitors did not inhibit a range of different proteases, thus confirming that these compounds did not inhibit LF promiscuously (see Supplementary Fig. 1 online).

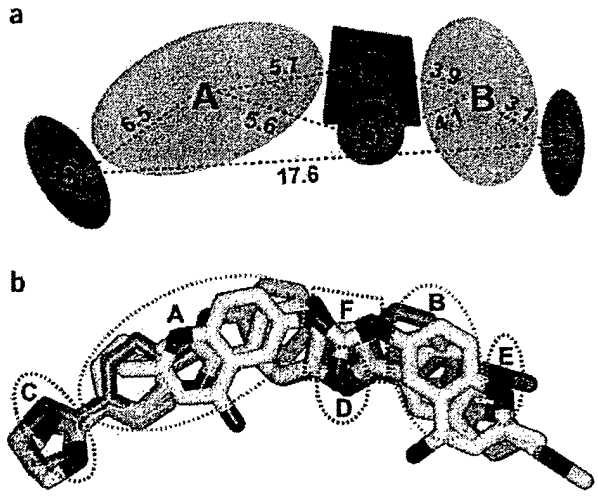
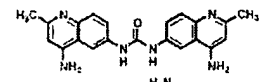
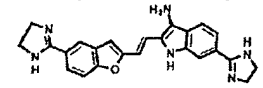
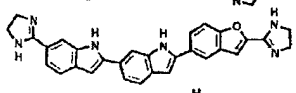
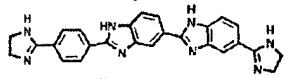
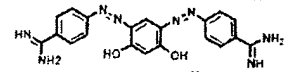
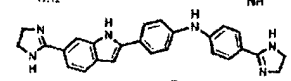
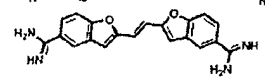


Figure 2 General pharmacophore model of the LF inhibitors. (a) Black dashed lines depict the distances between the various centroids of the pharmacophore centers. Green ellipses (A and B) are aromatic centers; red ellipses (C, D and E) are polar centers (hydrogen bond donors or acceptors); blue region (F) is a neutral linker that may include a variety of polar or hydrophobic groups. (b) Pharmacophoric overlap of LF inhibitors (stick rendering) and their correspondence to the general LF inhibitor pharmacophore shown in Figure 2a. The pharmacophoric overlap regions of compounds are highlighted in dashed lines (green, aromatic centers; blue, neutral (polar or hydrophobic groups acceptable) linker region; red, polar centers. For all structures: nitrogen, blue; oxygen, red. Carbon atoms for NSC 12155, yellow; for NSC 357756, magenta; for NSC 369721, green; for NSC 369728, light blue. The pharmacophore is based on the energy-refined X-ray conformation of NSC 12155 bound to LF. These data were combined with molecular docking studies of structurally related analogs (Table 1) from 3D database mining studies.

Table 1 Two-dimensional chemical representations of LF inhibitors with percent inhibition at a compound concentration of 20 μM, K_i values and type of inhibition

Structure	NSC number	% inhibition	K_i (μM)	Inhibition type
	12155	95	0.5 ± 0.18	Competitive
	357756	90	4.9 ± 1.7	Competitive
	369718	90	N.D.	N.D.
	369721	90	4.2 ± 0.21	Competitive
	359465	48	N.D.	N.D.
	377362	33	N.D.	N.D.
	240899	0	N.D.	N.D.

N.D., not determined.

Pharmacophoric features of anthrax LF inhibitors

We identified 19 compounds with >50% LF inhibition (at 20 μM inhibitor concentration) from the NCI Diversity Set screen. These included several organometallic and charged molecules. Here, we chose to concentrate on only relatively small organic compounds for structure-based studies, as these molecules are more likely to show therapeutic potential. The conformational spaces of two leads, NSC 12155 and NSC 357756, were subsequently explored to generate multiple pharmacophoric hypotheses, which were then used in 3D database mining studies to identify additional LF inhibitors. We carried out several iterations of this process, which consisted of 3D database mining of the entire NCI repository (as well as commercially available chemical repositories including the Available Chemicals Directory, MayBridge and BioByte) and subsequent biological testing, to identify new inhibitors. During this process >60 compounds were tested and most of them were inactive. However, six of the compounds, which showed a range of LF inhibitory potency, were used to develop and refine a consistent pharmacophore (Fig. 2a). A 3D superimposition of four of the most potent LF inhibitors (NSC 12155, NSC 357756, NSC 369718 and NSC 369721) (Fig. 2b) exhibits an excellent overlay of the polar heteroatoms and hydrophobic substituents of these molecules. The chemical structures of a range of identified LF inhibitors are shown in **Table 1**.

Kinetic studies

To determine the K_i values and types of inhibition mediated by the inhibitors (competitive, noncompetitive or uncompetitive), we determined kinetic constants of the peptide substrate and compared them with those obtained in the presence of different inhibitor concentrations. The K_m and V_{max} values for the LF-catalyzed hydrolysis of the peptide substrate were 19 μM and 1.1 μmol min^{-1} mg^{-1} of LF, respectively. NSC 12155, NSC 357756 and NSC 369721 showed competitive inhibition (Table 1), as they had no effect on the V_{max}, but $K_{m(app)}$ increased with inhibitor concentration (see Supplementary Fig. 2 online).

Anthrax LF–NSC 12155 cocrystal structure

The crystal structure of LF in complex with NSC 12155 (the most potent inhibitor) was determined at a resolution of 2.9 Å (electron density map, Fig. 3a). NSC 12155 binds to the catalytic site of LF with its urea moiety close to the catalytic Zn atom (within 4 Å). One quinoline ring shows strong electron density near the side chain of His690, suggesting a favorable π-stacking interaction between the histidine's side chain imidazole and the quinoline ring (Fig. 3b). Conversely, the second quinoline showed poor electron density, indicating that there is more rotational freedom about its quinoline-urea bond. Despite the overall lack of a strong positional preference for this quinoline, a more consistent density was detected near its amino substitution, indicating a slightly greater preference for a 'C-shaped' conformation of NSC 12155 when bound to LF. This is consistent with the pharmacophoric overlap shown in Figure 2b.

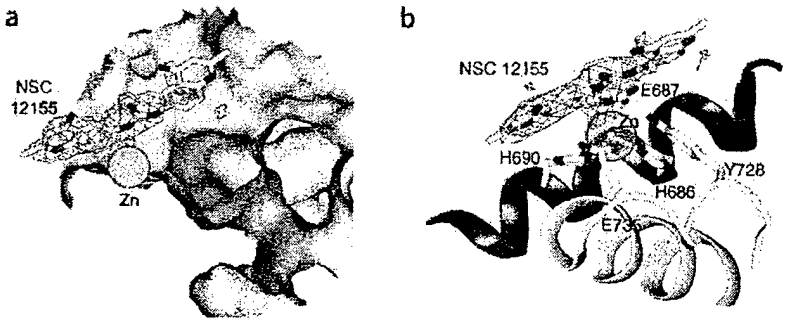


Figure 3 X-ray crystal structure of the LF–NSC 12155–Zn complex. The electron density surrounding NSC 12155 shown in these figures are $2F_o - F_c$ difference maps (see Methods) calculated at 2.9-Å resolution. **(a)** Detailed view of the electron density trace and overall model fit of NSC 12155. Molecular surface of LF colored by charge (red, negative; blue, positive), with Zn^{2+} (cyan), and the model of the inhibitor molecule NSC 12155 (yellow) in stick representation. The difference map, $2F_o - F_c$, is contoured at 1.1 σ level. **(b)** The inhibitor NSC 12155 bound in the active site of LF. The difference map, $2F_o - F_c$, is contoured at 1.0 σ. A portion of NSC 12155 appears nonrigid owing to a rotatable bond, and almost full electron density coverage is seen for this portion at a contour level of 0.6 σ. Inhibitor molecule (yellow), zinc-coordinating residues (H686, H690, E735) and catalytic residues (E687, Y728) are in stick representation. The Cα atoms of residues 680–694 (green, background) and 726–742 (beige, foreground) are in ribbon representation. The Zn^{2+} ion (cyan) is a lined sphere, and its hydrogen bonds with His686, His690 and Glu735 are represented as aligned small white spheres. These figures were prepared using SPOCK (http://mackerel.tamu.edu/spock/).

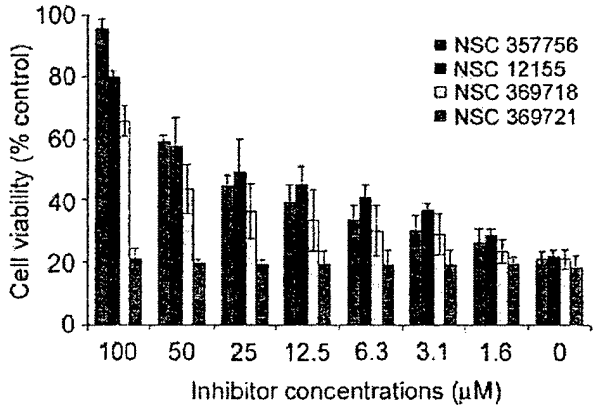


Figure 4 Efficacy of LF inhibitors in a cell-based toxicity assay. J774A.1 cells were pretreated with either DMSO control or various concentrations of inhibitors, and then incubated with anthrax lethal toxin. After 4 h, cell viability was determined with MTT dye.

Molecular docking studies

To further investigate whether the C conformation has an important role during the binding of NSC 12155 to LF, we used molecular docking to study the conformational preference of the freely rotating quinoline in the NSC 12155–LF model. Results from these analyses suggest that the NSC 12155 scaffold does prefer the planar C conformation to the 'L-shaped' conformation when bound to LF. This is further supported by the following: (i) quantum mechanical calculations at the level of density functional theory, as well as analysis of related crystal structures (data not shown), support a planar preference (either L or C shaped) for NSC 12155; (ii) rotation of the 'free' quinoline out of plane to its planar L conformation results in unfavorable hydrophobic-polar interactions between the amino groups of NSC 12155 and the side chain of Val675; (iii) in the planar C conformation, the urea oxo and quinoline amino substituents of NSC 12155 are more likely to engage in favorable intramolecular acid-base interactions; (iv) molecular docking studies of 32 substituted quinoline and urea derivatives (chemoinformatically mined from the NCI repository), which were inactive in the LF assay (data not shown), indicate that these scaffolds are either incapable of forming the preferred C conformation of NSC 12155 or lack features that would enable favorable binding; and (v) additional modeling studies of NSC 12155 indicate that the urea nitrogens are within range to form favorable acid-base interactions with the carboxylate of Glu687 (supported by X-ray data: distances of the urea nitrogens of NSC 12155 are 4.12 Å and 4.72 Å from OE1 and OE2 of Glu687, respectively).

Cytotoxicity assay

To determine the ability of the small molecule inhibitors to protect macrophages against LF, we pretreated the cells with NSC 12155, NSC 357756, NSC 369718 or NSC 369721 at concentrations ranging from 1 to 100 μM and further incubated them in the presence of anthrax lethal toxin. Cell viability was determined using MTT dye (Fig. 4). NSC 357756 showed 96% protection at 100 μM, whereas NSC 12155 and NSC 369718, the most potent of the LF inhibitors *in vitro*, showed lower protection at 100 μM. These three compounds showed some protection <25 μM, suggesting that they might be good leads against lethal toxin *in vivo*. Additionally, NSC 369721 was ineffective even at 100 μM in the cell-based toxicity assay. The moderate protection of these inhibitors is probably attributable to their limited ability to penetrate the macrophage cell membrane. The cell-based data will aid in the development of second-generation LF inhibitors.

DISCUSSION

Molecular docking studies of both inactive and active analogs of the compounds shown in Table 1 are consistent with the common pharmacophore (Fig. 2a) proposed in this study. For example, the amidine groups of NSC 240899 formed unfavorable steric and polar interactions when docked in the NSC 12155-binding site, which may explain this compound's complete lack of LF inhibition despite its structural similarity to NSC 357756. NSC 357756, NSC 369718 and NSC 369721 did not engage in unfavorable interactions when docked in the NSC 12155-binding site, supporting this hypothesis. However, the large size and solvent-exposed nature of the LF-binding groove also allows NSC 357756, NSC 369718 and NSC 369721 to assume several different binding modes near the enzyme's active site.

The X-ray structure of the LF–NSC 12155 complex and the extensive molecular docking studies with LF inhibitors also allow for the identification of favorable structural modifications that may enhance the potency of these compounds. For example, X-ray and molecular modeling studies of NSC 12155 indicate that the 0.5-μM K_i of this inhibitor could be improved by replacing one of the quinoline moieties with a pyrrole. Such a modification would provide an additional hydrogen bond with the carboxylate of Glu687. The planar C conformation of NSC 12155 could be stabilized by replacing its amino substituents with nitro groups, thus facilitating resonance throughout this scaffold. Additionally, our study in concert with Turk *et al.*[15] suggests that replacement of one of NSC 12155's quinoline rings with a tetra-aza-benzo[a] fluorene would enhance binding by placing additional molecular volume in the S1′ site of LF. Moreover, the deep S1′ pocket (visible in Fig. 3a, next to zinc) seems highly selective, such that a large hydrophobic ring structure would probably increase the affinity of an inhibitor for the LF active site.

In summary, these studies describe a first critical phase in generating therapeutically viable, small molecule (nonpeptidic) countermeasures for anthrax lethal toxin. During the next phase of inhibitor optimization, information obtained from the cell-based assay will guide the incorporation of structural components that will increase inhibitor bioavailability, while at the same time allowing for optimal binding affinity in the LF substrate-binding cleft.

METHODS

Diversity set. In brief, the NCI Diversity Set is a collection of 1,990 compounds chosen (from 71,756 open compounds in the NCI chemical repository with ≥1 g inventory) to cover a large, diverse range of molecular scaffolds and pharmacophore features, while also being relatively rigid (all compounds in the Diversity Set have five or fewer rotatable bonds, facilitating pharmacophore development and conformational sampling). For a detailed description of the Diversity Set compound selection and criteria see http://dtp.nci.nih.gov/branches/dscb/diversity_explanation.html.

Fluorescent plate-based assay. For high-throughput screening in 96-well plates, the reaction volume was 100 μl per well. Master mix containing 40 mM HEPES, pH 7.2, 0.05% (v/v) Tween 20, 100 μM $CaCl_2$ and 1 μg ml^{-1} of LF was added to each well containing 100 μM of NCI Diversity Set compound. The reaction was initiated by adding the optimized peptide substrate (MCA-KKVYPYPME[dnp]K amide), to a final concentration of 20 μM. Kinetic measurements were obtained every minute for 30 min using a fluorescent plate reader (Molecular Devices, Gemini XS). Excitation and emission maxima were 324 nm and 395 nm, respectively.

Table 2 Data collection summary of LF–NSC 12155–Zn complex crystal

Resolution range (Å)	25.0–2.90
Reflections	
Total	175,849
Unique	56,384
Completeness (%)[a]	99.5 (99.3)
R_{sym} (%)[a,b]	10.6 (49.8)
$I/\sigma I$[a]	11.7 (2.9)

[a]Values in parentheses are for the highest-resolution shell. [b]$R_{sym} = \Sigma|I - \langle I\rangle|/\Sigma\langle I\rangle$, where I is the observed intensity and $\langle I\rangle$ is the average intensity from multiple observations of symmetry-related reflections.

HPLC-based assay. An HPLC-based assay was used to validate the hits from the primary screen and eliminate the false positives obtained owing to fluorescence quenching. Reaction mix (30 µl total volume) containing 40 mM HEPES, pH 7.2, 0.05% (v/v) Tween 20, 100 µM $CaCl_2$, LF substrate (20 µM final concentration), with or without the inhibitor (20 µM final concentration), was incubated with LF (1 µg ml^{-1}) for 30 min at 30 °C. The reaction was stopped by adding 8 M guanidine hydrochloride in 0.3% (v/v) TFA. Substrate and products were separated on a Hi-Pore C18 column (Bio-Rad) using 0.1% (v/v) TFA (solvent A) and 0.1% (v/v) TFA + 70% (v/v) acetonitrile (solvent B). The column effluent was monitored at 365 nm, where the substrate and C-terminal cleavage products showed greater absorbance.

The HPLC-based assay was used for enzyme kinetic studies. Kinetic constants were obtained from plots of initial rates with seven concentrations of the substrate. For the best inhibitors, K_i and the type of inhibition were evaluated using seven different concentrations of the substrate ranging from 2 to 40 µM and four different concentrations of the inhibitor. K_i values for the competitive inhibitors were calculated using the equation $K_i = [I] / [(K_{int(app)} / K_m) - 1]$, where [I] is the inhibitor concentration[17]. K_i values in Table 1 are the averages ± s.d.

LF refinement and inhibitor docking. The structure of LF was energy-refined using the Discover (Accelrys) program's cff91 force field. Our strategy entailed using a step-down, template forced minimization procedure with the Zn coordination site fixed. This process was repeated until coordinates of the final model were within the experimentally determined X-ray crystallographic resolution. The inhibitor–enzyme structure coordinates were subsequently tether-minimized in the same manner as described above, and the final structure was subjected to hydropathic analysis using HINT (eduSoft).

Conformer generation. Conformational models of inhibitors were generated using Catalyst 4.7 (Accelrys). A 'best-quality' conformational search was used to generate conformers within 20 kcal mol^{-1} of the global energy minimum.

Data mining. Catalyst 4.7 (Accelrys) was used for all database mining. Briefly, the imidazole rings of NSC 357736 were used to form a three-dimensional search query (A.R.H. *et al.*, unpublished data). Subsequent molecular docking studies (see above) were used to suggest candidates for biological testing.

Quantum mechanical calculations. The conformations (L and C shaped) of NSC 12155 were fully optimized (until the norm of the gradient was $<5.0 \times 10^{-4}$) using DGauss (Oxford Molecular Group). Local spin density (LSD) correlation potentials were approximated by the Vosko-Wilk-Nusair method[18] and gaussian analytical functions were used as basis sets. LSD-optimized orbital basis sets of double ζ-split valence polarization quality[19] were used. In final optimizations, the BLYP exchange-correlation functional[20,21] was applied as a nonlocal gradient correction after each self-consistent field cycle.

Crystallization. Native, wild-type LF protein was crystallized using 13 mg ml^{-1} LF. Crystals were grown from 1.7 M $(NH_4)_2SO_4$, 0.2 M Tris-HCl, pH 7.5–8.0, 2 mM EDTA, using hanging-drop vapor diffusion[16]. Monoclinic crystals appeared after four days to two weeks, and were then harvested for experiments. The LF crystals belong to the monoclinic space group $P2_1$, with unit cell dimensions $a = 96.70$ Å, $b = 137.40$ Å, $c = 98.30$ Å, $\alpha = \gamma = 90°$, $\beta = 98°$, containing two molecules per asymmetric unit.

LF–inhibitor complexes. LF native crystals were harvested from the hanging drops in which they were grown, bathed in several rounds of fresh buffer without EDTA containing 1.9 M $(NH_4)_2SO_4$, 0.2 M Tris-HCl, pH 8.0, and left to soak in this solution for a further 30 min. These crystals were then used to obtain the protein–inhibitor–zinc complexes. All manipulations were done at room temperature (23–26 °C).

The LF–NSC 12155–Zn complex was obtained by soaking an individual native LF monoclinic $P2_1$ crystal in a solution of 1 mM $ZnSO_4$, 1.9 M $(NH_4)_2SO_4$, 0.2 M Tris-HCl, pH 8.0 for 5 min. The crystal was then transferred to a solution of 1.0 mM NSC 12155, 1% (v/v) DMSO, 1.9 M $(NH_4)_2SO_4$, 0.2 M Tris-HCl, pH 8.0 for 15 min. Finally, the crystal was transferred into a cryoprotectant solution of 1.0 mM NSC 12155, 2.4 M $(NH_4)_2SO_4$, 0.2 M Tris-HCl, pH 8.0, 2 mM EDTA, 25% (v/v) glycerol, and soaked at room temperature for 1 min. The crystal was then immediately mounted onto a cryoloop and flash-frozen in liquid nitrogen. All data were collected at 100 K.

Data collection. Datasets for the LF complexes were collected at the Stanford Synchrotron Radiation Laboratory (SSRL, Menlo Park, California, USA) on beamline 9-1 (wavelength = 0.983 Å). X-ray diffraction data were collected for the LF–NSC 12155–Zn complex to a resolution limit of 2.90 Å. Data collection statistics are shown in Table 2.

Structure solution and refinement. Collected data were processed in the HKL package[22]. Refinement and model building were done in CNS[23] and O[24], respectively. Using PDB entry 1J7N as the starting model, the model of LF alone was put through rigid body refinement and then minimization before the first initial maps were calculated for model building and further refinement. Excess electron density at 1.0 σ indicated the binding location of the inhibitor in the active site of LF. The model of the inhibitor was then built into this position and further refined in CNS[23]. The final *R*-factors were $R_{free} = 27.58\%$ and $R_{work} = 22.38\%$. The final model falls within or exceeds the limits of all the quality criteria of PROCHECK from the CCP4 suite[25].

Cytotoxicity assay. J774A.1 cells were preincubated with DMSO control or compounds for 30 min and then treated with PA (50 ng ml^{-1}) and LF (14 ng ml^{-1}). After 4 h incubation with the toxin, 25 µl of MTT (1 mg ml^{-1}) dye was added and the cells were further incubated for 2 h. The reaction was stopped by adding an equal volume of lysis buffer (20% (v/v) DMF and 20% (w/v) SDS, pH 4.7). Plates were incubated overnight at 37 °C and absorbance was read at 570 nm in a multiwell plate reader. Experiments were done in duplicate and repeated three independent times for each of the inhibitors tested. The results are the averages ± s.d.

Coordinates. The coordinates and structure factors for the LF–NSC 12155–Zn complex have been deposited in the Protein Data Bank (accession code 1PWP).

Note: Supplementary information is available on the Nature Structural & Molecular Biology website.

ACKNOWLEDGMENTS

This research was sponsored by the Medical Biological Defense Research Program, US Army Medical Research and Materiel Command, project no. 0242C012. We acknowledge the US National Cancer Institute for the allocation of computing time and staff support at the Advanced Biomedical Computing Center of the Frederick Cancer Research and Development Center. We thank the staff of the Stanford Synchrotron Radiation Laboratory (SSRL) for assistance during data collection. We also thank S. Leppla (US National Institutes of Health (NIH)) and D. Hsu for the LF protein preparation used in the crystal structure. Portions of this research were carried out at the SSRL, a national user facility operated by Stanford University on behalf of the US Department of Energy (DOE), Office of Basic Energy Sciences (BES). The SSRL Structural Molecular Biology Program is supported by the DOE, Office of Biological and Environmental Research, and by the NIH, National Center for Research Resources, Biomedical Technology Program, and the National Institute of General Medical Sciences.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 26 August; accepted 30 October 2003
Published online at http://www.nature.com/natstructmolbiol/

nature
structural &
molecular biology

The structural basis for substrate and inhibitor selectivity of the anthrax lethal factor

Benjamin E Turk[1,5], Thiang Yian Wong[2,5], Robert Schwarzenbacher[2], Emily T Jarrell[1], Stephen H Leppla[3], R John Collier[4], Robert C Liddington[2] & Lewis C Cantley[1]

Recent events have created an urgent need for new therapeutic strategies to treat anthrax. We have applied a mixture-based peptide library approach to rapidly determine the optimal peptide substrate for the anthrax lethal factor (LF), a metalloproteinase with an important role in the pathogenesis of the disease. Using this approach we have identified peptide analogs that inhibit the enzyme *in vitro* and that protect cultured macrophages from LF-mediated cytolysis. The crystal structures of LF bound to an optimized peptide substrate and to peptide-based inhibitors provide a rationale for the observed selectivity and may be exploited in the design of future generations of LF inhibitors.

Inhalational anthrax progresses rapidly to a highly fatal systemic infection[1]. The causative bacterium *Bacillus anthracis* secretes three plasmid-encoded toxin proteins that contribute to pathogenesis: protective antigen (PA), edema factor (EF) and lethal factor (LF)[2]. PA binds to a cell surface receptor and forms an oligomeric pore that translocates both EF and LF into the cytosol of target cells. The combination of PA and LF is known as lethal toxin (LeTx), and intravenous delivery of LeTx alone causes death in rodents[2,3]. In addition, *B. anthracis* strains deficient in either component of LeTx are greatly attenuated, suggesting an important role for the toxin in the disease[4]. As antibiotics alone typically fail against systemic anthrax unless administered at an early stage, LeTx has been proposed as a potential target for anthrax drugs to be used with antibiotics in combination therapy[1]. Several experimental approaches to LeTx neutralization based on inhibition of cellular LF uptake have shown efficacy in animal models[5,6].

LF is a zinc-dependent metalloproteinase that cleaves most MAP kinase kinase (MKK) enzymes at sites near their N termini[7–10]. Cleavage impairs the ability of the MKK to interact with and phosphorylate its downstream MAP kinase substrates by disrupting or removing a docking site known as the D-domain[11]. Inhibition of MAP kinase pathways by LF impairs dendritic cell and macrophage function and may help to establish infection[9,12]. Higher levels of toxin are cytotoxic specifically to macrophages and probably contribute to fatality later in the course of the disease[1,2,13,14]. Although the mechanisms by which MKK cleavage leads to macrophage cell death are not entirely known, p38 family MAP kinases seem to be required for survival of macrophages upon activation by bacterial endotoxins[15].

Efficient cleavage of MKKs requires interaction between an LF exosite that has not yet been characterized and a region in the MKK catalytic domain distal from the cleavage site[16]. However, mutation of residues surrounding the scissile bond in MKKs abolishes proteolysis, indicating that cleavage site recognition is also crucial to substrate selection by LF[7,15]. Accordingly, LF can cleave short peptides, and efficient substrates have been generated based on a consensus motif derived from MKK cleavage sites[17–19]. It is not clear, however, which positions surrounding the cleavage site are most critical for efficient catalysis, nor whether residues found in MKKs are optimal for cleavage by LF. Such information is important for the design of therapeutically useful small molecule LF inhibitors, as thus far only rather long (more than ten residues) peptide hydroxamates have been reported to specifically inhibit LF[19]. Here we take an unbiased approach to the discovery of LF substrates and inhibitors by selection from random pools of millions of peptides, and report the crystal structures of LF in complex with optimized substrates and small molecule peptide-based inhibitors.

RESULTS

Determination of the optimal peptide cleavage motif for LF

To gain insight into substrate recognition by LF and to facilitate the development of LF inhibitors, we applied a mixture-based peptide library approach that produces extended cleavage site motifs for proteases[20,21]. Initially we prepared a partially degenerate peptide mixture, acetyl-KKKPTPXXXXXAK (See Table 1 for explanation of nomenclature), in which we fixed six positions with the residues found N-terminal to the LF cleavage site in MKK-1 and followed them by a number of degenerate positions. Partial digestion of the library with LF followed by Edman sequencing of the mixture provided the specificity for the positions C-terminal to the cleavage site (Table 1).

[1]Division of Signal Transduction, Department of Medicine, Beth Israel Deaconess Medical Center, and Department of Cell Biology, Harvard Medical School, 330 Brookline Avenue, Boston, Massachusetts 02215, USA. [2]The Burnham Institute, 10901 North Torrey Pines Road, La Jolla, California 92037, USA. [3]National Institute of Allergy and Infectious Diseases, 9000 Rockville Pike, Bethesda, Maryland 20892, USA. [4]Department of Microbiology and Molecular Genetics, Harvard Medical School, 200 Longwood Avenue, Boston, Massachusetts 02115, USA. [5]These authors contributed equally to this work. Correspondence should be addressed to L.C.C. (lewis_cantley@hms.harvard.edu) or R.C.L. (rlidding@burnham.org).

Published online 29 December 2003; doi:10.1038/nsmb708

Table 1 LF cleavage site specificity and cleavage sites of known protein substrates

	Cleavage position									
	P6	P5	P4	P3	P2	P1	P1′	P2′	P3′	P4′
Consensus	R (2.1)	K (2.0)	K (2.0)	V (1.5)*	Y (3.1)	P	Y (3.0)	P (1.9)	N (1.4)	E (1.6)
	S (2.1)	R (1.9)	R (1.9)	P (1.5)*	R (1.6)		L (2.2)	Q (1.4)	M (1.3)	A (1.5)
	K (1.7)	S (1.7)	H (1.6)	F (1.4)*	F (1.4)		I (2.1)	R (1.4)		
	H (1.4)	H (1.5)	S (1.4)	A (1.4)*	L (1.3)		M (1.8)	K (1.3)		
							F (1.8)	G (1.3)		
							V (1.4)			
MKK-1	K	K	K	P	T	P	I	G	L	N
MKK-2	R	K	P	V	L	P	A	L	T	I
MKK-3	R	K	K	D	L	R	I	S	C	M
MKK-4	K	R	K	A	L	K	L	N	F	A
MKK-4	F	K	S	T	A	R	F	T	L	N
MKK-6	R	N	P	G	L	K	I	P	K	E
MKK-7	P	R	P	T	L	Q	L	P	L	A
MKK-7	P	R	H	M	L	G	L	P	S	T

Positions surrounding the scissile bond are defined as (...P3-P2-P1-P1′-P2′-P3′...) where cleavage occurs between the P1 and P1′ residues. Top: LF selectivity as determined using the peptide libraries acetyl-KKKPTPXXXXXAK (for the P1′–P4′ positions) and MXXXXXPYPMEDK(K-biotin) (for the P6–P2 positions). Selectivity values were determined by dividing the molar amount of a given residue within a sequencing cycle by the average molar amount of all residues within that cycle, so that a value of 1 is average and would thus indicate no selectivity. Only positive selections of ≥1.3 are shown. Values at the P3 position marked with an asterisk reflect the proportional increase of that residue from the previous cycle. Bottom: Residues present at positions surrounding the LF cleavage sites in MKK proteins.

To obtain selectivity information for sites N-terminal to the scissile bond, we constructed a secondary library, MXXXXXPYPMEDK (K-biotin), in which we fixed the residues most highly selected by LF at the primed positions. We also fixed proline at the P1 position, as an MKK-1 mutant bearing alanine at this position is not cleaved by LF[7]. Partial cleavage of this library was followed by removal of the undigested peptides and C-terminal fragments with immobilized avidin. Sequencing of the N-terminal fragments subsequently provided the specificity for LF at the unprimed positions (Table 1).

LF seems to be most selective at the P1′ position (immediately C-terminal to the scissile bond), where the enzyme requires a hydrophobic amino acid, and can accommodate both aliphatic and aromatic residues. Other features of the motif include a general selection for hydrophobic residues at the P2 position and an unusual selectivity for basic residues at multiple positions N-terminal to the cleavage site. Notably, sequence comparisons and mutagenesis studies have indicated that at least two basic residues and a downstream ΦXΦ sequence (where Φ indicates a hydrophobic amino acid and X any amino acid) are essential features of D-domains for mediating interactions with MAP kinases[22–24]. This similarity provides an evolutionary rationale for the targeting of these particular sites within the MKKs by LF: adaptive mutations in MKKs that would render them uncleavable would necessarily produce nonfunctional enzymes, thus making the acquisition of anthrax resistance unlikely.

Although general features of the selected consensus LF cleavage motif are reflected in the residues surrounding the cleavage sites within the MKKs (Table 1), specific aspects of the motif, such as the selection of tyrosine over other hydrophobic residues at the P1′ position, could not have been predicted based on consideration of known cleavage sites. Accordingly, a ten-residue peptide based on the consensus cleavage site (LF10) is cleaved ~50-fold more efficiently than an analogous MKK-1 cleavage site–spanning peptide (Table 2). We further substantiated the library selections by preparing additional peptides with alanine substitutions at various sites within the consensus. In each case, the substitution led to a substantial decrease in cleavage efficiency (Table 2). An extended 15-residue consensus peptide (LF15) provided a marked increase in cleavage efficiency over LF10, while maintaining favorable spectral properties (an eight-fold increase in fluorescence upon exhaustive cleavage). This peptide has the highest specificity constant of any LF peptide substrate thus far reported[17–19], allows detection of very low quantities of LF, and should therefore be useful in high-throughput screens for LF inhibitors.

Evaluation of peptide-based LF inhibitors

Substrate-derived inhibitors for metallo proteinases have been produced by incorporating a metal-chelating group either to the C terminus of a peptide corresponding to the unprimed positions, or to the N terminus of a peptide covering the primed positions[25,26]. As LF has substantial selectivity on either side of the scissile bond, we prepared both types of inhibitors and tested them for their ability to inhibit cleavage of the consensus peptide by LF. As in a previously reported study[19], we found that a relatively long C-terminal peptide hydroxamate is a potent LF inhibitor, whereas short peptide analogs such as acetyl-KVYP-hydroxamate inhibit the enzyme poorly (Table 3). Conversely, measurable inhibition was found with a small compound incorporating primed side residues, 2-thioacetyl-YPM-amide (SHAc-YPM, Table 3). This compound bears an N-terminal metal chelating group followed by a hydrophobic residue at the P1′ position, an arrangement shared by compounds previously reported to inhibit matrix metalloproteinases (MMPs)[27,28]. This relationship prompted us to test several similar MMP inhibitors for potency against LF. One such compound, GM6001 (3-(*N*-hydroxycarboxamido)-2-isobutyl-propanoyl-Trp-methylamide)[29], an N-terminal hydroxamic acid with a P1′ leucine mimetic, a P2′ tryptophan and a C-terminal methyl group, inhibited LF more potently than did the other compounds tested (Table 3 and data not shown). The enhanced potency of GM6001 over SHAc-YPM, despite the presence of predicted suboptimal residues, is probably attributable to the favorable substitution of the hydroxamic acid moiety for the thioacetyl group[28,30].

Table 2 Catalytic parameters for cleavage of substrate peptides by LF

Peptide	Sequence	k_{cat}/K_m (M^{-1} s^{-1})
MKK-1	Mca-KKPTPIQLN-Dnp	2,500 ± 800
LF10	Mca-KKVYPYPME-Dnp	130,000 ± 20,000
LF10-P5 Ala	Mca-AKVYPYPME-Dnp	7500 ± 500
LF10-P2 Ala	Mca-KKVAPYPME-Dnp	60,000 ± 10,000
LF10-P1′ Ala	Mca-KKVYPAPME-Dnp	22,000 ± 2,000
LF15	Mca-RRKKVYPYPME-Dnp-TIA	$4 \times 10^7 \pm 1 \times 10^7$

Residues in bold indicate substitutions to the consensus peptide. Substrate peptides contain N-terminal Mca (7-methoxycoumarin-4-acetyl) fluorescent groups and Dnp (2,4-dinitrophenyldiaminopropionic acid) quenching residues C-terminal to the cleavage site, allowing reaction progress to be followed fluorometrically by observing the increase in coumarin fluorescence upon cleavage (excitation 325 nm, emission 393 nm). For all peptides except LF15, the k_{cat}/K_m was determined by measuring the cleavage rate at 1 μM peptide (where $[S] << K_m$; [S] represents concentration of substrate). For the LF15 peptide, k_{cat} (3.4 s^{-1}) and K_m (85 nM) were determined individually by measuring the initial rate at various peptide concentrations. Values reflect the average of three separate determinations ± s.d.

Table 3 Potency of peptide-based LF inhibitors

Compound	K_i^{app} (μM)
Acetyl-KVYP-hydroxamate	>100
PLG-hydroxamate	>100
MKARRKKVYP-hydroxamate	0.0011 ± 0.0002
SHAc-YPM	11 ± 3
GM6001	2.1 ± 0.2

K_i^{app} values were determined by measuring inhibition of peptide cleavage (1 μM LF15 for the 10-mer hydroxamate or 1 μM LF10 for all other compounds) over a range of inhibitor concentrations. Values are the mean ± s.d. of three separate determinations, each done in triplicate.

Both SHAc-YPM and GM6001 inhibited cleavage of MKK proteins by LF *in vitro* with potency comparable to their ability to inhibit cleavage of the peptide substrate (Fig. 1a and data not shown). GM6001 also partially inhibited cleavage of MKKs in a LeTx-treated macrophage cell line (Fig. 1b). Notably, LF inhibition by GM6001 in cultured cells was sufficient to protect them from LeTx-induced cell death (Fig. 1c,d). Neither the thioacetyl compound nor the long C-terminal peptide hydroxamate was active in cell culture, presumably owing to poor cell permeability or metabolic instability (data not shown). We also found that the inhibitory potency of the C-terminal peptide hydroxamate (but not that of any of the other compounds) was substantially poorer when evaluated at physiological salt concentrations, which are much higher than for standard assay conditions for LF *in vitro* (data not shown). GM6001 could also prevent cell death when added as late as 3 h after LeTx, suggesting that it can protect cells subsequent to internalization of the toxin (Fig. 1e). These results indicate that small molecule metalloproteinase inhibitors provide a means to neutralize the biological activity of anthrax toxin.

Structures of LF in complex with peptides and inhibitors

To understand the molecular basis for substrate selectivity by LF and to guide further inhibitor design, we solved the X-ray crystal structures of LF in complex with a consensus peptide, LF20 (both in a zinc-free state and in an active site mutant with zinc), and with two of the inhibitors reported here, GM6001 and SHAc-YPM, both in the presence of zinc (Fig. 2a–c and Table 4). Crystals soaked in the MKARRKKVYP C-terminal hydroxamate showed additional electron density around the active site, but this was not interpretable as a single atomic model.

The LF20 peptide (MLARRKKVYPYPMEPTIAEG-amide) incorporates consensus residues (P5–P4′) surrounding the scissile bond based on the peptide library screen, flanked by residues of authentic MKK2. In the crystal structure of the zinc-free LF20 complex, nine peptide residues (from the P3 valine to the P6′ threonine) are defined by electron density; in the zinc-bound active site mutant, the peptide lies in the same location, and a further two residues at the N terminus are visible (lysines P5 and P4); whereas residues downstream of the cleavage site are in general less well defined, suggestive of partial cleavage. The peptide binds in an extended conformation, along the 40 Å–long substrate recognition groove (formed by domains II–IV) that was previously defined by soaking an MKK2-derived peptide into LF crystals[31] (Fig. 2a,d,e). However, the present complex structure is at substantially higher resolution than that of the earlier study, and, as expected, the LF20 binds more strongly than the MKK2 peptide. The new crystallographic data unequivocally demonstrate that the binding

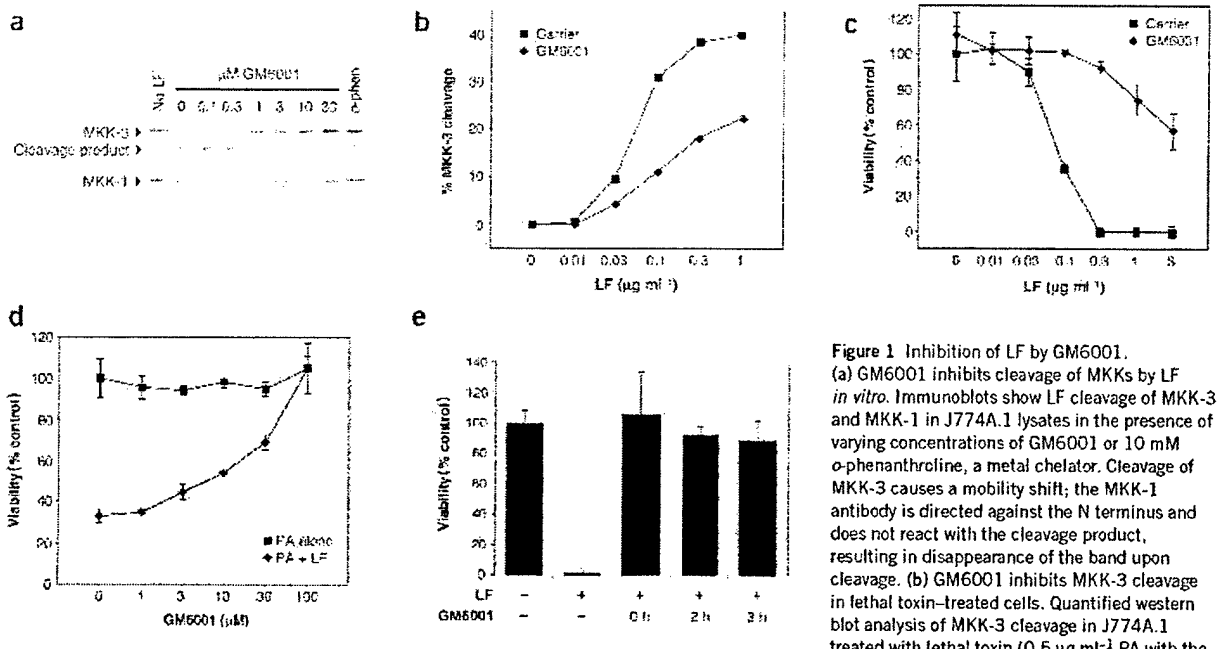


Figure 1 Inhibition of LF by GM6001. (a) GM6001 inhibits cleavage of MKKs by LF *in vitro*. Immunoblots show LF cleavage of MKK-3 and MKK-1 in J774A.1 lysates in the presence of varying concentrations of GM6001 or 10 mM *o*-phenanthroline, a metal chelator. Cleavage of MKK-3 causes a mobility shift; the MKK-1 antibody is directed against the N terminus and does not react with the cleavage product, resulting in disappearance of the band upon cleavage. (b) GM6001 inhibits MKK-3 cleavage in lethal toxin–treated cells. Quantified western blot analysis of MKK-3 cleavage in J774A.1 treated with lethal toxin (0.5 μg ml^{-1} PA with the indicated concentrations of LF) in the absence or presence of 100 μM GM6001. (c) Protection of J774A.1 cells from lethal toxin–mediated cell death by GM6001. Cell viability as determined by MTT assay after lethal toxin treatment in the presence of 100 μM GM6001 or 0.2% (v/v) DMSO carrier. (d) Dose-dependent neutralization of lethal toxin by GM6001. J774A.1 cell viability determined by MTT assay after treatment with lethal toxin (0.5 μg ml^{-1} PA + 0.3 μg ml^{-1} LF) or PA alone (0.5 μg ml^{-1}) in the presence of the indicated concentrations of GM6001. (e) GM6001 protects J774A.1 cells when added subsequent to LeTx. Cell viability is shown after treatment with PA alone (0.4 μg ml^{-1}) or PA with LF (25 ng ml^{-1}), with GM6001 added to 100 μM at the indicated time after toxin addition.

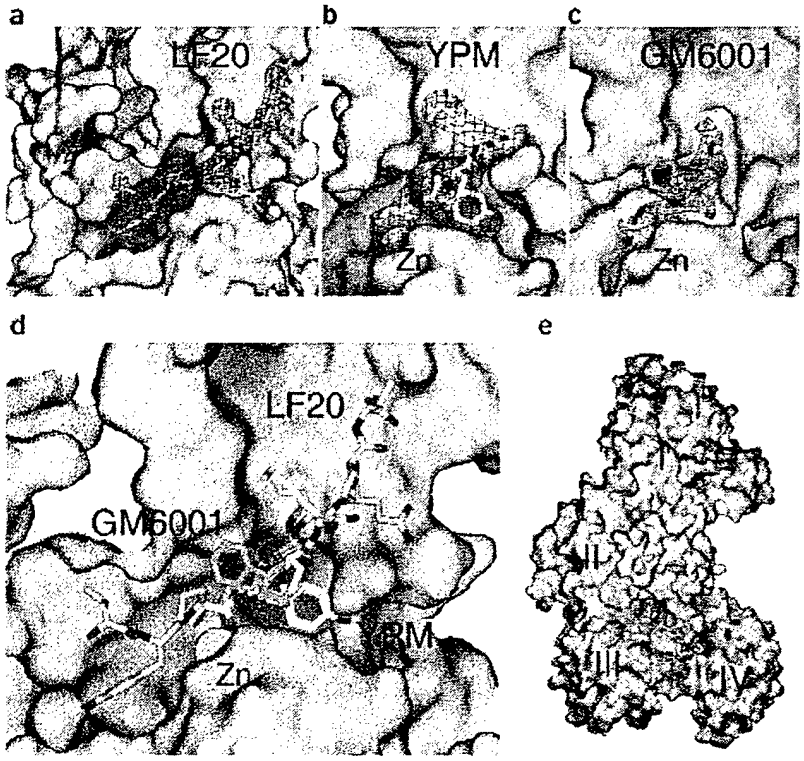


Figure 2 Structures of LF in complex with peptides and inhibitors. Molecular surface of LF is colored by charge (red, negative; blue, positive), with Zn^{2+} as a solid sphere (cyan) and the model of the peptide or inhibitor in ball-and-stick representation. The individual electron density surrounding each molecule is a $2F_o - F_c$ difference map calculated at the respective final resolution and contoured at 1.0 σ. (a) LF20 (yellow) in the absence of Zn^{2+}, resolution limit 2.85 Å. The model of bound LF20 shows the sequence VYPYPMEPT (residues 8–16 of the 20-residue-long LF20). This is the ordered region, and the electron density is clearly visible in difference maps ($2F_o - F_c$ and $F_o - F_c$) calculated from crystal X-ray diffraction data. (b,c) SHAc-YPM (white, labeled YPM), resolution limit 3.50 Å, and GM6001 (green), resolution limit 2.70 Å, respectively. Continuous electron density extends from the zinc atom to the metal-chelating moieties of the inhibitors (hydroxamate and thioacetyl, respectively). (d) The superposed individual complex structures of all three target molecules from a–c in the substrate-binding groove of LF, using the surface calculated for LF–LF20. The targets are all bound in the same N-to-C peptide orientation. (e) An overview of LF bound to the targets LF20, GM6001 and SHAc-YPM, superposed and colored as in d. The molecular surface was calculated from the LF–LF20 complex. The domains in LF are labeled I–IV. The catalytic site is in domain IV, where the zinc atom (not shown in this figure) is bound. These figures were prepared using SPOCK (http://mackerel.tamu.edu/spock/).

mode conforms to the canonical thermolysin substrate-binding mode[32]. The LF20 peptide is bound in a productive conformation, in contrast to that previously inferred from the LF–MKK2 structure[31], where the peptide is bound in a nonproductive mode (the reverse orientation and 6 Å distant from the active site). Therefore, the new complex structures, Protein Data Bank (PDB) entries 1PWV and 1PWW, supersede PDB entry 1JKY.

The ordered sequence of LF20 binds closely to the LF main chain and secondary structures surrounding the catalytic zinc-binding site. The P5 and P4 lysine residues lie close to a strongly acidic patch at the entrance to the active site, rationalizing the preference for basic residues at multiple positions upstream of the cleavage site. Residues P3–P1 form antiparallel β-sheet-like interactions with strand 4β3 of LF. The P2 tyrosine side chain occupies a fairly narrow hydrophobic pocket; this may explain the preference for tyrosine at this site. The P1′ tyrosine residue is buried within a deep hydrophobic S1′ pocket in LF, adjacent to the active site center. The pocket expands substantially on binding peptide (induced fit), including a ~3.5-Å shift of the main chain at Glu676 at the bottom of the pocket. Additionally, there is a ~3.0-Å shift of the side chain of Phe329, which is positioned along the substrate recognition groove, in close proximity to the active site and the bound peptide (this is also seen for all other bound ligands). The depth and plasticity of the S1′ cavity presumably allow the enzyme to accommodate large hydrophobic residues at the P1′ position; this explains why LF is most selective at this site.

The SHAc-YPM inhibitor shares three residues with the LF20 peptide downstream of the cleavage site, and the corresponding peptide electron density and derived model are markedly similar, with the P1′ tyrosine buried in the S1′ pocket (Fig. 2b,d,e). The thioacetyl moiety was modeled in a bidentate conformation[33,34] with the carbonyl oxygen atom and thiol sulfur atom directed toward the zinc. For the LF(E687C)–GM6001–Zn^{2+} complex (Fig. 2c–e), where LF(E687C) represents the LF E687C mutant, the peptide binds in a similar location. We modeled the hydroxamate moiety in the conventional bidentate planar conformation[27,32,33,35–37], with the carbonyl and hydroxyl oxygen atoms directed toward the zinc. The P1′ side chain is a leucine mimetic and binds in the S1′ pocket. The smaller side chain induces correspondingly less expansion of the S1′ pocket. The tryptophan side chain at the P2′ position makes no specific contacts with the protein, suggesting that it does not contribute to specificity.

DISCUSSION

The three independent LF-complex structures reported here indicate several common features essential for optimized substrate and inhibitor binding. The long hydrophobic substrate-binding groove and deep S1′ pocket adjacent to the catalytic Zn^{2+}-binding site seem to be the main determinants for strong target affinity. This strong hydrophobic selectivity has also been indicated by experimental data from nonpeptidic small molecule drug library screens of Panchal *et al.*[38] (this issue). These structures will enable the design of compounds with greater complementarity to the S1′ pocket and substrate recognition groove, combined with metal chelating groups spaced appropriately to allow for highly potent inhibition of LF.

Given the success of protease inhibition in the treatment of cardiovascular disease and AIDS, small molecule LF inhibitors would seem to be the most likely source for new drugs to treat anthrax. The possibility of encountering either naturally occurring or engineered antibiotic-resistant strains suggests that the availability of such

Table 4 Data collection summary for LF-complex crystals

	LF–LF20	LF(E687C)–LF20–Zn	LF–SHAc-YPM–Zn	LF(E687C)–GM6001–Zn
Data collection				
Space group	$P2_1$	$P2_1$	$P2_1$	$P2_1$
Cell dimensions (Å)				
a	96.70	96.70	96.70	96.70
b	137.40	137.40	137.40	137.40
c	98.30	98.30	98.30	98.30
Wavelength (Å)	1.07	0.98	1.08	0.97
Resolution range (Å)	50.0–2.85	30.0–2.80	30.0–3.50	50.0–2.70
Total reflections	96,701	94,088	91,831	255,861
Unique reflections	55,398	54,931	28,731	72,275
Completeness (%)[a]	92.2 (90.0)	86.8 (76.0)	90.8 (84.9)	99.6 (98.8)
R_{sym} (%)[a,b]	10.5 (46.6)	6.6 (40.9)	15.9 (45.1)	8.3 (48.0)
$I / \sigma I$[a]	6.7 (1.4)	12.2 (2.2)	7.4 (2.5)	15.6 (2.5)
Refinement statistics				
R_{work} (%)[b,c]	23.1	23.0	23.2	23.0
R_{free} (%)[b,c]	28.3	27.7	29.5	26.8

[a]Values in parentheses are for the highest-resolution shell. [b]$R_{sym} = \Sigma|I - \langle I\rangle| / \Sigma\langle I\rangle$, where I is the observed intensity and $\langle I\rangle$ is the average intensity from multiple observations of symmetry related reflections. [c]R-factor = $\Sigma||F_o| - |F_c|| / \Sigma|F_o|$; R_{work} represents reflections not in R_{free} set; R_{free} represents 5% of a random selection of data not used during refinement.

compounds would be crucial in minimizing potentially large numbers of deaths. The work described here creates many paths toward the production of such drugs, both by enabling the rapid screening of chemical libraries and by providing a structural basis for rational drug design. Our results suggest in particular that sizable libraries of MMP inhibitors already in existence are likely to contain additional LF inhibitors, perhaps with increased potency and specificity. This work also illustrates the utility of peptide libraries for both the rapid optimization of substrate peptides and the generation of lead compounds. Such methods should be generally applicable to any protease of interest as a therapeutic target.

METHODS

Peptide library methods. Cleavage site selectivity for LF was determined by modification of described methods[21]. Libraries were custom synthesized at the Tufts University Core Facility (Boston). Degenerate positions ('X') were prepared using isokinetic mixtures to produce equimolar amounts of the 19 proteogenic amino acids excluding cysteine. For determination of the primed side selectivity, the library acetyl-KKKPTPXXXXXAK (1 mM) was digested with LF[39] to 5–10% completion in a 10 µl reaction containing 20 mM HEPES, pH 7.4, 100 mM NaCl. The reaction products were analyzed by N-terminal peptide sequencing on an Applied Biosystems Procise 494 automated Edman sequencer. To determine the unprimed side selectivity, the library MXXXXXPYPMEDK(K-biotin) (20 µl at 1 mM) was digested to 5% completion as above, and quenched by adding an equal volume of 10 mM *o*-phenanthroline. The reaction products were incubated in batches with 500 µl avidin agarose (Sigma) in 500 µl of 25 mM ammonium bicarbonate with tumbling for 1 h, at which time the slurry was transferred to a column. The flowthrough and wash were combined, evaporated under reduced pressure and analyzed by Edman sequencing as described above.

Peptide cleavage assays. All peptides were synthesized at the Tufts University Core Facility except C-terminal peptide hydroxamates (Genemed Synthesis). Concentrations were determined based on the absorbance of the coumarin group (ε_{328} = 12,900 M^{-1} cm^{-1}) for the peptides and on tyrosine absorbance (ε_{280} = 1,200 M^{-1} cm^{-1}) for the inhibitors. Peptide cleavage assays were carried out in a Molecular Devices Spectramax Gemini XS fluorescence plate reader in black 96-well plates using LF10 digested to completion (which results in a 12-fold increase in fluorescence) as a standard. Reactions were run at 25 °C in 20 mM HEPES, pH 7.4, 0.1 mg ml^{-1} BSA (plus 1 mM DTT for assays of the thioacetyl inhibitor or 0.01% (v/v) Brij 35 for assays of the ten-residue hydroxamate inhibitor). For k_{cat} / K_m determinations, LF was used at 2–20 nM and the rates were determined from the linear range of the reaction progress curve (<10% substrate turnover). For the LF15 peptide, rates were determined in a continuous assay at varying substrate concentrations using a Photon Technology International Fluorescence system using 2 nM LF under the conditions described above, using the peptide at 1 µM digested to completion (eight-fold increase in fluorescence) as a standard. Data were corrected for the inner filter effect by measuring the quenching of an Mca-peptide standard at each substrate concentration. Data were fitted directly to the Michaelis-Menten equation. Peptide cleavage sites were confirmed by Edman sequencing of the reaction products.

Analysis of MKK cleavage. For *in vitro* MKK cleavage, J774A.1 cells were lysed in 0.5% (v/v) Igepal CA-630, 20 mM HEPES, pH 7.4, 100 mM NaCl, 1 mM DTT, 5% (v/v) glycerol, 1 mM PMSF, and 4 µg ml^{-1} each of leupeptin, pepstatin and aprotinin. LF was preincubated for 30 min at 25 °C with varying concentrations of inhibitor before the addition of J774A.1 cell lysate. After an additional 30 min the reaction was quenched by adding SDS-PAGE loading buffer. To analyze cleavage in cultured cells, J774A.1 cells in six-well plates were pretreated with GM6001 (CALBIOCHEM) or DMSO carrier alone (0.2% (v/v) final concentration in complete media) for 30 min at 37 °C before adding PA (to 0.5 µg ml^{-1}) and LF (to the indicated concentration). Cells were incubated at 37 °C for an additional 90 min, washed once with PBS and then lysed directly in SDS-PAGE loading buffer (100 µl per well) and boiled 10 min. Samples were fractionated by SDS-PAGE and transferred to PVDF membrane for immunoblotting with anti-MKK-3 (Santa Cruz Biotechnology C-19) or anti-MKK-1 N terminus (Upstate Biotechnology, catalog no. 06-269). MKK-3 cleavage was quantified using NIH Image software (http://rsb.info.nih.gov/nih-image/).

Lethal toxin assays. J774A.1 cells were plated in 96-well dishes at 3×10^5 cells per well and allowed to recover for 16 h, after which the medium was removed and replaced with fresh complete medium (100 µl per well) containing the indicated concentration of GM6001 or carrier alone (0.2% (v/v) DMSO). After 30 min, PA and/or LF were added to the indicated concentrations and incubation continued for an additional 4 h. To assay viability, 10 µl of 5 mg ml^{-1} MTT in PBS was added to each well, and incubation was continued for 2 h before aspirating the supernatant and extracting with 0.1 M HCl in isopropanol. Absorbance at 570 nm with a background correction at 690 nm was determined in an absorbance plate reader.

Crystallization. LF wild type and E687C active site mutant protein crystals were grown in 1.7 M $(NH_4)_2SO_4$, 0.2 M Tris-HCl, pH 8.0, 2 mM EDTA by the hanging-drop vapor diffusion method, at 25 ± 4 °C, using a protein concentration of 13 mg ml^{-1} (ref. 31). Cocrystals of LF with GM6001 grew under similar conditions. All crystals used are monoclinic, in space group $P2_1$, with unit cell dimensions a = 96.7 Å, b = 137.4 Å, c = 98.3 Å, α = 90°, β = 98.0°, γ = 90°, and contain two molecules per asymmetric unit. In general, similar features were observed at the two active sites, but the density for Molecule B was stronger.

LF–substrate and LF–inhibitor complexes. Native LF or LF E687C monoclinic $P2_1$ single crystals were harvested and bathed in several rounds of crystallization buffer prior to soaking in their respective target peptide or inhibitor solutions. Soaks were done at room temperature, 23 °C ± 2 °C. The treated crystals were then individually flash-frozen in liquid nitrogen. All data was collected was at 100 K, in a nitrogen cryostream.

The wild-type LF–LF20 peptide complex was obtained by soaking crystals in a solution of 10 mM LF20, 1.8 M $(NH_4)_2SO_4$, 0.2 M Tris-HCl, pH 8.0, 2 mM EDTA for 8 min. Each crystal was then transferred into a cryoprotectant solution of 10 mM LF20, 2.4 M $(NH_4)_2SO_4$, 0.2 M Tris-HCl, pH 8.0, 2 mM EDTA, 25% (v/v) glycerol, and bathed for a further 1 min before mounting in a cryoloop and flash-freezing. The LF(E687C)–LF20–Zn^{2+} crystal complex was first soaked in a solution of 1 mM $ZnSO_4$, 1.8 M $(NH_4)_2SO_4$, 0.2 M Tris-HCl, pH 8.0 for 5 min, followed by the treatment as described for the wild-type LF–LF20 complex.

The LF–SHAc-YPM inhibitor–Zn^{2+} complex was obtained by soaking crystals in 1 mM $ZnSO_4$, 1.8 M $(NH_4)_2SO_4$, 0.2 M Tris-HCl, pH 8.0 for 5 min; then in 5 mM SHAc-YPM, 1.8 M $(NH_4)_2SO_4$, 0.2 M Tris-HCl, pH 8.0 for a further 5 min; and then in 5 mM SHAc-YPM, 2.4 M $(NH_4)_2SO_4$, 0.2 M Tris-HCl, pH 8.0, 2 mM EDTA, 25% (v/v) glycerol for 1 min before mounting and freezing.

The LF–GM6001 and LF(E687C)–GM6001 inhibitor complex crystals were grown from a 1:2 molar ratio of LF to inhibitor and crystallized as for native. Crystals were soaked in 1 mM $ZnSO_4$, 1.8 M $(NH_4)_2SO_4$, 0.2 M Tris-HCl, pH 8.0 for 5 min, then in 0.1 mM GM6001 (0.7% (v/v) DMSO), 1.8 M $(NH_4)_2SO_4$, 0.2 M Tris-HCl, pH 8.0 for 2 min, and finally in 0.1 mM GM6001 (0.7% (v/v) DMSO), 2.4 M $(NH_4)_2SO_4$, 0.2 M Tris-HCl, pH 8.0, 2 mM EDTA, 25% (v/v) glycerol for <1 min before mounting and freezing. Using a LF(E687C)–GM6001 cocrystal, the LF(E687C)–GM6001–Zn^{2+} inhibitor complex crystal was also prepared with the method described here. No substantial differences in target binding or active site conformation between wild type or mutant LF–GM6001–Zn^{2+} complexes were observed (residue 687 is not involved directly in inhibitor or zinc binding). As the LF(E687C)–GM6001–Zn^{2+} complex gave higher-resolution data, this complex was used in further refinement.

Data collection. Data for the LF–LF20, LF(E687C)–LF20–Zn^{2+}, and LF–SHAc-YPM complexes were collected at the Stanford Synchrotron Radiation Laboratory (SSRL, Menlo Park, California, USA), on beamlines 1-5 (wavelength = 1.07 Å), 9-1 (wavelength = 0.98 Å) and 7-1 (wavelength = 1.08 Å). Data for the LF(E687C)–GM6001–Zn^{2+} complex were collected at the National Synchrotron Light Source (NSLS, Brookhaven, New York, USA) on beamline x12c (wavelength = 0.97 Å). X-ray diffraction data were collected for LF–LF20, LF(E687C)–LF20–Zn^{2+}, LF–SHAc-YPM–Zn^{2+}, and LF(E687C)–GM6001–Zn^{2+} to resolution limits of 2.85 Å, 2.80 Å, 3.50 Å and 2.70 Å, respectively.

Data processing and refinement. Crystallographic data were processed using the HKL package[40]. Refinement and model building were done in CNS[41] and O[42]. The high-resolution model of LF (PDB entry 1J7N)[31] was used as the starting model. The model was put through rigid body refinement and then minimization, and initial maps were calculated. Additional electron density at ≥1.0 σ in $2F_o - F_c$ and 2 σ in $F_o - F_c$ maps was clearly seen in the active site groove of LF for all cases. The model of the peptide or inhibitor with zinc was then built into this position and further refined in CNS[40]. Difference maps of the LF models, including peptide or inhibitor, and also omitting the peptide or inhibitor, were calculated in subsequent rounds of model rebuilding and refinement. Composite omit maps were also used. The final *R*-factors for each complex were as follows: LF–LF20 (Zn^{2+}-free), R_{free} = 28.3% and R_{work} = 23.1%; LF(E687C)–LF20–Zn^{2+}, R_{free} = 27.7% and R = 23.0%; LF–SHAc-YPM–Zn^{2+}, R_{free} = 29.5% and R = 23.2%; and LF(E687C)–GM6001–Zn^{2+}, R_{free} = 26.8% and R = 23.0%. The final models fall within or exceed the limits of all the quality criteria of PROCHECK from the CCP4 suite[43].

Coordinates. Coordinates and structure factors have been deposited in the Protein Data Bank (accession codes: 1PWQ, LF–YPM–Zn^{2+}; 1PWU, LF(E687C)–GM6001–Zn^{2+}; 1PWV, LF–LF20; 1PWW, LF(E687C)–LF20–Zn^{2+}).

ACKNOWLEDGMENTS

Thanks to P. Bartlett (University of California Berkeley), D. Tronrud and B. Matthews (University of Oregon) and B. Rupp (Lawrence Livermore National Laboratory) for pointing out the canonical binding mode for Zn metalloproteases and to E. Garman (University of Oxford) and A. Gonzalez (SSRL) for discussions on the crystallography. We also thank H. Robinson and S. Vaday for collecting data at the National Synchrotron Light Source (NSLS). Portions of this research were carried out at the Stanford Synchrotron Radiation Laboratory (SSRL), a national user facility operated by Stanford University on behalf of the US Department of Energy (DOE), Office of Basic Energy Sciences. The SSRL Structural Molecular Biology Program is supported by the DOE, Office of Biological and Environmental Research, and by the NIH, National Center for Research Resources, Biomedical Technology Program and the National Institute of General Medical Sciences. Data for this work were also collected at the NSLS, Brookhaven National Laboratory, which is supported by the DOE, Division of Materials Sciences and Division of Chemical Sciences, under contract no. DE-AC02-98CH10886. This work was supported by NIH (R.J.C., R.C.L. and L.C.C.), the US National Science Foundation (L.C.C.) and the US Department of the Army (DAMD17-03-1-0062 to L.C.C.). The US Army Medical Research Acquisition Activity, 820 Chandler Street, Fort Detrick, Maryland 21702-5014 is the awarding and administering acquisition office. The contents of this manuscript do not necessarily reflect the position or policy of the US government, and no official endorsement should be inferred. B.E.T. is a Leukemia and Lymphoma Society special fellow.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Structural & Molecular Biology* website for details).

Received 2 July; accepted 23 October 2003
Published online at http://www.nature.com/natstructmolbiol/

1. Dixon, T.C., Meselson, M., Guillemin, J. & Hanna, P.C. Anthrax. *New Engl. J. Med.* **341**, 815–826 (1999).
2. Duesbery, N.S. & Vande Woude, G.F. Anthrax toxins. *Cell. Mol. Life Sci.* **55**, 1599–1609 (1999).
3. Moayeri, M., Haines, D., Young, H.A. & Leppla, S.H. *Bacillus anthracis* lethal toxin induces TNF-α-independent hypoxia-mediated toxicity in mice. *J. Clin. Invest.* **112**, 670–682 (2003).
4. Pezard, C., Berche, P. & Mock, M. Contribution of individual toxin components to virulence of *Bacillus anthracis*. *Infect. Immun.* **59**, 3472–3477 (1991).
5. Sellman, B.R., Mourez, M. & Collier, R.J. Dominant-negative mutants of a toxin subunit: an approach to therapy of anthrax. *Science* **292**, 695–697 (2001).
6. Mourez, M. *et al.* Designing a polyvalent inhibitor of anthrax toxin. *Nat. Biotechnol.* **19**, 958–961 (2001).
7. Duesbery, N. *et al.* Proteolytic inactivation of MAP-kinase-kinase by anthrax lethal factor. *Science* **280**, 734–737 (1998).
8. Vitale, G. *et al.* Anthrax lethal factor cleaves the N-terminus of MAPKKs and induces tyrosine/threonine phosphorylation of MAPKs in cultured macrophages. *Biochem. Biophys. Res. Commun.* **248**, 706–711 (1998).
9. Pellizzari, R., Guidi-Rontani, C., Vitale, G., Mock, M. & Montecucco, C. Anthrax lethal factor cleaves MKK3 in macrophages and inhibits the LPS/IFNγ-induced release of NO and TNFα. *FEBS Lett.* **462**, 199–204 (1999).
10. Vitale, G., Bernardi, L., Napolitani, G., Mock, M. & Montecucco, C. Susceptibility of mitogen-activated protein kinase kinase family members to proteolysis by anthrax lethal factor. *Biochem. J.* **352**, 739–745 (2000).
11. Enslen, H. & Davis, R.J. Regulation of MAP kinases by docking domains. *Biol. Cell* **93**, 5–14 (2001).
12. Agrawal, A. *et al.* Impairment of dendritic cells and adaptive immunity by anthrax lethal toxin. *Nature* **424**, 329–334 (2003).
13. Friedlander, A.M. Macrophages are sensitive to anthrax lethal toxin through an acid-dependent process. *J. Biol. Chem.* **261**, 7123–7126 (1986).
14. Hanna, P.C., Acosta, D. & Collier, R.J. On the role of macrophages in anthrax. *Proc. Natl. Acad. Sci. USA* **90**, 10198–10201 (1993).
15. Park, J.M., Greten, F.R., Li, Z.W. & Karin, M. Macrophage apoptosis by anthrax lethal factor through p38 MAP kinase inhibition. *Science* **297**, 2048–2051 (2002).
16. Chopra, A.P., Boone, S.A., Liang, X. & Duesbery, N.S. Anthrax lethal factor proteolysis and inactivation of MAPK kinase. *J. Biol. Chem.* **278**, 9402–9406 (2003).
17. Hammond, S.E. & Hanna, P.C. Lethal factor active-site mutations affect catalytic activity *in vitro*. *Infect. Immun.* **66**, 2374–2378 (1998).
18. Cummings, R.T. *et al.* A peptide-based fluorescence resonance energy transfer assay for *Bacillus anthracis* lethal factor protease. *Proc. Natl. Acad. Sci. USA* **99**, 6603–6606 (2002).
19. Tonello, F., Seveso, M., Marin, O., Mock, M. & Montecucco, C. Screening inhibitors of anthrax lethal factor. *Nature* **418**, 386 (2002).
20. Songyang, Z. *et al.* SH2 domains recognize specific phosphopeptide sequences. *Cell* **72**, 767–778 (1993).
21. Turk, B.E., Huang, L.L., Piro, E.T. & Cantley, L.C. Determination of protease cleavage site motifs using mixture-based oriented peptide libraries. *Nat. Biotechnol.* **19**, 661–667 (2001).
22. Tanoue, T., Adachi, M., Moriguchi, T. & Nishida, E. A conserved docking motif in MAP kinases common to substrates, activators and regulators. *Nat. Cell Biol.* **2**, 110–116 (2000).
23. Enslen, H., Brancho, D.M. & Davis, R.J. Molecular determinants that mediate selective activation of p38 MAP kinase isoforms. *EMBO J.* **19**, 1301–1311 (2000).
24. Xu, B., Stippec, S., Robinson, F.L. & Cobb, M.H. Hydrophobic as well as charged residues in both MEK1 and ERK2 are important for their proper docking. *J. Biol. Chem.* **276**, 26509–26515 (2001).
25. Holmquist, B. & Vallee, B.L. Metal-coordinating substrate analogs as inhibitors of

metalloenzymes. *Proc. Natl. Acad. Sci. USA* **76**, 6216–6220 (1979).
26. Moore, W.M. & Spilburg, C.A. Purification of human collagenases with a hydroxamic acid affinity column. *Biochemistry* **25**, 5189–5195 (1986).
27. Gowravaram, M.R. *et al.* Inhibition of matrix metalloproteinases by hydroxamates containing heteroatom-based modifications of the P1′ group. *J. Med. Chem.* **38**, 2570–2581 (1995).
28. Baxter, A.D. *et al.* A novel series of matrix metalloproteinase inhibitors for the treatment of inflammatory disorders. *Bioorg. Med. Chem. Lett.* **7**, 897–902 (1997).
29. Grobelny, D., Poncz, L. & Galardy, R.E. Inhibition of human skin fibroblast collagenase, thermolysin, and *Pseudomonas aeruginosa* elastase by peptide hydroxamic acids. *Biochemistry* **31**, 7152–7154 (1992).
30. Levy, D.E. *et al.* Matrix metalloproteinase inhibitors: a structure-activity study. *J. Med. Chem.* **41**, 199–223 (1998).
31. Pannifer, A.D. *et al.* Crystal structure of the anthrax lethal factor. *Nature* **414**, 229–233 (2001).
32. Holmes, M.A. & Matthews, B.W. Binding of hydroxamic acid inhibitors to crystalline thermolysin suggests a pentacoordinate zinc intermediate in catalysis. *Biochemistry* **20**, 6912–6920 (1981).
33. Grams, F. *et al.* X-ray structures of human neutrophil collagenase complexed with peptide hydroxamate and peptide thiol inhibitors. Implications for substrate binding and rational drug design. *Eur. J. Biochem.* **228**, 830–841 (1995).
34. Gaucher, J.F. *et al.* Crystal structures of α-mercaptoacyldipeptides in the thermolysin active site: structural parameters for a Zn monodentation or bidentation in metalloendopeptidases. *Biochemistry* **38**, 12569–12576 (1999).
35. Dhanaraj, V. *et al.* X-ray structure of a hydroxamate inhibitor complex of stromelysin catalytic domain and its comparison with members of the zinc metalloproteinase superfamily. *Structure* **4**, 375–386 (1996).
36. Chen, L. *et al.* Crystal structure of the stromelysin catalytic domain at 2.0 Å resolution: inhibitor-induced conformational changes. *J. Mol. Biol.* **293**, 545–557 (1999).
37. Roswell, S. *et al.* Crystal structure of human MMP9 in complex with a reverse hydroxamate inhibitor. *J. Mol. Biol.* **319**, 173–181 (2002).
38. Panchal, R. *et al.* Identification of small molecule inhibitors of anthrax lethal factor. *Nat. Struct. Mol. Biol.* **11**, 67–72 (2004).
39. Roberts, J.E., Walters, J.W., Ballard, J.D. & Dietrich, W.F. Ltx1, a mouse locus that influences the susceptibility of macrophages to cytolysis caused by intoxication with *Bacillus anthracis* lethal factor, maps to chromosome 11. *Mol. Microbiol.* **29**, 581–591 (1998).
40. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
41. Brunger, A.T. *et al.* Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
42. Jones, T.A., Zou, J.Y., Cowan, S.W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
43. Bailey, S. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).

Crystal structure of an anthrax toxin-host cell receptor complex

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, CA 92037, U.S.A.*

[2]*Microbial Pathogenesis Section, National Institute of Allergy and Infectious Diseases, NIH, Bethesda, MD 20892, USA*

Anthrax toxin consists of three proteins: Protective Antigen (PA), Lethal Factor (LF) and Edema Factor (EF)[1]. The first critical step in the entry of the toxin into 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} pre-pore 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 receptor molecules, TEM8 and CMG2, bind to PA with high affinity and are required for 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 that mimics 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 quite different in the integrin domains, explaining the specificity of the interaction. CMG2 engages two domains of PA, and modeling of the receptor-bound PA_{63}

heptamer[6-8] suggests that the receptor acts as a pH-sensitive chaperone to ensure accurate and timely membrane insertion. The structure will provide new leads for the discovery of anthrax anti-toxins, and will aid in 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]. The PA monomer is a long slender molecule comprising four distinct domains. Two of these, 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 (1900 $Å^2$), consistent with the very high affinity (sub-nanomolar Kd) of this interaction[11]. The I domain adopts the "open" conformation, typical of integrin-ligand complexes[5,12]. PA mimics the ligand recognition mechanism of the integrins[5] by contributing an aspartic acid sidechain that completes the coordination sphere of the MIDAS magnesium ion, as predicted by mutagenesis[13,14] (Fig. 2). This single interaction contributes substantially to binding, since 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 explain the specificity of the interaction, as it does not distinguish between CMG2 and integrins. Specificity arises from two further 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 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 are two further loops from

CMG2 which make a number of specific polar interactions and salt-bridges (Figs. 3, 4a). Together with the MIDAS contact, CMG2 and PA domain IV bury 1300 $Å^2$ of surface area, a value very similar to two integrin-ligand interactions, which have affinities in the sub-micromolar range[5,12]. CMG2 and TEM8 are 60% identical in their I domains, and homology modeling 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 an identical fashion; however, the structure and sequence of the ridge are very different in integrins, explaining their weak binding.

The interaction between PA domain II and CMG2 was unexpected. 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). This additional contact rationalizes the very high affinity of the PA-CMG2 interaction. The pocket is adjacent to the MIDAS motif and is formed by two exposed tyrosines (119 and 158) and the β4-α4 loop, which line the sides of the pocket, and by a histidine at its base. The pocket is conserved in TEM8, but does not exist in the integrins I domains, thus providing further specificity. The importance of this loop was shown by systematic mutation of the PA molecule, which revealed 3 mutations in this loop that reduced toxicity by > 100-fold, including G342 at the tip of the β-hairpin that inserts into the pocket[15].

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, while 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[16]. 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 chaperone to prevent premature membrane insertion on the cell surface prior to endocytosis. The titration of histidines is implicated in triggering the conformational switch. The histidine at the base of the CMG2 pocket has no H-bonding partners, and is close to an arginine sidechain from the β3-β4 loop of PA. Protonation of this histidine provides a plausible trigger for the release of domain II from CMG2 in the acidified endosome. Moreover, the structure of the β3-β4 loop is pH sensitive, since it is ordered in crystals of PA grown at pH 7.5 (in the absence of receptor), but disordered in crystals grown at pH 6.0[6].

It is straightforward to model the 7:7 heptameric PA_{63}-CMG2 complex, since 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[11], and their N- and C-termini point downwards, towards the membrane. In the transition from pre-pore to pore, the 7 hairpin loops, one from each PA monomer[6,8], are predicted to create a 14-stranded membrane-spanning β-barrel. Assuming an α-hemolysin-like structure [17], 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[16].

Soluble versions of the CMG2 and TEM8 I domains protect against anthrax toxicity by acting as decoys[3,14], 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, while CMG2 is widely expressed in most tissues[18,19]. Anthrax toxin is being developed as an anti-tumour agent[20], and our structure will allow the design of PA molecules that bind better to TEM8 than to CMG2, thus minimizing the toxic side-effects from binding to CMG2 in normal tissues.

Methods

Protein expression and purification

PA was prepared as previously described[13]. The I domain of CMG2 was cloned as an N-terminal His-tag fusion in pET15b (Novagen) and expressed in *E. coli* strain BL21(DE3). Following 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 and flash-frozen in 150 mM NaCl, 20 mM TrisCl pH7.5, and comprises residues 40-218 of $CMG2^{386}$ (accession number AAK77222) plus an N-terminal extension of sequence GSHMLEDPRG as a result of the cloning strategy. The molecular weight was confirmed by MALDI-TOF mass spectrometry. To prepare the PA-CMG2 complex, PA was mixed at a final concentration of 4 mg/ml with a 3-fold molar excess of CMG2 and a 2-fold 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 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 x 10 x 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 cryoprotectant prior to diffraction analysis. They 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[21] (see Table 1). PA was positioned in the unit cell by Molecular Replacement (PDB ID code 1acc)[6] using MOLREP, and refined with REFMAC version 5.0[22]. 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)[23] was manually placed in the electron density. Model building was performed with O[24] and TURBOFRODO[25], and the solvent structure was built with ARP/wARP 6.0[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%, R_{WORK} = 20.7%) overall and (R_{FREE} = 37.2 %, R_{WORK} = 27.5%) in the outer resolution bin, with RMS deviations from ideal values of 0.017 Å for bond lengths and 1.65° for angles. Stereochemistry is excellent as assessed with PROCHECK[22], and the model is consistent with composite simulated annealing omit maps (3000°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; 2 Ca^{2+} ions in PA domain I; 2 Na^{+} ions; one PEG molecule; and one Mn^{2+} ion at the MIDAS site. 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]. PA domain 1 (residues 16-258) undergoes a small rotation as a consequence of crystal constraints when compared with the structure of isolated PA such that the rmsd 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 an rmsd of 0.60 with the isolated protein[23], while the C-terminal helix (residues 201-210) shifts downward by one helical turn.

1. Moayeri, 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 U S A.* **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 b-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 the α subunit of integrin CR3 (CD11b/CD18). *Cell* **80**, 631-635 (1995).

11. 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.* **Mar 04 [Epub ahead of print]** (2004).

12. 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).

13. 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).

14. 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).

15. Mourez, M. *et al.* Mapping dominant-negative mutations of anthrax protective antigen by scanning mutagenesis. *Proc Natl Acad Sci U S A.* **100**, 13803-13808 (2003).

16. Miller, C. J., Elliott, J. L. & Collier, R. J. Anthrax protective antigen: prepore-to-pore conversion. *Biochemistry* **38**, 10432-10441 (1999).

17. Song, L. *et al.* Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. *Science* **274**, 1859-1866 (1996).

18. Nanda, A. & St Croix, B. Tumor endothelial markers: new targets for cancer therapy. *Curr Opin Oncol.* **16**, 44-49 (2004).

19. Nanda, A. *et al.* TEM8 interacts with the cleaved C5 domain of collagen alpha 3(VI). *Cancer Res.* **64**, 817-820 (2004).

20. 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).

21. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzym.* **276**, 307-326 (1997).

22. Collaborative Computational Project, N. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr.* **D50**, 760-763 (1994).

23. 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 U S A.* **101**, 6367-6372 (2004).

24. 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.* **A47**, 110-119 (1991).

25. Roussel, A. & Cambillau, C. 86 (Silicon Graphics, Mountain View, CA, 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 Cryst.* **D54**, 905-921 (1998).

28. 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 the NIH and the DOD for financial support, and the DOE and staff at the SSRL for synchrotron access and support

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

Figure 1 Structure of the PA-CMG2 complex. Two orthogonal views are shown as ribbons. PA is coloured by domain (I-IV). CMG2 is in blue. The metal ion is shown as a magenta ball. All molecular graphics images were generated using the UCSF Chimera package[28] (http://www.cgl.ucsf.edu/chimera).

Figure 2 Comparison of the MIDAS motifs of the **a**, PA-CMG2 and **b**, integrin $\alpha 2\beta 1$–collagen[5] complexes. Coordinating side chains and two water molecules (ω) are shown as ball-and-stick. The metal is shown in blue. Carbon and oxygen atoms from CMG2 and integrin are dark blue and red, and numbered in black for CMG2. D683 from Domain IV of PA, and the analogous collagen glutamic acid, are in gold. Loops are shown as grey ribbons. Hydrogen bonds to the metal-bound waters are shown as dotted lines.

Figure 3 Intermolecular contacts between PA domains II and IV and CMG2. Contacting regions are colored blue and green for PA domain IV and CMG2, respectively. The β3-β4 loop, β2 and β3 strands and β2- β 3 loop of PA domain II, which are implicated in pore formation, are highlighted in red. The β2-β3 loop, which is disordered in monomeric PA, is shown as a dashed line. The MIDAS metal is labeled "M". The side chains of PA D683 and CMG2 H121 are shown as ball-and-stick in gold and cyan, respectively.

Figure 4 Key elements of the PA-CMG2 interaction **a**, Solvent-accessible surface (probe radius 1.4 Å) of the PA domain IV groove, with key side chains from three CMG2 loops (β1-α1, blue; β2-β3, red; α2-α3, green) shown as ball-and-stick (C: yellow, O: red, N: blue). The green loop forms the top of the groove. The MIDAS metal is labelled (M). **b**, Solvent-accessible surface (probe radius 1.0 Å) showing the CMG2 pocket into which the PA β3-β4 loop (red ribbon) inserts. The pocket is formed by three CMG2 sidechains (shown as ball-and-stick) and the β4-α4 loop (cyan).

Figure 5 Hypothetical model of the receptor-bound membrane-inserted PA pore. The PA_{63} heptamer (red) is based on the pre-pore crystal structure[6], with a hypothetical model of a membrane-spanning 14-stranded barrel[17] formed by rearrangement in each monomer of the segment shown in red in Figure 3. Seven copies of the CMG2 I domain bound to the heptamer are in blue. The 40 Å gap 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

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	
Rmerge (%)	17.6 (89.1)	
I/σ	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	
rmsd bond lengths (Å)	0.17	
rmsd 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.

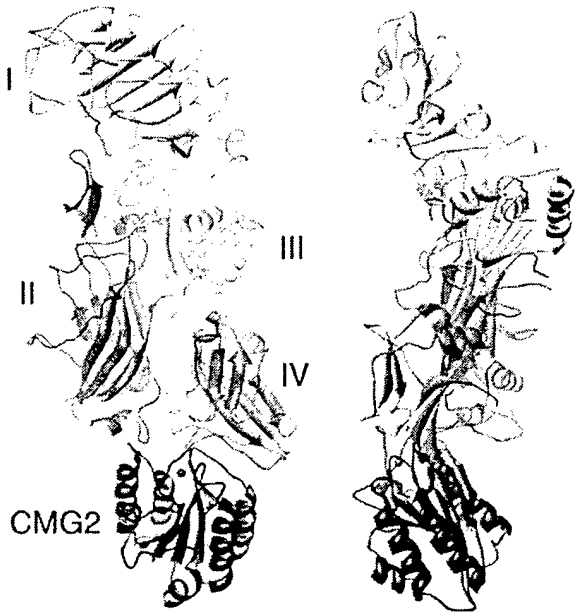


Figure 1

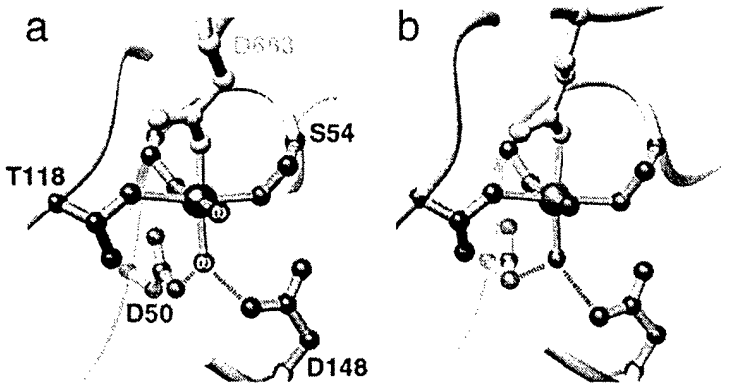


Figure 2

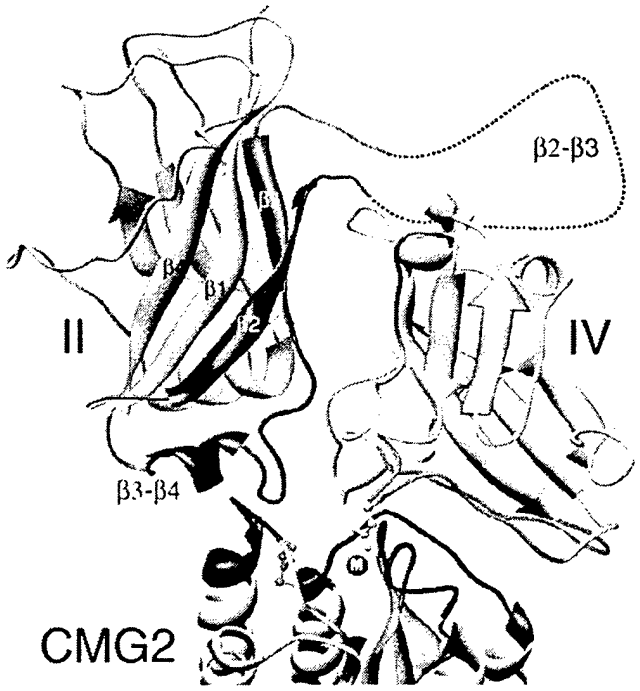


Figure 3

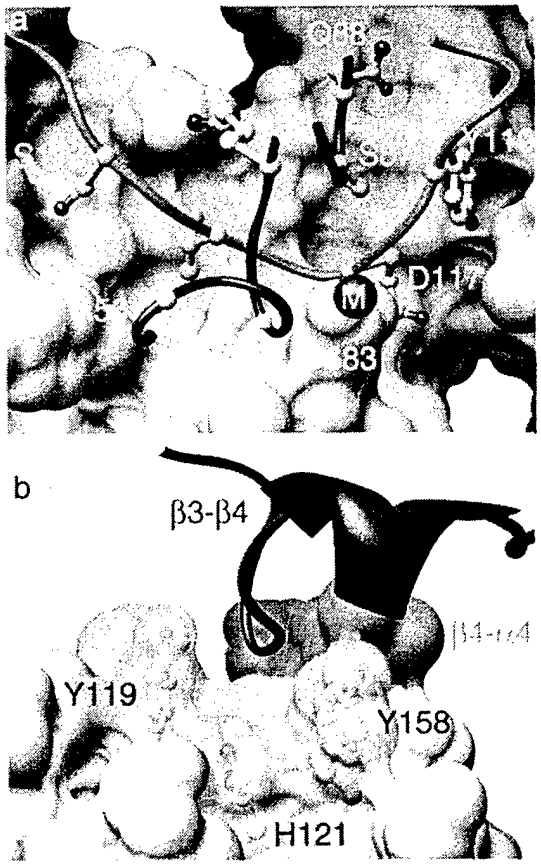


Figure 4

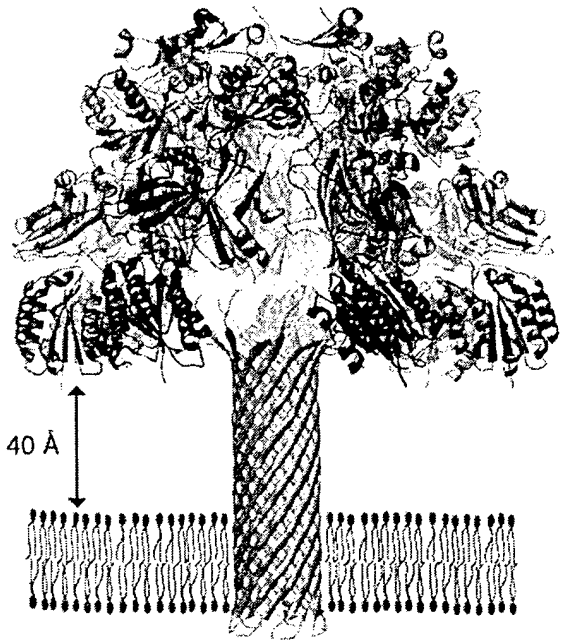


Figure 5